European Journal of Human Genetics (2021) 29:1446–1454
https://doi.org/10.1038/s41431-021-00835-8

ARTICLE

Colocalization analysis of polycystic ovary syndrome to identify potential disease-mediating genes and proteins

Jenny C. Censin[1,2] · Jonas Bovijn[1,2] · Michael V. Holmes[3,4,5] · Cecilia M. Lindgren[1,2,3,6,7]

Received: 21 October 2020 / Revised: 2 February 2021 / Accepted: 9 February 2021 / Published online: 4 March 2021

Abstract

Polycystic ovary syndrome (PCOS) is a common complex disease in women with a strong genetic component and downstream consequences for reproductive, metabolic and psychological health. There are currently 19 known PCOS risk loci, primarily identified in women of Han Chinese or European ancestry, and 14 of these risk loci were identified or replicated in a genome-wide association study of PCOS performed in up to 10,074 cases and 103,164 controls of European descent. However, for most of these loci the gene responsible for the association is unknown. We therefore use a Bayesian colocalization approach (Coloc) to highlight genes in PCOS-associated regions that may have a role in mediating the disease risk. We evaluated the posterior probabilities of evidence consistent with shared causal variants between 14 PCOS genetic risk loci and intermediate cellular phenotypes in one protein ($N = 3301$) and two expression quantitative trait locus datasets ($N = 31{,}684$ and $N = 80$–491). Through these analyses, we identified seven proteins or genes with evidence of a possibly shared causal variant for almost 30% of known PCOS signals, including follicle stimulating hormone and *ERBB3*, *IKZF4*, *RPS26*, *SUOX*, *ZFP36L2*, and *C8orf49*. Several of these potential effector proteins and genes have been implicated in the hypothalamic–pituitary–gonadal signalling pathway and provide an avenue for functional follow-up in order to demonstrate a causal role in PCOS pathophysiology.

Introduction

Polycystic ovary syndrome (PCOS) is a common endocrinopathy, affecting between 6 and 10% of women of reproductive age [1], with consequences for reproductive, metabolic, and psychological health [2, 3]. There is evidence of a clear genetic component [4], and genome-wide association studies (GWASs) have identified 19 risk loci [5–9]. Some of these risk loci are close to genes with a plausible connection to PCOS pathophysiology, including genes involved in for example insulin and hypothalamic–pituitary–gonadal (HPG) signalling (e.g., *INSR*, the insulin receptor gene and *FSHR*, the FSH-receptor gene) [3, 6–8]. However, for most PCOS-associated loci the mediating genes and their functional effects remain to be identified and/or confirmed [6].

One approach to improve biological understanding of a disease risk locus is through colocalization analysis of the disease and intermediate cellular phenotypes, such as gene expression and protein levels in different tissues [10]. Colocalization analysis quantifies the probability that two traits share the same causal variant, and can thereby highlight genes and proteins that may mediate the risk of a disease [10]. We

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41431-021-00835-8.

✉ Jenny C. Censin
jenny.censin@ndm.ox.ac.uk

[1] Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK

[2] Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

[3] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK

[4] Medical Research Council Population Health Research Unit at the University of Oxford, Nuffield Department of Population Health, University of Oxford, Oxford, UK

[5] Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, Big Data Institute Building, Roosevelt Drive, University of Oxford, Oxford, UK

[6] Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA

[7] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

therefore investigated the evidence of colocalization between 14 PCOS-associated loci identified in a recent GWAS in Europeans [6] together with one study with protein and two studies with expression quantitative trait loci (pQTL and eQTL, respectively). Our results highlight several genes and proteins linked to the HPG axis and follicular development, including e.g. FSH, *ZFP36L2*, and *RAD50*, that may be of particular interest for further functional follow-up.

Materials and methods

Polycystic ovary syndrome dataset

We obtained GWAS summary statistics for PCOS from Day et al. [6]. In their study, 14 genome-wide significant loci were identified in up to 10,074 cases and 103,164 controls of European ancestry (Fig. 1 and Table 1). Public summary statistics and single nucleotide polymorphism (SNP) estimates were available for (a) the 10,000 most robustly associated SNPs with estimates computed in the full sample and (b) for all SNPs with estimates computed in analyses excluding one of the cohorts (23andMe), resulting in a sample size of up to 4890 cases and 20,405 controls. We combined the two SNP summary statistics datasets to one dataset for use in the main analyses, with preference given to summary statistics computed using the full sample size. We then excluded SNPs found to be duplicated by position, missing relevant data, or indels.

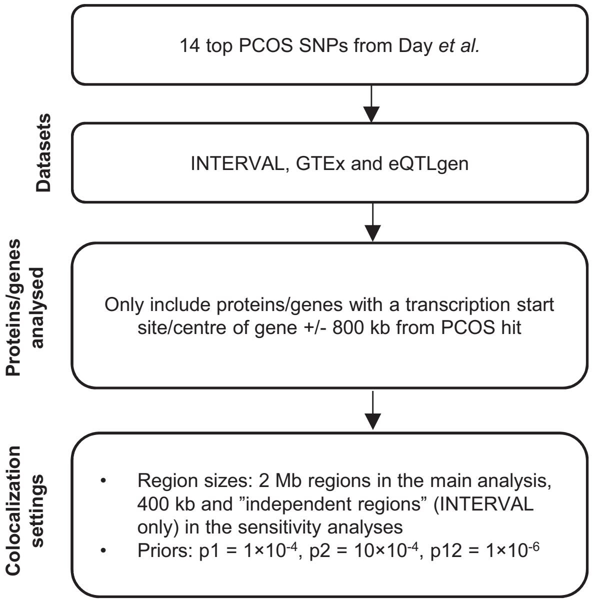

14 top PCOS SNPs from Day *et al.*
Datasets
INTERVAL, GTEx and eQTLgen
Proteins/genes analysed
Only include proteins/genes with a transcription start site/centre of gene +/- 800 kb from PCOS hit
Colocalization settings
• Region sizes: 2 Mb regions in the main analysis, 400 kb and "independent regions" (INTERVAL only) in the sensitivity analyses
• Priors: p1 = $1×10^{-4}$, p2 = $10×10^{-4}$, p12 = $1×10^{-6}$

Fig. 1 Study overview. Top SNPs associated with PCOS were extracted from Day et al. [6]. Colocalization was then performed between PCOS risk and both gene expression and protein levels, using data from INTERVAL [11, 12], GTEx [13], and eQTLgen [14]. Proteins and genes with a transcription start site or center of gene (depending on the gene/protein dataset) within 800 kb from a PCOS SNP were then analysed using the Bayesian colocalization software Coloc [10]. PCOS polycystic ovary syndrome; SNP single nucleotide polymorphism.

Protein and expression quantitative trait loci datasets

We used publicly available protein and expression genetic association data from the INTERVAL study [11, 12], the GTEx consortium [13], and the eQTLgen consortium [14]. pQTL data were taken from the INTERVAL study, which had performed GWASs for 2994 unique plasma proteins (3283 measured aptamers) in 3301 blood donors of European ancestry [11]. For GTEx, we used data from version 7, which contains *cis*-eQTL data for between 80 and 491 samples in 48 different tissues [13, 15]. Expression had been measured post-mortem, with ~85% of donors being of European ("White") ancestry in the whole sample [15]. Lastly, the eQTLgen consortium had performed *cis*- and *trans*-eQTL analysis in up to 31,684 individuals, predominantly of European ancestry [14]. Both *cis*-associations, containing SNPs within 1 Mb from the centre of the gene, and trans-associations, containing SNPs over 5 Mb from the centre of the gene, are publicly available [14]. For all these datasets, we then excluded SNPs that were duplicated by position, missing relevant data, or indels.

Colocalization analyses

We applied Coloc [10], a Bayesian test for colocalization, to evaluate the probability of evidence consistent with a shared causal signal between each PCOS risk loci and each p/eQTL (see Supplement). We performed colocalization using the coloc.abf() function in the Coloc R package, applying it to *cis*-genes using up to three different region sizes depending on QTL dataset.

For GTEx and eQTLgen, *cis*-association statistics were only available for SNPs within 1 Mb of the transcription start site and the centre of the gene, respectively [13, 14]. We therefore only analyzed genes and proteins with a transcription start site or centre of gene ± 800 kb of each top PCOS SNP (by *P* value) for all three QTL datasets, to ascertain that we had a sufficiently large region on both sides of the association peak to determine colocalization. We analyzed two different region sizes in GTEx and eQTLgen—the entire 2 Mb *cis*-region available in these datasets in the main analysis and a 400 kb region around the position of the top PCOS SNP as a sensitivity analysis. For GTEx, we only performed the analysis if the PCOS index SNP was present in the GTEx summary statistics for computational reasons. For colocalization analyses involving the protein data from the INTERVAL study [11], we evaluated three different region sizes—a 2 Mb region and a 400 kb region around the top SNP, as well

Table 1 Summary statistics for the 14 PCOS SNPs.

SNP	Full SNP name	Chr	Pos	EA	NEA	EAF	Odds ratio (95% CI)	*P*
rs2178575:G>A	NC_000002.11:g.213391766G>A	2	213391766	A	G	0.15	1.18 (1.13–1.23)	3.34e−14
rs11031005:T>C	NC_000011.9:g.30226356T>C	11	30226356	C	T	0.15	1.17 (1.12–1.23)	8.66e−13
rs804279:A>T	NC_000008.10:g.11623889A>T	8	11623889	A	T	0.26	1.14 (1.10–1.18)	3.76e−12
rs11225154:G>A	NC_000011.9:g.102043240G>A	11	102043240	A	G	0.09	1.20 (1.13–1.26)	5.44e−11
rs9696009:G>A	NC_000009.11:g.126619233G>A	9	126619233	A	G	0.07	1.22 (1.15–1.30)	7.96e−11
rs13164856:T>C	NC_000005.9:g.131813204T>C	5	131813204	T	C	0.73	1.13 (1.09–1.18)	1.45e−10
rs1784692:T>C	NC_000011.9:g.113949232T>C	11	113949232	T	C	0.82	1.15 (1.10–1.21)	1.88e−10
rs7563201:G>A	NC_000002.11:g.43561780G>A	2	43561780	G	A	0.55	1.11 (1.08–1.15)	3.68e−10
rs8043701:A>T	NC_000016.9:g.52375777A>T	16	52375777	T	A	0.18	1.14 (1.09–1.18)	9.61e−10
rs1795379:T>C	NC_000012.11:g.75941042T>C	12	75941042	C	T	0.76	1.12 (1.08–1.17)	1.81e−09
rs853854:T>A	NC_000020.10:g.31420757T>A	20	31420757	T	A	0.50	1.10 (1.07–1.14)	2.36e−09
rs2271194:A>T	NC_000012.11:g.56477694A>T	12	56477694	A	T	0.42	1.10 (1.07–1.14)	4.57e−09
rs10739076:A>C	NC_000009.11:g.5440589A>C	9	5440589	A	C	0.31	1.12 (1.07–1.16)	2.51e−08
rs7864171:G>A	NC_000009.11:g.97723266G>A	9	97723266	G	A	0.57	1.10 (1.06-1.13)	2.95e−08

Summary statistics for the PCOS-associated SNPs that were identified or replicated in Day et al. [6]. *Chr* chromosome, *CI* confidence interval, *EA* effect allele, *NEA* non-effect allele, *Pos* position, *SNP* single nucleotide polymorphism.

as the top SNP's "independent region" [16]. Independent regions were defined as the approximately independent regions of linkage disequilibrium in Europeans, as computed by Berisa et al. [17].

Coloc requires the assignment of prior probabilities for a SNP being associated with each trait (p1 and p2) and for a SNP being associated with both traits (p12). We set these prior probabilities to $p1 = 1 \times 10^{-4}$, $p2 = 1 \times 10^{-4}$, and $p12 = 1 \times 10^{-6}$, with the prior for p12 being more stringent than the default setting [10, 18].

Briefly, Coloc evaluates five different hypotheses. Hypothesis H_0, H_1 and H_2 correspond to situations without associated/causal SNPs in both the PCOS and the protein/gene dataset, H_3 to a situation where PCOS and the protein/gene have different associated/causal SNPs, and H_4 where PCOS and the protein/gene have evidence consistent with a shared associated/causal SNP, i.e., colocalization [10]. Since we performed colocalization as a hypothesis-generating approach, all analyses with a colocalization posterior probability $(PP) > 0.50$ were seen as having nominal evidence of colocalization and analyzed further. A colocalization PP just above >0.50 should be regarded with caution, and we set the threshold for strong evidence of colocalization at $PP \geq 0.75$. Power for detecting colocalization was computed as the sum of the PPs for hypothesis 3 (no colocalization) and hypothesis 4 (colocalization) [19].

Additional analyses

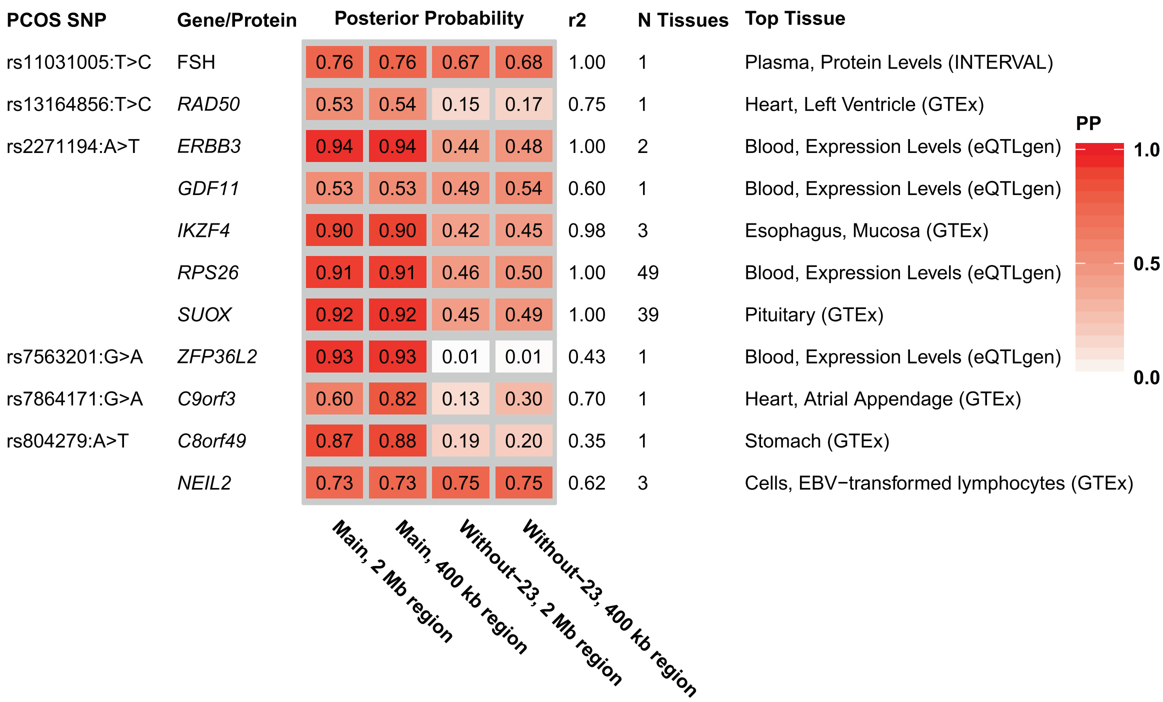
We followed up colocalizing regions with assessing phenome-wide association study (PheWAS) data for the top PCOS SNP using the Open Target Genetics platform [20]. The significance threshold for a PheWAS association on the Open Targets Genetics platform is approximately $P < 1 \times 10^{-5}$ (based on visual inspection of the plotted threshold [20]). We further corrected for the six SNPs we investigated and set the threshold to $P < 1.7 \times 10^{-6}$ (1×10^{-5} corrected for six SNPs).

We also performed a range of sensitivity analyses. Analyses were reconducted using the PCOS dataset where the 23andMe cohort had been excluded, to have roughly the same sample size for all SNPs. We also computed the PP of colocalization using HyPrColoc [16], a recently developed extension of Coloc [10] (see Supplement). For these analyses we used the larger region sizes of 2 Mb for all three QTL datasets, as well as the independent regions for INTERVAL (according to HyPrColoc recommendations [16]). Finally, we applied an experimental method, "interaction-Coloc" (see Supplement for detailed rationale, methods and results). In brief, we first identified other genes/proteins linked to the genes/proteins identified in the main analysis, based on data from protein–protein interaction experiments. We then performed colocalization analysis for these "linked" genes/proteins with PCOS risk under the assumption that evidence of colocalization for two linked genes/proteins strengthens the evidence for a role of affiliated pathways in PCOS pathophysiology. For this analysis, we used the default Coloc priors given the links to the genes/proteins identified in the main analysis.

Results

Colocalization highlights genes with a potential mediating role

We identified seven proteins and genes with strong evidence of colocalization $(PP \geq 0.75)$, including the protein FSH, and the genes *SUOX*, *ERBB3*, *IKZF4*, *RPS26*,



PCOS SNP	Gene/Protein	Posterior Probability				r2	N Tissues	Top Tissue
		Main, 2 Mb region	Main, 400 kb region	Without−23, 2 Mb region	Without−23, 400 kb region			
rs11031005:T>C	FSH	0.76	0.76	0.67	0.68	1.00	1	Plasma, Protein Levels (INTERVAL)
rs13164856:T>C	*RAD50*	0.53	0.54	0.15	0.17	0.75	1	Heart, Left Ventricle (GTEx)
rs2271194:A>T	*ERBB3*	0.94	0.94	0.44	0.48	1.00	2	Blood, Expression Levels (eQTLgen)
	GDF11	0.53	0.53	0.49	0.54	0.60	1	Blood, Expression Levels (eQTLgen)
	IKZF4	0.90	0.90	0.42	0.45	0.98	3	Esophagus, Mucosa (GTEx)
	RPS26	0.91	0.91	0.46	0.50	1.00	49	Blood, Expression Levels (eQTLgen)
	SUOX	0.92	0.92	0.45	0.49	1.00	39	Pituitary (GTEx)
rs7563201:G>A	*ZFP36L2*	0.93	0.93	0.01	0.01	0.43	1	Blood, Expression Levels (eQTLgen)
rs7864171:G>A	*C9orf3*	0.60	0.82	0.13	0.30	0.70	1	Heart, Atrial Appendage (GTEx)
rs804279:A>T	*C8orf49*	0.87	0.88	0.19	0.20	0.35	1	Stomach (GTEx)
	NEIL2	0.73	0.73	0.75	0.75	0.62	3	Cells, EBV−transformed lymphocytes (GTEx)

Fig. 2 Posterior probabilities for genes and proteins with any evidence of colocalization. In the main approach, we used a region size spanning 2 Mb, and 400 kb regions as a sensitivity analysis. In addition to the main PCOS dataset, we also performed colocalization analysis with a PCOS dataset where the 23andMe cohort had been excluded as a sensitivity analysis. r^2 is the linkage disequilibrium value between the top SNP in the main PCOS dataset and the top SNP in the tissue expression/protein dataset, using the 2 Mb region size. *N* tissues are the number of tissues where the colocalization PP > 0.5 in the main analysis. Only the results for the tissue with the highest posterior probability of colocalization in the main analysis are reported here (for full results and power calculations see Supplementary Tables 1–3). Gene-tissue combinations with a posterior probability of colocalization >0.50 were seen as having some evidence in favour of colocalization, whereas the threshold for strong evidence was set at ≥0.75. PCOS polycystic ovary syndrome, PP posterior probability, Without-23 the PCOS dataset where the 23andMe cohort had been excluded.

C8orf49, and *ZFP36L2* (Figs. 2–3 and Supplementary Tables 1–4; for a detailed description of genes see Supplement and Supplementary Figs. 1–10). In addition, four genes (*RAD50*, *GDF11*, *NEIL2*, and *C9orf3*) had nominal evidence of colocalization (PP > 0.50). Some of these genes and proteins, such as *RAD50*, had evidence of colocalization in only one tissue, whereas others, such as *RPS26* and *SUOX*, had evidence of colocalization in a large proportion of all tested tissues.

Regulatory annotations and associations with other traits

The colocalization results had highlighted circulating FSH as colocalizing at the rs11031005:T>C locus (PP = 0.76) (see Fig. 3). We found that the rs11031005:T>C C-allele was associated with both higher PCOS risk (OR 1.17, 95% CI 1.12–1.23, $P = 8.7 \times 10^{-13}$) and lower FSH levels (−0.166 standard deviations, SE = 0.035, $P = 2.0 \times 10^{-6}$) [11]. The locus was also associated with several traits related to female hormonal regulation in the PheWAS look-up, with the two traits showing the most robust associations being length of menstrual cycle ($P = 1.2 \times 10^{-42}$) and age at menopause ($P = 1.4 \times 10^{-15}$) (Supplementary Table 5; for results for the other loci see Supplementary Tables 6–10) [20, 21].

Other PCOS loci were colocalized with the expression levels of several genes. At the rs2271194:A>T locus, the results supported colocalization for four genes (*ERBB3* PP = 0.94, *IKZF4* PP = 0.90, *RPS26* PP = 0.91, and *SUOX* PP = 0.92), as well as nominal evidence (PP = 0.53) for *GDF11* (Fig. 2, Supplementary Figs. 2–6). The PheWAS of this variant highlighted associations with a range of different traits, including e.g. obesity, household income, and haematologic traits (Supplementary Table 6) [20]. Look-up of rs2271194:A>T and its proxies ($r^2 > 0.8$ in European-ancestry populations) in Haploreg [22] gave further evidence for a regulatory function in many different cell types, including e.g. immune cells, brain cells, and hepatocellular and cervical cancer cell lines.

Sensitivity analyses

We performed several sensitivity analyses. First, the number of SNPs included in an analysis can affect the PP of colocalization [18]. We therefore also conducted analyses

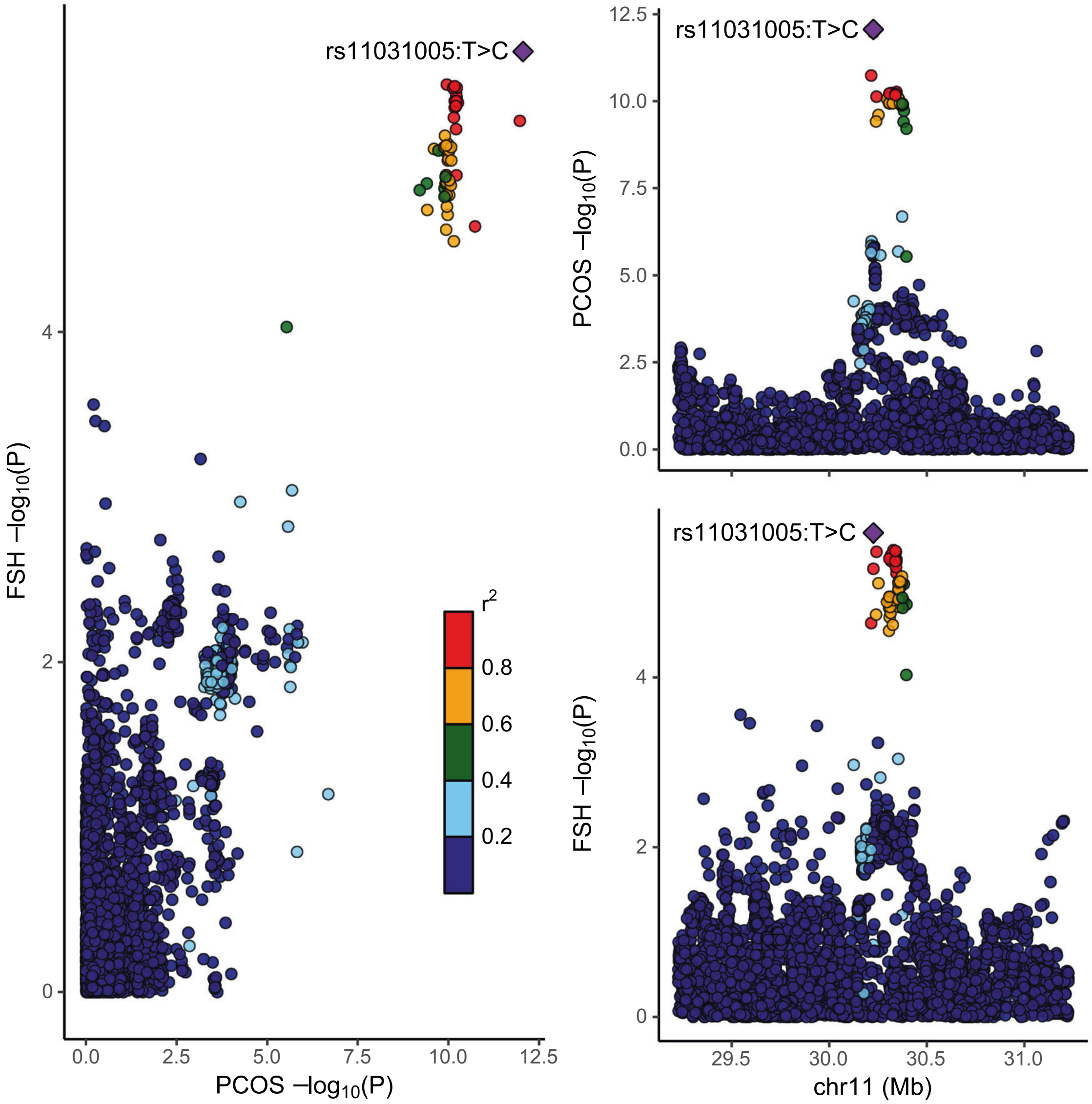


Fig. 3 Associations between genetic variants and PCOS risk, using the main PCOS dataset and 2 Mb region sizes for FSH protein levels in blood. In each plot, each dot is a genetic variant. The SNP with the most significant *P* value for PCOS is marked, with the other SNPs colour-coded according to linkage disequilibrium (r^2) in Europeans with the lead variant. SNPs with missing linkage disequilibrium information are also coded dark blue. In the left panel, −log10 *P* values for associations with PCOS risk are on the *x*-axes, and -log10 *P* values for associations with the protein levels on the *y*-axes. On the right panels, genomic positions are on the *x*-axes, and the *y*-axes show −log10 *P* values for PCOS on the upper panel and −log10 *P* values with the protein levels on the lower panel for the corresponding region. FSH follicle stimulating hormone; PCOS polycystic ovary syndrome; SNP single nucleotide polymorphism.

using a region size of 400 kb for all three e/pQTL datasets [10, 18], as well as approximately independent regions of linkage disequilibrium [17] in INTERVAL (performed in INTERVAL only since the other datasets did not provide genome-wide summary statistics) [16]. This sensitivity analysis supported the main findings; all SNP-gene/protein-tissue combinations with evidence of colocalization (PP > 0.5) in the main analysis had a PP > 0.5 regardless of region size using the main PCOS dataset (Fig. 4).

Second, Coloc uses SNP-associations to compute PPs [10], and association statistics are dependent on sample size. We therefore performed colocalization analyses using estimates from the PCOS dataset where the 23andMe cohort had been excluded, to obtain a similar sample size for all SNPs. This analysis generally had lower power (possible range 0–1, with a power >0.80 indicating strong power) to detect colocalization, and generally a correspondingly lower PP of colocalization (Supplementary Tables 1–3 and Fig. 4) [19].

Next, we performed colocalization analysis using the software HyPrColoc [16]. Results using HyPrColoc also provided evidence for colocalization for all gene/protein-

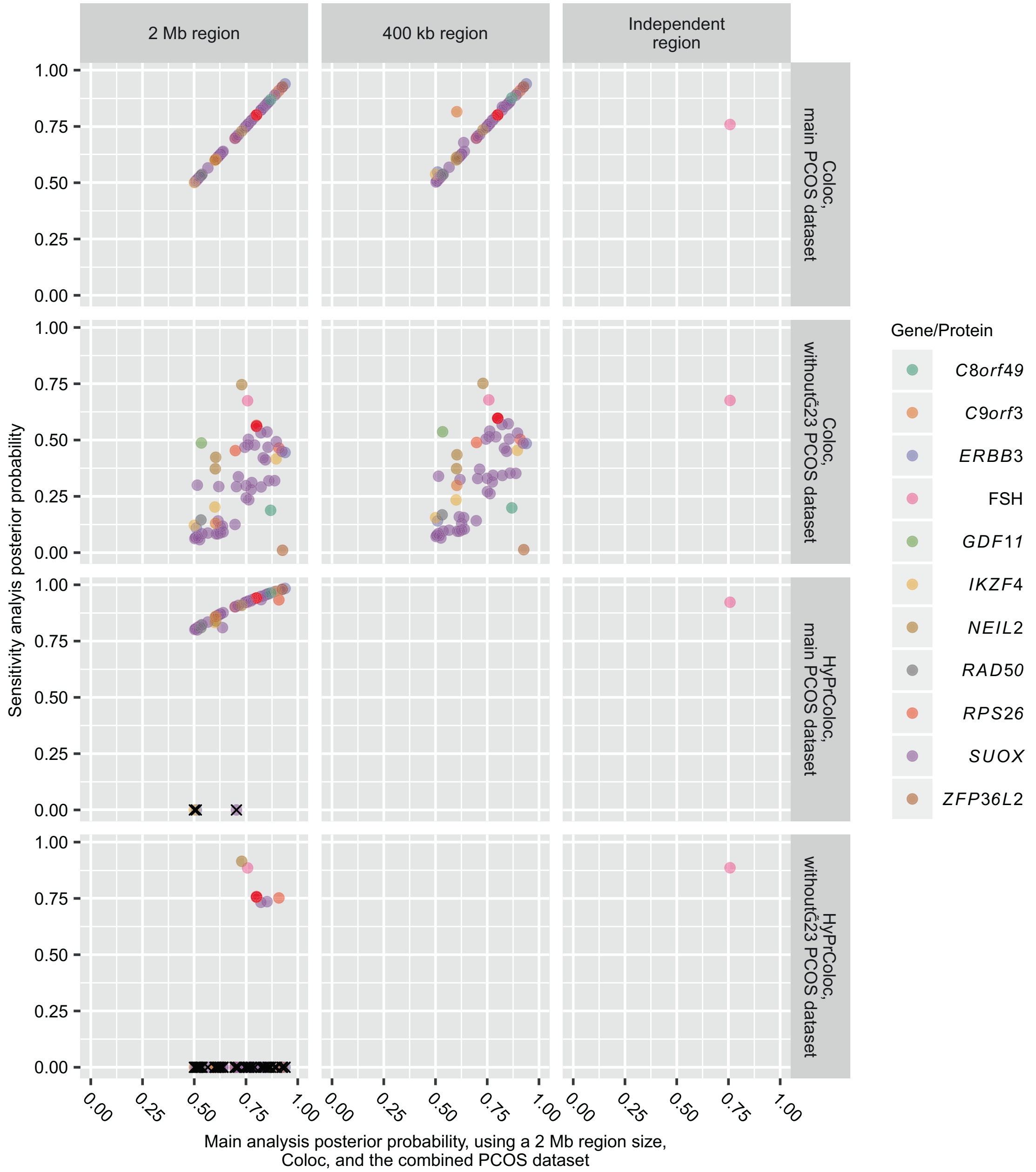


Fig. 4 Plot showing the posterior probabilities as computed in the sensitivity analyses compared to those obtained in the main analysis. Top left plot is a reference plot with the main results on both axes. Multiple genes on each plot represent different tissues and/or datasets. Method-region-dataset combinations missing on plot were not performed due to either limitations in data availability or recommendations for HyPrColoc settings (see "Materials and methods"). HyPrColoc analyses that are not colocalizing do not have a posterior probability computed; these were manually assigned a posterior probability of 0 and marked with "X" on the plot. Without-23, PCOS dataset where the 23andMe cohort had been excluded.

tissue combinations that were highlighted in the main analysis, except *ERBB3* in spinal cord, *IKZF4* in suprapubic skin, and *SUOX* in blood (but with evidence of colocalization between these genes and PCOS risk still present in other tissue types; Supplementary Tables 1, 2 and Fig. 4).

Finally, the experimental interaction-Coloc analyses provided suggestive evidence for two additional genes colocalizing with PCOS risk, highlighting their linked genes/proteins in the main analysis; FSH (as the linked gene coding for the FSH-receptor colocalized with PCOS risk at PP = 0.84) and *RAD50* (with its linked gene *UIMC1* colocalizing with PCOS risk at PP = 0.55) (see Supplement) [23].

Discussion

Using a Bayesian colocalization approach, our results highlight several genes and proteins that may have a role in

PCOS pathophysiology. We identify seven genes and proteins with strong evidence, and a further four genes and proteins with nominal evidence, of colocalization. Whereas potential mechanisms of action are unclear for some of the genes, half of the genes and proteins (FSH, *RAD50*, *ERBB3*, *RPS26*, and *ZFP36L2*) have links to the HPG axis and/or follicular development. We also find that a majority of the colocalizing genes are not the closest gene [6]. As the mediating genes for most of the genetic risk loci are still unclear [6], our results suggest genes with a higher likelihood of being involved in PCOS pathophysiology for functional follow-up.

The results highlighted FSH as a potential mediator at the rs11031005:T>C locus, corroborating the evidence of disruptions in gonadotropin signalling, specifically FSH and LH, contributing to PCOS pathophysiology [24]. FSH and LH are crucial hormones for follicular development and ovulation, and our PheWAS of the rs11031005:T>C locus showed an association with female reproductive traits [24]. LH and FSH share an alpha chain (encoded by *CGA* [25]), and disruption of *FSHB* has been associated with higher LH levels in humans [26]. It is thus possible that the PCOS association at the rs11031005:T>C locus may partly be caused by altered *FSHB* expression affecting LH levels. In addition, our results also implicated *ZFP36L2* at the rs7563201:G>A locus; another gene with links to gonadotropin signalling [27]. Female mice with a disruption in the *ZFP36L2* gene have disturbed oocyte maturation and ovulation, and its gene product has been implicated in regulation of LH-receptor levels [25, 27]. Previous studies have primarily suggested *THADA* to be the mediating gene at this locus [6, 28], but there was no evidence of *THADA* expression levels colocalizing with PCOS risk in any tissue in our study. We therefore suggest that *ZFP36L2* may be the mediating gene at the rs7563201:G>A locus and that the gene warrants further functional follow-up to evaluate a potential role in PCOS pathophysiology.

At the rs2271194:A>T locus, two of the colocalizing genes—*ERBB3* and *RPS26*—are likely candidates for mediating PCOS risk based on the literature, with both of them connected to the HPG axis. Gonadotropins have been shown to upregulate *ERBB3* expression and data suggest an important role in follicular development [29, 30]. The other gene, *RPS26*, has been implicated in DNA damage response and female fertility [25, 31, 32]. For example, oocyte-specific *Rps26*-knock-out mice have arrested oocyte growth, impaired follicle development, as well as poor response to gonadotropin stimulation [32]. Finally, we would like to highlight *RAD50* at the rs13164856:T>C locus. Female mice with disruptions in *RAD50* have reduced fertility [33] and the gene has been implicated in follicular development and oocyte development [34].

There are several strengths and limitations to our study. First, shared regulatory mechanisms between e.g. different genes and tissues can result in several gene/protein and tissue combinations colocalizing. However, it is unlikely that all of them are involved in disease development. The true mediating gene and tissue combination may not even have been investigated in the analyses, which may explain why some PCOS loci did not colocalize with any genes or proteins. In addition, we were surprised by the tissue types in which gene expression was colocalizing with PCOS. The disease is primarily thought of as being of hormonal and metabolic origin, wherefore we expected the results to highlight tissues types like the hypothalamus, the pituitary gland, and ovary and adipose tissue [2, 3]. Yet, many of the genes were colocalizing in seemingly unrelated tissue types, such as the heart (*RAD50* and *C9orf3)*. Still, for other genes, such as *RPS26*, even though the highest PP was achieved in the expression dataset with the highest sample size (whole blood in eQTLgen), the gene was also colocalizing with PCOS risk in ovary. Indeed, it is possible that the results in tissue types with many different cell types may have relatively lower PPs of colocalization if the disease is caused by changed expression in a single cell type in low abundance. Therefore, while colocalization can highlight genes, proteins, and tissues that are more likely to be involved in PCOS pathophysiology, results should be seen as hypothesis-generating rather than definitive evidence of a causal role. Second, if the causal SNP (or a proxy) is altering the coding sequence of a tested protein, it may become a false positive pQTL through changed aptamer binding. In our study, this could potentially result in rs11031005:T>C being a false pQTL for FSH, yet this is unlikely as the loci has previously been shown to associate with FSH levels using another protein quantification method [35]. Third, ancestral heterogeneity could potentially bias results due to different LD structure [10], even though all datasets primarily consisted of participants of European descent and we restricted our study to risk loci that were robust in a European-only GWAS [11, 13–15]. Fourth, the protein and expression datasets included both men and women [11, 13–15], whereas the PCOS GWAS [6] was performed in women only. If associations between genotypes and expression/protein levels differ between the sexes, it could bias results [36]. Fifth, Coloc also assumes a single causal variant per locus [10]. Accordingly, loci with multiple SNPs independently associated with either the disease or the intermediate trait may result in false negative colocalization results [10]. Still, Day et al. [6] did not report any multi-signal loci in the PCOS GWAS.

Finally, we performed a range of sensitivity analyses that largely supported the main results. We also presented an experimental method to nuance evidence, interaction-Coloc; however, we want to emphasize that the interaction-Coloc analyses should be interpreted with caution. The method has not been validated and is inherently limited by previously known interactions for each gene/protein, but we welcome future evaluations and developments of the method.

In summary, our results highlight potential mediating genes and proteins for almost a third of PCOS risk loci. Half of these genes and proteins have links to the HPG axis and follicular development, including the hormone FSH and the genes *ZFP36L2, ERBB3, RPS26*, and *RAD50*. In combination with previous studies that have indicated these genes as being involved in physiologic processes associated with PCOS, these genes may be of particular interest for further functional follow-up to assess if they have role in the disease development.

Data availability

Results are available in full in the supplementary tables and accessible at https://doi.org/10.6084/m9.figshare.13655444 (see Supplement for access to all other datasets used).

Acknowledgements We thank the PCOS Consortium. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. We thank the UK Biobank (http://www.ukbiobank.ac.uk/; application id 11867).

Funding This work was supported by funding from the Oxford Medical Research Council Doctoral Training Partnership (Oxford MRC DTP) and the Nuffield Department of Clinical Medicine, University of Oxford [17/18_MSD_1108275], to JCC, by funding from the Rhodes Trust, Clarendon Fund and the Medical Sciences Doctoral Training Centre, University of Oxford, to JB, by funding from the Medical Research Council to the unit that MVH works in, by a British Heart Foundation Intermediate Clinical Research Fellowship [FS/18/23/33512] and funding from the National Institute for Health Research Oxford Biomedical Research Centre to MVH, and by funding from the Li Ka Shing Foundation; the NIHR Biomedical Research Centre, Oxford; and NIH [5P50HD028138-27] to CML. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award [203141/Z/16/Z]. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Compliance with ethical standards

Conflict of interest JB has served as a consultant to the Bill and Melinda Gates Foundation Strategic Investment Fund. MVH has collaborated with Boehringer Ingelheim in research, and in accordance with the policy of the Clinical Trial Service Unit and Epidemiological Studies Unit (University of Oxford), did not accept any personal payment. CML has collaborated with Novo Nordisk and Bayer in research, and in accordance with the University of Oxford agreement, did not accept any personal payment.

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

References

1. Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Hum Reprod. 2016;31:2841–55.
2. Teede H, Deeks A, Moran L. Polycystic ovary syndrome: a complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan. BMC Med. 2010;8(Jun):41.
3. Azziz R, Carmina E, Chen Z, Dunaif A, Laven JSE, Legro RS, et al. Polycystic ovary syndrome. Nat Rev Dis Prim. 2016;2:16057.
4. Vink JM, Sadrzadeh S, Lambalk CB, Boomsma DI. Heritability of polycystic ovary syndrome in a Dutch twin-family study. J Clin Endocrinol Metab. 2006;91:2100–4
5. Hayes MG, Urbanek M, Ehrmann DA, Armstrong LL, Lee JY, Sisk R, et al. Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. Nat Commun. 2015;6:7502.
6. Day F, Karaderi T, Jones MR, Meun C, He C, Drong A, et al. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. PLoS Genet. 2018;14:e1007813.
7. Shi Y, Zhao H, Shi Y, Cao Y, Yang D, Li Z, et al. Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. Nat Genet. 2012;44:1020–5.
8. Chen Z-J, Zhao H, He L, Shi Y, Qin Y, Shi Y, et al. Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. Nat Genet. 2011;43:55–9.
9. Day FR, Hinds DA, Tung JY, Stolk L, Styrkarsdottir U, Saxena R, et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. Nat Commun. 2015;6(Sep):8464.
10. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:e1004383.
11. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558:73–9.
12. Di Angelantonio E, Thompson SG, Kaptoge S, Moore C, Walker M, Armitage J, et al. Efficiency and safety of varying the

frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. Lancet. 2017;390:2360–71.

13. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.
14. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. 2018. https://www.biorxiv.org/content/10.1101/447367v1.
15. GTEx Portal [Internet]. 2019. https://gtexportal.org/home/.
16. Foley CN, Staley JR, Breen PG, Sun BB, Kirk PDW, Burgess S, et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. 2019. https://www.biorxiv.org/content/10.1101/592238v1.
17. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 2016;32:283–5.
18. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. 2020;16: 838946.
19. Çalışkan M, Manduchi E, Rao HS, Segert JA, Beltrame MH, Trizzino M, et al. Genetic and epigenetic fine mapping of complex trait associated loci in the human liver. Am J Hum Genet. 2019;105:89–107.
20. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, et al. Open targets platform: new developments and updates two years on. Nucleic Acids Res. 2019;47:D1056–65.
21. UK Biobank—Neale lab [Internet]. Neale lab; 2019. http://www.nealelab.is/uk-biobank/
22. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40: D930–4.
23. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46:D649–55.
24. Li Y, Chen C, Ma Y, Xiao J, Luo G, Li Y, et al. Multi-system reproductive metabolic disorder: significance for the pathogenesis and therapy of polycystic ovary syndrome (PCOS). Life Sci. 2019;228:167–75.
25. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15.
26. Misgar RA, Wani AI, Bankura B, Bashir MI, Roy A, Das MFSH. β-subunit mutations in two sisters: the first report from the Indian sub-continent and review of previous cases. Gynecol Endocrinol. 2019;35:290–3.
27. Ball CB, Rodriguez KF, Stumpo DJ, Ribeiro-Neto F, Korach KS, Blackshear PJ, et al. The RNA-binding protein, ZFP36L2, influences ovulation and oocyte maturation. PLoS ONE. 2014;9: e97324.
28. Pau CT, Mosbruger T, Saxena R, Welt CK. Phenotype and tissue expression as a function of genetic risk in polycystic ovary syndrome. PLoS ONE. 2017;12:e0168870.
29. Chowdhury I, Branch A, Mehrabi S, Ford BD, Thompson WE. Gonadotropin-dependent neuregulin-1 signaling regulates female rat ovarian granulosa cell survival. Endocrinology. 2017;158: 3647–60.
30. Mukherjee A, Roy SK. Expression of ErbB3-binding protein-1 (EBP1) during primordial follicle formation: role of estradiol-17ß. PLoS ONE. 2013;8:e67068.
31. Cui D, Li L, Lou H, Sun H, Ngai S-M, Shao G, et al. The ribosomal protein S26 regulates p53 activity in response to DNA damage. Oncogene. 2014;33:2225–35.
32. Liu X-M, Yan M-Q, Ji S-Y, Sha Q-Q, Huang T, Zhao H, et al. Loss of oocyte Rps26 in mice arrests oocyte growth and causes premature ovarian failure. Cell Death Dis. 2018;9:1144.
33. Roset R, Inagaki A, Hohl M, Brenet F, Lafrance-Vanasse J, Lange J, et al. The Rad50 hook domain regulates DNA damage signaling and tumorigenesis. Genes Dev. 2014;28:451–62.
34. Inagaki A, Roset R, Petrini JHJ. Functions of the MRE11 complex in the development and maintenance of oocytes. Chromosoma. 2016;125:151–62.
35. Mbarek H, Steinberg S, Nyholt DR, Gordon SD, Miller MB, McRae AF, et al. Identification of common genetic variants influencing spontaneous dizygotic twinning and female fertility. Am J Hum Genet. 2016;98:898–908.
36. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. Science. 2020;369:eaba3066.