Colocalization highlights genes in hypothalamic–pituitary–gonadal axis as potentially mediating polycystic ovary syndrome risk

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

Affiliations

1. Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, OX3 7LF, UK
2. Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
3. NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, OX3 9DU, UK
4. Medical Research Council Population Health Research Unit at the University of Oxford, Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, 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, OX3 7LF, UK
6. Program in Medical and Population Genetics, Broad Institute, Cambridge, 02142, Massachusetts, USA
7. Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, OX3 9DU, UK

Corresponding author:
Name: Dr. Jenny C Censin
E-mail: jenny.censin@ndm.ox.ac.uk
Telephone: +44(0)1865 287835

Abstract

Polycystic ovary syndrome (PCOS) is a common disease in women with consequences for reproductive, metabolic and psychological health. Women with PCOS have disrupted signalling in the hypothalamic-pituitary-gonadal axis and studies have indicated that the disease has a large genetic component. While a recent genome-wide association study of PCOS performed in up to 10,074 cases and 103,164 controls of European decent identified 14 PCOS-associated regions, much of the disease pathophysiology remains unclear.

Here, we use a Bayesian colocalization approach to highlight genes that may have a potential role in PCOS pathophysiology and thus are of particular interest for further functional follow-up. We evaluated the posterior probabilities of shared causal variants between PCOS genetic risk loci and intermediate cellular phenotypes in one protein and two expression quantitative trait locus datasets, respectively. Sample sizes ranged from 80 to 31,684. In total, we identified seven proteins or genes with evidence of a shared causal variant for almost a third of PCOS signals, including follicle stimulating hormone (FSH) and the genes *ERBB3*, *IKZF4*, *RPS26*, *SUOX*, *ZFP36L2*, and *C8orf49*. Several of these genes and proteins have been implicated in the hypothalamic-pituitary-gonadal signalling pathway.

In summary, our results suggest potential effector proteins and genes for PCOS association signals. This highlights genes 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 6-10% of women of reproductive age (1). The disease has a heterogeneous clinical presentation (2–4), with consequences for reproductive, metabolic, and psychological health (2,3). Commonly, diagnosis is based on the Rotterdam criteria, which requires two out of three of 1) oligo- or anovulation, 2) signs of hyperandrogenism (clinical or biochemical), and 3) polycystic ovarian morphology, as well as exclusion of other diagnoses (3,4).

PCOS pathophysiology is still largely unclear (2), although one mechanism may be disrupted gonadotropin signalling that disturbs normal follicular development and ovulation (3). In healthy women of reproductive age, the pituitary gland secretes the gonadotropins luteinizing hormone (LH) and follicle-stimulating hormone (FSH) in response to pulsatile secretion of gonadotropin releasing hormone (GnRH) (5,6). These GnRH pulses are more frequent in women with PCOS (3,7). This changed secretion pattern causes an imbalance between LH and FSH, and a higher LH/FSH ratio (3,7–11), which may contribute to e.g. hyperandrogenism and disturbances in follicular maturation and ovulation (3,12). Other possible contributing factors that have been suggested include for example insulin resistance and inflammation (3,8). There is also evidence for a strong genetic component, with genetic factors suggested to explain 66% of the disease variance (13). Previous genome-wide association studies (GWAS) have highlighted risk loci 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) (3,14–18). However, for most PCOS-associated loci the mediating genes and their functional effects remain to be identified and/or confirmed (17,18).

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 (19). Therefore, to improve understanding of PCOS pathophysiology, we investigated the evidence of colocalization between 14 PCOS-associated loci identified in a recent GWAS in Europeans (18) and 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.

Results

Colocalization highlights genes with a potential mediating role

We extracted 14 PCOS risk loci from a recent GWAS of up to 10,074 cases and 103,164 controls of European ancestry (Fig 1, Table 1) (18). We assessed the evidence for colocalization (19) between these loci and pQTL data from INTERVAL and eQTL data from the Genotype-Tissue Expression project (GTEx) and eQTLgen (20–23). PCOS summary statistics based on the full sample (including up to 10,074 cases and 103,164 controls) were only available for the 10,000 most robustly associated single nucleotide polymorphisms (SNPs). For the other SNPs, summary estimates were based on analyses excluding the 23andMe cohort (up to 4,890 cases and 20,405 controls). We therefore used summary statistics based on a combined version of the available data, with preference given to SNP statistics including the 23andMe cohort (denoted "Combined" dataset) and a window size spanning 2 Mb around the most robustly associated PCOS SNP based on P-value (19,24,25).

We identified seven proteins and genes with evidence of colocalization (posterior probability (PP) ≥ 0.75), including the protein FSH, and the genes *SUOX*, *ERBB3*, *IKZF4*, *RPS26*, *C8orf49,* and *ZFP36L2* (26,27). In addition, four genes (*RAD50*, *GDF11*, *NEIL2*, *C9orf3*) showed nominal evidence of colocalization (PP > 0.50) (Fig 2 and Supplementary Tables 1-3; for a detailed description of genes not discussed below see the Supplement and Supplementary Figures 1-9). 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 (Supplementary Tables 1-2). However, tissue sample size seemed to influence the evidence of colocalization and a large proportion of the colocalizing gene-tissue combinations used blood expression data from eQTLgen (sample size up to 31,684), but many of these analyses did not surpass the colocalization threshold using the smaller GTEx blood expression dataset (sample size up to 369) (Supplementary Table 2).

Interaction-coloc analyses

Several genes and proteins had evidence of colocalization in some loci, which might be due to shared regulatory mechanisms (Fig 2). In addition, identification of true causal genes/proteins is dependent on tissue- and timepoint relevant QTL datasets, an inherent problem in colocalization analyses (19,28). We therefore suggest an exploratory approach, an "interaction-coloc"-analysis, to further query the evidence for each colocalizing gene/protein.

We reasoned that we could nuance the evidence of PCOS involvement for the colocalizing genes/proteins by assessing if other genes/proteins known to interact with them also had evidence of colocalization (Supplementary Figure 10). Specifically, if there is evidence of colocalization with PCOS for two genes/proteins known to interact with each other, this should in theory increase the likelihood of them and their affiliated pathway mediating the relationship with the disease (Fig 3). We therefore extracted protein-protein interaction data from Reactome (29) for the proteins and genes colocalizing (PP > 0.50) in our main analysis.

We then performed colocalization for these "interactors" (including both their genes and any protein products) with PCOS risk. Using this approach, we found evidence of colocalization for *FSHR* expression (interacting with FSH), and nominal evidence of colocalization for *RNF41* (interacting with *ERBB3)* and *UIMC1* (interacting with *RAD50*) expression (Fig 3-5 and Supplementary Table 4) (29).

Regulatory annotations and associations with other traits

Next, we analyzed phenome-wide associations (PheWAS) of the PCOS loci by characterizing their associations with other traits using public data (Supplementary Tables 5-10) (30). We also assessed regulatory evidence using Haploreg (31).

The colocalization results had highlighted circulating FSH as colocalizing at the rs11031005 locus (PP=0.76). We found that the rs11031005 C-allele was associated with both higher PCOS risk (OR 1.17, 95% CI 1.12-1.23, $P=8.7\times10^{-13}$) and lower FSH-levels (-0.166 standard deviations, standard error = 0.035, $P = 2.0\times10^{-6}$). In addition, rs11031005 was 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 ($P=1.2\times10^{-42}$) and age at menopause ($P=1.4\times10^{-15}$) (Supplementary Table 5) (30,32).

Other PCOS loci seemed more pleiotropic – at the rs2271194 locus, the results supported colocalization for four genes (*ERBB3, IKZF4, RPS26*, and *SUOX*), as well as nominal evidence for *GDF11* (Fig 2, Supplementary Figures 1-5). The PheWAS of this locus highlighted associations with a range of different traits, including e.g. obesity, hematologic, and social traits (Supplementary Table 6) (30). Look-up of the PCOS SNP and its proxies (r2>0.8 in Europeans) in Haploreg (31) gave further evidence for a regulatory function acting in a many different cell-types, including the presence of enhancer and promoter marks, location in DNase hypersensitivity sites, and binding of e.g. RNA polymerase II and transcription factors (31,33–37).

Sensitivity analyses and choice of priors

We performed several sensitivity analyses. Firstly, coloc uses SNP-associations to compute posterior probabilites (19), and association statistics are dependent on sample size. However, summary statistics for the entire PCOS sample (up to 10,074 cases and 103,164 controls) was only publicly available for the 10,000 most robustly associated SNPs. In contrast, full GWAS summary statistics were available for up to 4,890 cases and 20,405 controls (data based on analyses excluding the 23andMe cohort, denoted "Without-23" dataset). To ascertain similar sample sizes for all SNPs regardless of the strength of association, we therefore also performed colocalization using only the Without-23 PCOS dataset. Colocalization analyses using the Without-23 PCOS dataset generally had lower power (possible range 0-1, with a power >0.80 indicating strong power to determine colocalization) to detect colocalization, and generally a correspondingly lower PP of colocalization (Supplementary table 1-3) (24). For example, there was strong power and evidence for colocalization (power = 1.00 and PP = 0.93) between PCOS risk and expression of *ZFP36L2* at the rs7563201 locus using the Combined PCOS dataset, but considerably less power and PP using the Without-23 dataset (power = 0.28 and PP = 0.01).

Secondly, the number of SNPs included in the analysis can affect the PP of colocalization (25). We therefore also conducted analyses using a region size of +/- 200 kb for all three e/pQTL datasets (19,25), as well as approximately independent regions of linkage disequilibrium (38) in INTERVAL (performed in INTERVAL only since the other datasets did not provide genome-wide summary statistics) (39). In general, there was good consistency between all three window sizes (Figure 2, Supplementary tables 1-2).

Thirdly, we performed colocalization analysis using the software HyPrColoc to minimize the risk of software or coding errors (39). These results supported the main results (Supplementary Tables 1-2).

Finally, coloc requires specification of prior probabilities for both the likelihood that a SNP is associated with each trait (p1 and p2, respectively) and for the likelihood that a SNP is associated with both traits (p12). A previous study has shown that p1 = p2 = 1×10^{-4} is a reasonable setting in most scenarios, but the choice of p12 is more complex (25). We therefore decided to set p12 = 1×10^{-6} in the main analysis, corresponding to a stricter p12 than suggested (25) and stricter than the standard setting (19). For the interaction-coloc analyses, we used the standard coloc setting of p12 = 1×10^{-5}, given a hypothesized greater likelihood of colocalization in these analyses, as well as p12 = 1×10^{-6} as a sensitivity analysis (Figure 3, Supplementary Table 4).

Discussion

Our results highlight several genes and proteins that may have a role in PCOS development by using a Bayesian colocalization approach. We identify seven genes and proteins with strong and a further four genes and proteins with some evidence of colocalization, respectively. Several of these genes and proteins have links to the HPG axis and follicular development, further highlighting disruption of these processes as likely pathophysiological mechanisms in the disease. As the mediating genes for most of the genetic risk loci are still unclear (17,18), our results offer a potential to focus further functional follow-up studies on genes with a higher likelihood of being involved in PCOS pathophysiology.

Our results highlighted FSH (its beta-chain encoded by *FSHB*, located approximately 26 Kb from rs11031005 (33,40)) as a potential mediator at the rs11031005 locus. The results also implicated *ZFP36L2* at the rs7563201 locus. 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 (33,41). There is previous evidence for disruptions in gonadotropin signalling, specifically FSH and LH, being involved in PCOS pathophysiology (8,42). FSH and LH are crucial hormones for follicular development and ovulation (5,6,8), The two hormones share an alpha chain (encoded by *CGA* (33)), and disruption of *FSHB* has been associated with higher LH levels in both humans and mice (43,44). SNPs in the *FSHB* region have also been associated with levels of both LH and LH/FSH (45–47). It is thus possible that the PCOS association at the rs11031005 locus may partly be caused by altered *FSHB* expression affecting LH-levels, although the interaction-coloc evidence for involvement of the FSH-receptor also implies a direct role of FSH in the disease.

At the rs2271194 (at position 12:56477694 in GRCh 37 (48)) 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 (for a literature review of the other genes see Supplement). The gene *ERBB3* encodes a tyrosine-protein kinase receptor (Receptor tyrosine-protein kinase erbB-3) (33). *ERBB3* expression levels in granulosa cells vary over the estrous cycle in rats, with gonadotropins upregulating *ERBB3* expression and data suggesting an important role in follicular development (49,50). There was evidence of colocalization for *RNF41* (involved in regulation of Receptor tyrosine-protein kinase erbB-3 protein levels (33,51)) in the interaction-coloc analyses, but as the genes *ERBB3* and *RNF41* are in the same locus this cannot be regarded as additional evidence for *ERBB3*. The other likely candidate at the locus, *RPS26*, has been implicated in DNA damage response and female fertility (33,52,53). For example, oocyte-specific *Rps26*-knock-out mice have arrested oocyte growth, impaired follicle development, as well as poor response to gonadotropin stimulation (53), hence also implicating the HPG axis.

Another promising gene candidate is *RAD50*. The gene encodes DNA repair protein RAD50 (33), which together with MRE11 and another protein forms part of the MRE11 complex, which is involved in DNA damage response processes (54–59). Female mice with disruptions in the *Mre11* or *Rad50* genes have reduced fertility (55,59). It may be that the MRE complex affects oocyte elimination in the presence of DNA damage and thereby plays a part in follicular development and oocyte development (57). Even though our results only provided nominal evidence for involvement of *RAD50* in PCOS development, the evidence was strengthened by the interaction-coloc analyses that also gave nominal colocalization evidence for another gene (*UIMC1*) implicated in the same DNA repair processes as the MRE11 complex (33,60).

Importantly, 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 – indeed, the true mediating gene and tissue combination may not even have been investigated in the analyses. Therefore, while colocalization can highlight genes and proteins 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.

Whereas some genes exhibit more tissue-specific effects, others have similar effects in a range of tissues (20,61,62). We assessed colocalization using datasets including a wide range of tissue types (e.g. GTEx (20)) and datasets with large sample sizes (e.g. eQTLgen (21)), which should increase the chance of identifying colocalizing genes and proteins.

We also investigated if genes/proteins that may interact with the originally identified genes/proteins provided additional evidence of their involvement in PCOS pathophysiology. This is a novel approach, but whereas it in theory should provide a more independent confirmation of a gene/protein being involved in the disease, the results should be interpreted with caution. Some of the originally identified genes and proteins had many known interactors and others none, resulting in differing possibilities to identify colocalization. In addition, even though the interaction-coloc analysis delivered plausible results and presents a possible extension of colocalization methodology, it has not been validated.

There are also caveats with our study. Firstly, if the causal SNP (or a proxy) is altering the coding sequence of a tested protein, it may cause false positive results through changed aptamer binding. Secondly, ancestral heterogeneity could potentially bias results due to different LD-structure (19), even though all datasets primarily consisted of participants of European decent (20–22,63). Thirdly, the protein and expression datasets included both men and women (20–22,63), whereas the PCOS GWAS (18) was performed in women only. If associations between genotypes and expression/protein levels differ between the sexes, it could bias results. Finally, *coloc* assumes a single causal variant per locus (19). Accordingly, loci with multiple SNPs independently associated with either the disease or the intermediate trait risk may result in false negative results (19).

Conclusion

In summary, our results highlight potential mediating genes and proteins for almost a third of PCOS risk loci. Several 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.

Materials and Methods

Data on Polycystic ovary syndrome

We obtained GWAS summary statistics for PCOS from Day *et al.* (18). In the study, 14 genome-wide significant loci were identified in up to 10,074 cases and 103,164 controls of European ancestry. Public summary statistics were available for the full sample for the 10,000 most robustly associated SNPs, and for all SNPs from analyses excluding the 23andMe cohort (resulting in a sample size of up to 4,890 cases and 20,405 controls). To maximize power, we used a combined version of these two datasets as our main dataset (denoted "Combined" dataset), with preference given to data from the top

10,000 SNPs dataset. As a sensitivity analysis, we also performed all analyses using the all-SNP dataset where the 23andMe cohort had been excluded (denoted "Without-23" dataset), to have roughly the same sample size for all SNPs. We then excluded SNPs found to be duplicated by position, missing relevant data, or indels. Genetic variants were matched to rsIDs using the file "All_20180423.vcf.gz", available at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/ (48).

Quantitative trait loci datasets

We used publicly available protein and expression genetic association data from the INTERVAL study (22,23), the GTEx consortium (20), and the eQTLgen consortium (21).

pQTL data were taken from the INTERVAL study, which had performed GWASs for 2,994 unique plasma proteins (3,283 measured aptamers) in 3,301 blood donors of European ancestry (22). For GTEx, we used data from version 7, which contains cis-eQTL data for between 80-491 samples in 48 different tissues (20,63). Expression had been measured post-mortem, with ~85% of donors being of European ("White") ancestry in the whole sample (63). Lastly, the eQTLgen Consortium had performed cis- and trans-eQTL analysis in up to 31,684 individuals, predominantly of European ancestry (21). Both cis-associations, containing SNPs within 1 Mb from the center of the gene, and trans-assocations, containing SNPs over 5 Mb from the center of the gene, are publicly available (21). For all these datasets, we then excluded SNPs that were duplicated by position, missing relevant data, or indels.

Colocalization analyses

Coloc

We applied coloc (19), a Bayesian test for colocalization to evaluate the probability of a shared causal signal between each PCOS hit and each p/eQTL. 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. Gene positions and transcription start sites were determined using GRCh 37 and the biomaRt R package where needed (64,65).

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 (20,21). We therefore only included genes and proteins with a transcription start site or centre of gene +/- 800 kb of the 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 further analysed two different region sizes in GTEx and eQTLgen – the entire 2 Mb cis-region available in these datasets in the main analysis and +/- 200 kb of the top SNP as a sensitivity analysis. For GTEx, we only performed the analysis if the top SNP had been analyzed for computational reasons. For INTERVAL (22), we evaluated three different region sizes – +/- 1 Mb and +/- 200 kb of the top SNP, as well as the top SNP's "independent region" (19,24,39,66). Independent regions were defined as the approximately independent regions of linkage disequilibrium in Europeans, as computed by Berisa *et al.* (38).

We set the prior probabilities to p1 = 1×10^{-4}, p2 = 10×10^{-4}, and p12 = 1×10^{-6} (more stringent than default) (19,25). Minor allele frequencies from the PCOS dataset were used in all coloc analyses. For the number of cases and total sample size, we supplied 10,074 and 113,238 for the Combined dataset (albeit this would be smaller for the SNPs that were not in the top 10,000 SNPs dataset) and 4,890 and 25,295 for the sensitivity-analysis using the Without-23 PCOS dataset (the dataset with estimates

based on approximately equal sample sizes for each SNP). For INTERVAL and GTEx, we used the sample size reported for each tissue and dataset. For eQTLgen we supplied the average sample size for the included SNPs. As the eQTLgen summary statistics did not include effect estimates and standard errors, we let the coloc.abf() function approximate effect estimates from the P-values for this dataset (19).

Briefly, coloc evaluates the PP for five different hypotheses, which in this study correspond to:

- H_0: No causal association with either PCOS or the protein/gene
- H_1: Causal association with PCOS but not the protein/gene
- H_2: Causal association with the protein/gene but not PCOS
- H_3: Causal associations with both PCOS and the protein/gene, but with two separate causal SNPs
- H_4: Causal association with both PCOS and the protein/gene, with a shared causal SNP (19)

Studies use different thresholds to evaluate whether there is evidence of a shared causal variant (H_4), but the PP of colocalization can be seen as a numerical value of the certainty of the result (19,26,66–68). Since we performed colocalization as a hypothesis-generating approach, all analyses with a PP >0.50 were seen as having nominal evidence of colocalization and analyzed further. A PP just above >0.50 should be regarded with caution (19), and we set the threshold for strong evidence of colocalization at PP ≥0.75 (26,27). We also computed the power for detecting colocalization for the results with any evidence of colocalization as the sum of the PPs for hypothesis 3 (no colocalization) and hypothesis 4 (colocalization) (24).

HyPrColoc

To ascertain robustness, we also computed the posterior probability of colocalization using HyPrColoc (39), a recently developed extension of coloc (19). We used a similar approach as for coloc, but only using the larger region sizes of 1 Mb for all three QTL datasets, as well as the independent regions for INTERVAL. Default priors (prior.1 = 1×10^{-4} and prior.2 = 0.98) were used, whereas we set both the regional and alignment probability thresholds to 0.8 (more stringent than default) (39). As eQTLgen only provided Z-scores, we estimated betas and SEs using the formulas:

$$\hat{b} = z / \sqrt{2p(1-p)(n+z^2)}$$
$$SE = 1 / \sqrt{2p(1-p)(n+z^2)}$$

Where z is the Z-score, p is the minor allele frequency in the eQTLgen dataset and n is the sample size (21,69).

Protein-protein interaction follow-up analyses using coloc

To identify genes/proteins that interact with the primarily identified genes/proteins, we downloaded data with protein-protein interactions in humans (available at https://reactome.org/download/current/interactors/reactome.homo_sapiens.interactions.tab-delimited.txt) from Reactome (29). Genes listed as part of proteins interacting with any of our associated genes, and with ensembl gene identifiers, were extracted. For FSH, we only extracted interactions listed for the beta subunit (encoded by *FSHB*), since the alpha subunit (encoded by *CGA*) forms part of other hormones as well (70). We then extracted information of uniprot-identifiers, ensembl gene identifiers, gene positions and transcription start sites using GRCh 37 and the biomaRt R package to map between different datasets (64,65). We only included transcripts listed with a

numeric autosomal chromosome and with information available in biomaRt. We included SNPs within +/- 1 Mb from the average transcription start site in the colocalization analyses using INTERVAL dataset (22), for the other datasets all available SNPs were used. We then applied coloc (19), using +/- 1 Mb region sizes. As the genes and proteins in the interaction-coloc analyses already had evidence of protein-protein interactions with the genes identified in the main analyses, we considered the prior probability of colocalization higher and thus used a more lenient prior probability of colocalization than in the main analysis (p12 = 1×10^{-5}, which is the same as the default setting in coloc (19)).

PheWAS and in-silico investigations

We followed up colocalizing regions with assessing PheWAS data for the top PCOS SNP using the Open Target Genetics platform (30). The significance threshold for a PheWAS association on the Open Targets Genetics platform is approximately P<1×10^{-5} (based on visual inspection of the plotted threshold, which corresponds to a Bonferroni-correction of the number of investigated traits (30)). We further corrected for the six SNPs we investigated and set the threshold to P<1.7×10^{-6} (1×10^{-5} corrected for six SNPs). We also investigated the evidence for regulatory mechanisms for the colocalizing PCOS regions and the top SNP using Haploreg v4.1 (31).

Software

Analyses and plots were done using R versions 3.5.1 and 3.4.3 (71), bash version 4.1.2(2) (72), awk (73), and R packages coloc (19), hyprcoloc (39), LocusCompareR (74), tidyr (75), data.table (76), plyr (77), devtools (78), and ggplot2 (79).

Data availability

The PCOS GWAS summary statistics are available at https://www.repository.cam.ac.uk/handle/1810/283491 (18). The GTEx version 7 data are available at https://gtexportal.org/ (20). Effect allele frequencies for GTEx were taken from the files "GTEx_V7_cis_eqtl_summary.tar.gz (hg19)" (downloadable at http://cnsgenomics.com/software/smr/#DataResource). Independent regions as per Berisa *et al.* (38) can be accessed at https://bitbucket.org/nygcresearch/ldetect-data/downloads/. The summary statistics from the INTERVAL study is available at https://www.phpc.cam.ac.uk/ceu/proteins/ (22). Data from the eQTLgen consortia can be accessed at https://molgenis26.gcc.rug.nl/downloads/eqtlgen/cis-eqtl (21). Human protein-protein interactions from Reactome pathways is available at https://reactome.org/download/current/interactors/reactome.homo_sapiens.interactions.tab-delimited.txt. The PheWAS data were downloaded from the Open Targets Genetics website https://genetics.opentargets.org (30). In-silico functional investigations were done using Haploreg v4.1 at https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php (31). Individual-level data from UK Biobank cannot be shared publicly because of confidentiality but is available from the UK Biobank (https://www.ukbiobank.ac.uk/) for researchers who meet the criteria for access to confidential data. The UK Biobank has a Research Tissue Bank approval (Research Ethics Committee reference 16/NW/0274, this study's 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; WT-SSI/John Fell funds; the NIHR Biomedical Research Centre, Oxford; Widenlife; 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.

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).

Conflict of Interest Statement

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.

Figure Legends

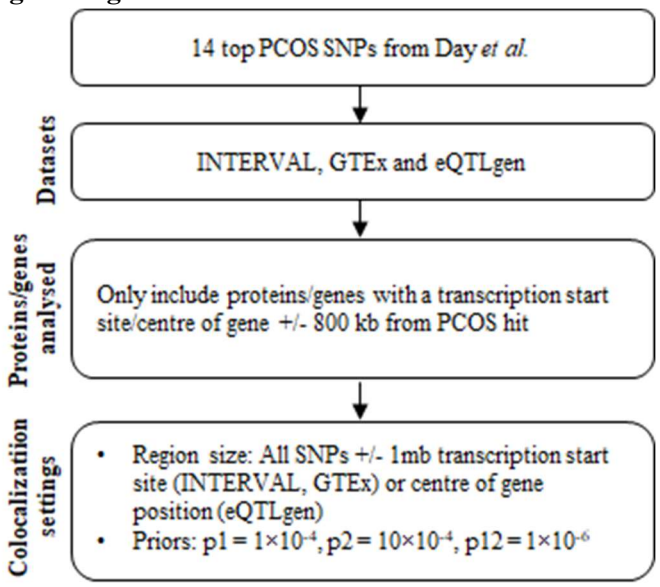

Figure 1. Study overview.

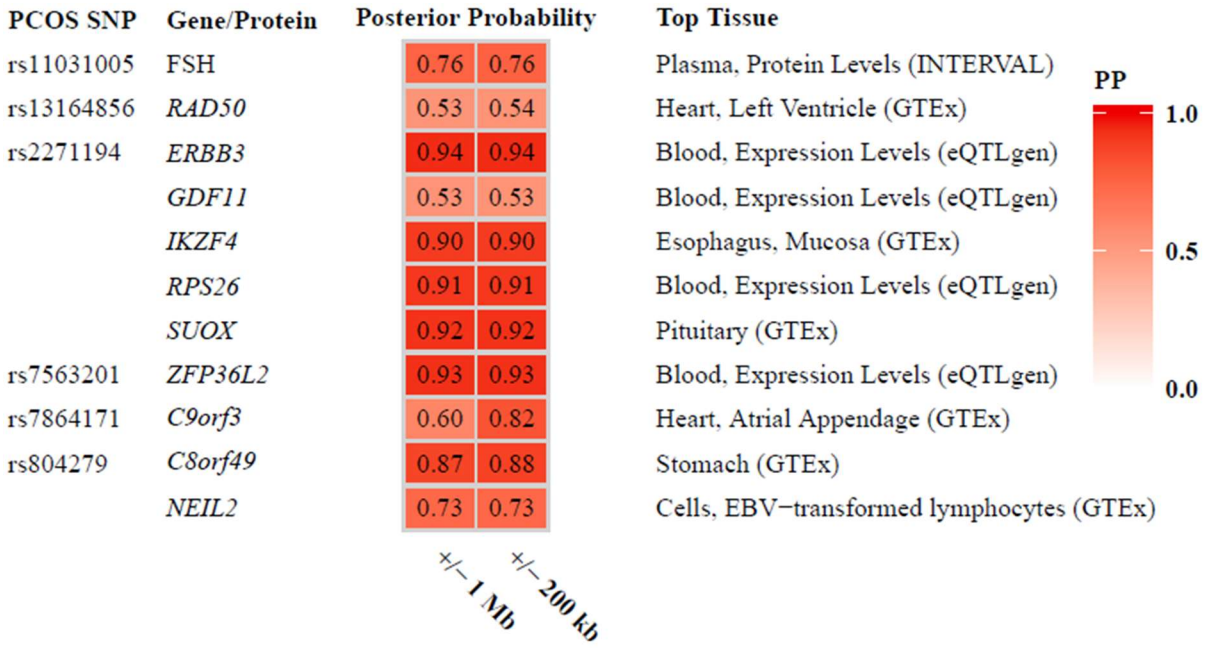

Figure 2. Posterior probabilities for genes and proteins with any evidence of colocalization. In the main approach, we used the Combined PCOS dataset and a region size spanning +/- 1 Mb. Only the results for the tissue with the highest posterior probability of colocalization in the main analysis are reported here (for full results and power calculations see Supplementary Table 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.

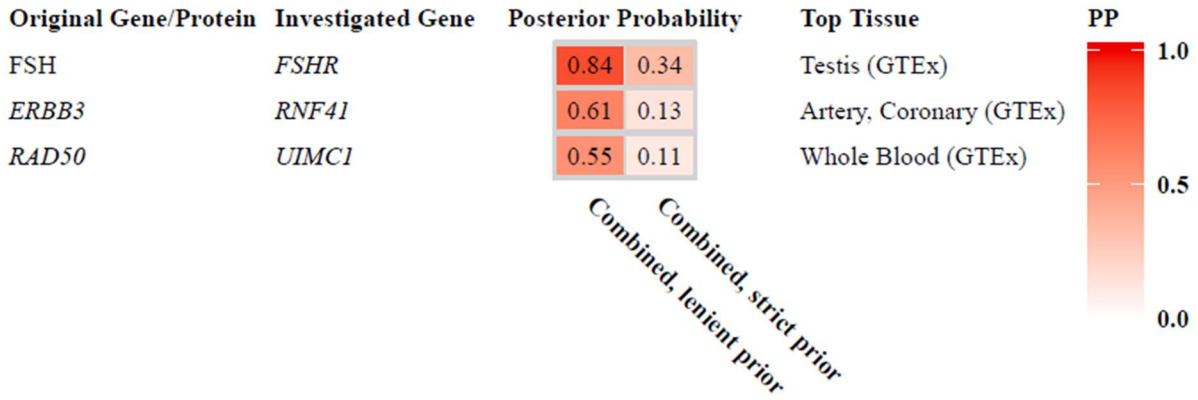

Figure 3. Posterior probabilities for genes with nominal evidence of colocalization in the interaction-coloc analyses.

In the main approach, we used the Combined PCOS dataset and a set the prior probability of colocalization to p12 = 1×10^{-5}. Sensitivity analyses included a more stringent prior probability of p12 = 1×10^{-6}. Note that *RNF41* – implicated in the same pathway as *ERBB3* – was also located in the rs2271194 locus. PP, posterior probability.

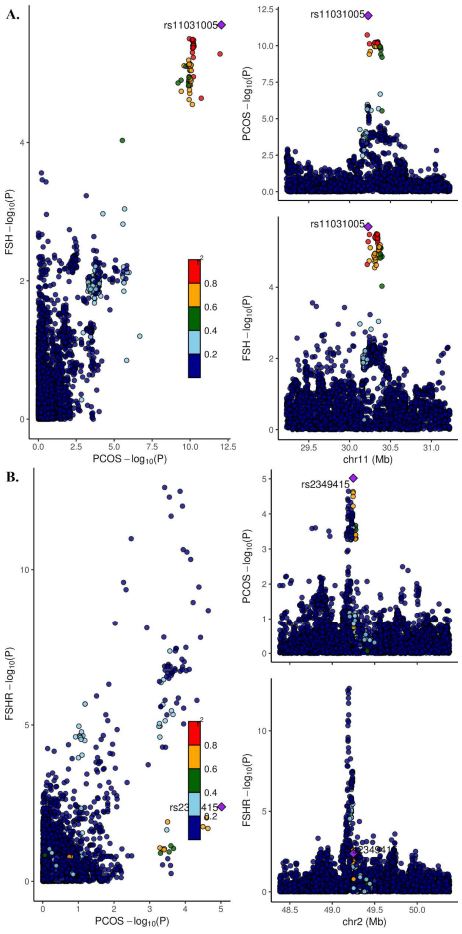

Figure 4A and 4B. Associations between genetic variants and PCOS risk, using Combined PCOS dataset, +/- 1 Mb region sizes for (A) FSH protein levels in blood (B) *FSHR* expression levels in testis

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 panels, -log10 P-values for associations with PCOS risk are on the x-axes, and -log10 P-values for associations with the protein/transcript 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/expression levels on the lower panel for the corresponding region. FSH, follicle stimulating hormone; PCOS, polycystic ovary syndrome; SNP, single nucleotide polymorphism.

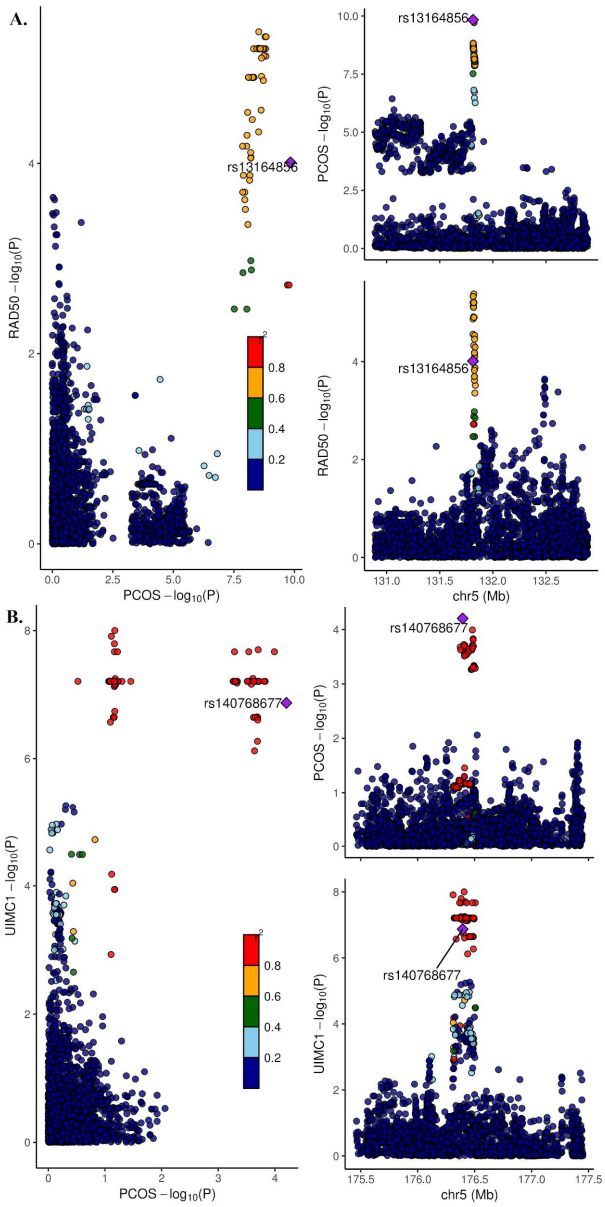

Figure 5A and 5B. Associations between genetic variants and PCOS risk, using Combined PCOS dataset, +/- 1 Mb region sizes for (A) *RAD50* expression levels in left ventricle of the heart (B) *UIMC1* expression levels in blood (GTEx)

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 panels, -log10 P-values for associations with PCOS risk are on the x-axes, and -log10 P-values

for associations with the expression 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 expression levels on the lower panel for the corresponding region. PCOS, polycystic ovary syndrome; SNP, single nucleotide polymorphism.

Tables

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

Table 1. Summary statistics for the top 14 single nucleotide polymorphisms associated with polycystic ovary syndrome from Day *et al.* (18)**.** SNP: single nucleotide polymorphism; Chr: chromosome; Pos: position (hg19); EA: effect allele; NEA: non-effect allele; EAF: effect allele frequency; CI: confidence interval; P: P-value.

Abbreviations

eQTL, expression quantitative trait locus
FSH, follicle-stimulating hormone
GnRH, gonadotropin-releasing hormone
GTEx, Genotype-Tissue Expression project
GWAS, genome-wide association study
HPG, hypothalamic-pituitary-gonadal
LH, luteinizing hormone
PCOS, polycystic ovary syndrome
PheWAS, phenome-wide association study
PP, posterior probability
pQTL, protein quantitative trait locus
SNP, single nucleotide polymorphism

References

1. Bozdag, G., Mumusoglu, S., Zengin, D., Karabulut, E. and Yildiz, B. O. (2016) The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. *Hum. Reprod.*, **31**, 2841–2855.
2. Teede, H., Deeks, A. and Moran, L. (2010) Polycystic ovary syndrome: a complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan. *BMC Med.*, **8**, 41.
3. Azziz, R., Carmina, E., Chen, Z., Dunaif, A., Laven, J. S. E., Legro, R. S., Lizneva, D., Natterson-Horowtiz, B., Teede, H. J. and Yildiz, B. O. (2016) Polycystic ovary syndrome. *Nat Rev Dis Primers*, **2**, 16057.
4. Rotterdam ESHRE/ASRM-Sponsored PCOS consensus workshop group (2004) Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). *Hum. Reprod.*, **19**, 41–47.
5. Blank, S. K., McCartney, C. R. and Marshall, J. C. (2006) The origins and sequelae of abnormal neuroendocrine function in polycystic ovary syndrome. *Hum. Reprod. Update*, **12**, 351–361.
6. Richards, J. S. and Pangas, S. A. (2010) The ovary: basic biology and clinical implications. *J. Clin. Invest.*, **120**, 963–972.
7. Krishnan, A. and Muthusami, S. (2017) Hormonal alterations in PCOS and its influence on bone metabolism. *J. Endocrinol.*, **232**, R99–R113.
8. Li, Y., Chen, C., Ma, Y., Xiao, J., Luo, G., Li, Y. and Wu, D. (2019) Multi-system reproductive metabolic disorder: significance for the pathogenesis and therapy of polycystic ovary syndrome (PCOS). *Life Sci.*, **228**, 167–175.
9. Taylor, A. E., McCourt, B., Martin, K. A., Anderson, E. J., Adams, J. M., Schoenfeld, D. and Hall, J. E. (1997) Determinants of abnormal gonadotropin secretion in clinically defined women with polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.*, **82**, 2248–2256.
10. Le, M. T., Le, V. N. S., Le, D. D., Nguyen, V. Q. H., Chen, C. and Cao, N. T. (2019) Exploration of the role of anti-Mullerian hormone and LH/FSH ratio in diagnosis of polycystic ovary syndrome. *Clin. Endocrinol.* , **90**, 579–585.
11. Yen, S. S., Vela, P. and Rankin, J. (1970) Inappropriate secretion of follicle-stimulating hormone and luteinizing hormone in polycystic ovarian disease. *J. Clin. Endocrinol. Metab.*, **30**, 435–442.
12. Fauser, B. C. and Van Heusden, A. M. (1997) Manipulation of human ovarian function: physiological concepts and clinical consequences. *Endocr. Rev.*, **18**, 71–106.
13. Vink, J. M., Sadrzadeh, S., Lambalk, C. B. and Boomsma, D. I. (2006) Heritability of polycystic ovary syndrome in a Dutch twin-family study. *J. Clin. Endocrinol. Metab.*, **91**, 2100–2104.
14. Shi, Y., Zhao, H., Shi, Y., Cao, Y., Yang, D., Li, Z., Zhang, B., Liang, X., Li, T., Chen, J., et al. (2012) Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nat. Genet.*, **44**, 1020–1025.
15. Chen, Z.-J., Zhao, H., He, L., Shi, Y., Qin, Y., Shi, Y., Li, Z., You, L., Zhao, J., Liu, J., et al. (2011) Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nat. Genet.*, **43**, 55–59.
16. Kosova, G. and Urbanek, M. (2013) Genetics of the polycystic ovary syndrome. *Mol. Cell. Endocrinol.*, **373**, 29–38.
17. McAllister, J. M., Legro, R. S., Modi, B. P. and Strauss, J. F., 3rd (2015) Functional genomics of PCOS: from GWAS to molecular mechanisms. *Trends Endocrinol. Metab.*, **26**, 118–124.
18. Day, F., Karaderi, T., Jones, M. R., Meun, C., He, C., Drong, A., Kraft, P., Lin, N., Huang, H., Broer, L., et al. (2018) Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.*, **14**, e1007813.
19. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C. and Plagnol, V. (2014) Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.*, **10**, e1004383.
20. GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

21. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018) Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*, 447367.
22. Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018) Genomic atlas of the human plasma proteome. *Nature*, **558**, 73–79.
23. Di Angelantonio, E., Thompson, S. G., Kaptoge, S., Moore, C., Walker, M., Armitage, J., Ouwehand, W. H., Roberts, D. J., Danesh, J. and INTERVAL Trial Group (2017) Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet*, **390**, 2360–2371.
24. Çalışkan, M., Manduchi, E., Rao, H. S., Segert, J. A., Beltrame, M. H., Trizzino, M., Park, Y., Baker, S. W., Chesi, A., Johnson, M. E., et al. (2019) Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. *Am. J. Hum. Genet.*, **105**, 89–107.
25. Wallace, C. (2019) Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *bioRxiv* **(2019)** , 838946.
26. Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., Sunyaev, S. R. and Cotsapas, C. (2017) Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.*, **49**, 600–605.
27. Franceschini, N., Giambartolomei, C., de Vries, P. S., Finan, C., Bis, J. C., Huntley, R. P., Lovering, R. C., Tajuddin, S. M., Winkler, T. W., Graff, M., et al. (2018) GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat. Commun.*, **9**, 5141.
28. Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B. and Eskin, E. (2016) Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.*, **99**, 1245–1260.
29. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
30. Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M., Faulconbridge, A., Hercules, A., McAuley, E., et al. (2019) Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res.*, **47**, D1056–D1065.
31. Ward, L. D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930-4.
32. UK Biobank — Neale lab. UK Biobank — Neale lab http://www.nealelab.is/uk-biobank/ (accessed Sep 10, 2019).
33. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
34. Gregor, P. D., Sawadogo, M. and Roeder, R. G. (1990) The adenovirus major late transcription factor USF is a member of the helix-loop-helix group of regulatory proteins and binds to DNA as a dimer. *Genes Dev.*, **4**, 1730–1740.
35. Hampsey, M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.*, **62**, 465–503.
36. ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
37. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
38. Berisa, T. and Pickrell, J. K. (2016) Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**, 283–285.
39. Foley, C. N., Staley, J. R., Breen, P. G., Sun, B. B., Kirk, P. D. W., Burgess, S. and Howson, J. M. M. (2019) A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *bioRxiv*, 592238.

40. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
41. Ball, C. B., Rodriguez, K. F., Stumpo, D. J., Ribeiro-Neto, F., Korach, K. S., Blackshear, P. J., Birnbaumer, L. and Ramos, S. B. V. (2014) The RNA-binding protein, ZFP36L2, influences ovulation and oocyte maturation. *PLoS One*, **9**, e97324.
42. van der Spuy, Z. M. and Dyer, S. J. (2004) The pathogenesis of infertility and early pregnancy loss in polycystic ovary syndrome. *Best Pract. Res. Clin. Obstet. Gynaecol.*, **18**, 755–771.
43. Misgar, R. A., Wani, A. I., Bankura, B., Bashir, M. I., Roy, A. and Das, M. (2019) FSH β-subunit mutations in two sisters: the first report from the Indian sub-continent and review of previous cases. *Gynecol. Endocrinol.*, **35**, 290–293.
44. Abel, M. H., Widen, A., Wang, X., Huhtaniemi, I., Pakarinen, P., Kumar, T. R. and Christian, H. C. (2014) Pituitary gonadotrophic hormone synthesis, secretion, subunit gene expression and cell structure in normal and follicle-stimulating hormone β knockout, follicle-stimulating hormone receptor knockout, luteinising hormone receptor knockout, hypogonadal and ovariectomised female mice. *J. Neuroendocrinol.*, **26**, 785–795.
45. Pau, C. T., Mosbruger, T., Saxena, R. and Welt, C. K. (2017) Phenotype and Tissue Expression as a Function of Genetic Risk in Polycystic Ovary Syndrome. *PLoS One*, **12**, e0168870.
46. Hayes, M. G., Urbanek, M., Ehrmann, D. A., Armstrong, L. L., Lee, J. Y., Sisk, R., Karaderi, T., Barber, T. M., McCarthy, M. I., Franks, S., et al. (2015) Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nat. Commun.*, **6**, 7502.
47. Ruth, K. S., Campbell, P. J., Chew, S., Lim, E. M., Hadlow, N., Stuckey, B. G. A., Brown, S. J., Feenstra, B., Joseph, J., Surdulescu, G. L., et al. (2016) Genome-wide association study with 1000 genomes imputation identifies signals for nine sex hormone-related phenotypes. *Eur. J. Hum. Genet.*, **24**, 284–290.
48. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
49. Chowdhury, I., Branch, A., Mehrabi, S., Ford, B. D. and Thompson, W. E. (2017) Gonadotropin-Dependent Neuregulin-1 Signaling Regulates Female Rat Ovarian Granulosa Cell Survival. *Endocrinology*, **158**, 3647–3660.
50. Mukherjee, A. and Roy, S. K. (2013) Expression of ErbB3-binding protein-1 (EBP1) during primordial follicle formation: role of estradiol-17ß. *PLoS One*, **8**, e67068.
51. Cao, Z., Wu, X., Yen, L., Sweeney, C. and Carraway, K. L., 3rd (2007) Neuregulin-induced ErbB3 downregulation is mediated by a protein stability cascade involving the E3 ubiquitin ligase Nrdp1. *Mol. Cell. Biol.*, **27**, 2180–2188.
52. Cui, D., Li, L., Lou, H., Sun, H., Ngai, S.-M., Shao, G. and Tang, J. (2014) The ribosomal protein S26 regulates p53 activity in response to DNA damage. *Oncogene*, **33**, 2225–2235.
53. Liu, X.-M., Yan, M.-Q., Ji, S.-Y., Sha, Q.-Q., Huang, T., Zhao, H., Liu, H.-B., Fan, H.-Y. and Chen, Z.-J. (2018) Loss of oocyte Rps26 in mice arrests oocyte growth and causes premature ovarian failure. *Cell Death Dis.*, **9**, 1144.
54. Grenon, M., Gilbert, C. and Lowndes, N. F. (2001) Checkpoint activation in response to double-strand breaks requires the Mre11/Rad50/Xrs2 complex. *Nat. Cell Biol.*, **3**, 844–847.
55. Theunissen, J.-W. F., Kaplan, M. I., Hunt, P. A., Williams, B. R., Ferguson, D. O., Alt, F. W. and Petrini, J. H. J. (2003) Checkpoint Failure and Chromosomal Instability without Lymphomagenesis in Mre11ATLD1/ATLD1 Mice. *Mol. Cell*, **12**, 1511–1523.
56. Paull, T. T. and Gellert, M. (1998) The 3′ to 5′ Exonuclease Activity of Mre11 Facilitates Repair of DNA Double-Strand Breaks. *Mol. Cell*, **1**, 969–979.
57. Inagaki, A., Roset, R. and Petrini, J. H. J. (2016) Functions of the MRE11 complex in the development and maintenance of oocytes. *Chromosoma*, **125**, 151–162.
58. Trujillo, K. M., Yuan, S. S., Lee, E. Y. and Sung, P. (1998) Nuclease activities in a complex of human recombination and DNA repair factors Rad50, Mre11, and p95. *J. Biol. Chem.*, **273**, 21447–21450.
59. Roset, R., Inagaki, A., Hohl, M., Brenet, F., Lafrance-Vanasse, J., Lange, J., Scandura, J. M., Tainer, J. A., Keeney, S. and Petrini, J. H. J. (2014) The Rad50 hook domain regulates DNA damage signaling and tumorigenesis. *Genes Dev.*, **28**, 451–462.

60. Mailand, N., Bekker-Jensen, S., Faustrup, H., Melander, F., Bartek, J., Lukas, C. and Lukas, J. (2007) RNF8 ubiquitylates histones at DNA double-strand breaks and promotes assembly of repair proteins. *Cell*, **131**, 887–900.
61. Fung, J. N., Mortlock, S., Girling, J. E., Holdsworth-Carson, S. J., Teh, W. T., Zhu, Z., Lukowski, S. W., McKinnon, B. D., McRae, A., Yang, J., et al. (2018) Genetic regulation of disease risk and endometrial gene expression highlights potential target genes for endometriosis and polycystic ovarian syndrome. *Sci. Rep.*, **8**, 11424.
62. Ongen, H., Brown, A. A., Delaneau, O., Panousis, N. I., Nica, A. C., GTEx Consortium and Dermitzakis, E. T. (2017) Estimating the causal tissues for complex traits and diseases. *Nat. Genet.*, **49**, 1676–1683.
63. GTEx Portal. GTEx Portal https://gtexportal.org/home/ (accessed Jun 28, 2019).
64. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **(2005)** , *21*, 3439–3440.
65. Durinck, S., Spellman, P. T., Birney, E. and Huber, W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **(2009)** , *4*, 1184–1191.
66. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., HIPSCI Consortium, Hale, C., Dougan, G. and Gaffney, D. J. (2018) Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.*, **50**, 424–431.
67. Graham, S. E., Nielsen, J. B., Zawistowski, M., Zhou, W., Fritsche, L. G., Gabrielsen, M. E., Skogholt, A. H., Surakka, I., Hornsby, W. E., Fermin, D., et al. (2019) Sex-specific and pleiotropic effects underlying kidney function identified from GWAS meta-analysis. *Nat. Commun.*, **10**, 1847.
68. Andaleon, A., Mogil, L. S. and Wheeler, H. E. (2018) Genetically regulated gene expression underlies lipid traits in Hispanic cohorts. *bioRxiv*, 507905.
69. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., et al. (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.
70. CGA glycoprotein hormones, alpha polypeptide [Homo sapiens (human)] - Gene - NCBI. CGA glycoprotein hormones, alpha polypeptide [Homo sapiens (human)] - Gene - NCBI https://www.ncbi.nlm.nih.gov/gene/1081 (accessed Jul 1, 2019).
71. R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.r-project.org/.
72. Free Software Foundation (2007) bash 4.1.2(2). bash 4.1.2(2) https://www.gnu.org/software/bash/.
73. Free Software Foundation (1989) GNU AWK 3.1.7. GNU AWK 3.1.7 https://www.gnu.org/software/gawk/manual/gawk.html.
74. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. and Montgomery, S. B. (2019) Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.*, 1.
75. Wickham, H. and Henry, L. (2019) tidyr: Easily Tidy Data with “spread()” and “gather()” Functions. tidyr: Easily Tidy Data with “spread()” and “gather()” Functions **(2019)** .
76. Dowle, M. and Srinivasan, A. (2019) data.table: Extension of `data.frame`. data.table: Extension of `data.frame` **(2019)** .
77. Wickham, H. (2011) The Split-Apply-Combine Strategy for Data Analysis. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software* **(2011)** , *40*, 1–29.
78. Wickham, H., Hester, J. and Chang, W. (2019) devtools: Tools to Make Developing R Packages Easier. devtools: Tools to Make Developing R Packages Easier **(2019)** .
79. Kassambara, A. (2018) ggpubr: “ggplot2” Based Publication Ready Plots. ggpubr: “ggplot2” Based Publication Ready Plots.