

ARTICLE

Received 19 Nov 2014 | Accepted 14 May 2015 | Published 18 Aug 2015

DOI: 10.1038/ncomms8502

OPEN

Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations

M. Geoffrey Hayes^{1,2,3,*}, Margrit Urbanek^{1,2,*}, David A. Ehrmann⁴, Loren L. Armstrong¹, Ji Young Lee¹, Ryan Sisk¹, Tugce Karaderi⁵, Thomas M. Barber⁶, Mark I. McCarthy^{5,7,8}, Stephen Franks⁹, Cecilia M. Lindgren^{5,10}, Corrine K. Welt¹¹, Evanthia Diamanti-Kandarakis¹², Dimitrios Panidis¹³, Mark O. Goodarzi¹⁴, Ricardo Azziz¹⁵, Yi Zhang^{16,17}, Roland G. James^{16,17}, Michael Olivier¹⁸, Ahmed H. Kissebah^{16,17,‡}, Reproductive Medicine Network[#], Elisabet Stener-Victorin¹⁹, Richard S. Legro^{20,**}, Andrea Dunaif^{1,2,**}

Polycystic ovary syndrome (PCOS) is a common, highly heritable complex disorder of unknown aetiology characterized by hyperandrogenism, chronic anovulation and defects in glucose homeostasis. Increased luteinizing hormone relative to follicle-stimulating hormone secretion, insulin resistance and developmental exposure to androgens are hypothesized to play a causal role in PCOS. Here we map common genetic susceptibility loci in European ancestry women for the National Institutes of Health PCOS phenotype, which confers the highest risk for metabolic morbidities, as well as reproductive hormone levels. Three loci reach genome-wide significance in the case-control meta-analysis, two novel loci mapping to chr 8p32.1 and chr 11p14.1, and a chr 9q22.32 locus previously found in Chinese PCOS. The same chr 11p14.1 SNP, rs11031006, in the region of the follicle-stimulating hormone B polypeptide (*FSHB*) gene strongly associates with PCOS diagnosis and luteinizing hormone levels. These findings implicate neuroendocrine changes in disease pathogenesis.

¹ Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois 60611, USA. ² Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois 60611, USA. ³ Department of Anthropology, Northwestern University, Evanston, Illinois 60208, USA. ⁴ Section of Endocrinology, Diabetes, and Metabolism, The University of Chicago, Chicago, Illinois 60637, USA. ⁵ Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. ⁶ Warwick Medical School, University of Warwick, Warwick CV4 7AL, UK. ⁷ Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LE, UK. ⁸ Oxford NIHR Biomedical Research Centre, Churchill Hospital, Headington OX3 7LE, UK. ⁹ Institute of Reproductive & Developmental Biology, Hammersmith Hospital, Imperial College London, London W12 0NN, UK. ¹⁰ Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. ¹¹ Division of Endocrinology, Metabolism and Diabetes, University of Utah, Salt Lake City, Utah 84112, USA. ¹² Endocrinology and Metabolism, University of Athens Medical School, Athens 115 27, Greece. ¹³ Division of Endocrinology and Human Reproduction, 2nd Department of Obstetrics and Gynecology, Aristotle University of Thessaloniki 54124, Greece. ¹⁴ Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA. ¹⁵ Departments of Obstetrics and Gynecology and Medicine, Medical College of Georgia, Georgia Regents University, Augusta, Georgia 30912, USA. ¹⁶ TOPS Obesity and Metabolic Research Center, Department of Medicine, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. ¹⁷ Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. ¹⁸ Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas 78256, USA. ¹⁹ Department of Physiology and Pharmacology, Karolinska Institutet, SE-171 77 Stockholm, Sweden. ²⁰ Department of Obstetrics and Gynecology, Penn State College of Medicine, Hershey, Pennsylvania 17033, USA. * These authors contributed equally to this work. ** These authors jointly supervised this work. ‡ Deceased. # A full list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to M.G.H. (email: ghayes@northwestern.edu).

Polycystic ovary syndrome (PCOS) is a common disorder of premenopausal women affecting 7–15% of this population worldwide¹. It is diagnosed by its reproductive phenotype of hyperandrogenism, ovulatory disturbances and polycystic ovarian changes. PCOS is also frequently associated with insulin resistance, pancreatic β -cell dysfunction and obesity that confer a significantly increased risk for type 2 diabetes (T2D)¹. The aetiology of PCOS is unknown. Anovulation and hyperandrogenism are a consequence primarily of disordered gonadotropin secretion¹. Increased LH stimulates ovarian testosterone (T) production while relative FSH deficiency results in arrest of ovarian folliculogenesis². There may also be constitutive increases in ovarian and adrenal androgen secretion¹. Increased T levels feedback on the hypothalamus contributing to disordered gonadotropin release³. Androgen administration during critical developmental windows produces phenocopies of PCOS in animal models suggesting that this mechanism may also be operative in humans⁴. Finally, insulin resistance and the resulting hyperinsulinemia contribute to both hyperandrogenism and ovulatory dysfunction¹.

In addition, there are abnormalities in the gonadotropin-independent development of preantral follicles in polycystic ovaries (PCO) resulting in an excess of primary follicles compared to normal ovaries². Decreased follicular atresia in PCO may contribute to this finding². However, 20–30% of women with regular menses and no reproductive symptoms have PCO morphology (PCOM)⁵. PCOM is more common in younger women⁵. Even in the absence of hormonal features of the syndrome, women with PCOM have increased sensitivity to exogenous gonadotropin administration⁶.

As the precise cause(s) of PCOS remains elusive, there has been substantial controversy surrounding the most appropriate diagnostic criteria¹. The first diagnostic criteria, the so-called National Institutes of Health (NIH) criteria, required clinical or biochemical evidence of hyperandrogenism and chronic anovulation with the exclusion of specific conditions that can present with similar features¹. The Rotterdam criteria⁷ added PCOM as a criterion and required two of three of these features for the diagnosis of PCOS. This modification resulted in multiple phenotypes of affected women, including those with PCOM and hyperandrogenism but no anovulation and those with anovulation and PCOM but no hyperandrogenism. The Androgen Excess Society criteria⁸ modified the Rotterdam criteria to require hyperandrogenism for the diagnosis of PCOS thereby removing the anovulation and PCOM phenotype. There is now general agreement that the NIH criteria identify the phenotype at greatest risk for insulin resistance and its associated metabolic features, whereas the other phenotypes have only minimal metabolic risk^{1,9}.

Familial aggregation of PCOS and its associated features^{10,11} as well as twin studies¹² have suggested that genetic factors play an important role in disease pathogenesis. Further, two or more phenotypes can be present in the same family suggesting some of these phenotypic differences reflect variable expression and/or penetrance of the same gene¹⁰. Several susceptibility loci for PCOS have been reproducibly mapped in family-based¹³ and genome-wide association studies^{14,15}. Previous PCOS genome-wide association studies (GWAS) have been limited to Asian cohorts diagnosed using the Rotterdam criteria^{14–16}. PCOS cohorts thus identified have multiple phenotypes of affected women⁸. Therefore, it has not been possible to determine whether the genetic architecture of PCOS varies by phenotype. Further, there have been no genome-wide quantitative trait analyses to assess potential genetic contributions to the hormonal features of the syndrome.

The aim of our study is to identify susceptibility loci for the NIH phenotype of hyperandrogenism and anovulation, which is

associated with high risk for insulin resistance and dysglycemia^{1,9}, and its quantitative reproductive hormone levels¹⁷. We perform a discovery GWAS (Stage 1) of 984 PCOS cases and 2,964 population controls followed by replication (Stage 2) in 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women. We follow this with a second replication (Stage 3) of the top 24 associations after a meta-analysis of stage 1 and 2 in a cohort of 217 PCOS cases and 1,335 population controls. Three loci reach genome-wide significance in the case-control meta-analysis of all three strata; two novel loci, chr 8p32.1 in the region of *GATA4* and *NEIL2* and chr 11p14.1 in the region of the follicle-stimulating hormone B polypeptide (*FSHB*) gene, and one previously found in Chinese PCOS¹⁵, chr 9q22.32 in the region of *c9orf3/FANCC*. The same chr 11p14.1 SNP, rs11031006, in the region of the *FSHB* gene also reaches genome-wide significance in the meta-analysis of the quantitative luteinizing hormone (LH) levels. These findings implicate gonadotropins in the pathogenesis of PCOS.

Results

Case-control analysis. Three loci were associated with PCOS in our cohort (Table 1) at a genome-wide significant threshold after Stage 3: the 8p32.1 *GATA4/NEIL2* locus, the 9q22.32 *c9orf3/FANCC* locus, and 11p14.1 *FSHB/ARL14EP* locus with sample-size weighted three-strata meta-analyses $P_{\text{meta-all}} = 8.0 \times 10^{-10}$; $P_{\text{meta-all}} = 4.6 \times 10^{-13}$; and $P_{\text{meta-all}} = 1.9 \times 10^{-8}$, respectively (Table 2; Figs 1a and 2a–c; Supplementary Data 1–2).

The 8p32.1 PCOS locus spanned a 6.6 region of extensive linkage disequilibrium ($r^2 > 0.2$) including the 3' end of *GATA4* and all of *NEIL2* into the intergenic region before *FDFT1*. The most strongly associated SNP (rs804279) was intergenic, approached genome-wide significance in the Stage 1 Discovery GWAS (A allele, OR = 0.74, 0.66–0.83 95%CI, logistic regression $P = 1.4 \times 10^{-6}$), and was nominally significant in the Stage 2 Metabochip Replication (A allele, OR = 0.82, 0.73–0.93 95%CI, logistic regression $P = 2.7 \times 10^{-3}$) for a sample-size weighted two-strata meta-analysis $P_{\text{meta}} = 1.9 \times 10^{-8}$. The additional UK Replication (Stage 3) was nominally significant (A allele, OR = 0.74, 0.58–0.94 95%CI, logistic regression $P = 0.013$) for a sample-size weighted three-strata meta-analysis $P_{\text{meta-All}} = 8.0 \times 10^{-10}$. The proportion of heritability explained by this SNP was 0.79% (ref. 18). After conditioning on the lead SNP in the region, the association was reduced several log orders. This pattern was consistent across all genome-wide significant loci suggesting that there was only one locus within each region of interest discussed below (Supplementary Fig. 1).

The 9q22.32 PCOS locus spanned a ~331.8 kb extensive linkage disequilibrium (LD) region that includes *c9orf3*, *FANCC* and a number of micro RNA genes. The most strongly associated SNP (rs10993397), in the large fifth intron of *c9orf3*, reached genome-wide significance in the Stage 1 Discovery GWAS (C allele, OR = 0.73, 0.66–0.81 95%CI, logistic regression $P = 2.2 \times 10^{-8}$), was also nominally significant in the Stage 2 Metabochip Replication (C allele, OR = 0.88, 0.79–0.98 95%CI, logistic regression $P = 2.4 \times 10^{-4}$) for a sample-size weighted two-strata meta-analysis $P_{\text{meta}} = 3.1 \times 10^{-11}$. This SNP was nominally significant (C allele, OR = 0.72, 0.58–0.89 95%CI, logistic regression $P = 3.9 \times 10^{-3}$) in the Stage 3 UK Replication for a sample-size weighted three-strata meta-analysis $P_{\text{meta-All}} = 4.6 \times 10^{-13}$. The proportion of heritability explained by this SNP was 0.79% (ref. 18). Rs10993397 was 61 kb upstream from the previously identified intronic association in a Chinese population¹⁵ but still within same the region of strong LD for both European and Chinese ancestry. To test the independence of the two loci, we conditioned the association of rs10993397 with

Table 1 | Demographic characteristics of PCOS cases and controls in the GWAS and replication.

Cohort	GWAS						Replication					
	Age (years)		BMI (kg m ⁻²)		Testosterone (ng dl ⁻¹)		Age (years)		BMI (kg m ⁻²)		Testosterone (ng dl ⁻¹)	
	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Controls
PCOS Family Study*	28 (24-32) n = 718	—	35.5 (29.1-41.4) n = 718	—	72 (59-91) n = 663	—	28 (24-32) n = 172	29 (24-33) n = 105	35.0 (28.3-41.5) n = 172	30.7 (23.3-35.0) n = 105	70 (59-93) n = 100	28 (17-35) n = 35
PPCOSI*	28 (26-30) n = 64	—	35.3 (27.4-42.1) n = 64	—	60 (44-82) n = 62	—	—	—	—	—	68 (49-86) n = 47	25 (19-32) n = 70
University of Chicago [†]	29 (25-34) n = 202	—	36.2 (31.8-42.1) n = 202	—	68 (52-93) n = 202	—	27 (23-30) n = 25	—	38.4 (32.4-42.7) n = 25	—	64 (50-85) n = 25	—
NUgene	—	46 (34-58) n = 2,964	—	25.0 (22.0-29.8) n = 2,964	—	—	—	—	—	—	—	—
Cedars-Sinai Medical Center [‡]	—	—	—	—	—	—	26 (22-31) n = 130	37 (27-42) n = 17	27.7 (22.5-33.9) n = 130	24.6 (21.5-31.5) n = 17	43 (31-62) n = 121	23 (17-29) n = 11
Massachusetts General Hospital	—	—	—	—	—	—	28 (24-32) n = 472	25 (22-31) n = 402	29.0 (23.9-37.0) n = 472	23.0 (21.1-25.0) n = 402	—	—
Medical College of Wisconsin*	—	—	—	—	—	—	38 (32-41) n = 452	—	31.9 (24.2-38.9) n = 452	—	—	19 (15-24) n = 452
PPCOSII*	—	—	—	—	—	—	28 (26-31) n = 265	—	35.1 (26.6-41.8) n = 265	—	50 (37-67) n = 265	—
University of Alabama at Birmingham [§]	—	—	—	—	—	—	27 (22-33) n = 193	32 (26-36) n = 91	34.1 (27.8-40.8) n = 193	23.7 (21.7-27.7) n = 91	77 (63-94) n = 193	44 (32-56) n = 81
University of Athens	—	—	—	—	—	—	23 (20-28) n = 542	26 (23-34) n = 164	24.5 (21.4-30.9) n = 542	22.0 (20.1-24.5) n = 164	72 (61-89) n = 542	31 (25-43) n = 163
Imperial College London/ University of Oxford [¶]	—	—	—	—	—	—	32 (28-36) n = 222	45 (45-45) n = 1,335	28.4 (23.2-36.5) n = 217	25.5 (23.0-29.4) n = 1,335	69 (53-86) n = 223	—

PCOS, polycystic ovary syndrome; BMI, body mass index; GWAS, genome-wide association studies. Data are expressed as median (25th-75th interquartile range). Hormone results are separated by assay method when multiple methods and labs were used within a cohort, however the normal range was identical.
 *Testosterone normal range < 59 ng dl⁻¹
 †Testosterone normal range < 71 ng dl⁻¹
 ‡Testosterone normal range < 35 ng dl⁻¹
 §Testosterone normal range < 85 ng dl⁻¹
 ||Testosterone normal range < 60 ng dl⁻¹
 ¶Testosterone normal range < 70 ng dl⁻¹

Table 2 | Genome-wide significant associations of PCOS and associated reproductive hormonal quantitative traits.

Trait	Variant	Chr	Mb	Gene or Eff Flanking Genes	Stage 1 (Discovery GWAS)				Stage 2 (MetaboChip replication)				Stage 1 + 2 meta-analysis P _{meta}	Stage 3 (UK replication)				Stage 1 + 2 + 3 meta-analysis P _{meta-all}	
					EAF	OR/ Beta ^{††}	95% CI / s.e.*	P	Gen/ Imp r ²	EAF	OR/ Beta ^{††}	95% CI / s.e.*		P	Gen/ Imp r ²	EAF	OR/ Beta ^{††}		95% CI / s.e.*
PCOS	rs804279	8	11.6	GATA4 / NEIL2	0.74	0.74	0.66-1.4×10 ⁻⁶	0.99	0.74	0.82	0.73-2.7×10 ⁻³	0.97	1.9×10 ⁻⁸	0.72	0.74	0.5	8-1.3×10 ⁻²	0.95	8.0×10 ⁻¹⁰
PCOS	rs10993397	9	97.7	C9orf3	0.44	0.73	0.66-2.2×10 ⁻⁸	1	0.44	0.88	0.79-2.4×10 ⁻⁴	1	3.1×10 ⁻¹¹	0.41	0.72	0.5	8-3.9×10 ⁻³	0.99	4.6×10 ⁻¹³
PCOS	rs11031006	11	30.2	KCNA4 / FSHB	0.15	1.41	1.24-7.8×10 ⁻⁸	Gen	0.16	1.18	1.02-6.3×10 ⁻³	Gen	4.3×10 ⁻⁹	0.15	1.12	0.8	3-4.5×10 ⁻¹	0.96	1.9×10 ⁻⁸
LH	rs11031006	11	30.2	KCNA4 / FSHB	0.20	3.19	0.53-1.8×10 ⁻⁹	Gen	0.21	2.19	0.657.7×10 ⁻⁴	Gen	8.5×10 ⁻¹²	0.15	0.60	0.12	1.8×10 ⁻⁶	0.96	2.7×10 ⁻¹⁶

CI, confidence interval; EAF, effect allele frequency in cases and controls combined; Eff, effect allele; Gen, genotyped; FSH, follicle-stimulating hormone; Imp r², Imputation r²; LH, luteinizing hormone; PCOS, polycystic ovary syndrome; P value, significance as assessed by logistic regressions (PCOS) or linear regressions (LH levels); P_{meta}, significance as assessed by sample-size weighted two-strata meta-analyses (PCOS Stage 1: 984 cases & 2,964 population control women; PCOS Stage 2: 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women (LH Stage 1: 645 PCOS cases; LH Stage 2: 398 PCOS cases). P_{meta-all}, significance as assessed by sample-size weighted three-strata meta-analyses (Stages 1 and 2 as described for P_{meta}; PCOS Stage 3: 217 PCOS cases and 1,335 population controls; LH Stage 3: 208 PCOS cases); OR, odds ratio.
 *OR and 95% CI for dichotomous trait PCOS; Beta and s.e. for quantitative trait LH.
 †Units: LH: mIU ml⁻¹.

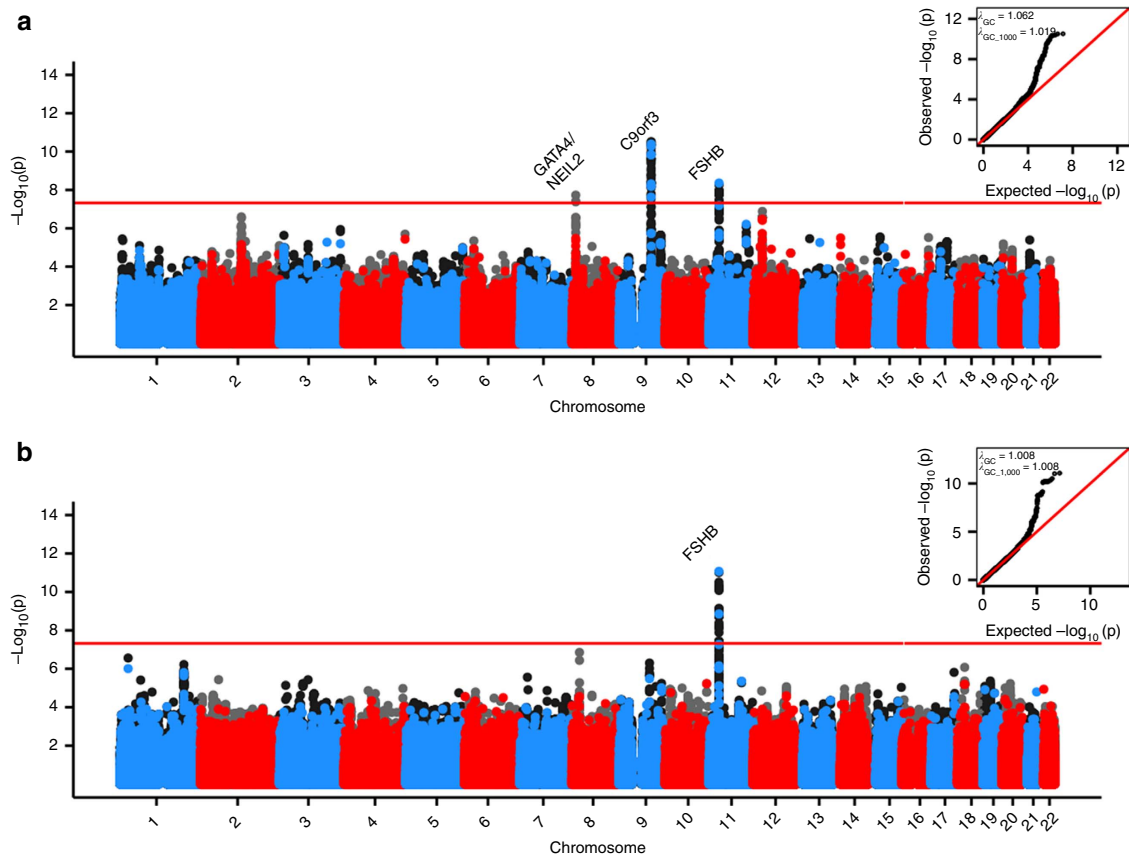


Figure 1 | Genome-wide association results for traits with genomewide significant hits in the meta-analysis of GWAS and replication phases.

Manhattan plots for **a**. PCOS, **b**. LH levels. Alternating blue and red colours indicate genotyped SNPs, and accompanying black and grey colours indicate imputed variants, on odd and even chromosomes, respectively. The red horizontal red line indicates genomewide significance. QQ plots and $\lambda_{GC}/\lambda_{GC1000}$ are inset in the upper right corner of each plot. For **(a)** PCOS, P values are from sample-size weighted two-strata meta-analysis of strata-specific logistic regression P values (Stage 1: 984 cases and 2,964 population control women; Stage 2: 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women). For **(b)** LH levels, P values are from sample-size weighted two-strata meta-analysis of strata-specific linear regression P values (Stage 1: 645 PCOS cases; Stage 2: 399 PCOS cases).

PCOS on the lead Chinese GWAS locus in this region (rs3802457). The association remained genome-wide significant after including the rs3802457 SNP as a covariate in the regression model of PCOS and rs10993397, suggesting that the European and Chinese ancestry specific loci in this region were independent of one another (Supplementary Data 3).

The 11p14.1 PCOS locus spanned a 129.5 kb extensive LD region beginning upstream of *FSHB* and continues through *ARL14EP* (also known as *C11orf46*). The most strongly associated SNP (rs11031006) for both PCOS approached genome-wide significance for PCOS in the Stage 1 Discovery GWAS (G allele, OR = 1.41, 1.24–1.62 95%CI, logistic regression $P = 7.8 \times 10^{-8}$) and was nominally significant in Stage 2 Metabochip Replication (G allele, OR = 1.18, 1.03–1.37 95%CI, logistic regression $P = 6.3 \times 10^{-3}$) for a sample-size weighted two-strata meta-analysis $P_{meta} = 4.3 \times 10^{-9}$. This SNP was not significant in the Stage 3 UK Replication but trended in the same direction as the Stages 1 and 2 (G allele, OR = 1.12, 0.83–1.50 95%CI, logistic regression $P = 0.45$) for a sample-size weighted three-strata meta-analysis $P_{meta-All} = 1.9 \times 10^{-8}$. The proportion of heritability explained by this SNP was 0.80% (ref. 18).

Quantitative trait analysis. The Stage 1 Discovery GWAS for reproductive hormone levels within the PCOS cases ($N = 645$ –957 depending on hormone trait) identified three loci

reaching genome-wide significance: the 7q33 *DGKI* locus with total T levels, linear regression $P = 1.3 \times 10^{-8}$; the 10q26.11 *CASC2* locus with FSH levels, linear regression $P = 3.2 \times 10^{-8}$; and the 11p14.1 *FSHB/ARL14EP* locus with LH levels, linear regression $P = 1.8 \times 10^{-9}$ (Table 2; Figs 1b and 2d; SupplementaryData 1–2, Supplementary Figs 2–5). None of these variants showed significant association in the control women (Supplementary Data 4). No genome-wide significant associations with dehydroepiandrosterone sulfate (DHEAS) or sex hormone binding globulin (SHBG) levels were observed (Supplementary Data 1; Supplementary Figs 6 and 7). The same 11p14.1 *FSHB/ARL14EP* SNP, rs11031006, that was most strongly associated with PCOS reached genome-wide significance with LH in the Stage 1 Discovery GWAS (G allele, Beta = 3.19, 0.53 s.e., mIU ml $^{-1}$, linear regression $P = 1.8 \times 10^{-9}$) and was nominally significant in the Stage 2 Metabochip Replication (G allele, Beta = 2.19, 0.65 s.e., mIU ml $^{-1}$, linear regression $P = 7.7 \times 10^{-4}$) for a sample-size weighted two-strata meta-analysis $P_{meta} = 8.5 \times 10^{-12}$. The Stage 3 UK Replication was also nominally significant (G allele, Beta = 0.6, 0.12 s.e., mIU ml $^{-1}$, linear regression $P = 1.8 \times 10^{-6}$) for a sample-size weighted three-strata meta-analysis $P_{meta-All} = 2.7 \times 10^{-16}$. The proportion of phenotypic variance explained by this SNP was 5.4%. Upon further investigation we determined this high proportion of phenotypic variance explained was largely influenced by several individuals with LH > 25 mIU ml $^{-1}$ most likely measured mid-

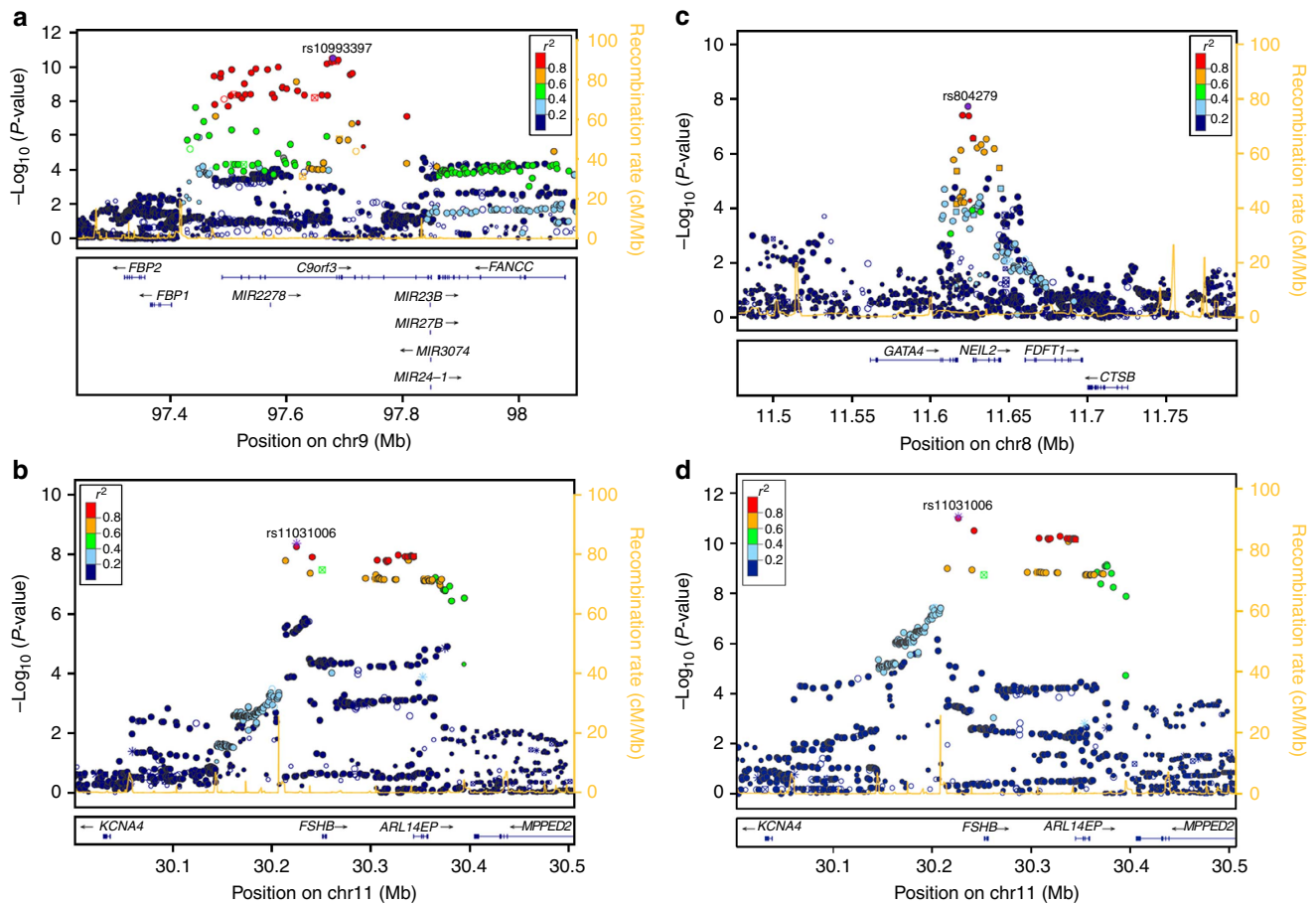


Figure 2 | Locuszoom plot of association results, linkage disequilibrium and recombination rates around the genome-wide significant loci. (a) Chr. 9 PCOS locus (*c9orf3*/*FANCC*), **(b)** Chr. 11 PCOS locus (*FSHB*/*ARL14EP*), **(c)** Chr. 8 PCOS locus (*GATA4*/*NEIL2*), **(d)** Chr. 11 LH locus (*FSHB*/*ARL14EP*). In each, the top panel reflects the meta-analysis results of the combined GWAS and replication phases. The LD estimates are colour coded as a heatmap from dark blue ($0 \geq r^2 > 0.2$) to red ($0.8 \geq r^2 > 1.0$). Recombination hotspots are indicated by the yellow lines (recombination rate in cM Mb⁻¹ from HapMap (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2008-03_rel22_B36/rates/)). The bottom panel shows the genes and their orientation for each region. For **(a–c)** PCOS loci, *P* values are from sample-size weighted two-strata meta-analysis of strata-specific logistic regression *P* values (Stage 1: 984 cases and 2,964 population control women; Stage 2: 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women). For **(d)** LH level locus, *P* values are from sample-size weighted two-strata meta-analysis of strata-specific linear regression *P* values (Stage 1: 645 PCOS cases; Stage 2: 399 PCOS cases).

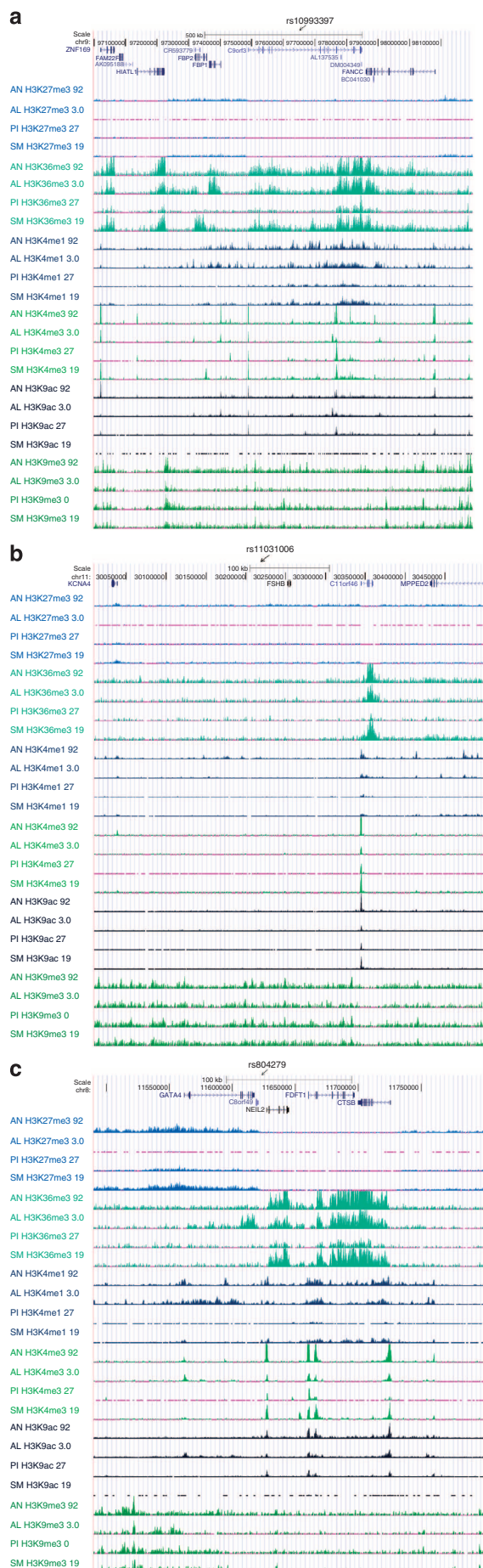
cycle. Trimming the model to only include individuals with LH < 25 mIU ml⁻¹ the association of rs11031006 with LH levels increased in significance (Stage 1 Discovery GWAS: Beta = 3.18, 0.53 s.e., mIU ml⁻¹, linear regression $P = 2.0 \times 10^{-9}$; Stage 2 Replication: Beta = 1.99, 0.42 s.e., mIU ml⁻¹, linear regression $P = 3.3 \times 10^{-6}$; sample-size weighted two-strata meta-analysis $P_{\text{meta}} = 3.4 \times 10^{-14}$), and the proportion of phenotypic variance explained was reduced to 2.3%.

The 7q33 T locus spans an 8.7 kb extensive LD region within *DGKI*. The most strongly associated SNP (rs706560) reached genome-wide significance in the Stage 1 Discovery GWAS (G allele, Beta = 33.29, 5.85 s.e., ng dl⁻¹, linear regression $P = 1.3 \times 10^{-8}$). The proportion of phenotypic variance explained by this SNP was 1.9%. The 10q26.11 FSH locus spans a 6 kb extensive LD region that includes *CASC2*. The most strongly associated variant (chr10:119956844:I) reached genome-wide significance in the Stage 1 Discovery GWAS (T allele, Beta = 0.27, 0.05 SE, log₁₀(mIU ml⁻¹), linear regression $P = 3.2 \times 10^{-8}$). The proportion of phenotypic variance explained by this variant was 0.1%. Neither of these SNPs passed imputation quality thresholds in the Stage 2 Metachip Replication cohort. Further, chr10:119956844:I was not significant in the Stage 3 UK Replication (TCA allele, Beta = 0.11, 0.47 s.e., ng dl⁻¹, linear

regression $P = 0.79$), and no longer reached genome-wide significance in the combined meta-analysis (sample-size weighted three-strata meta-analysis $P_{\text{meta-All}} = 7.3 \times 10^{-7}$). Likewise, rs706560 was not significant in the Stage 3 UK Replication (G allele, Beta = -0.71, 0.47 s.e., ng dl⁻¹, linear regression $P = 0.15$), and no longer reached genome-wide significance in the combined meta-analysis (sample-size weighted three-strata meta-analysis $P_{\text{meta-All}} = 6.4 \times 10^{-6}$).

Since BMI itself has a strong genetic component¹⁹, we explored how our results changed when BMI was removed from the model for the four genome-wide significant loci (Supplementary Data 5). The associations with all four loci (three with PCOS diagnosis, one with LH levels) were only modestly different (less than 1.3 log orders; two increase and two decrease, in linear regression significance) and all remained genome-wide significant. Further the odds ratios for the three PCOS loci changed only 0.01–0.02 in Phase 1 and 0.02–0.07 in Phase 2, which is equivalent to 5.9% change or less. The linear regression beta for the LH locus increased 11 and 22%, respectively, in Stage 1 and 2 when BMI is removed from the model, which suggests that BMI moderately exacerbates the effect of the 11p14.1 LH locus near *FSHB* on LH levels.

We performed an *in silico* look-up of the eQTLs for the three genome-wide significant loci using SCANDb (<http://>



www.scandb.org; Supplementary Data 6). The *c9orf3* locus did not have any eQTLs meeting nominal significance ($P < 1 \times 10^{-3}$) as calculated by the QTDT. The *GATA4/NEIL2* locus had one eQTL meeting nominal significance (*NMURI*; QTDT $P = 7.0 \times 10^{-5}$), and the *FSHB/ARL14EP* PCOS and LH level locus had four eQTLs meeting nominal significance (*C20orf20*, QTDT $P = 6.0 \times 10^{-5}$, *ARL6IP6*, QTDT $P = 7.0 \times 10^{-5}$; *CCDC69*, QTDT $P = 1.0 \times 10^{-4}$; and *AGPAT2*, QTDT $P = 1.0 \times 10^{-4}$). All of these were detected in European populations (CEU) only. None met a stringent threshold for genome-wide significance.

Transcriptional analysis. There was strong evidence for transcriptional activation as assayed by H3K36m3 and H3K4m3 signals in the chr 9q22.32 *c9orf3/FANCC* PCOS locus (Fig. 3a). Marks indicative of transcriptional activation localized to the 5' end of *FANCC* and to a ~100 kb region encompassing a 3' terminal transcript of *C9orf3* plus a number uncharacterized transcripts in all four ENCODE cell lines (adipose nuclei, adult liver, pancreatic islets and skeletal muscle). Consistent with these observations micro array analysis demonstrated robust *FANCC* expression in the liver, pancreas, and, at low levels, in skeletal muscle, and *c9orf3* expression in pancreatic islets and liver (GNF Expression Atlas 1 Human Data on Affy U95 Chips <http://genome.ucsc.edu>). The RNAseq tracks (<http://genome.ucsc.edu>; Fig. 4a) demonstrated virtually universal expression of *FANCC*, as well as *HIAT11*, although the latter is outside the bounds of extensive LD from the lead SNP in the region (Fig. 2a).

The chr 8p32.1 PCOS locus was transcriptionally active in adipose nuclei, adult liver, skeletal muscle and pancreatic islets. *NEIL2* and *FDFT1* had evidence for promoter activity via H3K4m3 and H3K9ac binding and evidence of actively transcribed regions via H3K36m3 signals (Fig. 3c). *FDFT1* was expressed at low levels in adipocytes and pancreatic islets and *GATA4* was expressed at low levels in liver and pancreatic islets in microarray analysis (GNF Expression Atlas 1 Human Data on Affy U95 Chips <http://genome.ucsc.edu>). RNAseq (Fig. 4c) showed absence of expression levels for *GATA4*, inconsistent expression levels of *NEIL2*, and high expression levels of *FDFT1* and *CTSB*, although the latter is outside the bounds of extensive LD from the lead SNP in the region (Fig. 2c).

The most promising functional candidate gene was the *FSHB/ARL14EP* PCOS and LH levels locus mapping to chr 11p14.1. *FSHB* showed no evidence for transcriptional activation in the selected ChipSeq ENCODE (Fig. 3b) or RNAseq cell lines (Fig. 4b), which do not include pituitary tissues. Microarray analysis detected *FSHB* expression predominantly in the pituitary, with low levels detected in adipocytes and the pancreas. Furthermore, *ARL14EP* had evidence for promoter activity (H3K4m3 and H3K9ac) in all four ENCODE cell lines and for actively transcribed genes (H3K36m3) in all cell lines except pancreatic islets. *ARL14EP* was expressed at universally high levels with RNAseq in the cell types shown in Fig. 4b but the microarray analysis detected *ARL14EP* only at low levels in

Figure 3 | ENCODE histone mark profile for genome-wide significant PCOS and LH level GWAS loci (<http://www.roadmapepigenomics.org/> data). Methylation marks in four PCOS relevant human tissue mapping to: (a) Chr. 9 PCOS locus (*c9orf3/FANCC*), (b) Chr. 11 PCOS and LH level locus (*FSHB/ARL14EP*) and (c) Chr. 8 PCOS locus (*GATA4/NEIL2*). Each is a schematic of the genomic region including diagrams of genes mapping to the region. ChipSeq signal tracks for each tissue (AN, adipose nuclei; AL, adult liver; PI, pancreatic islets; SM, skeletal muscle) for six histone marks (H3K27m3, H3K36m3, H3K4m1, H3K4m3, H3K9ac and H3K9m3) are labelled to the left of the panel.

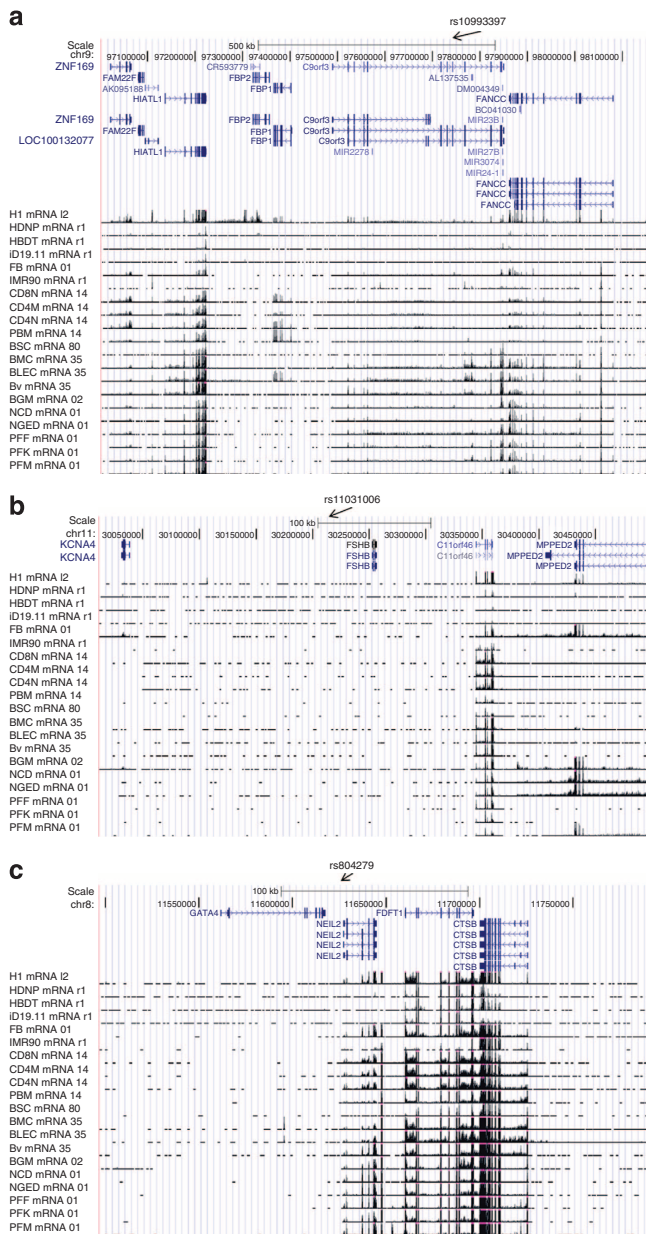


Figure 4 | ENCODE mRNA levels for genome-wide significant PCOS and LH level GWAS loci (<http://www.roadmapepigenomics.org/data>).

RNAseq levels in four PCOS relevant human tissue mapping to the (a) Chr. 9 PCOS locus (*c9orf3/FANCC*), (b) Chr. 11 PCOS and LH level locus (*FSHB/ARL14EP*) and (c) Chr. 8 PCOS locus (*GATA4/NEIL2*). Each is a schematic of the genomic region including diagrams of genes mapping to the region. Sources of RNA are labelled to the left of the panel.

adipocytes (GNF Expression Atlas 1 Human Data on Affy U95 Chips <http://genome.ucsc.edu>).

Histone mark and mRNA level ENCODE data for the unreplicated 7q33 T (*DGKI*) locus and 10q26.11 FSH (*CASC2*) locus are presented in Supplementary Figs. 8-9.

Discussion

PCOS is a complex genetic disease characterized by disordered gonadotropin secretion, hyperandrogenism, PCO and insulin resistance¹. Each of the biochemical features has been hypothesized to play a causal role in the development of PCOS^{1,4}. There are multiple affected phenotypes depending on the diagnostic criteria applied^{1,8}. We undertook this study to

investigate genetic loci associated with the NIH phenotype of PCOS that confers the highest risk for metabolic morbidities, including T2D and metabolic syndrome^{1,9}. Three loci reached genome-wide significance in the PCOS case-control meta-analysis, two novel loci, chr 8p32.1 *GATA4/NEIL2* and chr 11p14.1 *FSHB/ARL14EP*, and a chr 9q22.32 *c9orf3/FANCC* locus previously found in Chinese PCOS. The lead SNP at the same novel 11p14.1 *FSHB/ARL14EP* locus, rs11031006, was strongly associated with LH levels in addition to PCOS diagnosis. These findings suggest that genetic variation in FSH plays an important role in the aetiology of PCOS in European ancestry women with the NIH phenotype.

FSHB encodes the FSH beta polypeptide²⁰, a member of a family of glycoprotein hormones that includes LH, TSH and hCG, and consists of a common alpha subunit and a hormone-specific beta subunit²¹. FSH plays a central role in the regulation of ovarian folliculogenesis, which is disordered in PCOS², in part due to a relative decrease in FSH secretion. LH regulates theca cell T synthesis and increased ovarian androgen production in PCOS is LH-dependent¹. Adjusting for LH levels in the regression model between rs11031006 and PCOS negated the association (OR = 1.18, 1.02–1.37 95%CI, $P = 6.3 \times 10^{-3}$ without LH in the model; OR = 1.55, 1.01–2.39 95%CI, $P = 0.66$ with LH in the model; Supplementary Data 7), suggesting that the association with *FSHB* is mediated by LH. These findings suggest that variation in *FSHB* contributes to the cardinal gonadotropin secretory changes characteristic of PCOS²². Indeed, an *FSHB* promoter polymorphism has been associated with higher circulating LH and lower circulating FSH levels in men²³, analogous to characteristic changes in gonadotropin levels in PCOS¹. This polymorphism is also associated with reduced sperm counts²³. It remains possible that other genes in LD with the 11p14.1 signal, such as *ARL14EP*, which encodes ADP-ribosylation factor-like 14 effector protein that may function to control the movement of MHC class II-containing vesicles²⁴, contribute to PCOS pathogenesis. Additional fine-mapping and functional studies will be needed to resolve this issue.

The second novel locus associated with European PCOS was 8p32.1 in the region of *GATA4* and *NEIL2*. *GATA4* encodes a zinc-finger transcription factor that regulates gonadal development and the transcription of steroidogenic genes. Deletion of *GATA4* results in abnormal responses to exogenous gonadotropins and impaired fertility in mice²⁵. *GATA4* is expressed in PCOS ovarian follicles but this expression does not differ from control follicles²⁶. However, the expression of another member of this family, *GATA6*, is increased in PCOS theca cells²⁷. *NEIL2* encodes nei endonuclease VIII-like 2, a member of a class of DNA glycosylases involved in the repair of DNA damage²⁸. The 8p32.1 locus also encompasses the promoter region of *FDFT1* that encodes farnesyl-diphosphate farnesyl transferase, the first specific enzyme in the cholesterol biosynthesis pathway and thus also critical for testosterone biosynthesis. *FDFT1* is associated with nonalcoholic fatty liver disease²⁹.

We also replicated the 9q22.32 *c9orf3/FANCC* PCOS locus previously reported in Han Chinese¹⁵, although the Chinese and European loci were independent of one another (Supplementary Data 3). Multi-ethnic fine-mapping will be required to tease this apart further. One gene in this region, *c9orf3*, encodes a member of the M1 zinc aminopeptidase family, aminopeptidase O (AP-O)³⁰. AP-O is expressed in the pancreas, placenta, liver, testis and heart where it is postulated to play a role in proteolytic processing of biologically active peptides, for example, the cleavage of angiotensin III to generate angiotensin IV³⁰. *FANCC* encodes the protein for Fanconi anaemia complementation group C protein (FANCC) that functions in complexes to repair DNA cross-linking damage, and also suppresses apoptosis³¹.

Table 3 | Genetic association results of variants identified in previous PCOS GWAS.

Trait	SNP	Chr	Mb	Gene or Flanking Genes	Eff	Non-Eff	Asian GWAS			Stage 1 Discovery GWAS			Stage 2 Metabochip Replication					Stage 1 + 2 Meta-Analyses		
							OR	P value	Ref	EAF	OR	95%CI	P value	Gen/Imp	r ²	EAF	OR	95%CI	P value	Gen/Imp
PCOS	rs12468394	2	43.5	THADA	C	A	0.72	1.59 × 10 ⁻²⁰	14	0.5	0.88	0.80-0.98	7.61 × 10 ⁻⁰²	0.97	0.49	0.97	0.87-1.08	4.80 × 10 ⁻⁰³	Gen	1.51 × 10 ⁻⁰³
PCOS	rs12478601	2	43.7	THADA	C	T	0.72	3.48 × 10 ⁻²³	14	0.57	0.91	0.82-1.00	1.43 × 10 ⁻⁰¹	Gen	0.58	1	0.9-1.11	9.26 × 10 ⁻⁰²	1	2.76 × 10 ⁻⁰²
PCOS	rs13405728	2	48.9	LHCGR	A	G	0.71	7.55 × 10 ⁻²¹	14	0.06	0.74	0.59-0.94	3.65 × 10 ⁻⁰²	0.97	0.06	0.81	0.65-1.02	4.03 × 10 ⁻⁰²	0.93	3.50 × 10 ⁻⁰²
PCOS	rs2268361	2	49.2	FSHR	C	T	0.87	9.89 × 10 ⁻¹³	15	0.63	0.88	0.80-0.98	9.61 × 10 ⁻⁰²	Gen	0.6	0.84	0.76-0.94	1.76 × 10 ⁻⁰¹	0.93	3.19 × 10 ⁻⁰²
PCOS	rs4385527	9	97.6	C9orf3	G	A	0.84	5.87 × 10 ⁻⁹	15	0.41	0.77	0.70-0.86	1.53 × 10 ⁻⁰⁶	Gen	0.42	0.89	0.8-0.99	9.51 × 10 ⁻⁰⁴	Gen	6.39 × 10 ⁻⁰⁹
PCOS	rs3802457	9	97.7	C9orf3	G	A	0.77	5.28 × 10 ⁻¹⁴	15	0.02	0.65	0.45-0.95	3.50 × 10 ⁻⁰²	0.91	NA	NA	NA	NA	NA	3.50 × 10 ⁻⁰²
PCOS	rs10818854	9	126.4	DENND1A	G	A	1.51	9.40 × 10 ⁻¹⁸	14	0.05	1.26	1.02-1.56	1.77 × 10 ⁻⁰³	0.99	0.05	1.45	1.13-1.86	6.89 × 10 ⁻⁰²	0.95	3.66 × 10 ⁻⁰⁴
PCOS	rs10986105	9	126.5	DENND1A	T	G	1.47	6.90 × 10 ⁻¹⁵	14	0.05	1.3	1.04-1.63	4.05 × 10 ⁻⁰⁴	1	0.05	1.69	1.3-2.2	1.71 × 10 ⁻⁰³	0.98	2.30 × 10 ⁻⁰⁶
PCOS	rs1894116	11	102	YAP1	A	G	1.27	1.08 × 10 ⁻²²	15	0.1	1.15	0.97-1.35	1.46 × 10 ⁻⁰²	Gen	NA	NA	NA	NA	NA	1.46 × 10 ⁻⁰²
PCOS	rs705702	12	56.3	RAB5B/SUOX	A	G	1.27	8.64 × 10 ⁻²⁶	15	0.32	1.17	1.05-1.30	1.15 × 10 ⁻⁰²	0.98	0.32	0.97	0.86-1.09	6.09 × 10 ⁻⁰¹	0.96	2.36 × 10 ⁻⁰²
PCOS	rs2272046	12	66.2	HMG2	A	C	0.7	1.95 × 10 ⁻²¹	15	0.02	0.9	0.62-1.31	6.09 × 10 ⁻⁰¹	0.95	NA	NA	NA	NA	NA	6.09 × 10 ⁻⁰¹
PCOS	rs4784165	16	52.3	TOX3	T	G	1.15	3.64 × 10 ⁻¹¹	15	0.26	0.94	0.84-1.06	8.92 × 10 ⁻⁰¹	0.97	NA	NA	NA	NA	NA	8.92 × 10 ⁻⁰¹
PCOS	rs2059807	19	7.1	INSR	A	G	1.14	1.09 × 10 ⁻⁸	15	0.61	1.08	0.97-1.20	1.86 × 10 ⁻⁰¹	Gen	0.63	1.1	0.99-1.23	3.11 × 10 ⁻⁰¹	Gen	9.63 × 10 ⁻⁰²
PCOS	rs6022786	20	52.4	SUMO1P1	A	G	1.13	1.83 × 10 ⁻⁹	15	0.57	0.96	0.87-1.06	1.47 × 10 ⁻⁰¹	0.98	NA	NA	NA	NA	NA	1.47 × 10 ⁻⁰¹
PCOS	rs10505648	8	13.7	KHDRBS3	G	A	0.52	5.46 × 10 ⁻⁸	16	0.52	1.01	0.91-1.12	4.95 × 10 ⁻⁰¹	Gen	0.52	0.96	0.86-1.07	8.51 × 10 ⁻⁰¹	Gen	6.88 × 10 ⁻⁰¹

CI, confidence interval; EAF, effect allele frequency in cases and controls combined; Eff, effect allele; Gen, genotyped; Imp, Imputation r²; PCOS, polycystic ovary syndrome; P value, significance as assessed by logistic regression for Stage 1 and Stage 2; P_{meta}, significance as assessed by sample-size weighted two-strata meta-analyses (PCOS Stage 1: 984 cases & 2,964 population control women; PCOS Stage 2: 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women).

Further, we nominally replicated ($P < 0.05$), in the same direction, seven of the eleven previously identified Chinese PCOS SNPs (Table 3), including gonadotropin receptor genes *LH/CGR* and *FSHR*, as well as *THADA*, *DENND1A*, *YAP1*, and *RAB5B/SUOX*^{14,15}, exceeding the number of associations we would expect to replicate by chance alone. We did not replicate the previously identified 8q24.2 Korean PCOS locus (Table 3)¹⁶. However, this GWAS in Korean women with the Rotterdam PCOS phenotypes failed to replicate any of the European or Chinese PCOS loci at a significance level of 10^{-5} or higher¹⁶. The failure to replicate these loci could be due to a lack of power in the Korean study or to differences in the phenotypes studied, although both the Chinese and Korean GWAS used the Rotterdam diagnostic criteria.

Two quantitative hormone level loci, in addition to the 11p14.1 *FSHB/ARL14EP* LH locus, were genome-wide significant in the Stage 1 Discovery GWAS of quantitative hormone levels: the 7q33 T locus, in the region of *DGKI*, and the 10q26.11 FSH locus, in the region of *CASC2*. They failed quality control in the Stage 2 Metabochip Replication cohort and did not demonstrate significant associations in the Stage 3 UK Replication cohort. After inclusion of the Stage 2 and 3 replication data, the meta-analysis across all three strata no longer reached genome-wide significance for these two loci. Further, the locuszoom plot (Supplementary Fig. 5) for the 10q26.11 FSH locus (*CASC2*) indicates only a single variant in the region was strongly associated ($P < 10^{-4}$), potentially indicating poor imputation of this indel, although the imputation $r^2 > 0.9$. Therefore, these findings should be considered tentative, and will require further testing in additional cohorts. *DGKI*, diacylglycerol kinase iota, is a member of the type IV diacylglycerol kinase family³². *CASC2*, cancer susceptibility candidate 2, has been associated with endometrial cancer³³.

A GWAS of T and sex hormone binding globulin (SHBG) levels in men³⁴ found an association between variants in the *SHBG* gene and these quantitative hormone levels. We replicated the findings for rs12150660 (Supplementary Table 8) with both T and SHBG levels, but not at genome-wide significant thresholds. We could not attempt to replicate the associations with rs6258 as the genotype data for that SNP did not pass quality control thresholds in our study. As the mean and range of T levels in both PCOS and control women were much smaller (approximately one-tenth that in men), this may have limited the ability to detect associations identified in men³⁴.

Increased BMI is commonly associated with PCOS, particularly in US cohorts of European ancestry¹. Accordingly, we adjusted

for BMI in our model to reduce the chance of identifying BMI-related rather than PCOS-related loci. However, removing BMI from the model had minimal impact on our results. All four genome-wide significant loci (three PCOS diagnosis associated, one LH levels associated) were genome-wide significant with or without the inclusion of BMI as a covariate. Removing the BMI covariate led to evenly split number of increases and decreases in significance, and the effect sizes changed $< 6\%$. We also did not find any evidence for genome-wide significant associations with PCOS of variants in known obesity or type 2 diabetes genes. Collectively, these findings suggest that while BMI is associated with PCOS, genes involved in BMI have little impact on the genetics of PCOS.

The heritability of PCOS is $\sim 70\%$ in monozygotic twin studies¹². However, the portion of heritability accounted for by the three PCOS loci were small ($< 1\%$ each) as has been observed with other common complex diseases and traits³⁵. Although we now count fourteen replicated genome-wide significant PCOS loci across Chinese^{14,15} and European³⁶⁻³⁸ ancestry populations including those identified in this study, each has relatively small effect sizes that when taken collectively only account for a small portion of the heritability of PCOS as estimated from twin studies. Investigations of the genome-wide genetic architecture³⁹ of PCOS should increase this, but other variables (for example, structural variants, epigenetic factors, epistasis and so on) likely account for the remaining heritability.

We assessed the transcriptional potential of the three PCOS GWAS loci using three independent methods. Each of the three methods capture slightly different data on transcription potential and have different sensitivities to detect transcription^{40,41}. Data from the two sequencing-based data methods (ChIP-Seq and RNASeq) are more congruent with each other than either are with data from the expression arrays (GNF Expression Atlas 1 Human Data on Affy U95 Chips <http://genome.ucsc.edu>). Further, the results for any given tissue or cell line are derived from a single individual and thus do not take into account inter individual variation in gene expression (<http://genome.ucsc.edu>). With these caveats in mind, each of the genomic regions identified in our GWAS was shown to be transcriptionally active in at least one tissue that is affected in PCOS. These findings are consistent with these genomic regions harbouring PCOS susceptibility genes, although clearly further studies in tissues relevant to the PCOS reproductive phenotype are needed.

Our study was limited to women with the NIH phenotype of PCOS that is defined by endocrine features rather than PCOM.

We focused on this phenotype because it is the one associated with highest risk for insulin resistance and other metabolic disorders^{1,9}. Further, ~90% of women with NIH phenotype PCOS have PCOM^{1,42}. PCOM is also a common, age-related finding in otherwise reproductively normal women^{5,42}. Nevertheless, even in the absence of hormonal abnormalities, PCOM appear to have intrinsic abnormalities in folliculogenesis² and gonadotropin responsiveness⁶. PCOS cohorts diagnosed by Rotterdam criteria are enriched for the phenotypes with PCOM¹ and with less severe endocrine abnormalities^{1,8,9,43}. Accordingly, the inclusion of these additional phenotypes may have contributed to the differing results between the present and previous GWAS^{15,16}. In addition, the fact that we did not exclude PCOM in our control subjects, as was done in the previous Asian PCOS GWAS^{15,16}, may have limited our power to detect genetic variation associated with PCOM.

In conclusion, replicated GWAS in Chinese^{14,15} and European cohorts^{36–38} implicate genes modulating gonadotropin action, *LHCGR* and *FSHR*, and secretion, *FSHB*. These findings suggest that gonadotropins play an aetiologic role in PCOS pathogenesis, analogous to the insights into causative biologic pathways provided by T2D GWAS implicating β -cell genes and obesity GWAS implicating neuroendocrine genes modulating food intake⁴⁴. Further, these PCOS GWAS studies are complementary since the Chinese cohort contained additional Rotterdam phenotypes, whereas our cohort contained only the NIH phenotype with hyperandrogenism and chronic anovulation. The Chinese GWAS findings of associations with the genes encoding gonadotropin receptors, *LHCGR* and *FSHR*, suggest that ovarian gonadotropin action plays a role in all of the Rotterdam phenotypes, while our findings of associations of *FSHB* with PCOS and LH levels suggest that gonadotropin secretion drives the accompanying hormonal derangements of hyperandrogenism and anovulation. Further genetic analyses stratified by Rotterdam phenotype are warranted. In addition, there was a common 9q22.32 PCOS locus in the Chinese and European cohorts, suggesting shared susceptibility variants in these ethnically diverse populations. As human populations ancestral to Chinese and Europeans diverged from Africans ~60,000 years ago^{45,46}, the 9q22.32 region may harbour an evolutionarily conserved genetic susceptibility factor for PCOS⁴⁷. Studies in African ancestry PCOS populations should provide further insight into the evolutionary history of this disorder.

Methods

Subjects. Nine-hundred eighty-four PCOS cases and 2,964 population controls (Stage 1), followed by replication (Stage 2) in 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women were studied. An additional replication (Stage 3) of the top variant from each region with $P < 5 \times 10^{-6}$ ($N = 24$) was performed in 217 PCOS cases and 1,335 1958 British Birth Cohort (that is, controls) samples. The study was approved by the Institutional Review Board of Northwestern University Feinberg School of Medicine as well as by the Institutional Review Boards of the study investigators' institutions (Institutional Review Boards of Brigham and Women's Hospital, Carolinas Medical Centre, Cedars-Sinai Medical Centre, Pennsylvania State Milton S. Hershey Medical Centre, Magee-Womens Hospital, Medical College of Wisconsin, Massachusetts General Hospital, University of Pennsylvania, Stanford University, University of Alabama at Birmingham, University of Chicago, University of Colorado at Denver, University of Michigan, University of Medicine and Dentistry of New Jersey, University of Texas Health Science Centre, University of Texas Health Science Centre at San Antonio, University of Vermont, Virginia Commonwealth University, Wayne State University; the Research Ethics Committee of Athens University Medical School; and the North Thames Multicenter Research Ethics Committee). All subjects were Caucasians of European ancestry and gave written informed consent before study. Subjects providing samples for the NUGene DNA biorepository consented to the use of their deidentified clinical data and DNA samples by third party investigators⁴⁸.

PCOS subjects were ages 13–45 years (Table 1). Subjects in the discovery cohort (Supplementary Data 9) were recruited by the authors (A.D., R.S.L., D.A.E.) and the Reproductive Medicine Network (RMN) for the Pregnancy in Polycystic Ovary

Syndrome (PPCOSI) study⁴⁹. PCOS subjects (Supplementary Data 10) were recruited by the same authors, the RMN PPCOSII study⁵⁰, as well as by collaborations with PCOS research centres across the United States (C.K.W., M.O.G., R.A.) and in Greece (E.D.K., D.P.) for Stage 2 Replication and in the United Kingdom (T.B., M.M., S.F.) for Stage 3 Replication. PCOS cases fulfilled NIH criteria for PCOS¹. Cases fulfilling these criteria also meet the Rotterdam and Androgen Excess Society criteria for PCOS¹. The cases had hyperandrogenism (clinical and/or biochemical) and chronic anovulation (eight or fewer menses per year) with exclusion of other hyperandrogenic disorders in the differential diagnosis, such as non-classical adrenal 21-hydroxylase deficiency¹. Ovarian morphology was not a criterion for the diagnosis as it is a non-specific finding that does not correlate with the endocrine phenotype^{5,42}.

Control DNA samples for the discovery cohort came from women self-identified as 'white', age 18–97 years collected by the NUGene DNA bank⁴⁸. The samples were selected from women with data on body weight, height and age. Control women in the Stage 2 replication cohort were reproductively normal women, ages 15–45 years, with regular menses and no hirsutism who were enrolled by the following authors (A.D., R.S.L., C.K.W., E.D.K., D.P., M.O.G., R.A.) as control subjects for their PCOS studies; the control women had T levels within the normal range for the assay used by each author. Questionnaire-based reproductive histories, DNA and EDTA-plasma samples were obtained from 18- to 45-year-old women who were enrolled in the Metabolic Risk Complications of Obesity Genes project⁵¹. Women selected had regular menses and were not receiving oral contraceptive pills. They had normal total T levels determined using liquid chromatography-tandem mass spectrometry (LC-MS/MS)⁵². The details of recruitment and phenotyping of the discovery and replication cohort PCOS cases and the replication cohort control women have been reported previously (see references in Supplementary Data 11). Control women in the UK replication cohort (Stage 3) were from the 1958 British Birth Cohort (58BBC), and all were of the age 44–45 years at the time of DNA sample and phenotypic data collection (BMI and age).

Hormone level assays. The hormone levels used for subject phenotyping and in the quantitative trait analyses were measured as outlined in Supplementary Data 11. All analyses are adjusted for the study site at which the assay was performed, a central laboratory was used by the RMN and was, therefore, considered as one site. Further, some sites changed assay methodology so the analyses were also adjusted for the assay methodology. The normal range for each assay was established in the following number of premenopausal, reproductively normal, non-hirsute control women with regular menses: 98 Cedars Sinai Medical Centre⁵³; 100 University of Alabama at Birmingham⁵⁴; 108 University of Athens⁵⁵; 39 University of Chicago⁵⁶, and 75 United Kingdom¹¹.

The normal ranges for the PCOS Family Study were established as reported in 43 premenopausal reproductively normal women with 27–35 day menstrual cycles and no hirsutism at Penn State Milton S. Hershey Medical Centre¹⁰. Since 2006, the T assay for the PCOS Family Study has been performed at the University of Virginia Centre for Research in Reproduction Ligand Assay and Analysis Core using the same assay method (Supplementary Data 11). The normal ranges were validated in this assay and have been reassessed in an additional cohort of 209 premenopausal reproductively normal women with 27–35 day regular menses and no clinical symptoms of androgen excess to ensure that the range has not changed. The normal range for LC-MS/MS T assays used to select control women enrolled in the Metabolic Risk Complications of Obesity Genes project⁵¹ was also determined in this additional cohort of 209 control women. The assays for PPCOSI and PPCOSII were performed at the University of Virginia Centre for Research in Reproduction Ligand Assay and Analysis Core. Accordingly, the same normal range was used at these study sites as was used for the PCOS Family Study T assays performed in this core.

Stage 1 Discovery GWAS. PCOS cases and NUGene control DNA samples included in the GWAS phase were genotyped using the Illumina Omni Express (HumanOmniExpress-12v1_C) at the Centre for Inherited Disease Research (CIDR), Johns Hopkins University, Baltimore, MD.

GWAS Quality control (QC). Genotypic data that passed initial quality control at the genotyping centre were released to the PCOS study team who performed quality control procedures of the data following recommendations from the GENEVA⁵⁷ and eMERGE^{58,59} consortia, in which we were involved. Specifically, we removed poorly performing samples or SNPs based on misspecified sex, chromosomal anomalies, unintended sample duplicates, sample relatedness, low call rate, high number of Mendelian errors, departures from Hardy-Weinberg equilibrium, duplicate discordance, sex differences in heterozygosity and low minor allele frequencies as detailed in Supplementary Data 12 and 13.

Ancestry. To investigate population structure, we used principal components analysis (PCA) in EIGENSTRAT (<http://genenpath.med.harvard.edu/~reich/Software.htm>), essentially as described by Price *et al.*⁶⁰ We used PCA for two purposes: to identify population group outliers and to include as covariates in the statistical model used for association testing to account for population structure. Initially, we

analysed all unduplicated PCOS GWAS study samples separately, along with HapMap (Phase 3 CEU, CHB, JPT and YRI) to detect of population outliers among the study. We started with a set of 703,171 autosomal SNPs with missing call rate < 1%. From this pool of overlapping SNPs, we selected a subset through two rounds of LD pruning using the pair-wise genotypic correlation method in PLINK⁶¹ (<http://pngu.mgh.harvard.edu/~purcell/plink/>). In the first round, we removed SNPs in short-range LD using a window size of 50 SNPs with a five SNP offset and an r^2 threshold of 0.8. The second round of pruning removed SNPs in long-range LD using a window size equal to the median number. The r^2 threshold was set at 0.8 for the second round of pruning, after which 76,602 SNPs remained that were also found in HapMap. After exclusion of outlier subjects (34 GWAS phase and 37 replication phase samples > 3 s.d. values from the median of the first two principal components), we used the first two eigenvectors from the results in these analyses as covariates in the association tests to adjust for any possible population structure among the study subjects (Supplementary Fig. 10). The genomic inflation parameter λ_{GC1000} (calculated following de Bakker, *et al.*⁶² was sufficiently small (< 1.05 in each of the GWAS performed with the different phenotypes) to suggest that population stratification was adequately controlled in our model and that there was no inflation of false positives from this potential confounder.

Imputation. We performed genotype imputation using IMPUTE2⁶³ (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) and the 1000 Genomes reference panel⁶⁴. We used a 'cosmopolitan' 1000 Genomes reference panel for imputation consisting of an intersection of the AFR, EUR, and ASN panels. We first used the strand-checking utility of SHAPEIT v2 (ref. 65) (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html) to ensure consistent strand assignments between the reference data set and the QC cleaned and filtered data sets, and we subsequently corrected strand and/or removed SNPs where strandedness could not be resolved. We used a conservative allelic r^2 threshold of 0.9 to remove questionable imputed SNPs.

Association tests. The genotype call probabilities from the filtered IMPUTE2 output were used in a logistic regression model between PCOS case-control status and the genotypes probabilities under an additive model adjusting for BMI, and the first three principal components of ancestry. Similarly, we used a linear regression model between each of the quantitative reproductive hormone levels (treated as continuous variables) and the genotype probabilities under an additive model adjusting for BMI, the first three principal components of ancestry, assay site and assay method. FSH and SHBG were log-transformed to normalize before conducting the association tests. We used the frequentist approach in SNPTEST⁶⁶ v2.2.0 (https://mathgen.stats.ox.ac.uk/genetics_software/snpstest/snpstest.html) or ProbABEL⁶⁷ v0.4.3 to estimate the odds ratio and 95% confidence intervals (PCOS case/control status), or beta and s.e. values (hormone levels) for each regression model and assess the significance of the association between the SNP and the phenotype of interest.

Stage 2 Metachip replication. We attempted replication our top associations in a second set of 1,799 PCOS cases and 1,231 phenotyped controls by adding 12,921 SNPs (those with $P < 1 \times 10^{-5}$ in the GWAS phase, as well as candidate SNPs for ancillary studies) as custom content to the Metachip⁶⁸. Genotyping was performed at the Broad Institute Centre for Genotyping and Analysis (CGA), Cambridge, MA, who released 9,893 of the 12,921 add-on SNPs attempted. Genotyping QC, assessing ancestry, imputation to 1,000 genomes, and association tests followed that of the GWAS phase described above. To assess population substructure in the Stage 2 Replication, we analysed all unduplicated PCOS replication study samples separately, along with HapMap (Phase 3 CEU, CHB, JPT and YRI) to detect of population outliers among the study. We started with a set of 197,415 autosomal SNPs with missing call rate < 2%. From this pool of overlapping SNPs, we selected a subset through two rounds of LD pruning using the pair-wise genotypic correlation method in PLINK. In the first round we remove SNPs in short-range LD using a window size of 50 SNPs with a five SNP offset and an r^2 threshold of 0.8. The second round of pruning removed SNPs in long-range LD using a window size equal to the median number. The r^2 threshold was set at 0.8 for the second round of pruning after which 53,828 SNPs remained that were also found in HapMap. After exclusion of outlier subjects (> 2 s.d. values from the median of the first two principal components) we used the first two eigenvectors from the results in these analyses as covariates in the association tests to adjust for any possible population structure among the study subjects (Supplementary Fig. 11).

Stage 3 (UK) Replication. The Illumina Core Exome and Illumina Human 1.2 Million Duo Custom v1_A platforms were used to genotype the UK cases and controls, respectively. Overlapping variants ($N = 191,563$) were included in the quality control (QC) and imputation phases. As cases and controls were typed separately, stringent QC thresholds were applied before and after merging the case-control data. Pre-imputation QC included exclusion filters: (i) HWE P value < 0.0001 in cases and controls separately, and HWE P value < 0.01 in the merged data; (ii) SNP call rate < 95% in cases and controls separately; (iii) sample

call rate < 95% in cases and controls separately; (iv) differential SNP missingness between cases and controls in the merged data, P value < 0.01; (v) s.e. > 10 in pre-imputation logistic regression analysis; (vi) MAF < 1% in the final clean merged data. In addition, related individuals and ethnic outliers were detected and excluded using PLINK pairwise IBD estimation and multidimensional scaling analysis (MDS).

Pre-phasing and imputation were performed using SHAPEIT and IMPUTE2 with the 1000 Genomes March 2012 (v3) reference panel, respectively. SNPTEST was used for association analyses. Case-control and quantitative analyses included C1 from the PLINK MDS analysis as a covariate to adjust for population stratification. Variants with info score < 0.4, s.e. < 0 and s.e. > 10, and P -value < 0 were excluded from the association results.

Meta-analysis. The betas and s.e. values were combined across the GWAS and replication cohorts using meta-analysis under a fixed effects model weighting each strata by sample size. METAL⁶⁹ (<http://www.sph.umich.edu/csg/abecasis/metal/index.html>) calculates a z-statistic, which summarizes the magnitude and direction of effect for the association of a reference allele selected at each marker. After aligning the SNPTEST output from each of the four cohorts to the same reference allele, a weighted sum of the individual cohort results is used to calculate an overall z-statistic and P value. The square root of the cohort specific sample size is used as the proportional weight, and these squared weights sum to 1.

Predicting transcription activity of the GWAS loci. To investigate the transcriptional potential of the primary GWAS loci identified in the dichotomous and quantitative trait analyses, we examined the histone mark patterns and mRNA levels for each locus using ENCODE ChIP-Seq data, RNAseq (<http://genome.ucsc.edu>) and GNF Expression Atlas 1 Human Data on Affy U95 Chips (<http://genome.ucsc.edu>). A broad tissue distribution was tested for the expression array and the RNAseq. Multiple tissues contribute to the PCOS phenotype. Although there are no ChIP-Seq ENCODE results from any ovarian tissues, PCOS is also associated with substantial metabolic abnormalities¹. These abnormalities include defects in adipocyte, skeletal muscle and hepatic insulin action as well as pancreatic β -cell dysfunction¹. Our analysis included on metabolic tissues affected in PCOS (adipose nuclei, adult liver, pancreatic islets and skeletal muscle).

We present chromatin Immunoprecipitation sequencing (ChIP-Seq) results from Encyclopaedia of DNA Elements (ENCODE) Consortium Roadmap Epigenomics Consortium [(<http://www.roadmapepigenomics.org/data>) for six histone marks (H3K27m3, H3K36m3, H3K4m1, H3K4m3, H3K9ac and H3K9m3) in four human tissues that are potentially relevant to PCOS (adipose nuclei, adult liver, pancreatic islets and skeletal muscle)]. H3K27m3 histone mark is the tri-methylation of lysine 27 of the H3 histone protein and is associated with facultatively expressed genes (www.epigenetek.com). H3K36m3, tri-methylation of lysine 36 of H3, is associated with actively transcribed genes (www.epigenetek.com). H3K4m1, mono-methylation of lysine 4 of H3, is associated with enhancers and with DNA regions downstream of transcription starts (www.epigenetek.com). H3K4m3, tri-methylation of lysine 4 of H3, is associated with promoters that are active or poised to be activated (www.epigenetek.com). H3K9ac, acetylation of lysine 9 of H3, is found in actively transcribed promoters (www.epigenetek.com). H3K9m3, tri-methylation of lysine 9 of H3, in contrast is found in constitutively repressed genes (www.epigenetek.com). We considered a methylation mark to active if there is more than one peak with a height greater than $5 \times$ times background height.

References

1. Diamanti-Kandaraki, E. & Dunaif, A. Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. *Endocr. Rev.* **33**, 981–1030 (2012).
2. Franks, S., Stark, J. & Hardy, K. Follicle dynamics and anovulation in polycystic ovary syndrome. *Hum. Reprod. Update* **14**, 367–378 (2008).
3. Eagleson, C. A. *et al.* Polycystic ovarian syndrome: evidence that flutamide restores sensitivity of the gonadotropin-releasing hormone pulse generator to inhibition by estradiol and progesterone. *J. Clin. Endocrinol. Metab.* **85**, 4047–4052 (2000).
4. Abbott, D. H., Dumesic, D. A. & Franks, S. Developmental origin of polycystic ovary syndrome - a hypothesis. *J. Endocrinol.* **174**, 1–5 (2002).
5. Johnstone, E. B. *et al.* The polycystic ovary post-Rotterdam: a common, age-dependent finding in ovulatory women without metabolic significance. *J. Clin. Endocrinol. Metab.* **95**, 4965–4972 (2010).
6. Fauser, B. C., Diedrich, K. & Devroey, P. Predictors of ovarian response: progress towards individualized treatment in ovulation induction and ovarian stimulation. *Hum. Reprod. Update* **14**, 1–14 (2008).
7. Rotterdam ESHRE ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil. Steril.* **81**, 19–25 (2004).
8. Azziz, R. *et al.* Positions statement: criteria for defining polycystic ovary syndrome as a predominantly hyperandrogenic syndrome: an Androgen Excess Society guideline. *J. Clin. Endocrinol. Metab.* **91**, 4237–4245 (2006).

9. Fauser, B. C. *et al.* Consensus on women's health aspects of polycystic ovary syndrome (PCOS): the Amsterdam ESHRE/ASRM-Sponsored 3rd PCOS Consensus Workshop Group. *Fertil. Steril.* **97**, 28–38.e25 (2012).
10. Legro, R. S., Driscoll, D., Strauss, 3rd J. F., Fox, J. & Dunaif, A. Evidence for a genetic basis for hyperandrogenemia in polycystic ovary syndrome. *Proc. Natl Acad. Sci. USA* **95**, 14956–14960 (1998).
11. Franks, S. *et al.* Ovarian morphology is a marker of heritable biochemical traits in sisters with polycystic ovaries. *J. Clin. Endocrinol. Metab.* **93**, 3396–3402 (2008).
12. Vink, J. M., Sadrzadeh, S., Lambalk, C. B. & Boomsma, D. I. Heritability of polycystic ovary syndrome in a Dutch twin-family study. *J. Clin. Endocrinol. Metab.* **91**, 2100–2104 (2006).
13. Urbanek, M., Sam, S., Legro, R. S. & Dunaif, A. Identification of a polycystic ovary syndrome susceptibility variant in fibrillin-3 and association with a metabolic phenotype. *J. Clin. Endocrinol. Metab.* **92**, 4191–4198 (2007).
14. Chen, Z. J. *et al.* Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nat. Genet.* **43**, 55–59 (2011).
15. Shi, Y. *et al.* Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nat. Genet.* **44**, 1020–1025 (2012).
16. Lee, H. *et al.* Genome-wide association study identified new susceptibility loci for polycystic ovary syndrome. *Hum. Reprod.* **30**, 723–731 (2015).
17. Cui, L. *et al.* Genotype-phenotype correlations of PCOS susceptibility SNPs identified by GWAS in a large cohort of Han Chinese women. *Hum. Reprod.* **28**, 538–544 (2013).
18. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
19. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
20. Watkins, P. C. *et al.* DNA sequence and regional assignment of the human follicle-stimulating hormone beta-subunit gene to the short arm of human chromosome 11. *DNA* **6**, 205–212 (1987).
21. Albanese, C. *et al.* The gonadotropin genes: evolution of distinct mechanisms for hormonal control. *Recent Prog. Horm. Res.* **51**, 23–58 discussion 59–61 (1996).
22. Rebar, R. *et al.* Characterization of the inappropriate gonadotropin secretion in polycystic ovary syndrome. *J. Clin. Invest.* **57**, 1320–1329 (1976).
23. Tuttleman, F. *et al.* Combined effects of the variants FSHB -211G>T and FSHR 2039A>G on male reproductive parameters. *J. Clin. Endocrinol. Metab.* **97**, 3639–3647 (2012).
24. Paul, P. *et al.* A Genome-wide multidimensional RNAi screen reveals pathways controlling MHC class II antigen presentation. *Cell* **145**, 268–283 (2011).
25. LaVoie, H. A. The GATA-keepers of ovarian development and folliculogenesis. *Biol. Reprod.* **91**, 38 (2014).
26. Sander, V. A. *et al.* Alterations of folliculogenesis in women with polycystic ovary syndrome. *J. Steroid Biochem. Mol. Biol.* **124**, 58–64 (2011).
27. Wood, J. R., Ho, C. K. M., Nelson-Degrave, V. L., McAllister, J. M. & Strauss, J. F. The molecular signature of polycystic ovary syndrome (PCOS) theca cells defined by gene expression profiling. *J. Reprod. Immunol.* **63**, 51–60 (2004).
28. Bandaru, V., Sunkara, S., Wallace, S. S. & Bond, J. P. A novel human DNA glycosylase that removes oxidative DNA damage and is homologous to Escherichia coli endonuclease VIII. *DNA Repair* **1**, 517–529 (2002).
29. Chalasani, N. *et al.* Genome-wide association study identifies variants associated with histologic features of nonalcoholic fatty liver disease. *Gastroenterology* **139**, 1567–1576 (2010).
30. Diaz-Perales, A. *et al.* Identification of human aminopeptidase O, a novel metalloprotease with structural similarity to aminopeptidase B and leukotriene A4 hydrolase. *J. Biol. Chem.* **280**, 14310–14317 (2005).
31. Huang, F. *et al.* The Fanconi anemia group C protein interacts with uncoordinated 5 A and delays apoptosis. *PLoS ONE* **9**, e92811 (2014).
32. Ding, L., Traer, E., McIntyre, T. M., Zimmerman, G. A. & Prescott, S. M. The cloning and characterization of a novel human diacylglycerol kinase, DGK α . *J. Biol. Chem.* **273**, 32746–32752 (1998).
33. Baldinu, P. *et al.* Identification of a novel candidate gene, CASC2, in a region of common allelic loss at chromosome 10q26 in human endometrial cancer. *Hum. Mutat.* **23**, 318–326 (2004).
34. Ohlsson, C. *et al.* Genetic determinants of serum testosterone concentrations in men. *PLoS Genet.* **7**, e1002313 (2011).
35. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
36. Welt, C. K. *et al.* Variants in DENND1A are associated with polycystic ovary syndrome in women of European ancestry. *J. Clin. Endocrinol. Metab.* **97**, E1342–E1347 (2012).
37. Mutharasan, P. *et al.* Evidence for chromosome 2p16.3 polycystic ovary syndrome susceptibility locus in affected women of European ancestry. *J. Clin. Endocrinol. Metab.* **98**, E185–E190 (2013).
38. Goodarzi, M. O. *et al.* Replication of association of DENND1A and THADA variants with polycystic ovary syndrome in European cohorts. *J. Med. Genet.* **49**, 90–95 (2012).
39. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
40. Ombrello, M. J., Sikora, K. A. & Kastner, D. L. Genetics, genomics, and their relevance to pathology and therapy. *Best Pract. Res. Clin. Rheumatol.* **28**, 175–189 (2014).
41. Costa, V., Aprile, M., Esposito, R. & Ciccociola, A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.* **21**, 134–142 (2013).
42. Legro, R. S. *et al.* Polycystic ovaries are common in women with hyperandrogenic chronic anovulation but do not predict metabolic or reproductive phenotype. *J. Clin. Endocrinol. Metab.* **90**, 2571–2579 (2005).
43. Diamanti-Kandarakis, E. & Panidis, D. Unravelling the phenotypic map of polycystic ovary syndrome (PCOS): a prospective study of 634 women with PCOS. *Clin. Endocrinol.* **67**, 735–742 (2007).
44. Hirschhorn, J. N. Genomewide association studies-illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).
45. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
46. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
47. Azziz, R., Dumesic, D. A. & Goodarzi, M. O. Polycystic ovary syndrome: an ancient disorder? *Fertil. Steril.* **95**, 1544–1548 (2011).
48. McCarty, C. A. *et al.* The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).
49. Legro, R. S. *et al.* The Pregnancy in Polycystic Ovary Syndrome study: baseline characteristics of the randomized cohort including racial effects. *Fertil. Steril.* **86**, 914–933 (2006).
50. Legro, R. S. *et al.* Letrozole versus clomiphene for infertility in the polycystic ovary syndrome. *N. Engl. J. Med.* **371**, 119–129 (2014).
51. Kissebah, A. H. *et al.* Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the metabolic syndrome. *Proc. Natl Acad. Sci. USA* **97**, 14478–14483 (2000).
52. Torchen, L. C., Fogel, N. R., Brickman, W. J., Papanicolaou, R. & Dunaif, A. Persistent apparent pancreatic beta-cell defects in premenarchal PCOS relatives. *J. Clin. Endocrinol. Metab.* **99**, 3855–3862 (2014).
53. Knochenhauer, E. S. *et al.* Prevalence of the polycystic ovary syndrome in unselected black and white women of the southeastern United States: a prospective study. *J. Clin. Endocrinol. Metab.* **83**, 3078–3082 (1998).
54. Salameh, W. A. *et al.* Specificity and predictive value of circulating testosterone assessed by tandem mass spectrometry for the diagnosis of polycystic ovary syndrome by the National Institutes of Health 1990 criteria. *Fertil. Steril.* **101**, 1135–1141.e2 (2014).
55. Diamanti-Kandarakis, E. *et al.* A survey of the polycystic ovary syndrome in the Greek island of Lesbos: hormonal and metabolic profile. *J. Clin. Endocrinol. Metab.* **84**, 4006–4011 (1999).
56. Rosenfeld, R. L., Mortensen, M., Wroblewski, K., Littlejohn, E. & Ehrmann, D. A. Determination of the source of androgen excess in functionally atypical polycystic ovary syndrome by a short dexamethasone androgen-suppression test and a low-dose ACTH test. *Hum. Reprod.* **26**, 3138–3146 (2011).
57. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
58. Turner, S. *et al.* Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* Chapter 1, Unit1–19 (2011).
59. Zuvich, R. L. *et al.* Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet. Epidemiol.* **35**, 887–898 (2011).
60. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
61. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
62. deBakker, P. I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
63. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
64. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
65. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
66. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

67. Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134 (2010).
68. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).
69. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* **26**, 2190–2191 (2010).

Acknowledgements

This study was supported by US National Institutes of Health (NIH) grants R01 HD057223 (A.D.), P50 HD044405 (A.D.), U54 HD34449 (A.D.), R01 HD057450 (M.U.), and R01 HD056510 (R.S.L.), R01 HL075079 (D.A.E.), P50 HD057796 (D.A.E.), R01 HD065029 (C.K.W.), R01 HD29364 (R.A.), K24 HD01346 (R.A.), R01 DK071895 (A.H.K.), R01 DK65598 (A.H.K.), Wellcome Trust grants 090532 (M.I.M.), 098381 (M.I.M.) and 086596 (C.M.L.), MRC Programme grants G9710020 (S.F. and M.I.M.) and G0802782 (S.F.), as well as The University of Chicago Diabetes Research and Training Centre DK20595 (D.A.E.), Award 1-10-CT-57 from the American Diabetes Association (C.K.W.), and by TOPS (Take Off Pounds Sensibly) Club, Inc.. Partial funding for the clinical studies was provided by UL1 TR000150, UL1 RR033184, UL1 TR000430 and UL1 RR025758 from the National Centre for Advancing Translational Sciences. This study makes use of GWAS data generated by the Wellcome Trust Case-Control consortium 2 (WTCCC2) UK controls from the 1958 British Birth Cohort. The NIH Cooperative Multicenter Reproductive Medicine Network includes the following Members of the Steering Committee and Grant Numbers: Grants U10 HD27049 (Christos Coutifaris, University of Pennsylvania, Philadelphia, PA), U10 HD38992 (Richard S. Legro, Penn State College of Medicine, Hershey, PA), U10 HD055925 (Heping Zhang, Yale University, New Haven, CT), U10 HD39005 (Michael P Diamond, Wayne State University, Detroit, MI), U10 HD38998 (William D. Schlaff, University of Colorado, Denver, CO), U10 HD055936 (Gregory M. Christman, University of Michigan, Ann Arbor, MI), U10 HD055942 (Robert G Bryzski, University of Texas at San Antonio, San Antonio, TX), and U10 HD055944 (Peter R Casson, University of Vermont, Burlington VT), Chair of the Steering Committee: Nanette Santoro, University of Colorado, Denver CO and Program Officer, Esther Eisenberg, Eunice Kennedy Shriver

NICHD.) Some hormone assays were performed at the University of Virginia Centre for Research in Reproduction Ligand Assay and Analysis Core that is supported by U54 HD28934 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development.

Author contributions

M.G.H., M.U., D.A.E., R.S.L. and A.D. conceived and designed the study. M.G.H., L.L.A., J.Y.L. R.S., T.K. and C.M.L. performed experiments and statistical analyses. A.D., D.A.E., R.S.L., T.B., M.I.M., S.F., C.K.W., E.D.-K., D.P., M.O.G., R.A., RMN, Y.Z., R.G.J., M.O., A.H.K. and E.S.-V. recruited study subjects and measured or analysed phenotypic data. M.G.H., M.U. and A.D. wrote the paper. All authors critically reviewed and approved the paper.

Additional information

Accession codes: Genotype data have been deposited in dbGaP under the accession code phs000368.v1.p1.

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Hayes, M. G. *et al.* Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nat. Commun.* **6**:7502 doi: 10.1038/ncomms8502 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

Reproductive Medicine Network

Ruben Alvero²¹, Huiman X. Barnhart²², Valerie Baker²³, Kurt T. Barnhart²⁴, G. Wright Bates²⁵, Robert G. Brzycki²⁶, Bruce R. Carr²⁷, Sandra A. Carson²⁸, Peter Casson²⁹, Nicholas A. Cataldo²³, Gregory Christman³⁰, Christos Coutifaris²⁴, Michael P. Diamond³¹, Esther Eisenberg³², Gabriella G. Gosman³³, Linda C. Giudice³⁴, Daniel J. Haisenleder³⁵, Hao Huang³⁶, Stephen A. Krawetz³¹, Scott Lucidi³⁷, Peter G. McGovern³⁸, Evan R. Myers³⁹, John E. Nestler⁴⁰, Dana Ohl³⁰, Nanette Santoro²¹, William D. Schlaff²¹, Peter Snyder²⁴, Michael P. Steinkampf²⁵, J.C. Trussell⁴¹, Rebecca Usadi⁴², Qingshang Yan³⁶ & Heping Zhang³⁶

²¹Department of Obstetrics and Gynecology, University of Colorado, Aurora, Colorado 80045, USA. ²²Department of Biostatistics & Bioinformatics, Duke University Medical Centre, Durham, North Carolina 27710, USA. ²³Department of Obstetrics and Gynecology, Stanford University Medical Centre, Stanford, California 94304, USA. ²⁴Department of Obstetrics and Gynecology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA. ²⁵Department of Obstetrics and Gynecology, University of Alabama, Birmingham, Alabama 35249, USA. ²⁶Department of Obstetrics and Gynecology, University of Texas Health Science Centre, San Antonio, Texas 78229, USA. ²⁷Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ²⁸Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, Texas 77030, USA. ²⁹Department of Obstetrics and Gynecology, University of Vermont, Burlington, Vermont 05495, USA. ³⁰Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, Michigan 48109, USA. ³¹Department of Obstetrics and Gynecology, Wayne State University, Detroit, Michigan 48201, USA. ³²Fertility and Infertility Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Rockville, Maryland 20852, USA. ³³Department of Obstetrics Gynecology, and Reproductive Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. ³⁴Department of Obstetrics Gynecology, and Reproductive Sciences, University of California at San Francisco, San Francisco, California 94143, USA. ³⁵Ligand Core Laboratory, University of Virginia Center for Research in Reproduction, Charlottesville, Virginia 22908, USA. ³⁶Department of Biostatistics, Yale University School of Public Health, New Haven, Connecticut 06520, USA. ³⁷Department of Obstetrics and Gynecology, Virginia Commonwealth University, Richmond, Virginia 23298, USA. ³⁸Department of Obstetrics, Gynecology and Women's Health, University of Medicine and Dentistry of New Jersey, Newark, New Jersey 07103, USA. ³⁹Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, North Carolina 27710, USA. ⁴⁰Department of Internal Medicine, Virginia Commonwealth University School of Medicine, Richmond, Virginia 23298, USA. ⁴¹Department of Urology, State University of New York Upstate Medical University, Onondaga, New York 13210, USA. ⁴²Carolinas Medical Center, Charlotte, North Carolina 28204, USA.