Phenotypic clustering reveals distinct subtypes of polycystic ovary syndrome with novel genetic associations

Matthew Dapas[1], Frederick T. J. Lin[1], Girish N. Nadkarni[2], Ryan Sisk[1], Richard S. Legro[3], Margrit Urbanek[1,4,5], M. Geoffrey Hayes[1,4,6¶*], Andrea Dunaif[7,¶*]

[1] Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

[2]Division of Nephrology, Icahn School of Medicine at Mount Sinai, New York, NY

[3] Department of Obstetrics and Gynecology, Penn State College of Medicine, Hershey, PA

[4] Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

[5] Center for Reproductive Science, Northwestern University Feinberg School of Medicine, Chicago, IL

[6] Department of Anthropology, Northwestern University, Evanston, IL

[7] Division of Endocrinology, Diabetes and Bone Disease, Icahn School of Medicine at Mount Sinai, New York, NY

[¶] These authors jointly supervised this work.

* Corresponding author

E-mail:andrea.dunaif@mssm.edu (AD)

ghayes@northwestern.edu (MGH)

DISCLOSURE STATEMENT: The authors have nothing to disclose.

Abstract

Background

Polycystic ovary syndrome (PCOS) is a common, complex genetic disorder affecting up to 15% of reproductive age women worldwide, depending on the diagnostic criteria applied. These diagnostic criteria are based on expert opinion and have been the subject of considerable controversy. The phenotypic variation observed in PCOS is suggestive of an underlying genetic heterogeneity, but a recent meta-analysis of European ancestry PCOS cases found that the genetic architecture of PCOS defined by different diagnostic criteria was generally similar, suggesting that the criteria do not identify biologically distinct disease subtypes. We performed this study to test the hypothesis that there are biologically relevant subtypes of PCOS.

Methods and Findings

Unsupervised hierarchical cluster analysis was performed on quantitative anthropometric, reproductive, and metabolic traits in a genotyped discovery cohort of 893 PCOS cases and an ungenotyped validation cohort of 263 PCOS cases. We identified two PCOS subtypes: a "reproductive" group (21-23%) characterized by higher luteinizing hormone (LH) and sex hormone binding globulin (SHBG) levels with relatively low body mass index (BMI) and insulin levels; and a "metabolic" group (37-39%), characterized by higher BMI, glucose, and insulin levels with lower SHBG and LH levels. We performed a GWAS on the genotyped cohort, limiting the cases to either the reproductive or metabolic subtypes. We identified alleles in four novel loci that were associated with the reproductive subtype at genome-wide significance (*PRDM2/KAZN1*, $P=2.2\times10^{-10}$; *IQCA1*, $P=2.8\times10^{-9}$; *BMPR1B/UNC5C*, $P=9.7\times10^{-9}$; *CDH10,* $P=1.2\times10^{-8}$) and one locus that was significantly associated with the

metabolic subtype (*KCNH7/FIGN*, P=1.0×10^{-8}). We have previously reported that rare variants in *DENND1A*, a gene regulating androgen biosynthesis, were associated with PCOS quantitative traits in a family-based whole genome sequencing analysis. We classified the reproductive and metabolic subtypes in this family-based PCOS cohort and found that the subtypes tended to cluster in families and that carriers of rare *DENND1A* variants were significantly more likely to have the reproductive subtype of PCOS. Limitations of our study were that only PCOS cases of European ancestry diagnosed by NIH criteria were included, the sample sizes for the subtype GWAS were small, and the GWAS findings were not replicated.

Conclusions

In conclusion, we have found stable reproductive and metabolic subtypes of PCOS. Further, these subtypes were associated with novel susceptibility loci. Our results suggest that these subtypes are biologically relevant since they have distinct genetic architectures. This study demonstrates how precise phenotypic delineation can be more powerful than increases in sample size for genetic association studies.

Introduction

Understanding the genetic architecture of complex diseases is a central challenge in human genetics (1-3). Often defined according to arbitrary diagnostic criteria, complex diseases can represent the phenotypic convergence of numerous genetic etiologies (4-8). Recent studies in type 2 diabetes support the concept that there are disease subtypes with distinct genetic architecture (7,8). Identifying and addressing genetic heterogeneity in complex diseases could increase power to detect causal variants and improve treatment efficacy (9).

Polycystic ovary syndrome (PCOS) is a highly heritable, complex genetic disorder affecting up to 15% of reproductive-age women worldwide, depending on the diagnostic criteria applied (10). It is characterized by a variable constellation of reproductive and metabolic abnormalities (**Fig 1**). It is the leading cause of anovulatory infertility and a major risk factor for type 2 diabetes (T2D) in women (11). Despite these substantial morbidities, the etiology(ies) of PCOS remains unknown (12). Accordingly, the commonly-used diagnostic criteria for PCOS, the National Institutes of Health (NIH) criteria (13) and the Rotterdam criteria (14,15), are based on expert opinion, rather than mechanistic insights, and are designed to account for the diverse phenotypic presentations of PCOS. The NIH criteria require the presence of hyperandrogenism (HA) and chronic oligo/anovulation or ovarian dysfunction (OD) (13). The Rotterdam criteria include polycystic ovarian morphology (PCOM) and require the presence of at least two of these three key reproductive traits, resulting in three different affected phenotypes: HA and OD with or without PCOM, also known as NIH PCOS, as well as two additional non-NIH Rotterdam phenotypes, HA and PCOM, and OD and PCOM.

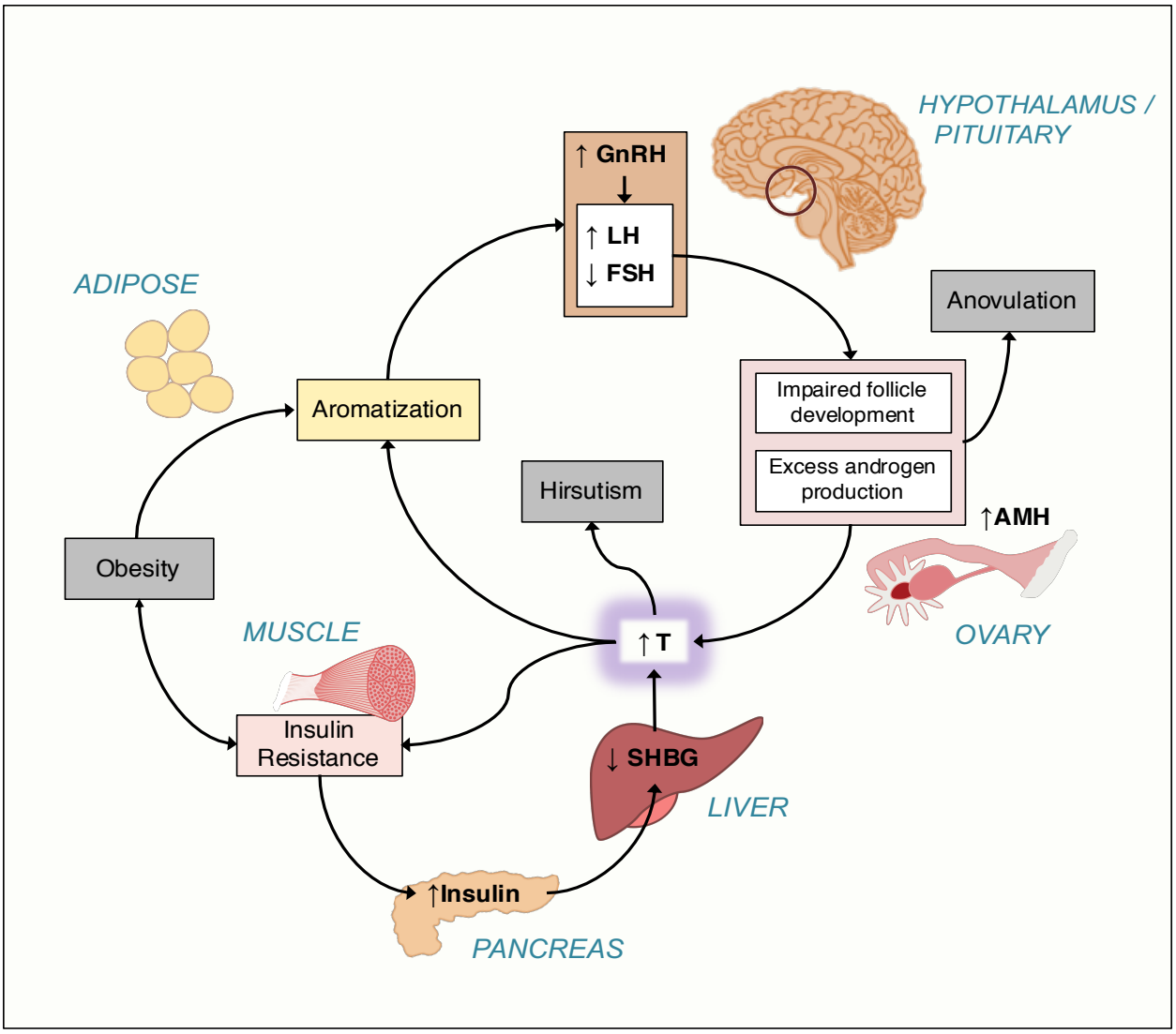


Fig 1. Pathophysiology of PCOS (reviewed in (10)). There is increased frequency of pulsatile gonadotropin-releasing hormone (GnRH) secretion from the arcuate nucleus of the hypothalamus that selectively increases luteinizing hormone (LH) secretion. LH stimulates ovarian theca cell testosterone (T) production. T is incompletely aromatized to estradiol by the adjacent granulosa cells because of relative follicle stimulating hormone (FSH) deficiency. There are constitutive increases in the activity of multiple steroidogenic enzymes in theca cells from women with PCOS, which contributes to increased T production. Increased adrenal androgen production may also be present in PCOS. T acts in the periphery to produce clinical signs of androgen excess, such as hirsutism, acne, and alopecia. T and androstenedione can also be aromatized extragonadally to estradiol and estrone, respectively, resulting in unopposed estrogen action on the endometrium. T feeds back on the hypothalamus to decrease the sensitivity to the normal feedback effects of estradiol and progesterone to slow GnRH pulse frequency. Anti-Müllerian hormone (AMH) levels are frequently increased in

PCOS; this hormone is secreted by small, growing preantral follicles, which are increased in PCOS. Recent studies suggest AMH acting through its cognate receptor on GnRH neurons in the arcuate nucleus contributes to the pathogenesis of PCOS (16,17).

PCOS is often associated with profound insulin resistance due to a unique defect in post-binding insulin-mediated signal transduction. Insulin is a co-gonadotropin that acts in synergy with LH to amplify the reproductive abnormalities of PCOS. Insulin signaling in the hypothalamus also appears to be important for ovulation. Insulin is a major negative regulator of hepatic synthesis of sex hormone-binding globulin (SHBG), the specific transport protein for T; only T which is not bound to SHBG is biologically active.

Genomewide association studies (GWAS) have considerably advanced our understanding of the pathophysiology of PCOS. These studies have implicated gonadotropin secretion (18) and action (19,20), androgen biosynthesis (19-21), metabolic regulation (21,22) and ovarian aging (22) in PCOS pathogenesis. A recent meta-analysis (21) of GWAS was the first study to investigate the genetic architecture of the diagnostic criteria. Only one of 14 PCOS susceptibility loci identified was significantly more strongly associated with the NIH phenotype compared to non-NIH Rotterdam phenotypes or to self-reported PCOS. These findings suggested that the genetic architecture of the phenotypes defined by the different PCOS diagnostic criteria was generally similar. Therefore, the current diagnostic criteria do not appear to identify genetically distinct disease subtypes.

It is possible to identify physiologically relevant complex disease subtypes through cluster analysis of phenotypic traits (7,23,24). Indeed, there have been previous efforts to subtype PCOS using unsupervised cluster analysis of its hormonal and anthropometric traits (25-28). However, there has been no validation that the resulting PCOS subtypes were biologically meaningful by testing their association with genetic variants, with other independent biomarkers, or with outcomes, such as therapeutic responses. In this study, we sought to 1) identify phenotypic subtypes of PCOS using an unsupervised clustering

129 approach on reproductive and metabolic quantitative traits from a large cohort of women
130 with PCOS, 2) validate those subtypes in a replication cohort, and 3) test whether the
131 subtypes thus identified were associated with distinct common genetic variants. As an
132 additional validation, we investigated the association of the subtypes with rare genetic
133 variants we recently identified in a family-based PCOS cohort (29).
134

135 # Methods

136 ## Subjects

137 This study used biochemical and genotype data from the previously published PCOS
138 GWAS, Hayes and Urbanek et al., 2015 (18), in which a discovery cohort (Stage 1) of 984
139 PCOS cases and 2,964 population controls were studied, followed by a replication cohort
140 (Stage 2) of 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women.
141 All cases were of European ancestry and each subject provided written informed consent
142 prior to the study (18). PCOS cases were ages 13-45 years and were diagnosed according to
143 the NIH criteria (10) of hyperandrogenism and chronic anovulation (eight or fewer menses
144 per year), excluding specific disorders of the adrenals, ovaries, or pituitary (30). Cases
145 fulfilling the NIH criteria also meet the Rotterdam criteria for PCOS (10). Two additional
146 PCOS cohorts were included in the present study who fulfilled the NIH criteria and were
147 phenotyped according to the same methods as the genotyped GWAS cohort. An independent,
148 ungenotyped cohort of 263 women with PCOS was used for clustering replication. A family-
149 based whole-genome sequencing cohort of 73 women with PCOS was used for investigating
150 subtype clustering in families and for rare variant analysis (29).

Population-based control DNA samples for the GWAS Stage 1 cohort were obtained from the NUgene biobank (31) from women of European ancestry, ages 18-97 years. Control women in the Stage 2 cohort were phenotyped reproductively normal women of European ancestry, ages 15–45 years, with regular menses and normal T levels, and who were not receiving contraceptive steroids for at least 3 months prior to study (18). T, DHEAS, SHBG, luteinizing hormone (LH), follicle-stimulating hormone (FSH), fasting glucose (Glu0), and fasting insulin (Ins0) levels were measured as previously reported (18).

Clustering

Clustering was performed in PCOS cases on eight adjusted quantitative traits: BMI, T, DHEAS, Ins0, Glu0, SHBG, LH, and FSH. There were 893 combined cases from both stages with complete quantitative trait data available for clustering (**Table S1**). Quantitative trait values were first $\log_e$-normalized and adjusted for age and assay method, which varied according to the different study sites where samples were collected (18), using a linear regression. An inverse normal transformation was then applied for each trait to ensure equal scaling. The normalized trait residuals were clustered using unsupervised, agglomerative, hierarchical clustering according to a generalization of Ward's minimum variance method (32,33) on Manhattan distances between trait values. Differences in adjusted, normalized trait values between subtypes were assessed using Kruskal-Wallis and pairwise Wilcoxon rank-sum tests corrected for multiple testing (Bonferroni). Cluster stability was assessed by computing the mean Jaccard coefficient from a repeated nonparametric bootstrap resampling (n=1000) of the dissimilarity matrix (34). Jaccard coefficients below 0.5 indicate that a cluster does not capture any discernable pattern within the data, while a mean

coefficient above 0.6 indicates that the cluster reflects a real pattern within the data (34). Cluster reproducibility was further assessed by repeating the clustering procedure in an independent cohort of 263 PCOS cases.

Association Testing

Stage 1 samples were genotyped using the Illumina OmniExpress (HumanOmniExpress-12v1_C) single nucleotide polymorphism (SNP) array. Stage 2 samples were genotyped using the Metabochip (35) with added custom variant content based on ancillary studies and the discovery results (18). The Stage 2 association replication in this study was therefore limited; many of the loci from Stage 1 were therefore not characterized in Stage 2. Low quality genotypic data were removed as described previously (18). SNPs were filtered according to minor allele frequency (MAF $\geq$0.01), Hardy-Weinberg equilibrium ($p \geq 1\times10^{-6}$), call rate ($\geq$0.99), minor allele count (MAC>5), Mendelian concordance, and duplicate sample concordance. Only autosomal SNPs were considered. Ancestry was evaluated using a principal component analysis (PCA) (36) on 76,602 LD-pruned SNPs (18). Samples with values >3 standard deviations from the median for either of the first two principal components (PCs) were excluded (34 in discovery; 37 in replication). Genotype data was phased using ShapeIT (v2.r790) (37) and then imputed to the 1000 Genomes reference panel (Phase3 v5) (38) using Minimac3 via the Michigan Imputation Server (39). Imputed SNPs with an allelic r^2 below 0.8 were removed from analysis.

Association testing was performed separately for Stage 1 and Stage 2 cohorts. Of the 893 combined cases from both stages included in the clustering analysis, 555 were from the Stage 1 cohort and 338 were from the Stage 2 cohort. 2,964 normal controls were used in

Stage 1, and 1,134 were used in Stage 2. Logistic regressions were performed using SNPTEST (40) for case-control status under an additive genetic model, adjusting for BMI and first three PCs of ancestry. P-values are reported as P_1 and P_2 for Stage 1 and Stage 2, respectively. Cases were limited to specific subtypes selected from clustering results. The betas and standard errors were combined across Stage 1 and Stage 2 cohorts for each subtype under a fixed meta-analysis model weighting each strata by sample size (41). Association test outputs were aligned to the same reference alleles and weighted z-scores were computed for each SNP. The square root of the cohort-specific sample size was used as the proportional weight. Meta-analysis P-values (P_{meta}) were adjusted for genomic inflation. Associations with P-values < 1.67×10^{-8} were considered statistically significant, based on the standard $P<5\times10^{-8}$ used in conventional GWAS adjusted for the three independent association tests performed.

Chromatin interactions

Neighboring chromatin interactions were investigated in intergenic loci using high-throughput chromatin conformation capture (Hi-C) data from the 3DIV database (42). Topologically associating domains (TADs) were identified using TopDom (43) with a window size of 20.

Identifying subtypes in PCOS families

Quantitative trait data from the affected women (n=73) in the family-sequencing cohort (29) were adjusted and normalized as described above. Subtype classifiers were modeled on the adjusted trait values and cluster assignments from the genotyped cohort. Several classification methods were compared using 10-fold cross-validation, including

support vector machine, random forest (RF), Gaussian mixed-model, and quadratic discriminant analysis (44). The classifier with the lowest error rate was then applied to the affected women in the family-sequencing cohort to identify subtypes of PCOS in the family data. Some of the probands from the family-based cohort were included in our previous GWAS (18). Therefore, there was some sample overlap between the training and test cohorts: of the 893 genotyped women used to identify the original subtype clusters, 47 were also probands in the family-based cohort. Differences between subtypes in the proportion of women with *DENND1A* rare variants were tested using the chi-square test of independence.

Results

PCOS subtypes

The clustering revealed two distinct phenotypic subtypes: 1) a group (23%) characterized by higher LH and SHBG levels with relatively low BMI and Ins0 levels, which we designated "reproductive", and 2) a group (37%) characterized by higher BMI and Glu0 and Ins0 levels with relatively low SHBG and LH levels, which we designated "metabolic" (**Fig 2**). The key traits distinguishing the reproductive and metabolic subtypes were BMI, insulin, SHBG, glucose, LH, and FSH, in order of importance according to relative pairwise Wilcoxon rank-sum test statistics (**Fig 3**). The remaining cases (40%) demonstrated no distinguishable pattern regarding their relative phenotypic trait distributions and were designated "indeterminate". The reproductive and metabolic subtypes clustered along opposite ends of the SHBG vs. Ins0/BMI axis, which was highly correlated with the first PC of the adjusted quantitative traits (**Fig 4**). The reproductive subtype was the most stable cluster, with a mean bootstrapped Jaccard coefficient ($\bar{\gamma}_C$) of 0.61, followed by the metabolic

subtype with $\bar{\gamma}_C$=0.55. The indeterminate group did not appear to capture any discernable pattern within the data ($\bar{\gamma}_C$=0.41) and was both overlapping and intermediate between the reproductive and metabolic subtypes on the SHBG vs. Ins0/BMI axis.

The clustering procedure was then repeated in an independent, non-genotyped cohort of 263 NIH PCOS cases diagnosed according to the same criteria as the genotyped cohort. The clustering yielded similar results, with a comparable distribution of reproductive (26%, $\bar{\gamma}_C$=0.57), metabolic (39%, $\bar{\gamma}_C$=0.46), and indeterminate clusters (35%, $\bar{\gamma}_C$=0.40) (**Fig 5**).

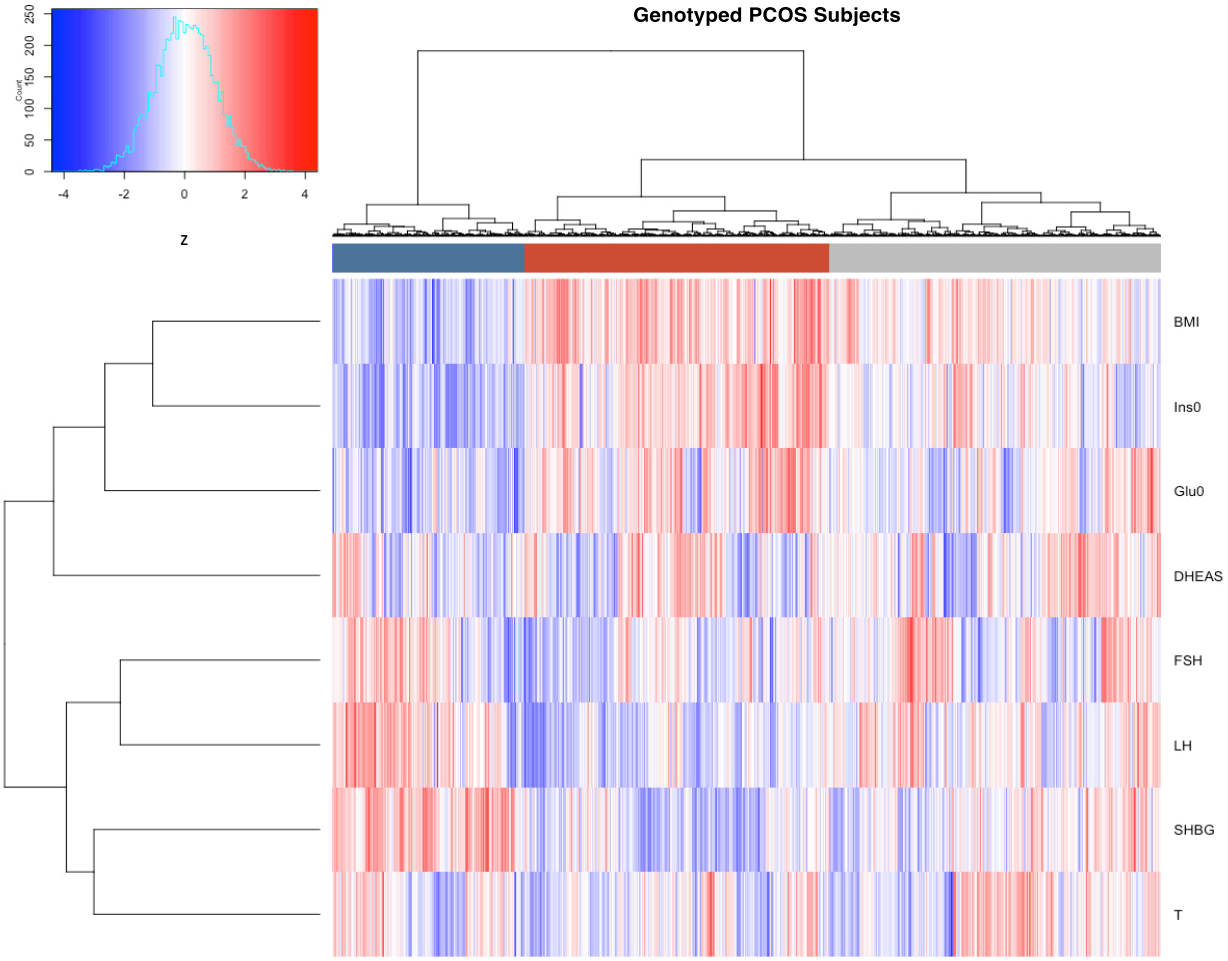


Fig 2. Hierarchical clustering of genotyped PCOS cases. Hierarchical clustering of 893 genotyped PCOS cases according to adjusted quantitative traits revealed two distinct phenotypic subtypes: a “reproductive” cluster, and a “metabolic” cluster; the remaining cases were designated as “indeterminate”. The reproductive, metabolic, and indeterminate clusters are shown in the color bar as dark blue, dark red, and grey, respectively. Heatmap colors correspond to trait z-scores, as shown in the frequency histogram where red indicates high values and blue indicates low values for each trait. BMI, body mass index; SHBG, sex hormone binding globulin; DHEAS, dehydroepiandrosterone sulfate; Glu0, fasting glucose; Ins0, fasting insulin; LH, luteinizing hormone; FSH, follicle-stimulating hormone.

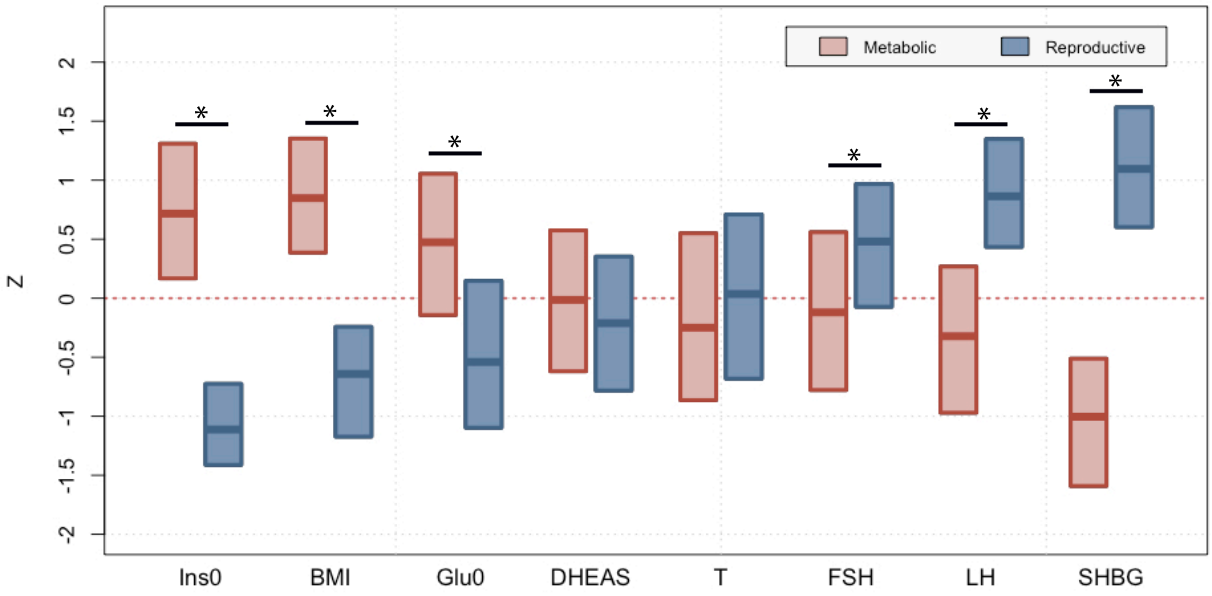


Fig 3. Phenotypic trait distributions in reproductive and metabolic subtypes. Median and interquartile ranges are shown for normalized, adjusted quantitative trait distributions of genotyped PCOS cases with reproductive or metabolic subtype. The figure illustrates the traits for which the subtypes differ significantly with an asterisk (*Bonferroni adjusted Wilcoxon, $P_{adj}<0.05$): Ins0, BMI, Glu0, FSH, LH and SHBG. BMI, body mass index; SHBG, sex hormone binding globulin; DHEAS, dehydroepiandrosterone sulfate; Glu0, fasting glucose; Ins0, fasting insulin; LH, luteinizing hormone; FSH, follicle-stimulating hormone.

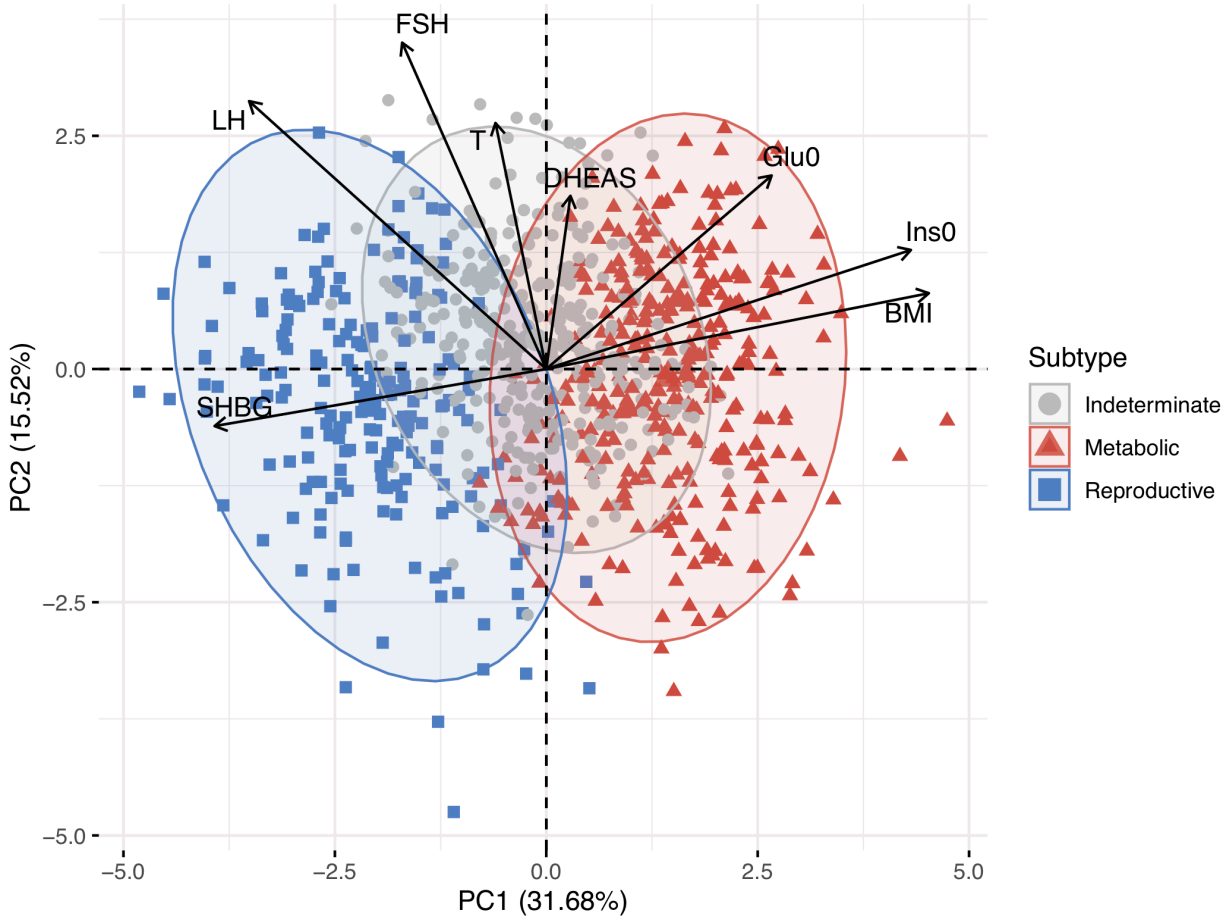
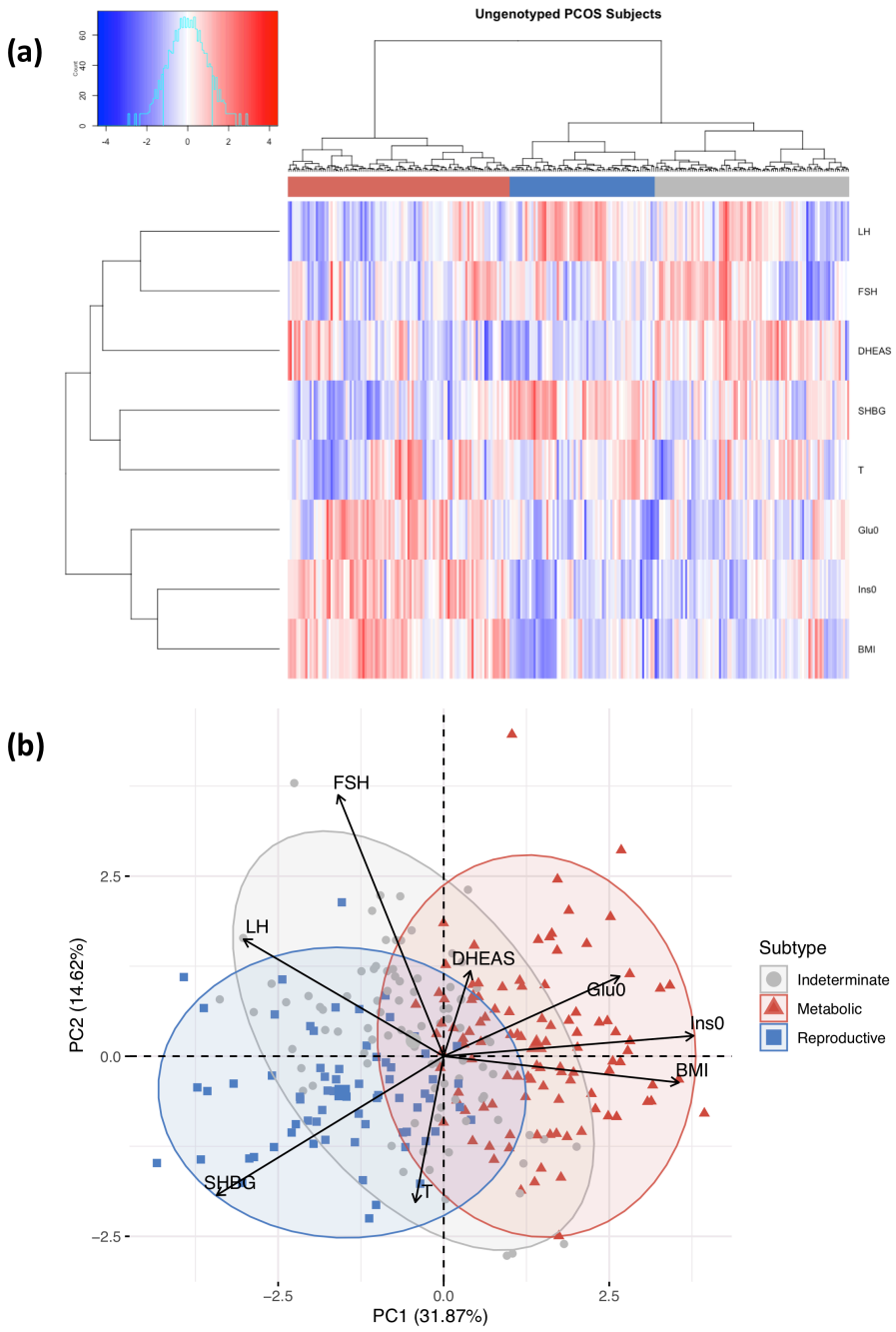


Fig 4. PCA plot of quantitative traits for genotyped PCOS cases. Genotyped PCOS cases are plotted on the first two principal components of the quantitative trait data and colored according to their identified subtype. Subtype clusters are shown as 95% concentration ellipses, assuming bivariate normal distributions. The relative magnitude and direction of trait correlations with the principal components are shown with black arrows.



277

278 **Fig 5. Clustering of ungenotyped PCOS cases.** (a) Hierarchical clustering of 263
279 ungenotyped PCOS cases according to adjusted quantitative traits replicate
280 reproductive (blue), metabolic (red), and unclassified (grey) clusters. Heatmap colors
281 correspond to trait z-scores. (b) PCA plot of ungenotyped PCOS cases replicate results
282 from genotyped cases. (a) Hierarchical clustering of 263 ungenotyped PCOS cases
283 according to adjusted quantitative traits replicate reproductive (blue), metabolic (red), and
284 unclassified (grey) clusters. Heatmap colors correspond to trait z-scores. (b) PCA plot of
285 ungenotyped PCOS cases replicate results from genotyped cases.

Subtype genetic associations

Genome-wide association testing identified alleles in four novel loci that were associated with the reproductive PCOS subtype at genome-wide significance (chr1 p36.21 *PRDM2/KAZN*, $P=2.23\times10^{-10}$; chr2 q37.3 *IQCA1*, $P=2.76\times10^{-9}$; chr4 q22.3 *BMPR1B/UNC5C*, $P=9.71\times10^{-9}$; chr5 p14.2-p14.1 *CDH10,* $P=1.17\times10^{-8}$) and one novel locus that was significantly associated with the metabolic subtype (chr2 q24.2-q24.3 *KCNH7/FIGN*, $P=1.03\times10^{-8}$). Association testing on the indeterminate cases replicated the 11p14.1 *FSHB* locus from our original GWAS (18) (**Table 1; Figs 6 and 7**).

The strongest association signal with the reproductive subtype appeared in an intergenic region of 1p36.21 579kb downstream of the *PRMD2* gene and 194kb upstream from the *KAZN* gene (**Fig 8a**). The lead SNP in the locus (rs78025940; OR=4.75, 2.82-7.98 95%CI, $P_1=2.16\times10^{-10}$, $P_{meta}=2.23\times10^{-10}$) was imputed ($r^2 = 0.91$) in Stage 1 only. The SNP was not genotyped in Stage 2. The lead genotyped SNP in the locus (rs16850259) was also associated with the reproductive subtype with genome-wide significance ($P_{meta}=2.14\times10^{-9}$) and was genotyped only in Stage 1 (OR=5.57, 3.24-9.56 95% CI, $P_1=2.08\times10^{-9}$). In ovarian tissue, the SNPs appear to be centrally located within a large 2Mb TAD stretching from the *FHAD1* gene to upstream of the *PDPN* gene (**Fig 9**).

The 2q37.3 locus spanned a 50kb region of strong LD overlapping the 5' end and promoter region of the *IQCA1* gene (**Fig 8b**). The SNP rs76182733 had the strongest association in this locus ($P_{meta}=2.76\times10^{-9}$) with the reproductive subtype. The signal was genotyped only in Stage 1 (OR=5.68, 3.00-10.78 95%CI, $P_1=2.69\times10^{-9}$) and was imputed with an imputation r^2 value of 0.84.

Table 1. Genome-wide significant associations with PCOS subtypes

					Stage 1 (Discovery)						Stage 2 (Replication)						
Chr	**Mb**	**Variant**	**Gene(s)**	**EA**	**EAF**	**β**	**OR**	**95% CI**	**P**	**Imp r^2**	**EAF**	**β**	**OR**	**95% CI**	**P**	**Imp r^2**	**P_{meta}**
1	*14.7*	rs78025940	*PRDM2/ KAZN*	A	0.02	3.02	4.75	2.82-7.98	2.16×10^{-10}	0.91	-		-	-		-	2.23×10^{-10}
2	*237.4*	rs76182733	*IQCA1*	G	0.01	3.79	5.68	3.00-10.78	2.67×10^{-9}	0.84	-		-	-		-	2.76×10^{-9}
4	*96.1*	rs17023134	*BMPR1B/ UNC5C*	G	0.05	1.62	3.02	2.06-4.42	1.40×10^{-8}	0.87	0.06	0.61	1.71	0.98-2.99	7.81×10^{-2}	0.83	9.71×10^{-9}
5	*24.7*	rs7735176	*CDH10*	A	0.01	3.80	5.09	2.62-9.86	1.14×10^{-8}	0.93	-		-	-		-	1.17×10^{-8}
2	*164.2*	rs55762028	*KCNH7/ FIGN*	C	0.01	5.05	1.86	0.92-3.75	9.17×10^{-9}	0.96	-		-	-		-	1.03×10^{-8}
11	*30.3*	rs10835638	*FSHB*	T	0.16	0.78	1.81	1.44-2.27	3.13×10^{-8}	0.98	0.17	0.77	2.01	1.49-2.70	2.67×10^{-5}	0.97	4.94×10^{-12}

Variant information and association statistics are shown for the most strongly associated SNP in each significant locus. Reproductive subtype loci are highlighted in blue, metabolic loci in red, indeterminate loci in grey. EA = effect allele; EAF: effect allele frequency in cases and controls combined; β = effect size from association regression; OR = odds ratio; CI = confidence interval; Imp r^2= imputation r^2 for imputed SNPs; P = stage-specific significance as assessed by logistic regression; P_{meta} = significance as assessed by sample-size weighted two-strata meta-analysis, adjusted for genomic inflation factor. Cases and controls by stage: Stage 1 = 201 metabolic, 123 lean, 231 indeterminate, 2964 controls; Stage 2 = 128 metabolic, 84 lean, 126 indeterminate, 1134 controls. NOTE: Not all SNPs were genotyped or imputed in both stages.

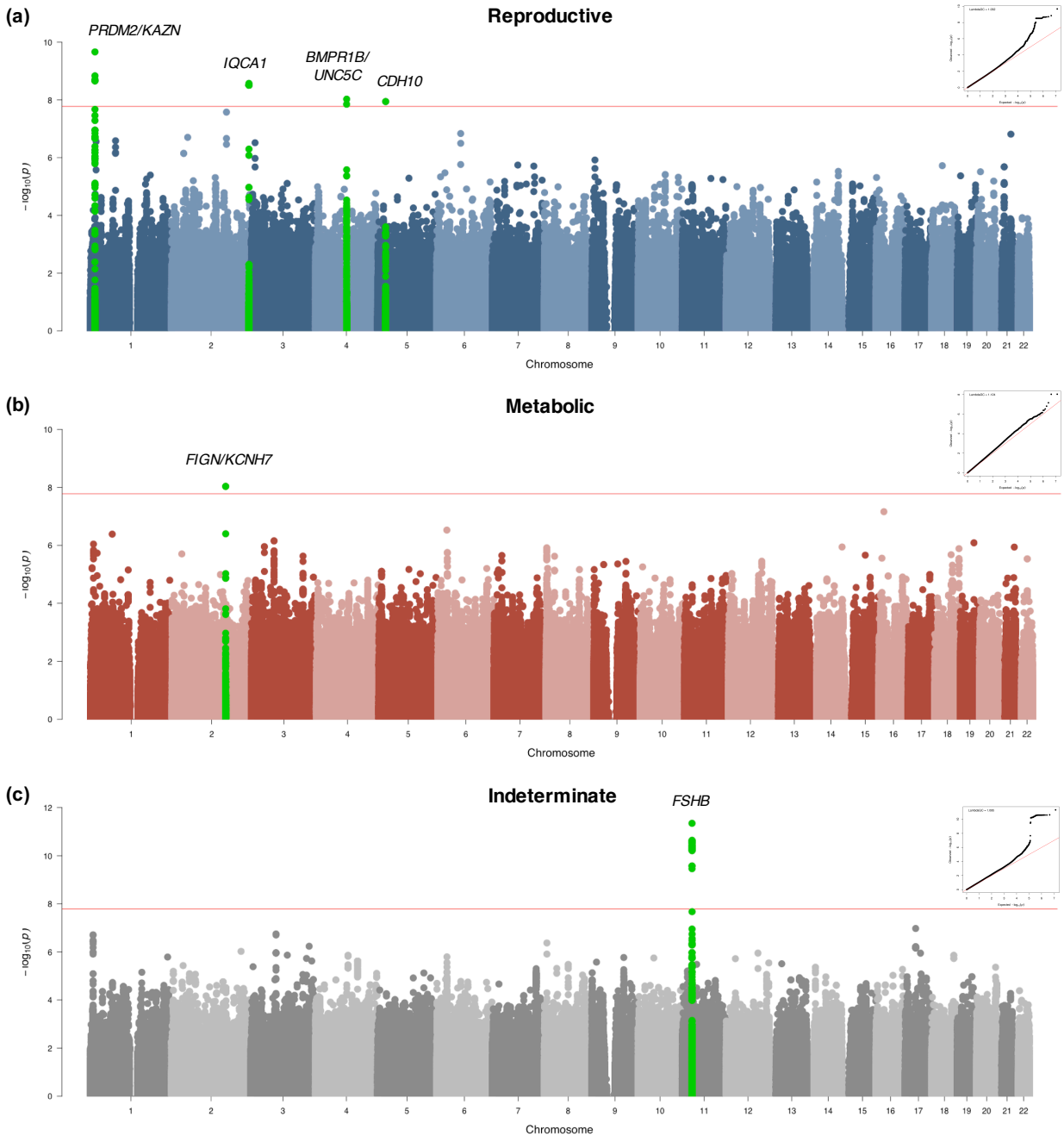


Fig 6. Genome-wide association results. Manhattan plots for (a) reproductive and (b) metabolic PCOS subtypes. The red horizontal line indicates genome-wide significance ($p \leq 2.5\times10^{-8}$). Genome-wide significant loci are colored in green and labeled according to nearby gene(s). QQ plots with genomic inflation factor, λ_{GC}, are shown adjacent to corresponding plots.

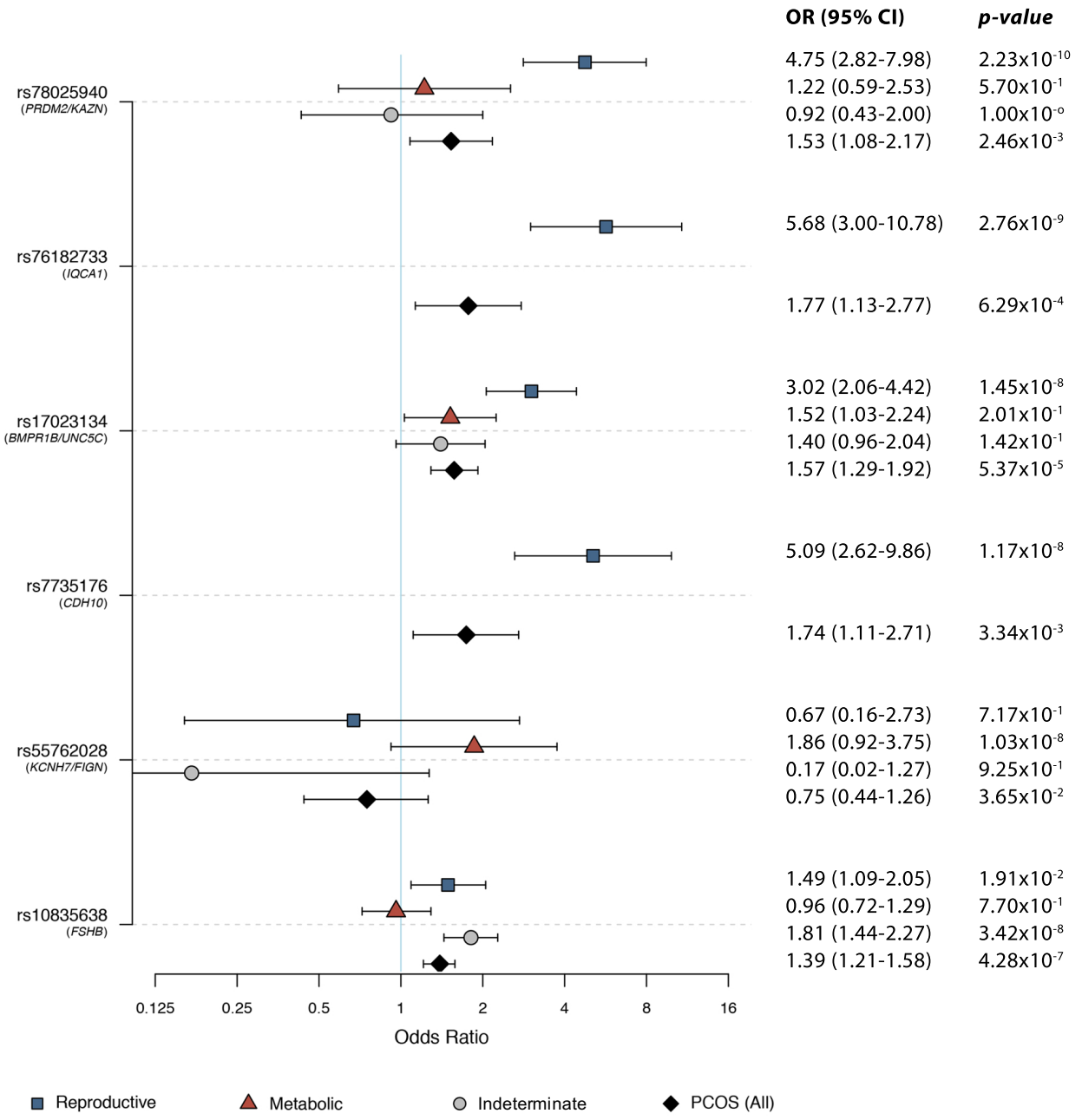


Fig 7. Risk allele odds ratios in PCOS and PCOS subtypes. Odds ratios with 95% confidence intervals and association P-values from the Stage 1 discovery cohort are shown for each subtype-specific novel risk allele identified in this study relative to the corresponding values for the other subtypes and for PCOS disease status in general (includes all subtypes). Some SNPs were not characterized in certain subtypes due to low allele counts or low imputation confidence.

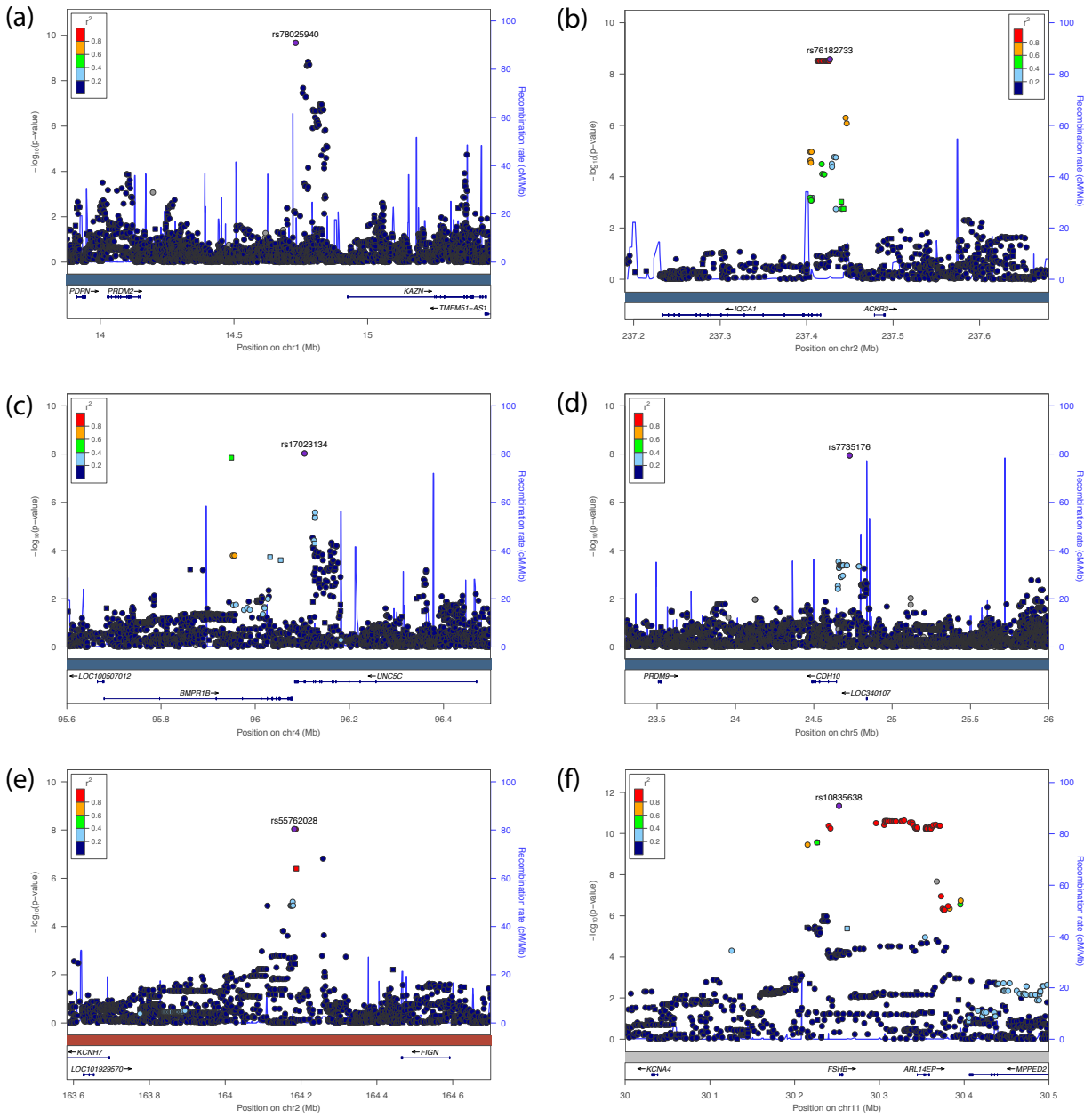
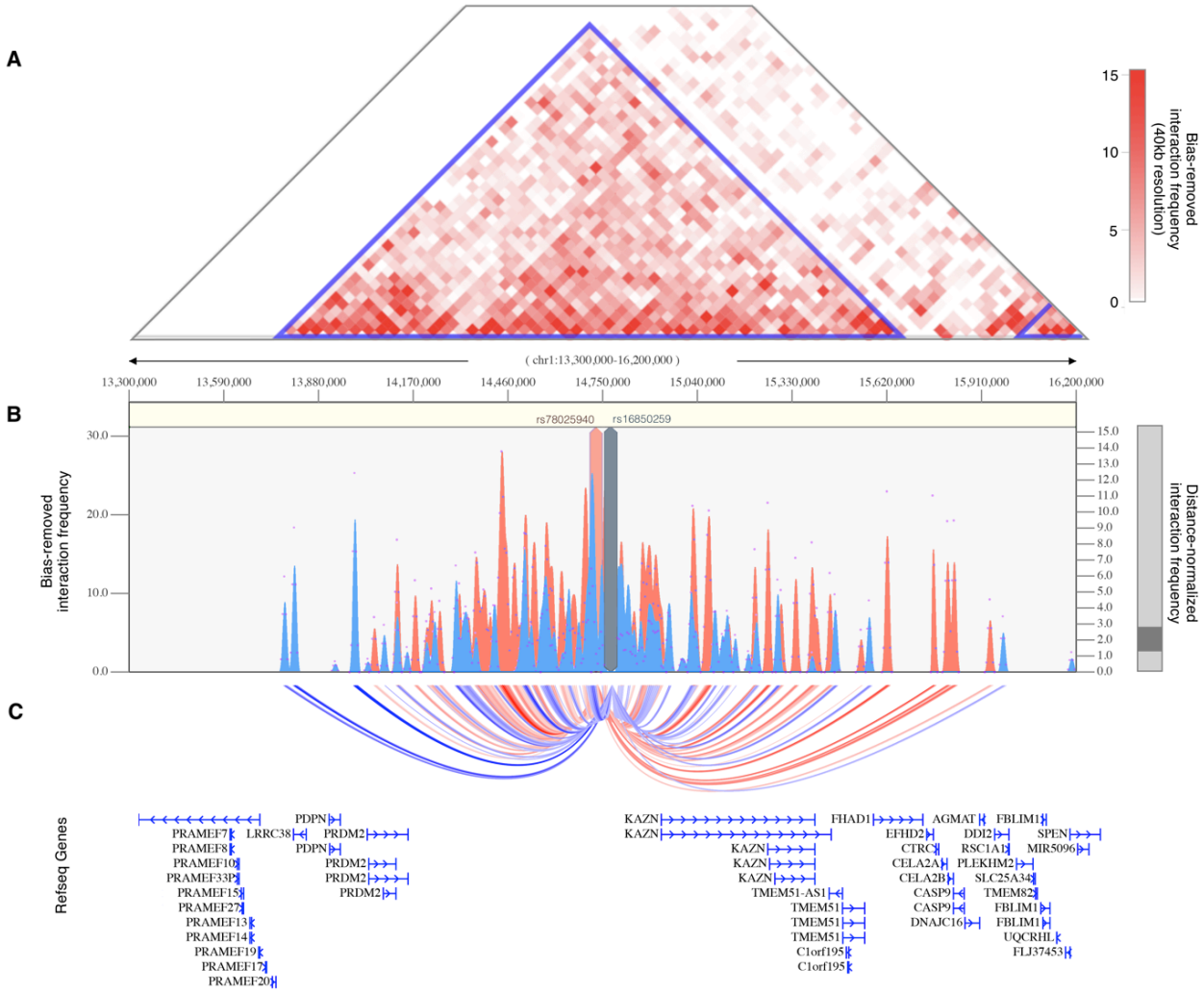


Fig 8. Regional association plots of genome-wide significant loci. Regional plots of association (left y-axis) and recombination rates (right y-axis) for the chromosomes (a) 1p36.21, (b) 2q37.3, (c) 4q22.3, (d) 5p14.2-p14.1, (e) 2p24.2-q24.3, and (f) 1p14.1 loci after imputation. The lead SNP in each locus is labeled and marked in purple. All other SNPs are color coded according to the strength of LD with the top SNP (as measured by r^2 in the European 1000 Genomes data). Imputed SNPs are plotted as circles and genotyped SNPs as squares.



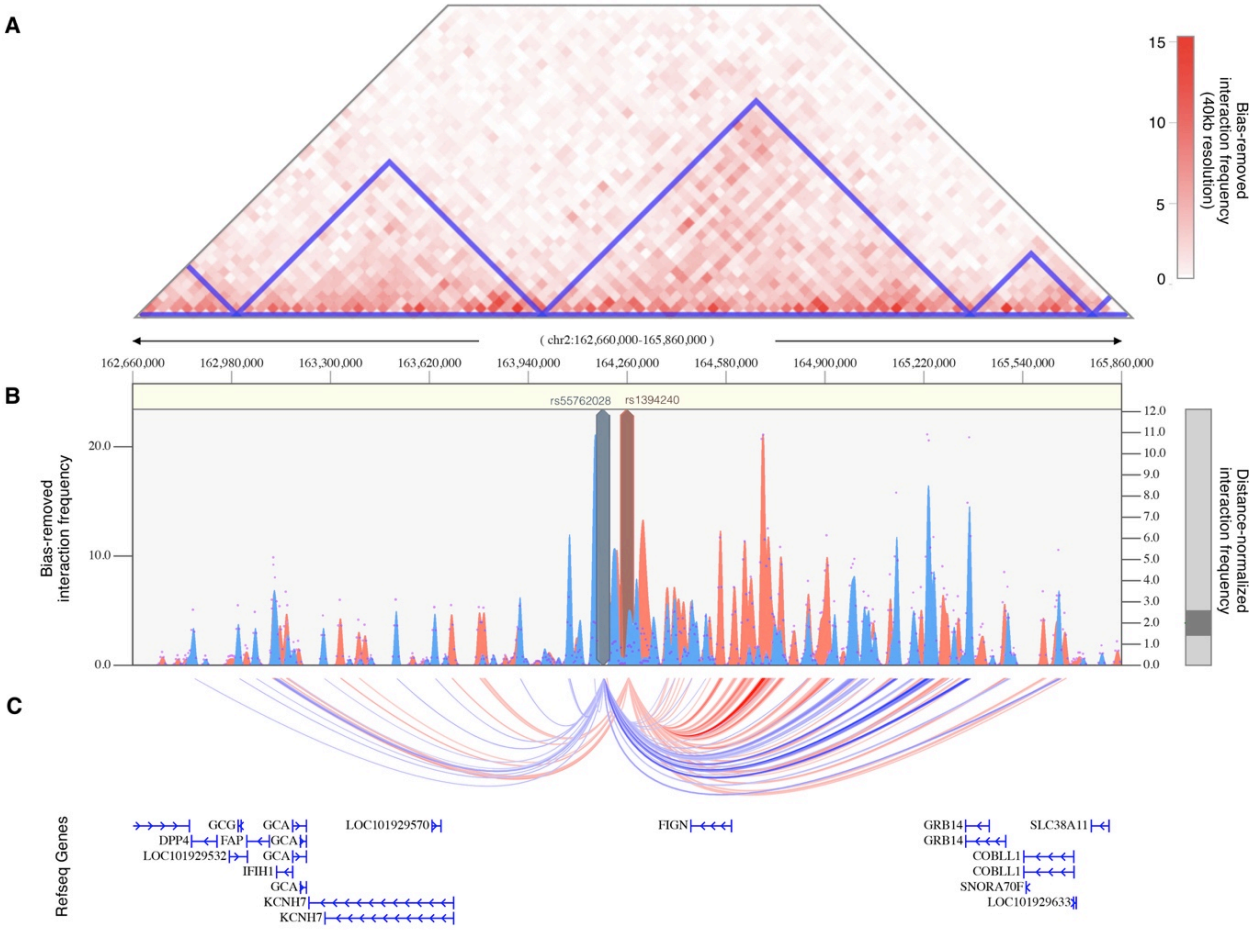
338

339 **Fig 9. Chromatin interaction map of *PRDM2*/*KAZN* locus.** (A) Shown is the
340 interaction frequency heatmap from chr1:13,300,000-16,200,000 in ovarian tissue. The
341 color of the heatmap indicates the level of normalized interaction frequencies with blue
342 triangles indicating topological association domains. (B) One-to-all interaction plots are
343 shown for the lead SNP (rs78025940; shown in red) and lead genotyped SNP
344 (rs16850259; shown in blue) as bait. Y-axes on the left and the right measure bias-
345 removed interaction frequency (red and blue bar graphs) and distance-normalized
346 interaction frequency (magenta dots), respectively. (C) The arc-representation of
347 significant interactions for distance-normalized interaction frequencies ≥ 2 is displayed
348 relative to the RefSeq-annotated genes in the locus.

The 4q22.3 locus spanned a 500kb region of LD including the 3' ends of both the *BMPR1B* and *UNC5C* genes (**Fig 8c**). The most strongly associated SNP (rs17023134; $P_{meta}=9.71\times10^{-9}$) in the locus was within an intronic region of *UNC5C*, and was associated with the reproductive subtype in the Stage 1 discovery cohort with genome-wide significance (OR=3.02, 2.06-4.42 95%CI, $P_1=1.40\times10^{-8}$), but was not significantly associated in the Stage 2 replication analysis (OR=1.71, 0.98-2.99 95%CI, $P_2=7.8\times10^{-2}$). The SNP was imputed with an r^2 of 0.87 and 0.83 in the Stage 1 and Stage 2 analyses, respectively. The most strongly associated genotyped SNP in the locus (rs10516957) was also genome-wide significant ($P_{meta}=1.46\times10^{-8}$) and was located in an intronic region of *BMPR1B*. The genotyped SNP was nominally associated with the reproductive subtype in both the Stage 1 (OR=2.42, 1.66-3.52 95%CI, $P_1=6.72\times10^{-6}$) and Stage 2 (OR=2.40, 1.51-3.82 95%CI, $P_2=4.7\times10^{-4}$) analyses with nearly identical effect sizes.

In the 5p14.2-p14.1 locus, 83kb upstream of the *CDH10* gene (**Fig 8d**), two adjacent SNPs (rs7735176, rs16893866) in perfect LD were equally associated with the reproductive subtype with genome-wide significance ($P_{meta}=1.17\times10^{-8}$). The SNPs were imputed in Stage 1 (OR=5.09, 2.62-9.86 95%CI, $P_1=1.14\times10^{-8}$) with an imputation r^2 of 0.93.

The single locus containing genome-wide significant associations with the metabolic subtype was in an intergenic region of 2q24.2-q24.3 roughly 200kb downstream from *FIGN* and 500kb upstream from *KCNH7* (**Fig 8e**). The lead SNP, rs55762028, was imputed in Stage 1 only (OR=1.86, 0.92-3.75 95%CI, $P_1=9.17\times10^{-9}$, $P_{meta}=1.03\times10^{-8}$). In pancreatic tissue, the lead SNPs appear to be located terminally within a 1.3Mb TAD encompassing the *FIGN* gene and reaching upstream to the *GRB14* gene (**Fig 10**).



371

372 **Fig 10. Chromatin interaction map of *KCHN7*/*FIGN* locus.** (A) Shown is the
373 interaction frequency heatmap from chr2:162,660,000 to 165,860,000 in pancreatic
374 tissue. The color of the heatmap indicates the level of normalized interaction
375 frequencies with blue triangles indicating topological association domains. (B) One-to-all
376 interaction plots are shown for the lead SNP (rs13401392; shown in blue) and 2nd-
377 leading SNP (rs1394240; shown in red) as bait. Y-axes on the left and the right measure
378 bias-removed interaction frequency (blue and red bar graphs) and distance-normalized
379 interaction frequency (magenta dots), respectively. (C) The arc-representation of
380 significant interactions for distance-normalized interaction frequencies ≥ 2 is displayed
381 relative to the RefSeq-annotated genes in the locus.

Association testing on the indeterminate cases replicated the genome-wide significant association in the 11p14.1 *FSHB* locus (**Fig 8f**) identified in our original GWAS (14). The lead SNP (rs10835638; $P_{meta}=4.94\times10^{-12}$) and lead genotyped SNP (rs10835646; $P_{meta}=2.75\times10^{-11}$) in this locus differed from the index SNPs identified in our original GWAS (rs11031006) and in the PCOS meta-analysis (rs11031005), but both of the previously identified index SNPs were also associated with the indeterminate subgroup with genome-wide significance in this study (rs11031006: $P_{meta}=2.96\times10^{-10}$; rs11031005: $P_{meta}=2.91\times10^{-10}$) and are in LD with the lead SNP rs10835638 ($r^2 = 0.59$) (38). The other significant signals from our original GWAS (18) were not reproduced in any of the subtype association tests performed in this study (**Table 2**).

Table 2. Previous GWAS association signals in PCOS subtypes

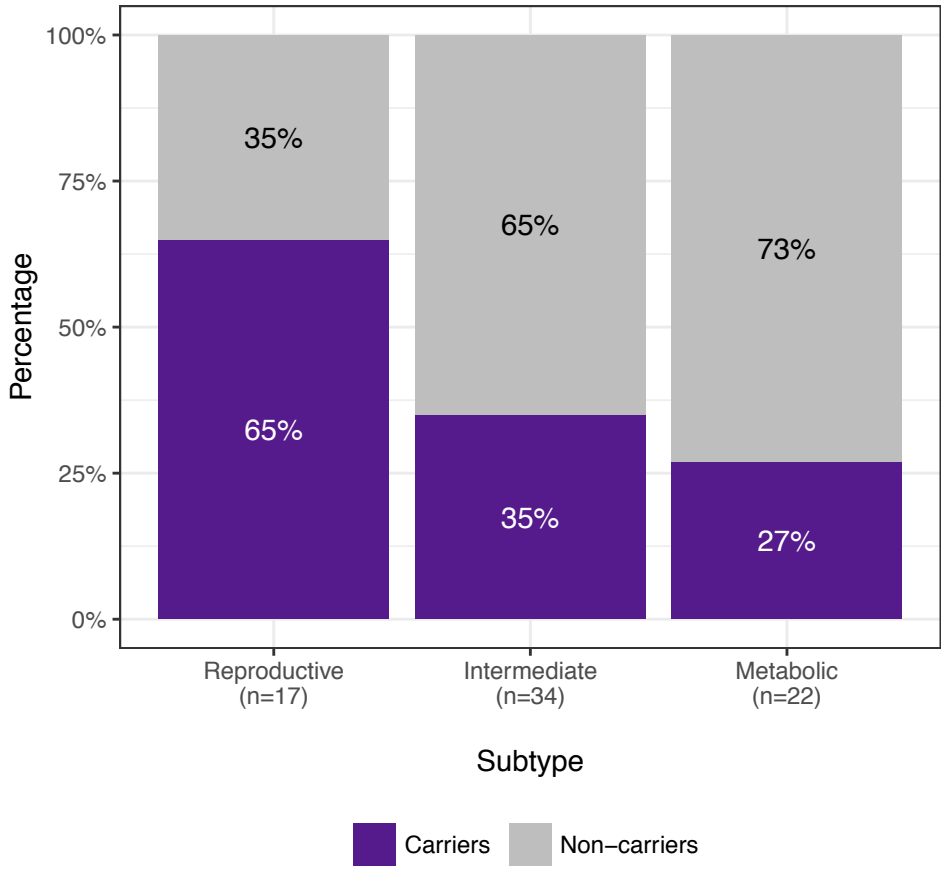
Variant	Locus	PCOS	Reproductive	Metabolic	Indeterminate
rs804279	*GATA4/NEIL2*	$P = 8.0 \times 10^{-10}$	$P = 2.4 \times 10^{-3}$	$P = 9.9 \times 10^{-2}$	$P = 3.1 \times 10^{-3}$
rs10993397	*C9orf3*	$P = 4.6 \times 10^{-13}$	$P = 2.3 \times 10^{-4}$	$P = 6.9 \times 10^{-5}$	$P = 1.1 \times 10^{-5}$
rs11031006	*FSHB*	$P = 1.9 \times 10^{-8}$	$P = 8.8 \times 10^{-6}$	$P = 6.6 \times 10^{-1}$	$P = 3.0 \times 10^{-10}$

Subtype-specific association statistics are shown for each of the SNPs that were significantly associated with PCOS in Hayes and Urbanek et al. (18). P = significance as assessed by sample-size weighted two-strata meta-analysis, adjusted for genomic inflation.

Subtypes in PCOS families

The RF classifier yielded the lowest overall subtype misclassification rate (13.2%) of the tested methods, according to 10-fold cross-validation of the genotyped cohort. Affected women from the family-based cohort were classified accordingly using a RF model. Seventy-three daughters of the 83 affected women from the family-based cohort had complete

403 quantitative trait data available for subtype classification. Seventeen (23.3%) were classified
404 as having the reproductive subtype of PCOS, and 22 (30.1%) were classified as having the
405 metabolic subtype. Of 14 subtyped sibling pairs, only 8 were concordantly classified
406 (57.1%); however, there was only one instance of the reproductive subtype and metabolic
407 subtype occurring within the same nuclear family as the remaining discordant pairs each
408 featured one indeterminate member. The proportion of affected women with one or more of
409 the previously-identified (29) deleterious, rare variants in *DENND1A* varied by subtype.
410 Women classified as having the reproductive subtype of PCOS were significantly more likely
411 to carry one or more of the *DENND1A* rare variants compared to other women with PCOS
412 (P=0.03; **Fig 11**). The distribution of affected women and *DENND1A* rare variant carriers are
413 shown relative to the adjusted quantitative trait PCs in **Fig 12**.
414



415

416 **Fig 11. *DENND1A* rare variant carriers by subtype.** The proportions of affected
417 women with *DENND1A* rare variants in families with PCOS are shown by classified
418 subtype. Women with the reproductive subtype were significantly more likely to
419 carry one or more of the *DENND1A* rare variants compared to other women with
420 PCOS (P=0.03)

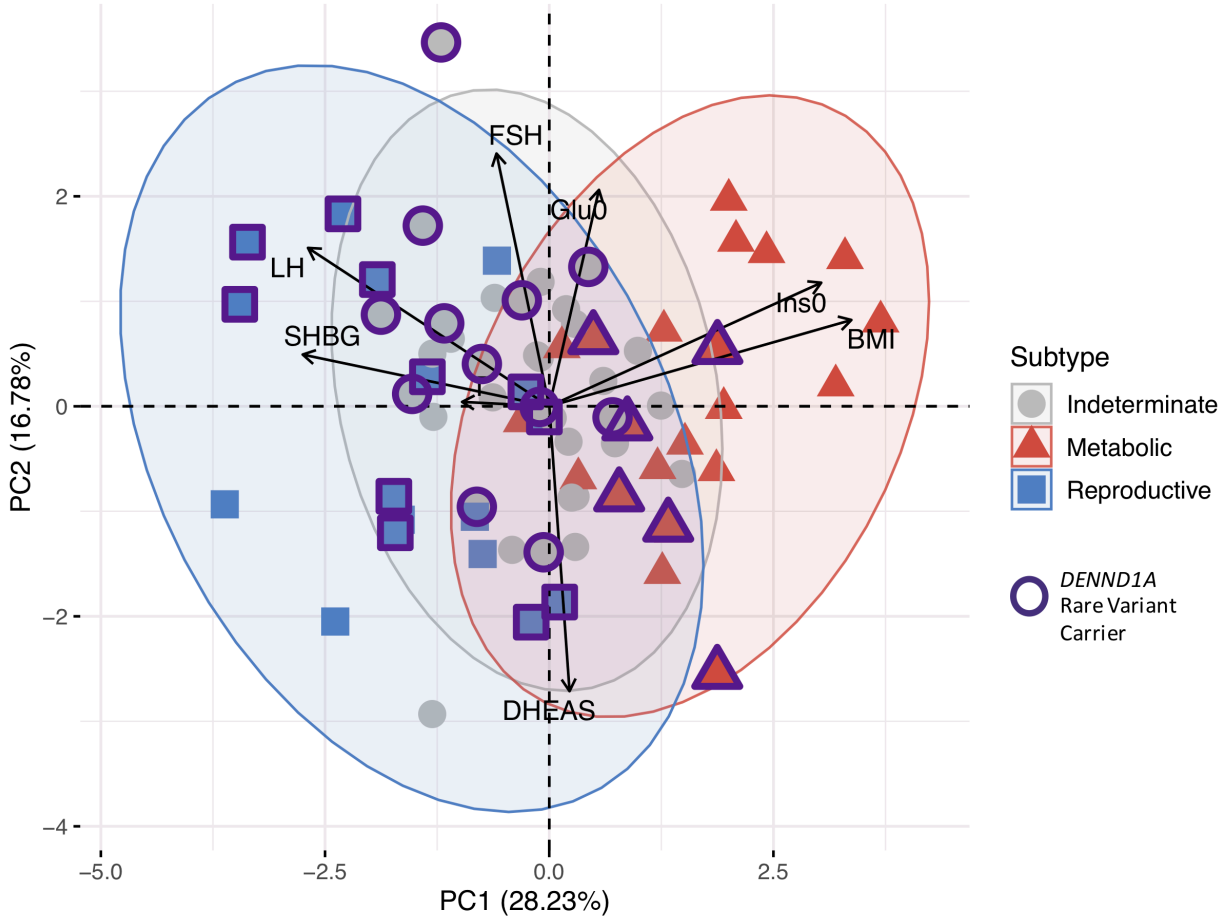


Fig 12. PCA of affected women in PCOS families showing *DENND1A* rare variant carriers. The distribution of affected women and *DENND1A* rare variant carriers are shown relative to their classified subtypes and their adjusted quantitative trait PCs in families affected by PCOS.

Discussion

It is becoming increasingly clear that common, complex traits, such as T2D, are a heterogeneous collection of disease subtypes (24,45-47). There is emerging evidence that these subtypes have different genetic architectures (24,47,48). Consistent with these concepts, we identified reproductive and metabolic subtypes of PCOS by unsupervised hierarchical cluster analysis of quantitative hormonal traits and BMI and found novel loci uniquely associated with

these subtypes with substantially larger effect sizes than those associated with PCOS disease status in GWAS (18-22). These findings suggest that these subtypes are both genetically distinct as well as more etiologically homogenous (9). Our findings are in contrast to the recent PCOS GWAS meta-analysis (21) that found that only one of 14 loci was uniquely associated with the NIH phenotype compared to non-NIH Rotterdam phenotypes. These latter findings suggest that the NIH and Rotterdam diagnostic criteria do not identify biologically distinct subtypes of PCOS. There have been previous efforts to subtype PCOS using unsupervised clustering (25-28), but no subsequent investigation into the biologic relevance of the resulting subtypes.

The key traits driving the subtypes were BMI, insulin, SHBG, glucose, LH, and FSH levels. The reproductive subtype was characterized by higher LH and SHBG levels with lower BMI and blood glucose and insulin levels. The metabolic subtype was characterized by higher BMI and glucose and insulin levels with relatively low SHBG and LH levels. The remaining 40% of cases had no distinguishable cluster-wide characteristics and mean trait values were between those of the reproductive and metabolic subtypes. The relative trait distributions and results of the PCAs (**Figs 3, 4, 5b**) showed the reproductive and metabolic subtypes as collections of subjects on opposite ends of a phenotypic spectrum with the remaining indeterminate subjects scattered between the two. Bootstrapping and clustering in an independent cohort revealed that the reproductive and metabolic subtypes were stable and reproducible. When the GWAS was repeated, novel susceptibility loci were associated with the reproductive and metabolic subtypes, suggesting that they had distinct genetic architecture. The indeterminate PCOS cases were associated with the reported locus at *FSHB*, but the association signal was stronger than that of our original GWAS (18),

455 suggesting that the indeterminate group was also more genetically homogenous after the
456 reproductive and metabolic subtypes were removed from the analysis.

457 Two of the loci associated with the reproductive subtype implicate novel biologic
458 pathways in PCOS pathogenesis. The association signal on chr1 appeared downstream of and
459 within the same TAD as the *PRDM2* gene (**Figs 8a, 9**), which is an estrogen receptor co-
460 activator (49) that is highly expressed in the ovary (50) and pituitary gland (51). In an
461 independent rare variant association study in PCOS families, *PRDM2* demonstrated the 5th
462 strongest gene-level association with altered hormonal levels in PCOS families (P=6.92×10^{-3})
463 out of 339 genes tested (29). PRDM2 binds with ligand bound estrogen receptor alpha (ERα)
464 to open chromatin at ERα target genes (49,52). PRDM2 also binds with the retinoblastoma
465 protein (53), which has been found to play an important role in follicular development in
466 granulosa cells (54,55).

467 The reproductive subtype association in the 4q22.3 locus overlapped with the
468 *BMPR1B* gene, which transcribes a type-I AMH receptor highly expressed in granulosa cells
469 and in GnRH neurons (16) that regulates follicular development (56). BMPR1B (bone
470 morphogenetic protein receptor type IB), also known as ALK6 (Activin Receptor-Like
471 Kinase 6), heterodimerizes with the TGF-β type-II receptors, including AMHR2, and binds
472 AMH and other BMP ligands to initialize TGF-β signaling via Smads 1/5/8 (57). *BMPR1B*
473 has been found to mediate the AMH response in ovine granulosa cells (58), and *BMPR1B*-
474 deficient mice are infertile and suffer from a variety of functional defects in the
475 ovary (59,60). One of the BMPR1B ligand genes, *BMP6*, had the 3rd strongest gene-level
476 association with altered hormonal levels (P=4.00×10^{-3}) out of 339 genes tested in our rare
477 variant association study in PCOS families (29). Collectively, these results make *BMPR1B* a

compelling candidate gene in PCOS pathogenesis. These findings also support our sequencing studies that have implicated pathogenic variants in the AMH signaling pathway in PCOS (61,62).

The nature of the potential involvement in PCOS is less clear for the other loci associated with the reproductive subtype. The 2q37.3 locus overlapped with the promoter region of the *IQCA1* gene. Its function in humans is not well characterized, but *IQCA1* is highly expressed in the pituitary gland (51). The 5p14.2-p14.1 locus overlapped the promoter region of the *CDH10* gene (Cadherin 10). *CDH10* is almost exclusively expressed in the brain (50), and is putatively involved in synaptic adhesions, axon outgrowth and guidance (63).

The lone significant association signal with the metabolic subtype was located in an intergenic region 200-280kb downstream of the *FIGN* gene, 490-570kb upstream of *KCNH7*. *KCNH7* encodes a voltage-gated potassium channel (alias ERG3). *KCNH7* is primarily expressed in the nervous system (64), but has been found in murine islet cells (65,66). *FIGN* encodes fidgetin, a microtubule-severing enzyme most highly expressed in the pituitary gland and ovary (50). A genetic variant in *FIGN* was found to reduce the risk of congenital heart disease in Han Chinese by modulating transmembrane folate transport (67,68). The TAD encompassing the association signal in this locus includes *FIGN* and extends upstream to the *GRB14* gene (**Fig 9**). *GRB14* plays an important role in insulin receptor signaling (69,70) and has been associated with T2D in GWAS (71). Given the various metabolic associations for the genes in this chromosomal region, it is plausible that causal variants in this locus could impact a combination of these genes.

Despite evidence linking neighboring genes to PCOS pathways in each of the aforementioned loci, it remains possible, of course, that other more distant genes in LD underlie the association signals. Causal variants are often up to 2 Mb away from the associated SNP, not necessarily in the closest gene (72). Fine-mapping and functional studies are needed in order to confirm the causal variants in each of these loci. In addition, the sample sizes for the subtype GWAS were small, some of the associations were based only on imputed SNPs in Stage 1, and a replication association study has not yet been performed. However, the aforementioned functional evidence for several of the loci—particularly for *PRDM2* and *BMPR1B*—support the validity of their associations. Also, the fact that one of the genes associated with the reproductive subtype, *PRDM2*, was associated with PCOS quantitative traits in our family-based analysis (29) does represent a replication of this signal by an independent analytical approach. Nevertheless, our genetic association results should be considered preliminary.

The effect sizes of the subtype alleles, particularly those associated with the reproductive subtype (Odds Ratio [OR] 3.02-5.68) (**Table 1**), were substantially greater than the effects (OR 0.70-1.51) observed for alleles associated with PCOS diagnosis in previous GWAS (18-22). In general, there is an inverse relationship between allele frequency and effect size (1) because alleles with larger phenotypic effects are subject to purifying selection and, therefore, occur less frequently in the population (73,74). Accordingly, in contrast to the common variants (Effect Allele Frequency [EAF]>0.05) associated with PCOS in previous GWAS (18-22), the alleles associated with the subtypes were all of low frequency (EAF 0.01-0.05; **Table 1**). However, given the limited cohort sizes in this study, the subtype association testing did not have adequate power to detect associations with more modest

effect sizes, such as those from our previous GWAS (18). It is also possible that the large effect sizes were somewhat inflated by the so-called "winners curse" (75,76), but they nonetheless suggest that the subtypes were more genetically homogeneous than PCOS diagnosis in general.

In applying a subtype classifier to our family-based cohort, we found twelve affected sibling pairs in which at least one of the daughters was classified with the reproductive or metabolic subtype. Six of these sibling pairs were classified with the same subtype. There was only one discordant pairing of the reproductive subtype with the metabolic subtype. This further suggests that the reproductive and metabolic subtypes are genetically distinct in their origins. The greater prevalence of *DENND1A* rare variant carriers observed in women with the reproductive subtype in the family-based cohort implicates this gene in the pathogenesis of this subtype. *DENND1A* is known to regulate androgen biosynthesis in the ovary (77,78); therefore, we would expect *DENND1A*-mediated PCOS to be more closely associated with the reproductive subtype of PCOS. However, we did not find an association between any *DENND1A* alleles and the reproductive subtype in the subtype GWAS, perhaps due to allelic heterogeneity or to our limited power to detect associations with more modest effect sizes.

We only studied women with PCOS as defined by the NIH diagnostic criteria. Future studies will investigate whether similar reproductive and metabolic clusters are present in non-NIH Rotterdam PCOS cases. In particular, it is possible that there will be no metabolic subtype in non-NIH Rotterdam PCOS cases since these phenotypes have minimal metabolic risk (79,80). Indeed, in a previous effort to identify phenotypic subtypes in Rotterdam PCOS cases (28), the cluster that most closely resembled the reproductive subtype represented the largest proportion of PCOS women at 44%, of whom only 78% met the NIH criteria for PCOS,

546 whereas the cluster that most closely resembled the metabolic subtype constituted only 12% of
547 the cohort, but 98% met the NIH diagnostic criteria. Due to the within-cohort normalization of
548 quantitative traits prior to clustering, our method is well-suited for identifying subsets of cases
549 that occupy either end of a phenotypic spectrum within different populations, but it therefore
550 may not be suitable for directly comparing subtype membership between different populations.
551 Our cohort included only women of European ancestry. It will be of considerable
552 importance to investigate whether subtypes are present in women with PCOS of other ancestries.
553 Women with PCOS of diverse races and ethnicities have similar reproductive and metabolic
554 features (81-83). However, there are differences in the severity of the metabolic defects due to
555 differences in the prevalence of obesity (84) and well as to racial/ethnic differences in insulin
556 sensitivity (85,86). Further, the susceptibility loci associated with subtypes in other ancestry
557 groups may differ since the low frequency and large effect size of the variants associated with
558 the reproductive subtype in our European cohort suggests these variants are of relatively recent
559 origin, and, therefore, may be population-specific (1,87,88).
560 In conclusion, using an unsupervised clustering approach featuring quantitative
561 hormonal and anthropometric data, we identified novel reproductive and metabolic subtypes
562 of PCOS with distinct genetic architectures. The genomic loci that were significantly
563 associated with either of these subtypes include a number of new, highly plausible PCOS
564 candidate genes. Moreover, our results demonstrate that precise phenotypic delineation,
565 resulting in more homogeneous subsets of affected individuals, can be more powerful and
566 informative than increases in sample size for genetic association studies. Our findings
567 indicate that further study into the genetic heterogeneity of PCOS is warranted and could
568 lead to a transformation in the way PCOS is classified, studied, and treated.

Acknowledgements

This study was supported by National Institutes of Health (NIH) Grants R01HD057223 (A.D.), P50 HD044405 (A.D.) and R01 HD085227 (A.D.). M.D. was supported by a Ruth L. Kirschstein National Research Service Award Institutional Research Training Grant, T32 DK007169. We thank the NIH Cooperative Multicenter Reproductive Medicine Network (https://www.nichd.nih.gov/research/supported/rmn) for recruiting some of the women with PCOS who participated who participated in the genomewide association study of Hayes and Urbanek et al. (18) and whose genotype data were used in this study . We also thank the following investigators for recruiting some of the control women who participated in the genomewide association study of Hayes and Urbanek et al. (18) and whose genotype data were used in this study: Dimitrios Panidis, MD, PHD (Aristotle University of Thessaloniki, Greece); Mark O. Goodarzi (Cedars-Sinai Medical Center, Los Angeles, CA); Corrine K. Welt, MD (University of Utah School of Medicine, Salt Lake City, UT; formerly of Massachusetts General Hospital, Boston, MA); Ahmed H. Kissebah (deceased, Medical College of Wisconsin, Milwaukee, WI); Ricardo Azziz, MD (State University of New York, NY; formerly of University of Alabama at Birmingham, AL); and Evanthia Diamanti-Kandarakis, MD, PhD (University of Athens Medical School, Greece).

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
2. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177-86.
3. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell. 2018;173(7):1573-80.
4. Ringman JM, Goate A, Masters CL, Cairns NJ, Danek A, Graff-Radford N, et al. Genetic heterogeneity in Alzheimer disease and implications for treatment strategies. Curr Neurol Neurosci Rep. 2014;14(11):499.
5. Flint J, Kendler KS. The genetics of major depression. Neuron. 2014;81(3):484-503.
6. von Coelln R, Shulman LM. Clinical subtypes and genetic heterogeneity: of lumping and splitting in Parkinson disease. Curr Opin Neurol. 2016;29(6):727-34.
7. Ahlqvist E, Storm P, Karajamaki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018;6(5):361-9.
8. Udler MS, Kim J, von Grotthuss M, Bonas-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. 2018;15(9):e1002654.
9. Saria S, Goldenberg A. Subtyping: What it is and its role in precision medicine. IEEE Intelligent Systems. 2015;30(4):70-5.
10. Diamanti-Kandarakis E, Dunaif A. Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. Endocr Rev. 2012;33(6):981-1030.
11. Rubin KH, Glintborg D, Nybo M, Abrahamsen B, Andersen M. Development and Risk Factors of Type 2 Diabetes in a Nationwide Population of Women With Polycystic Ovary Syndrome. J Clin Endocrinol Metab. 2017;102(10):3848-57.
12. Dunaif A. Perspectives in Polycystic Ovary Syndrome: From Hair to Eternity. J Clin Endocrinol Metab. 2016;101(3):759-68.
13. Zawadski JKD, A. Diagnostic criteria for polycystic ovary syndrome; towards a rational approach. Polycystic Ovary Syndrome. Boston, Massachussets: Blackwell Scientific; 1992. p. 377-84.
14. Rotterdam EA-SPcwg. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Hum Reprod. 2004;19(1):41-7.
15. Rotterdam EA-SPCWG. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertil Steril. 2004;81(1):19-25.
16. Cimino I, Casoni F, Liu X, Messina A, Parkash J, Jamin SP, et al. Novel role for anti-Mullerian hormone in the regulation of GnRH neuron excitability and hormone secretion. Nat Commun. 2016;7:10055.

17. Tata B, Mimouni NEH, Barbotin AL, Malone SA, Loyens A, Pigny P, et al. Elevated prenatal anti-Mullerian hormone reprograms the fetus and induces polycystic ovary syndrome in adulthood. Nat Med. 2018;24(6):834-46.
18. Hayes MG, Urbanek M, Ehrmann DA, Armstrong LL, Lee JY, Sisk R, et al. Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. Nat Commun. 2015;6:7502.
19. Chen ZJ, Zhao H, He L, Shi Y, Qin Y, Shi Y, et al. Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. Nat Genet. 2011;43(1):55-9.
20. Shi Y, Zhao H, Shi Y, Cao Y, Yang D, Li Z, et al. Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. Nat Genet. 2012;44(9):1020-5.
21. Day F, Karaderi T, Jones MR, Meun C, He C, Drong A, et al. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. PLoS Genet. 2018;14(12):e1007813.
22. Day FR, Hinds DA, Tung JY, Stolk L, Styrkarsdottir U, Saxena R, et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. Nat Commun. 2015;6:8464.
23. Castaldi PJ, Dy J, Ross J, Chang Y, Washko GR, Curran-Everett D, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. Thorax. 2014;69(5):415-22.
24. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015;7(311):311ra174.
25. Tzeng CR, Chang YC, Chang YC, Wang CW, Chen CH, Hsu MI. Cluster analysis of cardiovascular and metabolic risk factors in women of reproductive age. Fertil Steril. 2014;101(5):1404-10.
26. Dewailly D, Alebic MS, Duhamel A, Stojanovic N. Using cluster analysis to identify a homogeneous subpopulation of women with polycystic ovarian morphology in a population of non-hyperandrogenic women with regular menstrual cycles. Hum Reprod. 2014;29(11):2536-43.
27. Daan NM, Koster MP, de Wilde MA, Dalmeijer GW, Evelein AM, Fauser BC, et al. Biomarker Profiles in Women with PCOS and PCOS Offspring; A Pilot Study. PLoS One. 2016;11(11):e0165033.
28. Huang CC, Tien YJ, Chen MJ, Chen CH, Ho HN, Yang YS. Symptom patterns and phenotypic subgrouping of women with polycystic ovary syndrome: association between endocrine characteristics and metabolic aberrations. Hum Reprod. 2015;30(4):937-46.
29. Dapas M, Sisk R, Legro RS, Urbanek M, Dunaif A, Hayes MG. Family-based quantitative trait meta-analysis implicates rare noncoding variants in DENND1A in polycystic ovary syndrome. J Clin Endocrinol Metab. 2019.

30. Legro RS, Driscoll D, Strauss JF, 3rd, Fox J, Dunaif A. Evidence for a genetic basis for hyperandrogenemia in polycystic ovary syndrome. Proc Natl Acad Sci U S A. 1998;95(25):14956-60.
31. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4:13.
32. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif. 2014;31(3):274-95.
33. Strauss T, von Maltitz MJ. Generalising Ward's Method for Use with Manhattan Distances. PLoS One. 2017;12(1):e0168288.
34. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data An. 2007;52(1):258-71.
35. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 2012;8(8):e1002793.
36. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.
37. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10(4):e1004234.
38. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
39. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7.
40. 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. 2007;39(7):906-13.
41. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.
42. Yang D, Jang I, Choi J, Kim MS, Lee AJ, Kim H, et al. 3DIV: A 3D-genome Interaction Viewer and database. Nucleic Acids Res. 2018;46(D1):D52-D7.
43. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44(7):e70.
44. Venables WN, Ripley BD, Venables WN. Modern applied statistics with S. 4th ed. New York: Springer; 2002. xi, 495 p. p.
45. Tuomi T, Santoro N, Caprio S, Cai M, Weng J, Groop L. The many faces of diabetes: a disease with increasing heterogeneity. Lancet. 2014;383(9922):1084-94.
46. Skyler JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes. 2017;66(2):241-55.

47. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018.
48. Udler MS, Kim J, von Grotthuss M, Bonàs-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLOS Medicine. 2018;15(9):e1002654.
49. Di Zazzo E, De Rosa C, Abbondanza C, Moncharmont B. PRDM Proteins: Molecular Mechanisms in Signal Transduction and Transcriptional Regulation. Biology (Basel). 2013;2(1):107-41.
50. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
51. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462-70.
52. Carling T, Kim KC, Yang XH, Gu J, Zhang XK, Huang S. A histone methyltransferase is required for maximal response to female sex hormones. Mol Cell Biol. 2004;24(16):7032-42.
53. Liu L, Shao G, Steele-Perkins G, Huang S. The retinoblastoma interacting zinc finger gene RIZ produces a PR domain-lacking product through an internal promoter. J Biol Chem. 1997;272(5):2984-91.
54. Andreu-Vieyra C, Chen R, Matzuk MM. Conditional deletion of the retinoblastoma (Rb) gene in ovarian granulosa cells leads to premature ovarian failure. Mol Endocrinol. 2008;22(9):2141-61.
55. Yang QE, Nagaoka SI, Gwost I, Hunt PA, Oatley JM. Inactivation of Retinoblastoma Protein (Rb1) in the Oocyte: Evidence That Dysregulated Follicle Growth Drives Ovarian Teratoma Formation in Mice. PLoS Genet. 2015;11(7):e1005355.
56. Reader KL, Haydon LJ, Littlejohn RP, Juengel JL, McNatty KP. Booroola BMPR1B mutation alters early follicular development and oocyte ultrastructure in sheep. Reprod Fertil Dev. 2012;24(2):353-61.
57. Shimasaki S, Moore RK, Otsuka F, Erickson GF. The bone morphogenetic protein system in mammalian reproduction. Endocr Rev. 2004;25(1):72-101.
58. Estienne A, Pierre A, di Clemente N, Picard JY, Jarrier P, Mansanet C, et al. Anti-Mullerian hormone regulation by the bone morphogenetic proteins in the sheep ovary: deciphering a direct regulatory pathway. Endocrinology. 2015;156(1):301-13.
59. Yi SE, LaPolt PS, Yoon BS, Chen JY, Lu JK, Lyons KM. The type I BMP receptor BmprIB is essential for female reproductive function. Proc Natl Acad Sci U S A. 2001;98(14):7994-9.
60. Sugiura K, Su YQ, Eppig JJ. Does bone morphogenetic protein 6 (BMP6) affect female fertility in the mouse? Biol Reprod. 2010;83(6):997-1004.
61. Gorsic LK, Dapas M, Legro RS, Hayes MG, Urbanek M. Functional Genetic Variation in the Anti-Mullerian Hormone Pathway in Women with Polycystic Ovary Syndrome. J Clin Endocrinol Metab. 2019.

62. Gorsic LK, Kosova G, Werstein B, Sisk R, Legro RS, Hayes MG, et al. Pathogenic Anti-Mullerian Hormone Variants in Polycystic Ovary Syndrome. J Clin Endocrinol Metab. 2017;102(8):2862-72.

63. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33(Database issue):D501-4.

64. Shi W, Wymore RS, Wang HS, Pan Z, Cohen IS, McKinnon D, et al. Identification of two nervous system-specific members of the erg potassium channel gene family. J Neurosci. 1997;17(24):9423-32.

65. Hardy AB, Fox JE, Giglou PR, Wijesekara N, Bhattacharjee A, Sultan S, et al. Characterization of Erg K+ channels in alpha- and beta-cells of mouse and human islets. J Biol Chem. 2009;284(44):30441-52.

66. Muhlbauer E, Bazwinsky I, Wolgast S, Klemenz A, Peschke E. Circadian changes of ether-a-go-go-related-gene (Erg) potassium channel transcripts in the rat pancreas and beta-cell. Cell Mol Life Sci. 2007;64(6):768-80.

67. Wang D, Chu M, Wang F, Zhou A, Ruan M, Chen Y. A Genetic Variant in FIGN Gene Reduces the Risk of Congenital Heart Disease in Han Chinese Populations. Pediatr Cardiol. 2017;38(6):1169-74.

68. Wang D, Wang F, Shi KH, Tao H, Li Y, Zhao R, et al. Lower Circulating Folate Induced by a Fidgetin Intronic Variant Is Associated With Reduced Congenital Heart Disease Susceptibility. Circulation. 2017;135(18):1733-48.

69. Desbuquois B, Carre N, Burnol AF. Regulation of insulin and type 1 insulin-like growth factor signaling and action by the Grb10/14 and SH2B1/B2 adaptor proteins. FEBS J. 2013;280(3):794-816.

70. Kasus-Jacobi A, Perdereau D, Auzan C, Clauser E, Van Obberghen E, Mauvais-Jarvis F, et al. Identification of the rat adapter Grb14 as an inhibitor of insulin actions. J Biol Chem. 1998;273(40):26026-35.

71. Zhao W, Rasheed A, Tikkanen E, Lee JJ, Butterworth AS, Howson JMM, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. Nat Genet. 2017;49(10):1450-7.

72. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? Nucleic Acids Res. 2016;44(13):6046-54.

73. Maher MC, Uricchio LH, Torgerson DG, Hernandez RD. Population genetics of rare variants and complex diseases. Hum Hered. 2012;74(3-4):118-28.

74. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A. 2011;108(44):18026-31.

75. Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet. 2001;69(6):1357-69.

76. Kraft P. Curses--winner's and otherwise--in genetic epidemiology. Epidemiology. 2008;19(5):649-51; discussion 57-8.

77. McAllister JM, Modi B, Miller BA, Biegler J, Bruggeman R, Legro RS, et al. Overexpression of a DENND1A isoform produces a polycystic ovary syndrome theca phenotype. Proc Natl Acad Sci U S A. 2014;111(15):E1519-27.

78. Tee MK, Speek M, Legeza B, Modi B, Teves ME, McAllister JM, et al. Alternative splicing of DENND1A, a PCOS candidate gene, generates variant 2. Mol Cell Endocrinol. 2016;434:25-35.

79. Moran L, Teede H. Metabolic features of the reproductive phenotypes of polycystic ovary syndrome. Hum Reprod Update. 2009;15(4):477-88.

80. Fauser BC, Tarlatzis BC, Rebar RW, Legro RS, Balen AH, Lobo R, 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. 2012;97(1):28-38 e25.

81. Carmina E, Koyama T, Chang L, Stanczyk FZ, Lobo RA. Does ethnicity influence the prevalence of adrenal hyperandrogenism and insulin resistance in polycystic ovary syndrome? Am J Obstet Gynecol. 1992;167(6):1807-12.

82. Guo M, Chen ZJ, Eijkemans MJ, Goverde AJ, Fauser BC, Macklon NS. Comparison of the phenotype of Chinese versus Dutch Caucasian women presenting with polycystic ovary syndrome and oligo/amenorrhoea. Hum Reprod. 2012;27(5):1481-8.

83. Louwers YV, Lao O, Fauser BC, Kayser M, Laven JS. The impact of self-reported ethnicity versus genetic ancestry on phenotypic characteristics of polycystic ovary syndrome (PCOS). J Clin Endocrinol Metab. 2014;99(10):E2107-16.

84. Essah PA, Nestler JE, Carmina E. Differences in dyslipidemia between American and Italian women with polycystic ovary syndrome. J Endocrinol Invest. 2008;31(1):35-41.

85. Dunaif A, Sorbara L, Delson R, Green G. Ethnicity and polycystic ovary syndrome are associated with independent and additive decreases in insulin action in Caribbean-Hispanic women. Diabetes. 1993;42(10):1462-8.

86. Engmann L, Jin S, Sun F, Legro RS, Polotsky AJ, Hansen KR, et al. Racial and ethnic differences in the polycystic ovary syndrome metabolic phenotype. Am J Obstet Gynecol. 2017;216(5):493 e1- e13.

87. Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, et al. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. Proc Natl Acad Sci U S A. 2004;101(52):18075-80.

88. McCarthy MI. The importance of global studies of the genetics of type 2 diabetes. Diabetes Metab J. 2011;35(2):91-100.