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

Using biochemical and genotype data from a previously published PCOS genome-wide association study (GWAS), we investigated whether there were reproducible phenotypic subtypes of PCOS with subtype-specific genetic associations. Unsupervised hierarchical cluster analysis was performed on quantitative anthropometric, reproductive, and metabolic traits in a genotyped cohort of 893 PCOS cases (median and interquartile range [IQR]: age = 28 [25–32], body mass index [BMI] = 35.4 [28.2–41.5]). The clusters were replicated in an independent
Conclusions

In conclusion, we have found reproducible reproductive and metabolic subtypes of PCOS. Furthermore, these subtypes were associated with novel, to our knowledge, susceptibility loci. The signs and symptoms of PCOS are heterogeneous, which suggests that the etiology may differ among subsets of women with PCOS. Elucidating the genetic mechanisms of PCOS could result in improved diagnosis and treatment.

Author summary

Why was this study done?

- Polycystic ovary syndrome (PCOS) is one of the most common endocrine disorders in women of reproductive age.
- The signs and symptoms of PCOS are heterogeneous, which suggests that the etiology may differ among subsets of women with PCOS.
- Elucidating the genetic mechanisms of PCOS could result in improved diagnosis and treatment.

What did the researchers do and find?

- A clustering analysis of 893 women with PCOS, using reproductive and metabolic quantitative traits, was performed to identify subsets of affected women with similar hormonal profiles.
- There were distinct reproductive and metabolic “subtypes” of women with PCOS.
- Novel genetic variants were uniquely associated with each of the PCOS subtypes.

What do these findings mean?

- Our results suggest that there are distinct forms of PCOS that are associated with different underlying biological mechanisms.
- Women with PCOS may be poorly served by being grouped under a single diagnosis because PCOS subtypes may differ in responses to therapy and in long-term outcomes.


Academic Editor: Jenny E. Myers, University of Manchester, UNITED KINGDOM

Received: August 22, 2019; Accepted: May 13, 2020; Published: June 23, 2020

Copyright: © 2020 Dapas et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: This study used data that we collected previously for our PCOS GWAS (Hayes and Urbanek et al. Nat Commun 6:7502, 2015) [19]. Stage 1 genotype data have been deposited in the database of Genotypes and Phenotypes (dbGaP) under the accession code phs000368.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000368.v1.p1). The following variables are provided along with the genotype data: SUBJID (de-identified), Case_Control (disease status), Sex, Age, Height, Weight, BMI, Race, Ethnicity. Subtype classifications for these subjects are provided in S5 Table. This study used additional array and whole-genome sequencing
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 under the same clinical diagnosis [4–8]. Recent studies in type 2 diabetes (T2D) 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 [11–13]. It is the leading cause of anovulatory infertility and a major risk factor for T2D in young women [14]. Despite these substantial morbidities, the etiology (or etiologies) of PCOS remains unknown [15]. Accordingly, the commonly used diagnostic criteria for PCOS, the National Institutes of Health (NIH) criteria [16] and the Rotterdam criteria [17,18], 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 ovulatory dysfunction (OD) [16]. The Rotterdam criteria include polycystic ovarian morphology (PCOM) and require the presence of at least 2 of these 3 key reproductive traits, resulting in 3 different affected phenotypes: HA and OD with or without PCOM, also known as NIH PCOS, as well as 2 additional non-NIH Rotterdam phenotypes, HA and PCOM and OD and PCOM.

Genome-wide association studies (GWAS) have considerably advanced our understanding of the pathophysiology of PCOS. These studies have implicated gonadotropin secretion [19] and action [20,21], androgen biosynthesis [20–22], metabolic regulation [22–23], and ovarian aging [23] in PCOS pathogenesis. A recent meta-analysis [22] 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 were 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 [5,24,25]. Indeed, there have been previous efforts to subtype PCOS using unsupervised cluster analysis of its hormonal and anthropometric traits [26–29]. 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 approach on reproductive and metabolic quantitative traits from a large cohort of women with PCOS, 2) validate those subtypes in an independent cohort, and 3) test whether the subtypes thus identified were associated with distinct common genetic variants. As an additional validation, we investigated the association of the subtypes with rare genetic variants we recently identified in a family-based PCOS cohort [30].

Methods

Subjects

This study used biochemical and genotype data from our previously published PCOS GWAS, Hayes and Urbanek and colleagues [19], in which a discovery sample (Stage 1) of 984 unrelated PCOS cases and 2,964 population controls was studied, followed by a replication sample (Stage 2) of 1,799 PCOS cases and 1,231 phenotyped reproductively normal control women. All cases were of European ancestry. The present study began as an exploratory analysis to test the hypothesis that subtypes existed within the PCOS GWAS cohorts. Further analyses were performed once subtypes were identified. This study is reported according to the Strengthening the Reporting of Genetic Association Studies (STREGA) guideline (S1 Checklist). The study was approved by the Institutional Review Board of Northwestern University Feinberg School of Medicine, and each subject provided written informed consent prior to the study [19].

All PCOS cases were aged 13–45 years and were diagnosed according to the NIH criteria [10] of hyperandrogenism and chronic anovulation (8 or fewer menses per year), excluding specific disorders of the adrenal glands, ovaries, or pituitary gland [31]. Cases fulfilling the NIH criteria also meet the Rotterdam criteria for PCOS [10]. The GWAS cohorts included in the cluster analysis, the PCOS Family Study and Pregnancy in PCOS II (PPCOSII) study [19] (S1 Table), had complete data for the following traits: body mass index (BMI), testosterone (T), sex hormone binding globulin (SHBG), dehydroepiandrosterone sulfate (DHEAS), luteinizing hormone (LH), follicle-stimulating hormone (FSH), fasting insulin (Ins0), and fasting glucose (Glu0). Complete data for these quantitative traits were not available in the other GWAS cohorts because of differences in phenotyping protocols [19] (S1 Table). Two additional NIH PCOS cohorts with complete quantitative trait data were included in the present study. An ungenotyped cohort of 263 cases was used for clustering replication. A family-based whole-genome sequencing cohort of 73 PCOS cases was investigated to assess subtype clustering in families and for rare variant analysis [30].

Contraceptive steroids had been stopped at least 3 months prior to screening for the PCOS Family Study, ungenotyped, and whole-genome sequencing PCOS cohorts. Elevated T, non-SHBG bound T, and/or DHEAS levels were documented in all PCOS cases prior to enrollment in these cohorts. PPCOSII was a randomized clinical trial of letrozole versus clomiphene citrate for infertility in PCOS [32]. The PCOS cases in this study had contraceptive steroids discontinued at least 2 months prior to their baseline phenotyping visit. Since the PCOS women in this trial were seeking fertility, the majority were not on recent contraceptive steroid therapy. Therefore, it is unlikely that recent contraceptive steroid use altered T or SHBG levels in the PCOS cases included in the cluster analysis.
adjusted and normalized as described above. Subtype classifiers were modeled on the adjusted

Identifying subtypes in PCOS families

associating domains (TADs) were identified using TopDom [37] chromatin conformation capture (Hi-C) data from the 3DIV database [38]. Neighboring chromatin interactions were investigated in intergenic loci using high-throughput association testing was performed separately for Stage 1 and Stage 2 samples. Of the 893 combined cases from both stages included in the clustering analysis, 555 were from the Stage 1 sample, and 338 were from the Stage 2 sample. In Stage 1, 2,964 normal controls were used, 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 3 PCs of ancestry. P-values are reported as P<0.05, 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) on 76,602 linkage disequilibrium (LD)-pruned SNPs [19]. Samples with values >3 standard deviations from the median for either of the first 2 principal components (PCs) were excluded (34 in discovery; 37 in replication). Genotype data were phased using ShapeIT (v2.r790) and then imputed to the 1000 Genomes reference panel (Phase3 v5) and then imputed to the Michigan Imputation Server [41]. Imputed SNPs with an allelic ρ below 0.8 were removed from analysis [42].

Association testing was performed separately for Stage 1 and Stage 2 samples. Of the 893 combined cases from both stages included in the clustering analysis, 555 were from the Stage 1 sample, and 338 were from the Stage 2 sample. In Stage 1, 2,964 normal controls were used, and 1,134 were used in Stage 2. Logistic regressions were performed using SNPTEST for case–control status under an additive genetic model, adjusting for BMI and first 3 PCs of ancestry. P-values are reported as P₁ and P₂ 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 samples for each subtype under a fixed meta-analysis model weighting each strata by sample size [44]. Association test outputs were aligned to the same reference alleles and weighted z-scores were computed for each SNP. The square roots of each sample size were used as the proportional weights. 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 × 10⁻⁸ used in conventional GWAS adjusted for the 3 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 [45]. Topologically associating domains (TADs) were identified using TopDom [46] 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 [30] were adjusted and normalized as described above. Subtype classifiers were modeled on the adjusted
trait values and cluster assignments from the genotyped clustering cohort. Several classification methods were compared using 10-fold cross-validation, including support vector machine (SVM), random forest (RF), Gaussian mixed model (GMM), and quadratic discriminant analysis (QDA) \[47\]. 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 \[19\]. Therefore, there was some sample overlap between the training and test data: 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

Clustering was first performed in a cohort of 893 genotyped PCOS cases (\textit{Table 1, S3 Table}). The clustering revealed 2 distinct phenotypic subtypes: 1) a group (23\%, 207/893) characterized by higher LH and SHBG levels with relatively low BMI and Ins0 levels, which we designated "reproductive," and 2) a group (37\%, 329/893) characterized by higher BMI and Glu0 and Ins0 levels with relatively low SHBG and LH levels, which we designated "metabolic" (Fig 1). The key traits distinguishing the reproductive and metabolic subtypes were BMI, insulin, SHBG, glucose, LH, and FSH, in order of importance according to relative unpaired Wilcoxon rank–sum test statistics (Fig 2). The remaining cases (40\%, 357/893) demonstrated no distinguishable pattern regarding their relative phenotypic trait distributions and were designated "indeterminate" (S4 Table, S5 Table). The reproductive and metabolic subtypes clustered along opposite ends of the SHBG versus Ins0/BMI axis, which was highly correlated with the first PC of the adjusted quantitative traits (Fig 3). The reproductive subtype was the most stable cluster, with a mean bootstrapped Jaccard coefficient \((J)\) of 0.61, followed by the metabolic subtype with \((J) = 0.55\). The indeterminate group did not appear to capture any discernable pattern within the data \((J = 0.41)\) and was both overlapping and intermediate between the reproductive and metabolic subtypes on the SHBG versus Ins0/BMI axis.
Fig 3. PCA plot of quantitative traits for genotyped PCOS clustering cohort.

Genotyped PCOS cases are plotted on the first 2 PCs of the adjusted 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 PCs are shown with black arrows.

BMI, body mass index; DHEAS, dehydroepiandrosterone sulfate; FSH, follicle-stimulating hormone; Glu0, fasting glucose; Ins0, fasting insulin; LH, luteinizing hormone; PC, principal component; PCA, principal component analysis; PCOS, polycystic ovary syndrome; SHBG, sex hormone binding globulin; T, testosterone.

https://doi.org/10.1371/journal.pmed.1003132.g003

Table 1. Quantitative traits in cluster analysis PCOS cohorts by assay method.

https://doi.org/10.1371/journal.pmed.1003132.t001

The clustering procedure was then repeated in an independent, ungenotyped cohort of 263 NIH PCOS cases diagnosed according to the same criteria as the genotyped clustering cohort (Table 1). The clustering yielded similar results, with a comparable distribution of reproductive (26%, 68/263, χ² = 0.57), metabolic (39%, 104/263, χ² = 0.46), and indeterminate clusters (35%, 91/263, χ² = 0.40) (Fig 4).

Subtype genetic associations

Genome-wide association testing was performed for each of the 3 subtypes: reproductive, metabolic, and indeterminate (Table 2). We identified alleles in 4 novel, to our knowledge, loci that were associated with the reproductive PCOS subtype at genome-wide significance (chromosome [chr] 1 p36.21 PRDM2/KAZN, P = 2.23 × 10⁻¹⁰; chr2 q37.3 IQCA1, P = 2.76 × 10⁻⁹; chr4 q22.3 BMPR1B/UNC5C, P = 9.71 × 10⁻⁹; chr5 p14.2–p14.1 CDH10, P = 1.17 × 10⁻⁸) and one novel, to our knowledge, locus that was significantly associated with the metabolic subtype (chr2 q24.2–q24.3 KCNH7/FIGN, P = 1.03 × 10⁻⁸). Association testing on the indeterminate subtype replicated the 11p14.1 FSHB locus from our original GWAS [19] (Table 3; Figs 5 and 6).
Fig 5. Genome-wide association results.
Manhattan plots for (a) reproductive, (b) metabolic, and (c) indeterminate PCOS subtypes. The red horizontal line indicates genome-wide significance \((P \leq 1.67 \times 10^{-8})\). Genome-wide significant loci are colored in green and labeled according to nearby gene(s). Quantile–quantile plots with genomic inflation factor, \(\lambda_G\), are shown adjacent to corresponding Manhattan plots. PCOS, polycystic ovary syndrome. 
https://doi.org/10.1371/journal.pmed.1003132.g005

Fig 6. Risk allele ORs in PCOS and PCOS subtypes.
ORs with 95% CIs and association P-values from the Stage 1 discovery sample are shown for each subtype-specific 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 because of low allele counts or low imputation confidence. CI, confidence interval; OR, odds ratio; PCOS, polycystic ovary syndrome; SNP, single nucleotide polymorphism. 
https://doi.org/10.1371/journal.pmed.1003132.g006

Table 2. Demographic characteristics of GWAS subtypes and controls. 
https://doi.org/10.1371/journal.pmed.1003132.t002

Table 3. Genome-wide significant associations with PCOS subtypes. 
https://doi.org/10.1371/journal.pmed.1003132.t003

The strongest association signal with the reproductive subtype appeared in an intergenic region of 1p36.21 579 kb downstream of the PRMD2 gene and 194 kb upstream from the KAZN gene (Fig 7A). The lead SNP in the locus (rs78025940; odds ratio [OR] = 4.75, 2.82–7.98 95% confidence interval [CI], \(P_1 = 2.16 \times 10^{-10}\)). The meta-analysis \(P\)-value \(P_{\text{meta}} = 2.23 \times 10^{-10}\) 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 \(P_{\text{meta}} = 2.14 \times 10^{-8}\) and was genotyped only in Stage 1 \(OR = 5.57, 3.24–9.56 95\% CI, P_1 = 2.08 \times 10^{-3}\). In ovarian tissue, the SNPs appear to be centrally located within a large 2 Mb TAD stretching from the FHAD1 gene to upstream of the PDPN gene (Fig 8).
Fig 7. 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. LD, linkage disequilibrium; SNP, single nucleotide polymorphism.
https://doi.org/10.1371/journal.pmed.1003132.g007

The 2q37.3 locus spanned a 50-kb region of strong LD overlapping the 5′ end and promoter region of the \( IQCA1 \) gene (Fig 7B). The SNP rs76182733 had the strongest association in this locus (\( P_{\text{meta}} = 2.76 \times 10^{-8} \)) with the reproductive subtype. The signal was genotyped only in Stage 1 (OR = 5.68, 3.00–10.78 95% CI, \( P = 2.69 \times 10^{-8} \)) and was imputed with an imputation \( r^2 \) value of 0.84.

The 4q22.3 locus spanned a 500-kb region of LD, including the 3′ ends of both the \( BMPR1B \) and \( UNC5C \) genes (Fig 7C). The most strongly associated SNP (rs17023134; \( P = 9.71 \times 10^{-6} \)) in the locus was within an intronic region of \( UNC5C \) and was associated with the reproductive subtype in the Stage 1 discovery sample with genome-wide significance (OR = 3.02, 2.06–4.42 95% CI, \( P_1 = 1.40 \times 10^{-8} \)) but was not significantly associated in the Stage 2 replication analysis (OR = 1.71, 0.98–2.99 95% CI, \( P_2 = 7.8 \times 10^{-2} \)). 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 (\( P_{\text{meta}} = 1.46 \times 10^{-8} \)) 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, \( P_1 = 6.72 \times 10^{-3} \)) and Stage 2 (OR = 2.40, 1.51–3.82 95% CI, \( P_2 = 4.7 \times 10^{-4} \)) analyses with nearly identical effect sizes.

In the 5p14.2–p14.1 locus, 83 kb upstream of the \( CDH10 \) gene (Fig 7D), 2 adjacent SNPs (rs7735176, rs16893866) in perfect LD were equally associated with the reproductive subtype with genome-wide significance (\( P_{\text{meta}} = 1.17 \times 10^{-8} \)). The SNPs were imputed in Stage 1 (OR = 5.09, 2.62–9.86 95% CI, \( P_1 = 1.14 \times 10^{-8} \)) with an imputation \( r^2 \) value 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 200 kb downstream from \( FIGN \) and 500 kb upstream from \( KCNH7 \) (Fig 7E). The lead SNP, rs55762028, was imputed in Stage 1 only (OR = 1.86, 0.92–3.75 95% CI, \( P_1 = 9.17 \times 10^{-3} \), \( P_{\text{meta}} = 1.03 \times 10^{-8} \)). In pancreatic tissue, the lead SNPs appear to be located terminally within a 1.3-Mb TAD encompassing the \( FIGN \) gene and reaching upstream to the \( GRB14 \) gene (Fig 9).
Association testing on the indeterminate cases replicated the genome-wide significant association in the 11p14.1 FSHB locus (Fig 7F) identified in our original GWAS (14). The lead SNP (rs10835638; $P_{\text{meta}} = 4.94 \times 10^{-12}$) and lead genotyped SNP (rs10835646; $P_{\text{meta}} = 2.75 \times 10^{-11}$) 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: $P_{\text{meta}} = 2.96 \times 10^{-10}$; rs11031005: $P_{\text{meta}} = 2.91 \times 10^{-10}$) and are in LD with the lead SNP rs10835638 ($r^2 = 0.59$) [40]. The other significant signals from our original GWAS [19] were not reproduced in any of the subtype association tests performed in this study (Table 4).

Subtypes in PCOS families

The RF classifier yielded the lowest mean subtype misclassification rate (13.2%) compared to the SVM (13.6%), GMM (17.0%), and QDA (18.1%) models, according to 10-fold cross-validation of the genotyped clustering cohort. Affected women from the family-based cohort were classified accordingly using an RF model. Seventy-three daughters of the 83 affected women from the family-based cohort had complete quantitative trait data available for subtype classification. Seventeen (23.3%) were classified as having the reproductive subtype of PCOS, and 22 (30.1%) were classified as having the metabolic subtype. Of 14 subtyped sibling pairs, only 8 were concordantly classified (57.1%); however, there was only one instance of the reproductive subtype and metabolic subtype occurring within the same nuclear family because the remaining discordant pairs each featured one indeterminate member. The proportion of affected women with one or more of the previously identified [39] deleterious, rare variants in DENND1A varied by subtype. Women classified as having the reproductive subtype of PCOS were significantly more likely to carry one or more of the DENND1A rare variants compared to other women with PCOS ($P = 0.03$; Fig 10). The distribution of affected women and DENND1A rare variant carriers are shown relative to the adjusted quantitative trait PCs in Fig 11.
cells and in gonadotropin-releasing hormone (GnRH) neurons which transcribes a type-I anti-Müllerian hormone (AMH) receptor highly expressed in granulosa cells. The reproductive subtype association in the 4q22.3 locus overlapped with the same TAD as the PRDM2 gene (Figs 2, 3 and 4B). Our findings are in contrast to the recent PCOS GWAS meta-analysis [19–23] 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 [26–29], but no subsequent investigation into the biologic relevance of the resulting subtypes using genetic association analyses.

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 the mean trait values were between those of the reproductive and metabolic subtypes. The relative trait distributions and results of the PCAs (Figs 2, 3 and 4B) 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, different 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 [19], suggesting that the indeterminate group was also more genetically homogenous after the reproductive and metabolic subtypes were removed from the analysis.

Two of the loci associated with the reproductive subtype implicate novel biologic pathways in PCOS pathogenesis. The association signal on chr1 appeared downstream of and within the same TAD as the PRDM2 gene (Figs 7A and 8), which is an estrogen receptor coactivator [50] that is highly expressed in the ovary [51] and pituitary gland [52]. In an independent rare variant association study in PCOS families, PRDM2 demonstrated the fifth strongest gene-level association with altered hormonal levels in PCOS families (P = 6.92 × 10\(^{-3}\)) out of 339 genes tested [30]. PRDM2 binds with ligand-bound estrogen receptor alpha (ER\(\alpha\)) to open chromatin at ER\(\alpha\) target genes [50,53]. PRDM2 also binds with the retinoblastoma protein [54], which has been found to play an important role in follicular development in granulosa cells [55,56].

The reproductive subtype association in the 4q22.3 locus overlapped with the BMPR1B gene, which transcribes a type-I anti-Müllerian hormone (AMH) receptor highly expressed in granulosa cells and in gonadotropin-releasing hormone (GnRH) neurons [57] that regulates follicular...
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 [52]. The 5p14.2–p14.1 locus overlapped the promoter region of the CDH10 gene. CDH10 is almost exclusively expressed in the brain [51] and is putatively involved in synaptic adhesions, axon outgrowth, and guidance [56].

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 [74]. 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. Further, the fact that one of the genes associated with the reproductive subtype, PRDM2, was associated with PCOS quantitative traits in our family-based analysis [30] 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 (OR 3.02–5.68) (Table 3), were substantially greater than the effects (OR 0.70–1.51) observed for alleles associated with PCOS diagnosis in previous GWAS [19–23]. 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 [75, 76]. Accordingly, in contrast to the common variants (effect allele frequency [EAF] > 0.05) associated with PCOS in previous GWAS [19–23], the alleles associated with the subtypes were all of low frequency (EAF 0.01–0.05; Table 3). 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 “winner’s curse” [77, 78], 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 12 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 [79, 80]; 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 because of allelic heterogeneity or 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 because these phenotypes have minimal metabolic risk [81, 82]. Indeed, in a previous effort to identify phenotypic subtypes in Rotterdam PCOS cases...
[29], 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, whereas the cluster that most closely resembled the metabolic subtype constituted only 12% of the total sample, but 98% met the NIH diagnostic criteria. Furthermore, trait distributions may vary among women with PCOS from different geographic locations, such as in some of the sites excluded from our analysis because of incomplete quantitative trait data. For example, European PCOS cases have a lower prevalence of obesity compared to US PCOS cases [83]. Because of the within-cohort normalization of quantitative traits prior to clustering, our method is well-suited for identifying subsets of cases within populations, but therefore, it may not be suitable for directly comparing subtype membership between populations.

Our clustering cohorts included only US-based women of European ancestry. It will be of considerable importance to investigate whether subtypes are present in women with PCOS of other ancestries and geographic regions. Women with PCOS of diverse races and ethnicities have similar reproductive and metabolic features [84–86]. However, there are differences in the severity of the metabolic defects due to differences in the prevalence of obesity [89], as well as to racial/ethnic differences in insulin sensitivity [87–88]. Furthermore, the susceptibility loci associated with subtypes in other ancestry groups may differ because the low frequency and large effect size of the variants associated with the reproductive subtype in our European cohort suggests these variants are of relatively recent origin and therefore may be population-specific [1,89,90].

While the bootstrapping and clustering in an independent cohort demonstrated that the subtypes were reproducible, the Jaccard scores were relatively modest, with only the reproductive subtype yielding a mean Jaccard coefficient $\geq 0.6$. At least part of this outcome was likely due to the fact that all traits were fitted to a normal distribution using an inverse normal transformation prior to clustering. This transformation was done in order to prevent outliers from dictating cluster formations but also likely resulted in greater cluster overlap. Consequently, the metabolic and reproductive clusters we identified appear to represent opposite ends of a phenotypic spectrum with imperfect delineation. This spectrum, however, aligns with the known pathophysiology of PCOS and is bolstered by our genetic association findings. Our approach, therefore, appears to be a more reliable way of identifying subgroups of PCOS cases who have been noted in the literature [91] but have previously been defined using only a single trait like BMI [92–95] or by diagnostic criteria that do not reflect the genetic heterogeneity of the disorder [92]. Perhaps future studies that use clustering to identify reproductive and metabolic subtypes in PCOS can omit nondistinguishing traits such as DHEAS and T in an effort to reduce noise and improve subtype delineation and reproducibility.

In conclusion, using an unsupervised clustering approach featuring quantitative hormonal and anthropometric data, we identified reproductive and metabolic subtypes of PCOS that appeared to have distinct genetic architecture. The genomic loci that were significantly associated with either of these subtypes include a number of new, to our knowledge, highly plausible PCOS candidate genes. Moreover, our results demonstrate that precise phenotypic delineation, resulting in more homogeneous subsets of affected individuals, can be more powerful and informative than increases in sample size for genetic association studies. Our findings indicate that further study into the genetic heterogeneity of PCOS is warranted and could lead to a transformation in the way PCOS is classified, studied, and treated.

Supporting information

S1 Checklist_ STREGA checklist.
https://doi.org/10.1371/journal.pmed.1003132.s001
(DOCX)

S1 Table_. GWAS cohorts used in cluster analysis.
The cohorts from the Hayes, Urbanek, and colleagues PCOS GWAS and corresponding numbers of samples that were included in the clustering analysis are shown by GWAS cohort, adapted from Hayes, Urbanek, and colleagues. [19]: Table 1 and Supplemental Data Tables 9 and 10. GWAS, genome-wide association study; PCOS, polycystic ovary syndrome.
https://doi.org/10.1371/journal.pmed.1003132.s002
(DOCX)

S2 Table_. Age and BMI distributions for subjects excluded from cluster analysis.
Median age and BMI values are shown with the 25th and 75th percentiles for the subjects included in the cluster analysis and for those from the same GWAS cohorts who were excluded because of missing quantitative trait data. Distributions were compared using unpaired Wilcoxon rank–sum tests. P-values are unadjusted. BMI, body mass index; GWAS, genome-wide association study.
https://doi.org/10.1371/journal.pmed.1003132.s003
(XLSX)

S3 Table_. Assays used to measure quantitative traits.
Assays used to measure quantitative trait levels are listed by trait, then by site and methodology combination. Unless otherwise noted, kits were used per the manufacturer's instructions.

*Calibrated to WHO 1st International Standard #95/560.  
Diagnostic Products Corporation (DPC) (Los Angeles, CA, USA) [Note: In April 2006, DPC was acquired by Siemens Medical Solutions USA, Inc. (Malvern, PA, USA)].  
Diagnostic Systems Laboratories, Inc. (DSL) (Webster, TX, USA) [Note: In October 2005, DSL was acquired by Beckman Coulter (Brea, CA, USA)].  
Siemens Medical Solutions USA, Inc. (Malvern, PA, USA).  
Beckman Coulter, Inc. (Brea, CA, USA) [Note: In June 2011, Beckman Coulter was acquired by Danaher Corporation (Washington, DC, USA)].  
Linco Research, Inc. (St. Charles, MO, USA).  
American Laboratory Products Company (ALPCO) (Salem, NH, USA).  
BWH, Brigham and Women’s Hospital; DHEAS, dehydroepiandrosterone sulfate; ELISA, enzyme-linked immunosorbent assay; FSH, follicle-stimulating hormone; GO, Glucose; G0, fasting glucose; HMC, Pennsylvania State Milton S. Hershey Medical Center; IRMA, immunoradiometric assay; I0, fasting insulin; LH, luteinizing hormone; NU, Northwestern University; RIA, radioimmunoassay; SHBG, sex hormone binding globulin; T, testosterone; UVA, University of Virginia.

https://doi.org/10.1371/journal.pmed.1003132.s004

**S4 Table.** Quantitative traits in genotyped PCOS cohort by cluster.

Median trait values are shown with 25th and 75th percentiles for each clustering subtype. Details for each assay method are provided in **S3 Table**. BMI, body mass index; DHEAS, dehydroepiandrosterone sulfate; FSH, follicle-stimulating hormone; G0, fasting glucose; I0, fasting insulin; LH, luteinizing hormone; N, total number; PCOS, polycystic ovary syndrome; SHBG, sex hormone-binding globulin; T, testosterone.

https://doi.org/10.1371/journal.pmed.1003132.s005

**S5 Table.** Subtypes of Stage 1 GWAS samples.

Subtypes are provided for each of the 555 Stage 1 GWAS samples included in the clustering analysis according to their dbGaP SUBJIDs. dbGaP, database of Genotypes and Phenotypes; GWAS, genome-wide association study; SUBJID, subject ID.

https://doi.org/10.1371/journal.pmed.1003132.s006

**Acknowledgments**

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 in the GWAS of Hayes and Urbanek and colleagues [19] 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 GWAS of Hayes and Urbanek and colleagues [19] 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, USA); Corrine K. Welt, MD (University of Utah School of Medicine, Salt Lake City, UT, USA; formerly of Massachusetts General Hospital, Boston, MA, USA); Ahmed H. Kissebah (deceased, Medical College of Wisconsin, Milwaukee, WI, USA); Ricardo Azziz, MD (State University of New York, NY, USA; formerly of University of Alabama at Birmingham, AL, USA); and Evanthia Diamanti-Kandarakis, MD, PhD (University of Athens Medical School, Greece).

**References**

   [View Article](#) [PubMed/NCBI](#) [Google Scholar](#)

   [View Article](#) [PubMed/NCBI](#) [Google Scholar](#)

   [View Article](#) [PubMed/NCBI](#) [Google Scholar](#)

   [View Article](#) [PubMed/NCBI](#) [Google Scholar](#)

   [View Article](#) [PubMed/NCBI](#) [Google Scholar](#)


29. 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. pmid:25662806


62. Sugiura K, Su YQ, Eppig JJ. Does bone morphogenetic protein 6 (BMP6) affect female


   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar

   View Article • PubMed/NCBI • Google Scholar