Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hierarchical Clustering Identifies Reproducible Polycystic Ovary Syndrome Subtypes with Ancestry-Specific Genetic Associations

Kelly Brewer
Icahn School of Medicine at Mount Sinai
Ryan Sisk
Icahn School of Medicine at Mount Sinai
Hyejin Lee
Ewha Womans University School of Medicine
Chun Li
University of Southern California
Andrea Dunaif
andrea.dunaif@mssm.edu

Icahn School of Medicine at Mount Sinai

Research Article

Keywords: PCOS subtypes, Genome-wide association study, Hierarchical clustering, Trans-ancestry meta-analysis

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11676895/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Journal: Genome Medicine

Title: Hierarchical Clustering Identifies Reproducible Polycystic Ovary Syndrome Subtypes with Ancestry-Specific Genetic Associations

Authors: Kelly Brewer[1], Ryan Sisk[1], Hyejin Lee[2], Chun Li[3], Andrea Dunaif[1]

Affiliations: [1]Icahn School of Medicine at Mount Sinai, New York, NY, USA; [2]Department of Internal Medicine, Ewha Womans University School of Medicine, Seoul, Korea; [3]Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA.

Author list footnotes:

Corresponding author: andrea.dunaif@mssm.edu

Keywords: PCOS subtypes, Genome-wide association study, Hierarchical clustering, Trans-ancestry meta-analysis

ABSTRACT

Background

Complex diseases often represent the phenotypic convergence of multiple genetic etiologies under one diagnosis. Resolving this heterogeneity can increase power for genetic discovery and delineate causal pathways. PCOS diagnostic criteria are based on expert opinion, and the resulting phenotypes appeared to share similar genetic architecture. However, hierarchical clustering of quantitative traits has resolved discrete reproductive and metabolic subtypes with distinct genetic architectures; cases without stable cluster membership form a background subtype. Because the subtypes are defined from phenotypic traits alone, finding that they carry distinct genetic signals provides an orthogonal confirmation of clustering validity that internal metrics and trait-correlated data cannot. Here we test whether these subtypes are reproducible across ancestries, identify susceptibility loci, and validate the clustering method.

Methods

We performed hierarchical clustering of eight reproductive and metabolic traits in 1,449 European-ancestry and 418 Korean PCOS cases. We conducted overall and subtype-specific genome-wide association studies in the European cohort (991 cases, 3,877 controls) and the Korean cohort (417 cases, 926 controls), followed by trans-ancestry meta-analysis. We compared hierarchical and k-means clustering in 874 European cases using internal stability and separation metrics, with subtype-specific association as orthogonal validation.

Results

Hierarchical clustering reproduced the reproductive and metabolic subtypes in both cohorts. Before subtyping, the European cohort confirmed two established loci and identified one novel signal, whereas the Korean cohort yielded no genome-wide significant associations. Subtype-specific analysis identified additional loci that were only nominally significant before subtyping. Trans-ancestry meta-analysis identified two robust, subtype-restricted signals with concordant cross-ancestry effects. Remaining subtype-specific loci had minor allele

counts below the pre-specified threshold and are reported as provisional signals requiring replication. Hierarchical clustering with eight traits produced more stable subtyping with subtype-specific genetic associations; k-means yielded slightly greater geometric separation but no associated loci.

Conclusions

Phenotypically defined PCOS subtypes carry distinct genetic architectures that replicate across ancestries. Trans-ancestry meta-analysis identifies shared, subtype-restricted risk with concordant effects. Because the subtypes are defined solely on phenotypic traits, their distinct genetic associations provide orthogonal validation, supporting hierarchical clustering as a generalizable strategy for resolving complex trait heterogeneity.

Words 339 (350)

INTRODUCTION

Polycystic ovary syndrome (PCOS) is the most common endocrine disorder in premenopausal women and a major cause of anovulatory infertility and increased cardiometabolic risk[1, 2]. Although the etiology of PCOS remains unknown, it is recognized as a complex disorder arising from the interplay between genetic susceptibility and environmental factors[2]. Disease classification has long balanced lumping clinically similar disorders under one diagnosis against splitting them into distinct entities. McKusick recognized that similar phenotypes often arise from different genetic causes and that delineating these entities is a prerequisite for understanding etiology[3]. Current PCOS diagnostic criteria, based on expert opinion[1], define phenotypes with largely shared genetic architecture[4, 5], indicating that these classifications do not capture biologically relevant differences.

This heterogeneity complicates patient classification, gene discovery, and targeted treatment[6]. Furthermore, controversy regarding PCOS diagnostic criteria contributes to diagnostic delays and high rates of patient dissatisfaction[7]. An international renaming group recently designated the disorder polyendocrine metabolic ovarian syndrome (PMOS) without changing the diagnostic criteria or evaluating replicated evidence for genetically distinct subtypes[6, 8, 9]; we therefore retain the term PCOS.

Many complex diseases pose the same problem. They likely represent the convergence of multiple genetic etiologies under a single clinical diagnosis[10-15]. Unsupervised clustering can resolve this heterogeneity by identifying structure within high-dimensional trait data, revealing biologically and genetically distinct subtypes across type 2 diabetes (T2D), obesity, and depression[2, 13, 16-18]. Splitting cases into more homogeneous subgroups can also increase power for genetic association[19]. We previously applied hierarchical clustering (HC) to quantitative reproductive and metabolic traits in our well-characterized European (EUR) ancestry PCOS cohort and identified reproducible reproductive, metabolic, and background subtypes that differed in both phenotypic profile and genetic architecture, with subtype-specific genome-wide association study (GWAS) uncovering loci not detected in overall case-control analysis[6]. However, 4 of 6 loci could not be tested for replication because the lead SNPs were neither genotyped nor imputable on the Metabochip, the cardiometabolic custom array used for replication[20].

Subtype-specific genetic association provides a stringent test of clustering validity because germline genotype is randomly allocated at gamete formation, fixed at conception, and therefore causally prior to the disease process and the phenotypic variation used to define subtypes[21, 22]. Distinct genetic associations among subgroups defined solely by phenotypic traits provide orthogonal evidence supporting etiologic heterogeneity[6, 23]. This distinguishes genetic validation from internal cluster indices, which merely quantify geometric compactness or separation and will partition data even in the absence of true

subtypes, and from clinical outcomes or multiomic data, which are themselves correlated with the traits used for clustering[24, 25].

In the current study, we expanded our analysis in a larger EUR cohort using a denser imputation panel (TOPMed), which enabled imputation of nearly all previously significant lead SNPs, including all but one previously reported subtype lead SNP in both discovery and replication samples[6]. We added a previously reported Korean (KOR) cohort[26] to test whether the subtypes are present across ancestries. Comparing HC with k-means, only HC produced subtypes with distinct genetic architectures, providing orthogonal validation of the clustering method. Subtype-specific GWAS in the KOR cohort identified ancestry-specific loci not detected in the overall analysis[26], and trans-ancestry meta-analysis confirmed the established *C9orf3/FANCC* and *FSHB* loci in the metabolic and background subtypes, respectively, with concordant effects across ancestries. Together, these findings show that HC-based subtyping captures biologically meaningful variation and identifies susceptibility loci masked by PCOS diagnostic heterogeneity.

METHODS

Study Cohorts. This study used biochemical and genotype data from our previously reported European-ancestry (EUR) and Korean (KOR) PCOS case-control cohorts[20, 26]. All subjects were unrelated. Cases were 13–45 years old and fulfilled National Institutes of Health (NIH) diagnostic criteria of hyperandrogenism and chronic anovulation, defined as eight or fewer menses annually, after exclusion of known disorders of the ovaries, adrenals, and pituitary[1, 27]. These criteria also fulfill Rotterdam and Androgen Excess-PCOS Society Criteria for PCOS[1].

The following eight quantitative phenotypic traits were selected a priori based on biological knowledge of PCOS to capture its core reproductive and metabolic pathophysiology[2, 6, 27]: body mass index (BMI), luteinizing hormone (LH), follicle stimulating hormone (FSH), testosterone (T), dehydroepiandrosterone sulfate (DHEAS), sex hormone binding globulin (SHBG), fasting glucose (Glu0), and fasting insulin (Ins0). We included only PCOS cases with complete data for all eight traits. Eight hundred seventy-four EUR PCOS cases also had AMH (anti-Müllerian hormone) levels available. The EUR cohort included 99 additional cases (66 discovery, 33 replication) who had assays for phenotypic traits completed after the publication of our previous subtyping study[6]. We used both phenotyped and population-based control samples. Phenotyped control women in both cohorts were aged 15–45 years, with regular menses and normal T levels, and were not receiving contraceptive steroids for at least 3 months prior to the study [20]. Population-based control DNA samples (n=2951) for the EUR GWAS discovery sample were obtained from the NUgene biobank[28] from women of European ancestry, ages 18-97 years. The levels of LH, FSH, T, DHEAS, SHBG, Glu0, Ins0, and AMH were measured as previously reported[20]. The studies were approved by the Institutional Review Boards of Icahn School of Medicine at Mount Sinai and Ewha Womans University Medical Center. We obtained written informed consent from each participant before the study.

Clustering

We clustered 1,449 EUR and 418 KOR cases using the eight phenotypic traits listed above; in the 874 EUR cases with available AMH measurements, we repeated clustering with AMH as a ninth trait. Each quantitative trait value was log-normalized and adjusted for age and assay method, which varied by study site[20], using linear regression, after which an inverse normal transformation was applied to ensure equal scaling across traits. The normalized trait residuals were clustered by unsupervised agglomerative hierarchical clustering, using a generalization of Ward's minimum variance method[24, 29] applied to Manhattan distances

between trait values. This approach is widely used for quantitative traits and was previously applied to identify PCOS subtypes[6]. To determine the optimal clustering method, cases were also clustered using the k-means algorithm (Figure S4). Both clustering methods were run with two, three, or four groups (k = 2, 3, or 4) using eight or nine traits; only the 874 cases with all nine traits were included in the comparative analysis.

The optimal number of clusters was determined in both clustering approaches using ConsensusClusterPlus[30] by evaluating stability using pairwise consensus proportions, which measure how often two items are assigned to the same cluster across resampling iterations. Differences in adjusted, normalized trait values between subtypes were assessed using Kruskal–Wallis and unpaired Wilcoxon rank–sum tests corrected for multiple testing (Bonferroni). Average silhouette scores were computed using the silhouette function from the R cluster package[31].

Genotyping and imputation

The discovery cohort samples of European ancestry were genotyped using the Illumina OmniExpress (HumanOmniExpress 12v1_C; San Diego, CA, USA) single nucleotide polymorphism (SNP) array. The replication cohort samples were genotyped using the Metabochip[32] with added custom variant content based on ancillary studies and the discovery results[20]. The KOR cohort samples were genotyped using the Illumina HumanOmni1-Quad v1 array. SNPs were filtered according to minor allele frequency (MAF ≥ 0.01), Hardy–Weinberg equilibrium (P ≥ 1 × 10-6), call rate (≥ 0.975), minor allele count (MAC > 5), Mendelian concordance, and duplicate sample concordance. Only autosomal SNPs were considered. Ancestry was evaluated using principal component analysis (PCA)[33] on LD-pruned SNPs[20]. 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 were phased using Eagle v2.4 (v2.r790) and then imputed to the TOPMed imputation panel (r2) using the Minimac4 via the TOPMed Imputation Server[34]. The TOPMed reference panel includes over 97k samples and over 308 million variants built from high-quality whole genome sequencing data[35], which shows improved imputation quality for rarer variants compared to other reference panels[36]. Imputed SNPs with an allelic r^2 below 0.3 were removed from the analysis[37].

Ancestry and relatedness

Genetic PCs were calculated for all cohorts using Plink v1.9[38]. The first five principal PCs explained greater than 80% of genetic variation and were used as covariates in all analyses. Relatedness was assessed using KING[39]. Samples that were determined to be related up to the 4th degree were excluded from analysis (14 in discovery, 49 in replication, 1 in KOR cohort).

Functional annotation

All variants from the ancestry-specific and trans-ancestry meta-analyses were annotated using FUMA v1.6.1[40]. The SNP2GENE function was used to define lead SNPs and the independent significant SNPs in linkage disequilibrium (LD) with them ($r^2 \leq 0.6$). Independent significant SNPs were defined by a condFDR < 0.05 and LD $r^2 \leq 0.6$, and lead SNPs by a minimum LD threshold of $r^2 \leq 0.1$; SNPs that did not reach nominal significance ($P \leq 0.05$) in the replication cohort were excluded. SNP2GENE was then used to map SNPs to their nearest genes. We predicted functional consequences from the Combined Annotation-Dependent Depletion (CADD) score, the RegulomeDB score (probability of regulatory function), and the minimum chromatin state (predicted transcription/regulatory effects).

Pathway enrichment analysis

Genes mapped to genome-wide significant loci were tested for pathway enrichment using the KEGG 2021 Human and Reactome 2024 gene sets on the Enrichr platform[41, 42]. No pathway remained significant after correction for multiple testing; pathways reaching nominal significance ($P < 0.05$) with biological relevance to PCOS are reported as prioritized and interpreted as hypothesis-generating.

Gene expression analysis

To assess whether candidate genes were differentially expressed in PCOS-relevant tissues, publicly available transcriptomic datasets comparing PCOS cases and matched controls were queried from the NCBI Gene Expression Omnibus (GEO)[43]. We extracted expression data for GWAS-mapped genes and calculated log fold change, beta coefficient, t-statistic, and nominal P value. Genes with $P < 0.05$ were considered differentially expressed.

Cross-phenotype SNP analysis

Population-specific GWAS summary statistics for reproductive and metabolic traits were obtained from the NHGRI-EBI GWAS Catalog[44] and, for the East Asian (EAS) population, primarily from a comprehensive multi-trait Korean GWAS[45], because EAS summary statistics for these traits were otherwise limited (Table S3). Results were available for 17 traits in the EUR population and 31 in the EAS population; for EAS, 28 traits were obtained from the multi-trait Korean study and the remainder from the GWAS Catalog. Lead SNPs from our cohorts were queried against the corresponding population-specific summary statistics. Statistical significance was defined by Bonferroni correction for the number of traits tested ($P < 0.0029$ for 17 EUR traits; $P < 0.0016$ for 31 EAS traits). Because several reproductive and metabolic traits are correlated, this correction is conservative, as it treats each trait as an independent test.

Association testing

Association testing was performed separately for the EUR and KOR cohorts. Of the 1,449 EUR cases used for clustering, 620 were genotyped on the OMNI array (discovery) and 371 on the Metabochip (replication); of the 418 KOR cases used for clustering, 417 were analyzed for association after one sample was excluded due to relatedness. A two-stage discovery–replication GWAS was conducted in the EUR cohort. The KOR cohort, previously diagnosed by Rotterdam criteria[26], which encompass the NIH criteria[1], was restricted to NIH PCOS cases and analyzed for genome-wide association. Unless specified otherwise, the following steps were applied to both cohorts. Case-control logistic regression was performed in SNPTEST[46] under an additive genetic model, adjusting for the first five ancestry PCs. Association outputs underwent quality control in GWASinspector[47]. P value-based fixed-effects meta-analyses were carried out using METAL weighted by sample size[48], first across the EUR discovery and replication samples and then across the EUR and KOR cohorts for trans-ancestry analysis. For subtype GWAS, cases were restricted to the subtypes defined by hierarchical clustering. Lead variants were annotated to the nearest gene, unless another nearby gene showed stronger PCOS-relevant functional evidence. We report effect allele frequency (EAF) for lead variants, as the effect allele is the PCOS-increasing allele. Linkage disequilibrium (LD) with the previously reported variants was determined using LDlink platform's LDproxy tool[49] based on the European populations of the 1000 Genomes Project reference panel.

Because the calibration of single-marker tests depends on minor allele count (MAC) rather than allele frequency[50], we report stage-specific effect allele frequency, MAC, and imputation r^2 for every lead variant in each cohort and stage in which it was testable, and flag variants that could not be imputed (Table 2). Classical tests deviate from expected null behavior when the minor allele is observed fewer than approximately 80 times, largely independent of total sample size[51]. We therefore designated lead variants with stage-specific MAC < 80 as provisional and interpreted them as requiring independent replication, regardless of the observed P value. Loci were designated robust only when nominally

significant with concordant direction of effect in both stages and MAC ≥ 80 in each. To confirm that findings were not artifacts of low-count calibration, all tests were rerun with Firth bias-corrected logistic regression in PLINK 2.0[38], which remains calibrated for rare variants and unbalanced case–control designs[50]; Firth P values are reported in Table 2.

RESULTS

PCOS subtypes

Hierarchical clustering reproduced the reproductive and metabolic subtypes in both the 1,449 EUR and 418 KOR PCOS cases with complete data for the eight reproductive and metabolic traits (Table 1, Figure 1). Cases lacking stable membership in either cluster comprised the residual background subtype. In the EUR cohort, 26% (n = 368) of cases belonged to the reproductive subtype, 53% (n = 771) to the metabolic subtype, and 21% (n = 310) to the background subtype; the corresponding proportions in KOR were 20% (n = 84), 31% (n = 129), and 49% (n = 205).

Because the eight traits defined the clusters, their differences across subtypes characterize rather than validate the subtypes; subtype-specific genetic associations described below provide validation. Age did not differ by subtype in either cohort. The subtype-defining contrasts were concordant across ancestries. BMI showed the largest separation, highest in the metabolic and lowest in the reproductive subtype in both cohorts (all pairwise $P < 0.001$). Glu0 and Ins0 were also highest in the metabolic subtype ($P < 0.001$). LH and SHBG were highest in the reproductive and lowest in the metabolic subtype in both cohorts ($P < 0.001$).

Three traits showed ancestry-specific subtype patterns. T was highest in the reproductive subtype in KOR but did not differ among EUR subtypes. DHEAS was highest in the background subtype in EUR and in the reproductive subtype in KOR. FSH was highest in the background subtype in EUR but similarly elevated in the reproductive and background subtypes in KOR. Pairwise comparisons were based on Z scores to account for differences among assay methods; significance values are reported in Table 1.

GWAS

Overall GWAS. Before subtyping, three loci reached genome-wide significance in the EUR GWAS. The 9q22.32 signal near *C9orf3*/*FANCC* (rs12235316, P = 9.12×10^{-12}) was in linkage disequilibrium (LD) with the previously reported rs10993397 and rs7864171 variants (r^2 = 1.00 and 0.98, respectively)[4, 20]. The 11p14.1 association near *FSHB* (rs10835649, P = 4.70×10^{-10}) was in LD with the original rs11031006 signal (r^2 = 0.83)[20]. A novel locus at 10q21.2 near *ANK3* also achieved genome-wide significance (rs4400747, P = 3.55×10^{-8}; Figure S3). The other association from the original EUR GWAS, rs804279 near *GATA4*/*NEIL2*, remained nominally significant (P = 3.12×10^{-4}). No locus reached genome-wide significance in the overall KOR analysis, consistent with the previous report[26].

Reproductive subtype GWAS. Four loci reached genome-wide significance in the EUR reproductive subtype and two in KOR (Table 2; Figures 2a and 3a). EUR associations were 1p22.2 (rs72728759) near *GBP5* (P = 3.64×10^{-8}), 3q13.32 (rs114040582) near *IGSF11* (P = 1.13×10^{-8}), 5q23.1 (rs74679007) in *SEMA6A* (P = 2.31×10^{-8}), and 12q12 (rs10785467) near *TMEM117* (P = 2.52×10^{-8}). KOR associations were 1p36.21 (rs117222858) near *KAZN* (P = 9.07×10^{-10}) and 12p12.1 (rs113200232) near *SSPN* (P = 3.44×10^{-8}). Before subtyping, all four EUR signals were nominally significant in the overall EUR GWAS. In the subtype analysis, each was nominally significant in both the discovery and replication samples, with a concordant direction of effect. Both KOR signals were nominally significant in the overall KOR GWAS; the KOR cohort was analyzed in a single stage.

TMEM117 (rs10785467, EAF 0.099) exceeded the pre-specified MAC threshold of 80 in both the discovery (MAC 615) and replication (MAC 135) stages with concordant direction of effect and is designated robust. The remaining five lead SNPs are low-frequency, with stage-specific MAC below the threshold: *GBP5* (EAF 0.013, replication MAC 23), *IGSF11* (EAF

0.028, replication MAC 48), and *SEMA6A* (EAF 0.022, replication MAC 30) in EUR, and *KAZN* (EAF 0.014, MAC 29) and *SSPN* (EAF 0.028, MAC 57) in KOR. These five loci are designated provisional and require replication in independent samples.

Cross-cohort assessment was constrained by imputation. The *GBP5*, *IGSF11*, and *SEMA6A* lead SNPs were not imputed in the KOR reproductive subtype; the *TMEM117* lead SNP was imputed but not associated (P = 0.59). The KOR *KAZN* lead SNP (rs117222858) was not imputed in EUR. A distinct *KAZN* variant, rs78025940, reached genome-wide significance in our prior reproductive subtype EUR discovery sample ($P = 2.23 \times 10^{-10}$) but was not imputable in the prior replication cohort. With TOPMed imputation, it was testable in both stages (discovery $P = 5.59 \times 10^{-7}$; replication P = 0.229) and did not reach genome-wide significance in meta-analysis ($P = 1.84 \times 10^{-4}$). The two cohorts' *KAZN* lead SNPs are not in LD.

Metabolic subtype GWAS. Two loci reached genome-wide significance in the EUR metabolic subtype and one in KOR (Table 2; Figures 2b and 3b). EUR associations were 9q22.32 (rs10761370) at the established *C9orf3*/*FANCC* locus ($P = 1.21 \times 10^{-8}$) and 1q42.13 (rs114194538) near *PSEN2* ($P = 4.03 \times 10^{-8}$). The KOR association was 13q32.3 (rs150521566) near *PCCA* ($P = 4.01 \times 10^{-8}$). Before subtyping, rs10761370 was genome-wide significant in the overall EUR analysis ($P = 4.92 \times 10^{-8}$), and rs114194538 was nominally associated ($P = 7.95 \times 10^{-5}$). rs150521566 was not associated in the overall KOR analysis (P = 0.169).

rs10761370 is common (EAF 0.472), exceeds the pre-specified MAC threshold of 80 at every stage, and was nominally significant with concordant direction of effect in both the discovery and replication samples; it is designated robust. rs114194538 (EAF 0.016) and rs150521566 (EAF 0.018) fall below the MAC threshold and are designated provisional. rs10761370 approached nominal significance in the KOR metabolic subtype (P = 0.052). Cross-cohort comparison of the remaining metabolic subtype signals was not possible because rs114194538 could not be imputed in KOR and rs150521566 could not be imputed in EUR.

Background subtype GWAS. Two loci reached genome-wide significance in the EUR background subtype and two in KOR (Table 2; Figures 2c and 3c). EUR associations were 11p14.1 (rs10835649) near *FSHB* ($P = 8.49 \times 10^{-10}$) and 2p16.3 (rs113839416) near *NRXN1-DT* ($P = 3.99 \times 10^{-8}$). KOR associations were 7q21.13 (rs12669476) near *ZNF804B* ($P = 6.59 \times 10^{-9}$) and 17q23.2 (rs11868532) near *TBX2* (OR 1.93, $P = 1.55 \times 10^{-10}$). Before subtyping, all four loci were nominally significant in their respective overall GWAS. rs10835649 additionally reached genome-wide significance in the overall EUR GWAS ($P = 4.70 \times 10^{-10}$), and rs11868532 was nominally associated in the overall KOR GWAS ($P = 1.87 \times 10^{-7}$). Within the EUR subtype analysis, both lead SNPs were nominally significant in the discovery and replication samples with concordant direction of effect. rs10835649 is common (EAF 0.160), exceeded the pre-specified MAC threshold of 80 in both the discovery (MAC 992) and replication (MAC 350) stages, and is designated robust. rs11868532 is common (EAF 0.404) and exceeds the MAC threshold (MAC 914), but, like all KOR subtype-specific signals, has not been replicated in an independent cohort. rs113839416 (EAF 0.017, replication MAC 42) and rs12669476 (EAF 0.018, MAC 41) fall below the MAC threshold and are designated provisional.

rs10835649 was nominally associated in the KOR background subtype ($P = 3.82 \times 10^{-3}$), and rs11868532 was imputed in EUR but not associated (P = 0.587). Cross-cohort assessment was otherwise constrained by lack of imputation for rs113839416 in KOR and rs12669476 in EUR. These findings are consistent with both ancestry-specific and shared genetic contributions to the background subtype.

Comparison to previous subtype GWAS. Our previous EUR subtype GWAS[6] identified six genome-wide significant loci in the discovery sample (four reproductive, one metabolic, one background) in a cohort overlapping ~90% with the present study. Only two loci were evaluable in the prior replication sample because the remaining lead SNPs were neither genotyped nor imputed on the Metabochip[20, 32]. With TOPMed imputation, all previously identified SNPs were imputed in both the discovery and replication samples (Table S1) except the metabolic subtype lead SNP rs55762028 near *KCNH7*/*FIGN*. The four reproductive subtype signals, rs78025940 near *KAZN-AS1*, rs76182733 near *IQCA1*, rs17023134 near *BMPR1B*/*UNC5C*, and rs7735176 near *CDH10*, were nominally significant in the current meta-analysis ($P = 1.84 \times 10^{-4}$ to 0.022) but did not reach genome-wide significance. The background subtype lead SNP rs10835638 near *FSHB* was imputed in both stages in the previous and current analyses and remained genome-wide significant.

Clustering approach comparison. Having established that the subtypes carry distinct genetic architectures, we next asked whether that signal depends on the clustering method or the trait set. We compared hierarchical clustering (HC), which aggregates individuals based on pairwise similarity, and k-means (Km), which partitions individuals by minimizing within-cluster variance. Both were applied to 874 EUR PCOS cases with complete data on nine traits (BMI, T, SHBG, DHEAS, LH, FSH, Ins0, Glu0, and AMH; Table S5). Cluster

stability was assessed using ConsensusClusterPlus[30] across k = 2–4 using eight or nine traits, the ninth being AMH[52], a marker of follicle number implicated in reproductive subtype biology[53].

Across methods, the most stable solution was obtained with k = 3 using eight traits (pairwise consensus proportion: HC 0.975, Km 0.935). Stability was reduced with k = 2 (HC 0.879, Km 0.878) and did not improve with inclusion of AMH (HC 0.953, Km 0.927). No stable structure was observed at k = 4 for either method (range 0.346–0.659). Km produced modestly greater cluster separation than HC (average silhouette width 0.12 versus 0.09).

To determine whether the clusters captured disease-relevant structure, we performed subtype GWAS as orthogonal validation using the full genotyped cohort with the stable k=3 and eight trait methods. HC-defined subtypes yielded genome-wide significant, robust loci for two of three groups (metabolic: *C9orf3*/*FANCC*, rs10761370, $P = 1.21 \times 10^{-8}$; background: *FSHB*, rs10835649, $P = 8.49 \times 10^{-10}$; Table 2). We observed no genome-wide significant associations for Km-derived subtypes despite their slightly higher silhouette values, indicating that improved geometric separation does not translate into biological discrimination. Neither inclusion of AMH nor increasing k improved stability or yielded genetically supported subtypes. We restricted this comparison to the EUR cohort, the only cohort with AMH measurements and sufficient sample size for method comparison; the KOR cohort was too small for this analysis. These data support HC with eight traits as the preferred approach for resolving genetically distinct PCOS subtypes and underscore the necessity of orthogonal validation when evaluating clustering performance.

Trans-ancestry meta-analysis. Trans-ancestry meta-analysis across 991 EUR and 417 KOR PCOS cases identified two genome-wide significant loci, both designated robust in the EUR analysis (Table 2, Figure 4): rs10761370 at *C9orf3*/*FANCC* in the metaolic subtype ($P = 2.35 \times 10^{-9}$) and rs10835649 at *FSHB* in the background subtype ($P = 1.23 \times 10^{-11}$). Neither signal reached genome-wide significance in the KOR subtype GWAS, consistent with limited power in the smaller cohort. However, effect estimates were directionally concordant across ancestries (*C9orf3/FANCC*: OR 1.42 in EUR and 1.05 in KOR; *FSHB*: OR 2.01 in EUR and 1.85 in KOR). No provisional locus reached genome-wide significance in the trans-ancestry meta-analysis.

Functional annotation and mapping. To identify likely functional and causal variants at genome-wide significant loci, we performed SNP-level annotation using SNP2GENE within FUMA v1.6.1. Of the 159 candidate SNPs in strong LD ($r^2 \geq 0.6$) with lead SNPs, most were non-coding: 55% intronic and 38% intergenic (Table S6). Locus architecture varied substantially in extent and in the density of supporting variants. In the EUR reproductive subtype, rs72728759 lies within a 655-kb LD region encompassing the *GBP* gene family, whereas rs114040582, rs74679007, and rs10785467 are intronic to *IGSF11*, *SEMA6A*, and *TMEM117*, respectively. In the KOR reproductive subtype, rs117222858 lies upstream of *KAZN* within a 769-kb region of low LD containing few supporting SNPs, and rs113200232 lies within a 419-kb LD region that includes *SSPN*, *BHLHE41*, *RASSF8*, *MIR4302*, and *ITPR2-AS1*.

In the EUR background subtype, rs113839416 is intronic to *NRXN1-DT* and is not in LD with previously reported 2p16.3 PCOS signals near *LHCGR* and *FSHR* identified in Chinese (rs2268361, rs2349415) and EUR (rs13004711) cohorts, indicating that it tags a distinct locus (Figure S1)[5, 54]. The 11p14.1 lead SNP rs10835649 is intronic to *ARL14EP-DT* within a 412-kb LD region that includes *FSHB*, *ARL14EP*, and *MPPED2*, and is in strong LD with rs10835638 ($r^2 = 0.86$, EUR), the background subtype lead SNP in our prior GWAS

using a less dense imputation panel[6], and with rs11031006 (r^2 = 0.83, EUR), the original *FSHB* signal reported in PCOS GWAS[20, 55]. In the KOR background subtype, rs12669476 lies within a 584-kb region of weak LD containing *ZNF804B* and *TEX47*, and rs11868532 lies within a 1,178-kb region containing *TBX2*, *TBX4*, *C17orf82*, *BRIP1*, *BCAS3*, and *NACA2* (Figure S2).

The breadth of these LD regions and the sparse variant support at several loci, most notably *KAZN* and *ZNF804B*, constrain gene assignment. Except where noted, genes were assigned by physical proximity to the lead SNP, and the causal gene at each provisional locus remains undetermined. None of the lead SNPs exceeded the conventional deleteriousness threshold (CADD > 12.37). Four candidate SNPs in strong LD with rs10835649 (*FSHB*) had elevated CADD scores (19.91, 13.59, 13.30, and 12.69), suggesting putative regulatory function (Table S6). Three SNPs had a RegulomeDB score of 2b, consistent with regulatory DNA and possible transcription factor binding effects: two in LD with rs10835649 (*FSHB*) and one in LD with rs10785467 (*TMEM117*). All three were located in open chromatin (minChromState ≤ 5).

These predictions are supported by direct experimental evidence at the same locus. rs10835649 is in strong LD with rs11031006 (r^2 = 0.83), which lies within a conserved 450-bp element 26 kb upstream of *FSHB* that enhances *FSHB* transcription in reporter assays, shows gonadotrope-restricted open chromatin and enhancer-associated histone marks in mouse pituitary, and whose minor allele increases transcription through enhanced SF1 binding[56]. The same element confers activin and GnRH responsiveness through a SMAD binding site[56].Thus, the strongest in silico functional evidence was observed at *FSHB*, the only robust background subtype locus, where computational predictions converged with direct experimental evidence of regulatory function.

Gene expression. To assess whether candidate genes are differentially expressed in PCOS-relevant tissues, we queried a publicly available adipose stem cell dataset (GEO GSE124226). *PSEN2* was downregulated in adipose stem cells from PCOS cases relative to controls ($\log_2$ fold change = –0.61, P = 0.0012). *PSEN2* is a provisional locus, and this single-dataset, uncorrected comparison is hypothesis-generating rather than confirmatory. No other candidate gene showed differential expression.

Pathway enrichment. We evaluated pathway enrichment among genes assigned to genome-wide significant loci using KEGG 2021 Human and Reactome Pathways 2024 (Enrichr). No pathway survived correction for multiple testing (all adjusted P > 0.05); nominally enriched terms are reported for completeness (Table S7). These included propionyl-CoA catabolism (*PCCA*; P = 0.008), interferon and NOD-like receptor signaling (*FANCC*, *GBP3*, *GBP4*; P = 0.011–0.036), NOTCH signaling (*PSEN2*; P = 0.013–0.050), glycoprotein and peptide hormone biosynthesis (*FSHB*; P = 0.016–0.020), and gluconeogenesis and the pentose phosphate pathway (*FBP1*; P = 0.042 and 0.048). Because most terms were driven by a single gene, and several of those genes derive from provisional loci, these results should not be interpreted as evidence of pathway-level convergence.

Cross-phenotype associations. To assess pleiotropy, we queried lead SNPs against population-specific GWAS summary statistics for 17 EUR and 31 EAS traits, applying Bonferroni thresholds of P < 0.0029 and P < 0.0016, respectively (Tables S2–S4; Figure 5). In the EUR cohort, associations were restricted to the two robust loci. rs10835649 (*FSHB*) was associated with total T ($P = 2.5 \times 10^{-8}$), age at menarche ($P = 1.7 \times 10^{-15}$), bioavailable T ($P = 1.1 \times 10^{-7}$), FSH (P = 0.001), LH (P = 0.002), and SHBG (P = 0.003), a profile consistent with its established role in gonadotropin regulation and reproductive timing[55, 57]. rs10761370 (*C9orf3*/*FANCC*) was associated with systolic blood pressure (P = 0.002), with nominal associations for diastolic blood pressure, total T, and FSH. The blood pressure

association is consistent with the proposed role of *C9orf3* in the renin-angiotensin system and prior association of *C9orf3* variation with hypertension in diabetes[58]. None of the provisional EUR lead SNPs (rs72728759, rs114040582, rs74679007, rs10785467, rs114194538, rs113839416) showed any association surviving correction.

In the KOR cohort, rs11868532 (*TBX2*) exceeded the Bonferroni-corrected EAS threshold ($P < 0.0016$) for creatinine ($P = 6.03 \times 10^{-24}$), diastolic blood pressure ($P = 6.8 \times 10^{-6}$), BMI ($P = 1.0 \times 10^{-4}$), and systolic blood pressure ($P = 6.0 \times 10^{-4}$). Nominal associations ($P < 0.05$) were observed with T2D ($P = 0.002$), albumin, weight, total bilirubin, BUN, breast cancer, total cholesterol, and age at menarche (Table S3; Figure 5). rs113200232 (*SSPN*) and rs12669476 (*ZNF804B*) showed nominal associations with metabolic and reproductive traits only. rs117222858 and rs150521566 were absent from the KOR GWAS dataset and could not be evaluated. The strongest cross-phenotype profile in either cohort belongs to rs11868532 (*TBX2*). Although the breadth of its metabolic and cardiovascular associations is consistent with biological pleiotropy, cross-phenotype association does not constitute replication of a disease signal, which requires demonstration in the same phenotype[59, 60]; the *TBX2*–PCOS association itself remains unreplicated.

DISCUSSION

We replicate the distinct genetic architecture of PCOS subtypes and extend this framework across ancestries. Hierarchical clustering of the same quantitative reproductive and metabolic traits reproduced the reproductive, metabolic, and background subtypes in an independently ascertained KOR cohort[26]. Reproductive and metabolic axes of PCOS heterogeneity have also been identified by independent methods: hierarchical clustering in cohorts diagnosed by NIH or Rotterdam criteria, each yielding an intermediate group resembling the background subtype[61, 62]; k-means clustering in Chinese women[52]; and clustering of PCOS susceptibility loci by their trait associations[63, 64]. TOPMed

imputation[35] enabled evaluation of nearly all lead variants from our prior subtype GWAS[6], and subtype-specific association identified additional loci that were only nominally significant in the overall case-control analyses and revealed two shared signals with concordant effects across EUR and KOR cases: *C9orf3/FANCC* in the metabolic subtype and *FSHB* in the background subtype. Because germline genotype is causally prior to and independent of the traits used for clustering, these subtype-restricted associations provide orthogonal evidence that the phenotypic partitions reflect etiologic heterogeneity rather than statistical structure alone[21, 23]. The genetic results also discriminated between clustering methods: hierarchical clustering yielded subtype-specific associations, whereas k-means yielded none despite marginally greater geometric separation.

The same traits drove each subtype in both ancestries: LH and SHBG in the reproductive subtype, and BMI, Glu0, and Ins0 in the metabolic subtype, consistent with applying the same clustering algorithm and trait set across cohorts. However, ancestry-specific subtype patterns emerged. T was highest in the reproductive subtype in KOR but did not differ significantly among EUR subtypes; DHEAS was highest in the background subtype in EUR but in the reproductive subtype in KOR; and FSH was highest in the background subtype in EUR but similarly high in reproductive and background subtypes in KOR. Subtype membership also differed. In the EUR cohort, ~50% were in the metabolic subtype, ~25% in the reproductive, and ~20% in the background subtype. In KOR, ~50% belonged to the background subtype, ~20% to the reproductive subtype, and ~30% to the metabolic subtype. Because the subtypes were defined by the eight clustering traits, differences across those traits characterize the subgroups rather than validate them. However, we have previously reported in an independent Rotterdam-criteria cohort that the reproductive and metabolic subtypes differed in held-out measures including AMH, antral follicle count, LDL cholesterol, and blood pressure[61].

GWAS prior to subtyping confirmed previously reported findings in both ancestries[20, 26]. In the EUR cohort, the 9q22.32 signal near *C9orf3/FANCC* and the 11p14.1 association near *FSHB* reached genome-wide significance, together with a novel locus at 10q21.2 near *ANK3*; no genome-wide significant associations were detected in KOR. Subtype-specific analysis identified additional genome-wide significant loci in both cohorts, demonstrating the increased power gained by resolving phenotypic heterogeneity before genetic association testing[6, 19, 65]. Because several lead variants had stage-specific minor allele counts below the pre-specified threshold for reliable calibration of classical single-marker tests[66], we consider these provisional signals requiring independent replication rather than established loci. Their larger apparent effect sizes relative to our prior common variant findings (EUR OR range 1.42 to 3.57; KOR 1.93 to 7.23) are consistent with the expected inverse relationship between allele frequency and effect size[67-69] and may also reflect the modest cohort sizes, particularly in the KOR sample.

The trans-ancestry meta-analysis identified two robust, subtype-restricted signals: *C9orf3/FANCC* (9q22.32) in the metabolic subtype and *FSHB* (11p14.1) in the background subtype. Both reached genome-wide significance in the EUR analysis and showed a concordant direction of effect in KOR cases (*C9orf3/FANCC* OR 1.42 EUR, 1.05 KOR; *FSHB* OR 2.01 EUR, 1.85 KOR), and both correspond to the most consistently associated loci in prior PCOS GWAS across European and Han Chinese populations[4, 20, 54]. rs10835649 (*FSHB*) is in strong linkage disequilibrium with the previously reported rs10835638 and rs11031006, and rs10761370 (*C9orf3/FANCC*) with rs10993397 and rs7864171, indicating that our lead variants tag the same underlying risk signals[4, 20]. Sharing of these two signals across populations with different linkage disequilibrium structure, on a common haplotype background, is consistent with ancestral risk alleles predating the divergence of Eurasian populations[70-72], although formal tests of allele age and selection were beyond the scope of this study.

The loci identified in the EUR and KOR subtype analyses differed substantially, largely reflecting allele-frequency differences between populations. Several KOR lead variants were absent or extremely rare in European reference data and could not be imputed in the EUR cohort, and several EUR signals showed the converse. Such divergence is expected, because rarer variants tend to be more ancestry-specific[73] and risk-allele frequencies frequently differ markedly across ancestries[74]. Ancestry-specific analysis was therefore essential for detecting population-relevant signals that a combined analysis would dilute, whereas the trans-ancestry meta-analysis isolated the shared, most strongly associated loci. Together, these complementary analyses support the need for larger, more diverse cohorts to identify and validate ancestry-specific variants.

The two signals shared across ancestries map to loci with established and distinct biology. At 11p14.1, rs10835649 is in strong LD with rs11031006, the variant originally associated with PCOS and LH levels[20]. rs11031006 lies in a conserved enhancer 26 kb upstream of *FSHB*, where its minor allele increases FSHB transcription by enhancing steroidogenic factor 1 binding[75]. *FSHB* is among the most pleiotropic loci in reproductive genetics; its alleles are associated with circulating FSH, age at menarche and menopause, age at first birth, number of births, and dizygotic twinning[55, 76], as well as menstrual cycle length and T[57, 77]. In our cross-phenotype analysis, rs10835649 was associated with T ($P = 2.5 \times 10^{-8}$) and age at menarche ($P = 1.7 \times 10^{-15}$), and with FSH, LH, and SHBG ($P \leq 0.003$). These data place the *FSHB* signal in the hypothalamic–pituitary–ovarian axis, with effects on reproductive timing across the lifespan.

The 9q22.32 signal lies in an extended LD region spanning *C9orf3* (*AOPEP*) and *FANCC*[20], and the effector gene is unresolved. *C9orf3* encodes aminopeptidase O, which

generates angiotensin IV from angiotensin III[78]. A *C9orf3* variant is associated with plasma renin activity and hypertension in T2D[58], and renin–angiotensin activity correlates with insulin resistance in PCOS[79]. *FANCC* encodes a Fanconi anemia core-complex protein required for DNA interstrand crosslink repair. Defects in this pathway impair germ-cell development and cause primary ovarian insufficiency[80], and *FANCC* was prioritized at this locus in the largest PCOS GWAS, which linked PCOS susceptibility to reproductive longevity[5]. In our data, rs10761370 was associated with systolic blood pressure ($P = 0.002$) and nominally with diastolic blood pressure, T, and FSH. Its restriction to the metabolic subtype and its blood pressure association favor a cardiometabolic mechanism through *C9orf3*, although a *FANCC* effect on ovarian function cannot be excluded. Resolving the effector gene will require trans-ancestry fine-mapping. Among the provisional signals, the KOR *TBX2* lead variant is associated with blood pressure in Europeans[81] and with creatinine in Korean biobank data, and *KAZN*, which recurs across the EUR and KOR analyses with independent lead variants, has been linked to age at menarche and blood pressure[82, 83]. *PSEN2* was downregulated in adipose stem cells from women with PCOS. These annotations are hypothesis-generating and require functional validation.

Hierarchical clustering with eight traits produced the most stable solution and was the only method to yield subtype-specific genome-wide associations, whereas k-means produced marginally greater geometric separation but no associated loci (Results; Table 2). This discordance demonstrates the limitation of internal validity metrics: measures of compactness and separation will partition data even when discrete biological subtypes are absent[84]. Genotype therefore provides a more stringent, biologically independent test of clustering validity. We also reproduced the reproductive and metabolic subtypes in an independent European cohort meeting Rotterdam criteria[61], showing that they are not specific to the original NIH-defined cohort. A recent study identified four PCOS subtypes by k-means clustering and assessed them by re-clustering additional cohorts and predicting cluster labels from the clustering traits, but did not test them against a non-phenotypic

reference such as germline genotype[52]. Reproducible partitioning does not establish etiologic distinctness. In the present study, only hierarchical clustering yielded subtype-specific genetic associations, even though k-means achieved marginally greater geometric separation.

Data-driven approaches are increasingly used to resolve heterogeneity in complex diseases, including T2D, obesity, and psychiatric disorders[85]. However, the resulting subtypes often reproduce inconsistently across cohorts and ancestries[86-88] and few have been shown to differ in genetic architecture[23]. Much of this inconsistency reflects how subtypes are validated: they are frequently evaluated on the same phenotypic dimensions used to define them, or by internal clustering metrics that partition continuous variation whether or not biologically distinct groups exist[84]. Genetic heterogeneity, in which subtypes differ in their associated germline variants, has been proposed as the more stringent standard[23], but only one of the two common genetic strategies supplies it. Partitioned polygenic scores group disease-associated variants by their trait associations[13]. Because genotype defines these subgroups, genetic differences between them arise by construction and cannot validate them. The alternative assigns cases to discrete subtypes from phenotype alone and then tests each subtype by GWAS, as in T2D[16] and our previous PCOS study[6]. We apply this design and extend it by showing that both the subtypes and their associated loci replicate across two ancestries.

The principal limitations of our study concern statistical power, ancestral diversity, and functional resolution. The relatively small cohorts, particularly KOR, reduced power to detect ancestry-specific genome-wide associations and replicate low-count variants, for which sensitivity depends strongly on case number[66]. Accordingly, provisional loci with stage-specific minor allele counts below the calibration threshold require replication in

independent, ancestry-matched samples before they can be considered established. We examined only two ancestries, both of Eurasian origin. Because risk-allele frequencies and effect sizes can differ across more divergent populations, extending this subtype framework globally will require larger and more diverse samples[74], consistent with recent multi-ancestry PCOS efforts[89]. Functional interpretation also remains preliminary because our annotations are computational and correlative, most lead variants are non-coding, and defining effector genes and mechanisms will require fine-mapping and experimental validation[40]. Such validation is tractable in PCOS, as demonstrated by our mapping regulatory activity at a PCOS risk locus to *DENND1A*-dependent T production[90]; applying the same strategy to the subtype-specific loci identified here is a priority for future work.

Nevertheless, this study establishes the cross-ancestry validity and genetic utility of PCOS subtyping. Hierarchical clustering identified the reproductive and metabolic subtypes and background group in both EUR and KOR. Subtype-specific association revealed loci that remained only nominally significant in conventional case-control analyses. Trans-ancestry meta-analysis identified shared, subtype-restricted signals at *FSHB* and *C9orf3/FANCC* with concordant effects across cohorts. Together, these results demonstrate that the subtypes are not specific to a single population and that resolving phenotypic heterogeneity before association testing recovers genetic signals diluted in aggregate case-control analyses.

CONCLUSIONS

McKusick conceptualized genetic nosology more than half a century ago as a choice between lumping and splitting[3]. In complex disorders such as PCOS, expert opinion has largely determined that choice, grouping biologically heterogeneous patients under a single diagnosis[1]. The 2026 renaming of PCOS illustrates the persistence of this approach: the International Renaming Group changed the name but retained the diagnostic criteria, without evaluating replicated evidence for genetically distinct subtypes[8, 9].

Our study provides a data-driven framework for disease classification. We defined subgroups from quantitative traits rather than diagnostic convention, replicated them across ancestries, and identified genetic associations attenuated by the aggregate analysis, which offers greater power for homogeneous cases but reduced power for heterogeneous cases due to the noise introduced[19]. Because germline genotype is fixed at conception and precedes the phenotypes used to define these groups[21], subtype-specific associations provide biological validation that statistical separation alone cannot establish[91]. This approach uses phenotype to define candidate disease classes and the genome to determine whether they reflect distinct biology, replacing opinion with evidence as the basis for when to lump and when to split. Reproductive and metabolic subtypes have now been identified by complementary approaches[52, 63, 64] and, as shown here, replicate across ancestries with distinct genetic architectures, indicating that PCOS comprises biologically distinct disorders that must be rethought rather than merely renamed. Such genetically validated classification provides a foundation for precision medicine by identifying patient groups in which interventions can ultimately be matched to disease mechanisms[92].

ABBREVIATIONS

PCOS	Polycystic ovary syndrome
T2D	Type 2 diabetes
HC	Hierarchical clustering
EUR	European
GWAS	Genome-wide association study
KOR	Korean
BMI	Body mass index
LH	Luteinizing hormone
FSH	Follicle stimulating hormone
T	Testosterone
DHEAS	Dehydroepiandrosterone sulfate
SHBG	Sex hormone binding globulin
Glu0	Fasting glucose
Ins0	Fasting insulin
AMH	Anti-mullerian hormone
SNP	Single nucleotide polymorphism
MAF	Minor allele frequency
MAC	Minor allele count
PCA	Principal component analysis
PCs	Principal components
LD	Linkage disequilibrium
CADD	Combined Annotation-Dependent Depletion
EAS	East Asian
EAF	Effect allele frequency
OR	Odds ratio
Km	K-means

SUPPLEMENTARY INFORMATION

Additional file 1: Supplemental Figures S1-S4 and Tables S1-S5 & S7. Supplementary figures showing locuszoom plots for lead snps, pre-subtyping Manhattan plots, k-means analysis. Supplementary tables containing lookup of previously reported subtype PCOS signals, variant associations with reproductive and metabolic traits, cluster comparison, and enriched biological pathways.

Additional file 2: Table S6. Table containing functional annotation of candidate SNPs in loci associated with each subtype of PCOS.

DECLARATIONS

Ethics approval and consent to participate

These studies were approved by the institutional review boards of Icahn School of Medicine at Mount Sinai and Ewha Womans University Medical Center. Written informed consent was obtained from all study participants for both studies.

Consent for publication

Not applicable

Availability of data and materials

Requests for access to data presented here can be made to the corresponding author. Discovery genotype data have previously been uploaded to the database of Genotypes and Phenotypes (dbGaP) under accession code phs000368.v1.p1. Consent forms signed by participants limited the use of data to studies of the genetics of PCOS. Aggregated data and quantitative trait data are available from the corresponding author upon reasonable request and with approved data sharing agreements and relevant local IRB approval.

R code used to perform clustering has been uploaded to the public GitHub repository github.com/mdapas/PCOS_phenotype_clustering published by Dapas et al.[6] previously.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by National Institutes of Health (NIH) Grants P50 HD044405 (AD), R01 HD057223 (AD), R01 HD085227 (AD), and R01 HD100812 (AD). Korean study data was supported in part by the grant from the Korea Centers for Disease Control and Prevention (2009-E00591-00) and the Ewha Global Top5 Grant 2013 of Ewha Womans University.

Authors' contributions

All analysis was performed by K.B. and R.S. including QC of the European and Korean datasets. A.D. conceived the original idea and supervised the project. A.D. and K.B. drafted the manuscript. K.B., R.S., C.L., A.D. contributed to the interpretation of the results. H.L. was responsible for the study in Korea. All authors participated in the review and revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

None.

Tables

Table 1. Descriptive Characteristics of Cohorts

	EUR									KOR								
	n	**Pre-subtyping** Median (25-75)	**Reproductive** Median (25-75)	**Metabolic** Median (25-75)	**Background** Median (25-75)	P-value Rep-Met	P-value Rep-Bac	P-value Met-Bac	Overall P-value	*n*	**Pre-subtyping** Median (25-75)	**Reproductive** Median (25-75)	**Metabolic** Median (25-75)	**Background** Median (25-75)	P-value Rep-Met	P-value Rep-Bac	P-value Met-Bac	Overall P-value
Age (y)	1449	28 (24-32)	28 (25-31)	28 (25-32)	28 (24-31)	0.99	0.99	0.99	0.58	418	23 (19-27)	24 (20-27)	23 (19-27)	23 (20-27)	0.31	0.59	0.99	0.25
BMI (kg/m²)	1449	35.1 (28.4-41.4)	25.4 (22.2-29.2)	40.3 (35.5-45.3)	32.3 (27.8-36.5)	**<0.001**	**<0.001**	**<0.001**	**<0.001**	418	23.0 (20.5-26.8)	20.2 (19.0-21.5)	28.5 (26.1-31.3)	21.9 (20.4-23.7)	**<0.001**	**<0.001**	**<0.001**	**<0.001**
T (ng/dL)						0.99	0.21	0.07	0.07						**<0.001**	**<0.001**	0.74	**<0.001**
Method 1	837	73 (60-91)	74 (61-90)	70 (57-91)	76 (66-92)					418	75 (64-87)	90 (82-101)	72 (57-87)	72 (63-78)				
Method 2	587	51 (37-70)	46 (33-73)	54 (38-71)	53 (42-68)													
Method 3	25	68 (52-93)	82 (51-109)	62 (52-87)	68 (55-77)													
SHBG (nmol/L)						**<0.001**	**<0.001**	**<0.001**	**<0.001**						**<0.001**	**<0.001**	**<0.001**	**<0.001**
Method 1	740	53 (34-80)	87 (67-126)	40 (28-55)	67 (43-92)					418	41 (28-62)	71 (58-82)	27 (18-37)	41 (32-57)				
Method 2	60	27 (18-43)	40 (34-54)	20 (14-23)	43 (27-61)													
Method 3	34	32 (21-45)	59 (43-106)	22 (17-32)	48 (44-58)													
Method 4	108	28 (21-40)	53 (36-65)	23 (18-28)	33 (24-39)													
Method 5	25	14 (8-19)	20 (14-24)	11 (7-16)	18 (14-19)													
Method 6	3	12 (11-36)	60 (60-60)	11 (10-11)														
Method 7	479	29 (20-44)	51 (37-70)	22 (17-30)	32 (24-46)													
DHEAS (ng/mL)						0.99	**0.003**	**0.002**	**<0.001**						**<0.001**	**<0.001**	0.99	**<0.001**
Method 1	837	2103 (1509-2856)	2080 (1512-2958)	2097 (1424-2788)	2183 (1694-2958)					418	2370 (1910-3060)	2810 (2288-3318)	2260 (1760-3060)	2320 (1760-2990)				
Method 2	108	1983 (1319-2809)	2143 (1523-2796)	1745 (1106-2751)	2018 (1551-2830)													
Method 3	25	1610 (1110-2310)	1680 (1403-1873)	1310 (645-2348)	1890 (1560-3040)													
Method 4	479	1500 (947-2215)	1335 (896-1825)	1570 (893-2275)	1675 (1298-2535)													
Glu0 (mg/dL)						**<0.001**	**<0.001**	0.99	**<0.001**						**<0.001**	**0.003**	**<0.001**	**<0.001**
Method 1	294	91 (86-96)	85 (82-89)	93 (88-98)	94 (90-99)					418	86 (80-90)	81 (77-87)	89 (85-94)	85 (80-90)				

Method 2	481	88 (84-95)	82 (79-86)	91 (85-99)	89 (86-95)													
Method 3	15	93 (87-95)	86 (82-88)	95 (94-98)	94 (93-94)													
Method 4	479	85 (78-92)	78 (73-86)	88 (81-95)	87 (81-92)													
Method 5	25	96 (89-98)	86 (80-89)	98 (97-100)	95 (90-96)													
Method 6	155	86 (80-92)	83 (78-84)	87 (80-93)	88 (82-96)													
Ins0 (µU/mL)						**<0.001**	**<0.001**	**<0.001**	**<0.001**						**<0.001**	**<0.001**	**<0.001**	**<0.001**
Method 1	12	19 (14-51)	9 (8-9)	49 (21-61)	15 (14-17)					418	9 (7-13)	7 (5-8)	14 (10-19)	8 (6-10)				
Method 2	815	22 (15-34)	13 (10-15)	31 (24-42)	19 (14-25)													
Method 3	25	14 (9-39)	9 (6-9)	36 (16-52)	11 (9-18)													
Method 4	479	13 (5-23)	3 (2-5)	21 (16-33)	9 (6-13)													
Method 5	93	21 (14-30)	13 (10-15)	30 (22-41)	18 (15-22)													
Method 6	25	21 (16-27)	13 (12-13)	25 (20-31)	18 (16-19)													
LH (mIU/mL)						**<0.001**	**0.01**	**<0.001**	**<0.001**						**<0.001**	**<0.001**	**<0.001**	**<0.001**
Method 1	695	12 (8-18)	17 (9-23)	8 (6-12)	17 (14-22)					418	12 (7-16)	18 (14-24)	8 (4-11)	12 (8-15)				
Method 2	95	13 (9-17)	18 (11-24)	11 (7-14)	15 (13-20)													
Method 3	40	9 (6-14)	10 (8-15)	7 (5-9)	16 (13-24)													
Method 4	115	10 (6-14)	14 (10-16)	7 (5-9)	12 (9-15)													
Method 5	25	5 (4-11)	11 (8-15)	4 (3-5)	10 (9-12)													
Method 6	479	9 (6-13)	12 (8-18)	8 (6-10)	14 (10-18)													
FSH (mIU/mL)						**0.003**	**<0.001**	**<0.001**	**<0.001**						**<0.001**	0.99	**<0.001**	**<0.001**
Method 1	695	9 (8-11)	9 (7-11)	9 (7-10)	10 (9-12)					418	5 (4-6)	6 (4-6)	5 (4-5)	6 (5-6)				
Method 2	94	3 (3-4)	4 (3-4)	3 (3-4)	4 (3-5)													
Method 3	41	2.5 (2-3)	3 (2-3)	2 (2-3)	3 (2-5)													
Method 4	115	6 (5-7)	6 (5-7)	5 (4-6)	6 (5-8)													
Method 5	25	5 (3-6)	4 (4-6)	4 (3-5)	5 (4-6)													
Method 6	479	6 (5-7)	6 (5-7)	6 (5-7)	7 (6-8)													

For pairwise comparison, Z-scores were used to adjust for different assay methods; bold, significant after Bonferroni correction; BMI, body mass index; T, testosterone; SHBG, sex-hormone binding globulin; DHEAS, dehydroepiandrosterone sulfate; Glu0, glucose; Ins0, insulin; LH, luteinizing hormone; FSH, follicle-stimulating hormone; Rep, Reproductive; Met, Metabolic; Bac, Background; y, year.

Table 2. Statistically significant association signals in the meta-analyses and individual cohorts

rsID		rs72728759	rs114040582	rs74679007	rs10785467	rs117222858	rs113200232	rs114194538	rs10761370	rs150521566	rs113839416*	rs10835649	rs12669476	rs11868532
Chromosome		1	3	5	12	1	12	1	9	13	2	11	7	17
Position		89301921	119092733	117277611	43915630	13995494	26105641	226878653	95044644	100353793	51753545	30307628	88581112	61418722
Other Allele		C	T	C	C	T	T	C	A	A	C	A	T	T
Effect Allele		T	C	T	T	G	C	T	T	G	T	G	G	C
Candidate Gene		*GBP5*	*IGSF11*	*SEMA6A*	*TMEM117*	*KAZN*	*SSPN*	*PSEN2*	*C9orf3/FANCC*	*PCCA*	*NRXN1*	*FSHB*	*ZNF804B*	*TBX2*
Novel PCOS hit?		Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	Yes	No
EUR Discovery Imp r^2		0.91	0.97	0.95	0.99	-	0.98	0.90	0.99	-	0.97	1.00	-	0.96
EUR Replication Imp r^2		0.50	0.77	0.51	0.68	-	0.84	0.72	0.98	-	0.85	0.93	-	0.84
KOR Imp r^2		-	-	-	0.89	0.72	0.89	-	0.94	0.77	-	0.93	0.93	0.94
Pre-Subtyping														
EUR Meta-analysis	**P_{meta}**	4.94E-05	1.43E-04	2.86E-03	3.46E-05	-	3.61E-02	7.96E-05	**4.93E-08**	-	3.58E-04	**4.69E-10**	-	1.62E-01
KOR	**P**	-	-	-	3.44E-01	3.46E-02	2.31E-03	-	4.00E-02	1.69E-01	-	3.79E-04	6.13E-07	1.87E-07
Reproductive														
EUR Discovery 144 Cases 2951 Controls	**EAF**	0.01	0.03	0.02	0.10	-	0.06	0.01	0.46	-	0.02	0.16	-	0.38
	MAC	83	172	139	615	-	341	78	2836	-	96	972	-	2352
	Beta	2.007	1.967	2.057	0.884	-	0.028	0.054	0.255	-	0.271	0.324	-	0.002
	SE	0.50	0.37	0.42	0.20	-	0.27	0.53	0.12	-	0.48	0.16	-	0.13
	P	6.79E-05	7.51E-08	8.14E-07	9.75E-06	-	9.18E-01	9.20E-01	3.84E-02	-	5.71E-01	4.57E-02	-	9.86E-01
	P_{Firth}	1.12E-04	1.06E-07	2.16E-06	1.62E-05	-	9.74E-01	7.91E-01	3.80E-02	-	4.10E-01	2.06E-02	-	8.87E-01
EUR Replication 95 Cases 926 Controls	**EAF**	0.01	0.02	0.01	0.07	-	0.07	0.01	0.46	-	0.02	0.17	-	0.39
	MAC	23	48	30	135	-	138	27	948	-	41	356	-	796
	Beta	2.962	1.050	1.678	1.115	-	-0.511	0.994	0.241	-	0.858	0.573	-	-0.015
	SE	0.72	0.50	0.64	0.32	-	0.30	0.66	0.16	-	0.54	0.21	-	0.16
	P	3.74E-05	3.58E-02	8.51E-03	4.88E-04	-	9.10E-02	1.30E-01	1.29E-01	-	1.13E-01	5.78E-03	-	9.28E-01
	P_{Firth}	1.42E-04	2.17E-02	1.08E-02	1.03E-03	-	1.21E-01	6.70E-02	1.30E-01	-	5.97E-02	6.65E-03	-	9.46E-01
EUR Meta-analysis	**P_{meta}**	**3.64E-08**	**1.13E-08**	**2.31E-08**	**2.52E-08**	-	4.52E-01	4.01E-01	1.07E-02	-	2.00E-01	1.89E-03	-	9.76E-01
KOR 83 Cases 926 Controls	**EAF**	-	-	-	0.10	0.01	0.03	-	0.37	0.02	-	0.05	0.01	0.38
	MAC	-	-	-	192	29	57	-	1266	31	-	98	22	758
	Beta	-	-	-	-0.148	4.313	2.700	-	0.122	0.174	-	1.193	0.038	0.048
	SE	-	-	-	0.27	0.70	0.49	-	0.17	0.66	-	0.37	0.77	0.17
	P	-	-	-	5.89E-01	**9.07E-10**	**3.44E-08**	-	4.65E-01	7.92E-01	-	1.10E-03	9.61E-01	7.76E-01
	P_{Firth}	-	-	-	6.64E-01	5.16E-07	6.55E-07	-	4.07E-01	6.07E-01	-	1.30E-03	6.73E-01	7.53E-01
Trans-ancestry Meta-analysis	**P_{meta}**	-	-	-	2.00E-06	-	7.61E-02	-	9.03E-03	-	-	2.30E-05	-	9.21E-01
Metabolic														
EUR Discovery 327 Cases 2951 Controls	**EAF**	0.01	0.03	0.02	0.10	-	0.05	0.02	0.46	-	0.02	0.16	-	0.38
	MAC	81	166	136	640	-	356	99	3041	-	103	1031	-	2464
	Beta	0.249	0.104	0.395	0.258	-	-0.259	1.509	0.498	-	-0.098	0.358	-	0.231
	SE	0.42	0.32	0.36	0.16	-	0.22	0.36	0.10	-	0.37	0.13	-	0.10
	P	5.54E-01	7.48E-01	2.73E-01	1.10E-01	-	2.29E-01	2.40E-05	8.35E-07	-	7.94E-01	7.63E-03	-	2.59E-02
	P_{Firth}	4.78E-01	8.57E-01	4.30E-01	1.45E-01	-	3.80E-01	4.67E-06	7.66E-05	-	3.80E-01	1.42E-01	-	3.29E-02
	EAF	0.01	0.02	0.01	0.07	-	0.06	0.02	0.47	-	0.02	0.17	-	0.38
	MAC	23	52	25	151	-	146	35	1062	-	42	387	-	866

EUR Replication Metabolic 199 Cases 926 Controls	**Beta**	1.076	0.581	-0.904	0.343	-	-0.476	1.994	0.391	-	0.807	0.190	-	0.216
	SE	0.62	0.44	0.70	0.24	-	0.27	0.55	0.14	-	0.54	0.18	-	0.14
	P	8.52E-02	1.88E-01	1.94E-01	1.61E-01	-	8.30E-02	2.63E-04	4.21E-03	-	1.32E-01	2.81E-01	-	1.12E-01
	P_{Firth}	1.72E-02	1.17E-01	2.94E-01	1.20E-02	-	2.96E-02	2.67E-03	9.44E-03	-	6.48E-01	1.40E-01	-	2.18E-01
EUR Meta-analysis	**P_{meta}**	1.67E-01	6.45E-03	7.73E-01	3.67E-02	-	5.57E-02	**4.03E-08**	**1.21E-08**	-	5.92E-01	4.41E-03	-	6.45E-03
KOR 129 Cases 926 Controls	**EAF**	-	-	-	0.10	0.01	0.02	-	0.37	0.02	-	0.04	0.01	0.38
	MAC	-	-	-	201	24	51	-	1321	39	-	93	30	807
	Beta	-	-	-	0.237	0.246	0.115	-	0.398	4.338	-	0.244	2.586	0.434
	SE	-	-	-	0.34	0.77	0.60	-	0.20	0.79	-	0.47	0.77	0.20
	P	-	-	-	4.84E-01	7.48E-01	8.49E-01	-	5.21E-02	**4.01E-08**	-	6.01E-01	7.82E-04	2.80E-02
	P_{Firth}	-	-	-	8.53E-01	3.91E-02	6.76E-02	-	7.88E-01	1.86E-03	-	7.63E-01	5.01E-04	5.30E-02
Trans-ancestry Meta-analysis Metabolic	**P_{meta}**	-	-	-	2.90E-02	-	1.02E-01	-	**2.35E-09**	-	-	5.32E-03	-	6.44E-04
Background														
EUR Discovery 149 Cases 2951 Controls	**EAF**	0.01	0.03	0.02	0.10	-	0.06	0.01	0.46	-	0.02	0.16	-	0.38
	MAC	76	155	131	599	-	341	76	2828	-	105	992	-	2351
	Beta	0.308	-0.293	0.734	0.237	-	-0.015	-0.497	0.151	-	2.043	0.859	-	0.069
	SE	0.54	0.39	0.43	0.20	-	0.27	0.56	0.13	-	0.46	0.16	-	0.13
	P	5.66E-01	4.57E-01	8.53E-02	2.44E-01	-	9.55E-01	3.73E-01	2.28E-01	-	8.60E-06	1.45E-07	-	5.87E-01
	P_{Firth}	3.27E-01	6.39E-01	7.46E-02	2.03E-01	-	9.29E-01	4.78E-01	5.24E-01	-	1.10E-05	5.29E-07	-	6.99E-01
EUR Replication 77 Cases 926 Controls	**EAF**	0.01	0.02	0.01	0.06	-	0.07	0.01	0.46	-	0.02	0.17	-	0.39
	MAC	17	47	23	122	-	137	25	931	-	42	350	-	792
	Beta	0.199	1.256	-1.138	0.000	-	-0.447	0.913	0.318	-	2.083	0.722	-	-0.272
	SE	0.78	0.55	0.82	0.35	-	0.34	0.82	0.18	-	0.63	0.23	-	0.18
	P	7.98E-01	2.31E-02	1.65E-01	1.00E+00	-	1.90E-01	2.67E-01	7.12E-02	-	1.02E-03	1.55E-03	-	1.24E-01
	P_{Firth}	3.49E-01	1.71E-02	3.46E-01	7.23E-01	-	2.03E-01	2.86E-01	8.43E-02	-	3.21E-03	1.11E-03	-	1.14E-01
EUR Meta-analysis	**P_{meta}**	5.32E-01	6.34E-01	4.19E-01	3.11E-01	-	4.86E-01	8.22E-01	5.24E-02	-	**3.99E-08**	**8.49E-10**	-	7.73E-01
KOR 205 Cases 926 Controls	**EAF**	-	-	-	0.10	0.01	0.02	-	0.37	0.01	-	0.05	0.02	0.40
	MAC	-	-	-	229	20	53	-	1429	33	-	113	41	914
	Beta	-	-	-	0.327	-0.657	0.214	-	0.201	-0.185	-	0.735	2.450	0.734
	SE	-	-	-	0.18	0.58	0.36	-	0.11	0.47	-	0.25	0.42	0.11
	P	-	-	-	7.27E-02	2.54E-01	5.57E-01	-	7.96E-02	6.92E-01	-	3.82E-03	**6.59E-09**	**1.55E-10**
	P_{Firth}	-	-	-	9.41E-02	4.73E-01	3.77E-01	-	1.29E-01	7.51E-01	-	2.97E-03	4.56E-07	**1.57E-09**
Trans-ancestry Meta-analysis	**P_{meta}**	-	-	-	5.42E-02	-	7.31E-01	-	1.13E-02	-	-	**1.23E-11**	-	6.52E-03

Association results for lead SNPs reaching genome-wide significance ($P \leq 5 \times 10^{-8}$) in the overall or subtype-specific analyses of the EUR and KOR cohorts and in the trans-ancestry meta-analysis. For each variant, the effect allele frequency (EAF) and minor allele count (MAC) are reported separately for each analysis stage, because the calibration of single-marker association tests depends on MAC rather than allele frequency. Variants with stage-specific MAC below the pre-specified threshold of 80 are designated provisional and are interpreted as signals requiring replication in independent samples, irrespective of the observed P value; where a variant falls below the threshold in one stage only, the designation follows the lower count. Variants are designated robust when the stage-specific MAC is ≥ 80 in every stage and the association is nominally significant with a concordant direction of effect in both the discovery and replication samples. The KOR cohort was analyzed in a single stage and therefore cannot meet the two-stage criterion; KOR variants are classified by MAC alone. *Located on the same cytoband (2p16.3) as previously reported *FSHR* (rs2268361) in Shi et al.[54] but not in linkage disequilibrium; represents a distinct signal (Figure S1). Positions are on Build 38. Abbreviations: EAF, effect allele frequency; Beta, effect size from association regression; SE, standard error; MAC, minor allele count; P, stage-specific significance as assessed by logistic regression; P_{Firth}, significance after Firth's penalized logistic regression; Imp r^2, imputation quality; P_{meta}, significance as assessed by meta-analysis.

FIGURE LEGENDS

Figure 1. Comparison of hierarchical clustering in EUR and KOR PCOS cases. EUR (top) and KOR (bottom) principal component analysis (PCA) of adjusted quantitative trait data, with PCOS cases plotted on the first two principal components (PCs). Each case is colored according to its identified subtype, with 95% concentration ellipses indicating the subtype clusters. The black arrows represent the relative magnitude and direction of trait correlations with the PCs. Box plots illustrate the median and interquartile ranges (IQR) for each normalized variable, with Z scores displayed on the Y-axis. Subtypes are represented by color: reproductive (blue), metabolic (red) and background (gray).

Figure 2. EUR subtyping genome-wide association meta-analysis results. Manhattan plots for reproductive (a), metabolic (b), and background (c) PCOS subtypes from EUR discovery and replication cohorts. The red horizontal line indicates genome-wide significance ($P < 5 \times 10-8$). Genome-wide significant loci are colored in green and labeled according to nearest gene(s).

Figure 3. KOR subtyping genome-wide association results. Manhattan plots for reproductive (a), metabolic (b), and background (c) PCOS subtypes from KOR cohort. The red horizontal line indicates genome-wide significance ($P < 5 \times 10-8$). Genome-wide significant loci are colored in green and labeled according to nearest gene(s).

Figure 4. Trans-ancestry subtyping genome-wide association meta-analysis results. Manhattan plots for reproductive (a), metabolic (b), and background (c) PCOS subtypes from EUR and KOR cohorts. The red horizontal line indicates genome-wide significance ($P < 5 \times 10-8$). Genome-wide significant loci are colored in green and labeled according to nearest gene(s).

Figure 5. Heatmap illustrating SNP-trait associations in European (EUR) and Korean (KOR) cohorts. The associations between identified SNPs and relevant traits are depicted with color coding representing the direction and strength (Z-scores): positive (red) or negative (blue). ALT, alanine transaminase; BMI, body mass index; BP, blood pressure; BUN, blood urea nitrogen; DHEAS, dehydroepiandrosterone sulfate; FSH, follicle-stimulating hormone; HDLC, high-density lipoprotein cholesterol; LH, luteinizing hormone; Menarche, Age at menarche; SHBG, sex hormone-binding globulin.

1. Chang S, Dunaif A: **Diagnosis of polycystic ovary syndrome: Which criteria to use and when?** *Endocrinol Metab Clin North Am* 2021, **50:**11-23.

2. Dapas M, Dunaif A: **Deconstructing a syndrome: Genomic insights into PCOS causal mechanisms and classification.** *Endocr Rev* 2022, **43:**927-965.

3. McKusick VA: **On lumpers and splitters, or the nosology of genetic disease.** *Perspect Biol Med* 1969, **12:**298-312.

4. Day F, Karaderi T, Jones MR, Meun C, He C, Drong A, Kraft P, Lin N, Huang H, Broer L, et al: **Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria.** *PLoS Genet* 2018, **14:**e1007813.

5. Moolhuijsen LME, Zhu J, Mullin BH, Pujol-Gualdo N, Actkins KV, Mack JA, Rao H, Trivedi B, Kentistou KA, Zhao Y, et al: **Genomic analyses implicate hormonal and metabolic dysregulation in polycystic ovary syndrome.** *Nat Genet* 2026, **58:**1040-1050.

6. Dapas M, Lin FTJ, Nadkarni GN, Sisk R, Legro RS, Urbanek M, Hayes MG, Dunaif A: **Distinct subtypes of polycystic ovary syndrome with novel genetic associations: An unsupervised, phenotypic clustering analysis.** *PLoS Med* 2020, **17:**e1003132.

7. Gibson-Helm M, Teede H, Dunaif A, Dokras A: **Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome.** *J Clin Endocrinol Metab* 2017, **102:**604-612.

8. Teede HJ, Khomami MB, Morman R, Laven JSE, Joham AE, Costello MF, Patil M, Rees DA, Berry L, Cree MG, et al: **Polyendocrine metabolic ovarian syndrome, the new name for polycystic ovary syndrome: a multistep global consensus process.** *Lancet* 2026, **407:**2329-2339.

9. Dunaif A, Ehrmann DA: **Renaming PCOS to polyendocrine metabolic ovarian syndrome: still an impediment to progress.** *J Clin Endocrinol Metab* 2026, **111:**e2439-e2440.

10. Ringman JM, Goate A, Masters CL, Cairns NJ, Danek A, Graff-Radford N, Ghetti B, Morris JC, Dominantly Inherited Alzheimer N: **Genetic heterogeneity in Alzheimer disease and implications for treatment strategies.** *Curr Neurol Neurosci Rep* 2014, **14:**499.

11. Flint J, Kendler KS: **The genetics of major depression.** *Neuron* 2014, **81:**484-503.

12. von Coelln R, Shulman LM: **Clinical subtypes and genetic heterogeneity: of lumping and splitting in Parkinson disease.** *Curr Opin Neurol* 2016, **29:**727-734.

13. Udler MS, Kim J, von Grotthuss M, Bonas-Guarch S, Cole JB, Chiou J, Anderson CD, METASTROKE and the ISGC, Boehnke M, Laakso M, 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:**e1002654.

14. Ahlqvist E, Storm P, Karajamaki A, Martinell M, Dorkhan M, Carlsson A, Vikman P, Prasad RB, Aly DM, Almgren P, 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:**361-369.

15. Misra S, Wagner R, Ozkan B, Schon M, Sevilla-Gonzalez M, Prystupa K, Wang CC, Kreienkamp RJ, Cromer SJ, Rooney MR, et al: **Precision subclassification of type 2 diabetes: a systematic review.** *Commun Med (Lond)* 2023, **3:**138.

16. Mansour Aly D, Dwivedi OP, Prasad RB, Karajamaki A, Hjort R, Thangam M, Akerlund M, Mahajan A, Udler MS, Florez JC, et al: **Genome-wide association analyses highlight etiological differences underlying newly defined subtypes of diabetes.** *Nat Genet* 2021, **53:**1534-1542.

17. Nguyen TD, Harder A, Xiong Y, Kowalec K, Hagg S, Cai N, Kuja-Halkola R, Dalman C, Sullivan PF, Lu Y: **Genetic heterogeneity and subtypes of major depression.** *Mol Psychiatry* 2022, **27:**1667-1675.

18. Wesolowska-Andersen A, Brorsson CA, Bizzotto R, Mari A, Tura A, Koivula R, Mahajan A, Vinuela A, Tajes JF, Sharma S, et al: **Four groups of type 2 diabetes contribute to the etiological and clinical heterogeneity in newly diagnosed individuals: An IMI DIRECT study.** *Cell Rep Med* 2022, **3:**100477.

19. Traylor M, Markus H, Lewis CM: **Homogeneous case subgroups increase power in genetic association studies.** *Eur J Hum Genet* 2015, **23:**863-869.

20. Hayes MG, Urbanek M, Ehrmann DA, Armstrong LL, Lee JY, Sisk R, Karaderi T, Barber TM, McCarthy MI, Franks S, et al: **Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations.** *Nat Commun* 2015, **6:**7502.

21. Sekula P, Del Greco MF, Pattaro C, Kottgen A: **Mendelian randomization as an approach to assess causality using observational data.** *J Am Soc Nephrol* 2016, **27:**3253-3265.

22. Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, VanderWeele TJ, Higgins JPT, Timpson NJ, Dimou N, et al: **Strengthening the reporting of observational studies in epidemiology using mendelian randomization: The STROBE-MR statement.** *JAMA* 2021, **326:**1614-1621.

23. Dahl A, Zaitlen N: **Genetic influences on disease subtypes.** *Annu Rev Genomics Hum Genet* 2020, **21:**413-435.

24. Murtagh F, Legendre P: **Ward's Hierarchical Agglomerative Clustering Method: Which algorithms implement Ward's Criterion?** *J Classif* 2014, **31:**274-295.

25. Rintala TJ, Fortino V: **COPS: A novel platform for multi-omic disease subtype discovery via robust multi-objective evaluation of clustering algorithms.** *PLoS Comput Biol* 2024, **20:**e1012275.

26. Lee H, Oh JY, Sung YA, Chung H, Kim HL, Kim GS, Cho YS, Kim JT: **Genome-wide association study identified new susceptibility loci for polycystic ovary syndrome.** *Hum Reprod* 2015, **30:**723-731.

27. 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:**14956-14960.

28. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, 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.

29. Strauss T, von Maltitz MJ: **Generalising Ward's Method for use with Manhattan distances.** *PLoS One* 2017, **12:**e0168288.

30. Wilkerson MD, Hayes DN: **ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.** *Bioinformatics* 2010, **26:**1572-1573.

31. Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.** *J Comput Appl Math* 1987, **20:**53-65.

32. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J, et al: **The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits.** *PLoS Genet* 2012, **8:**e1002793.

33. 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:**904-909.

34. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al: **Next-generation genotype imputation service and methods.** *Nat Genet* 2016, **48:**1284-1287.

35. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al: **Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.** *Nature* 2021, **590:**290-299.

36. Kowalski MH, Qian H, Hou Z, Rosen JD, Tapia AL, Shan Y, Jain D, Argos M, Arnett DK, Avery C, et al: **Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations.** *PLoS Genet* 2019, **15:**e1008500.

37. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: **MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.** *Genet Epidemiol* 2010, **34:**816-834.

38. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**7.

39. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM: **Robust relationship inference in genome-wide association studies.** *Bioinformatics* 2010, **26:**2867-2873.

40. Watanabe K, Taskesen E, van Bochoven A, Posthuma D: **Functional mapping and annotation of genetic associations with FUMA.** *Nat Commun* 2017, **8:**1826.

41. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC Bioinformatics* 2013, **14:**128.

42. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.** *Nucleic Acids Res* 2016, **44:**W90-97.

43. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207-210.

44. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Gunes O, Hall P, Hayhurst J, et al: **The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource.** *Nucleic Acids Res* 2023, **51:**D977-D985.

45. Nam K, Kim J, Lee S: **Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits.** *Cell Genom* 2022, **2:**100189.

46. 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:**906-913.

47. Ani A, van der Most PJ, Snieder H, Vaez A, Nolte IM: **GWASinspector: comprehensive quality control of genome-wide association study results.** *Bioinformatics* 2021, **37:**129-130.

48. Willer CJ, Li Y, Abecasis GR: **METAL: fast and efficient meta-analysis of genomewide association scans.** *Bioinformatics* 2010, **26:**2190-2191.

49. Machiela MJ, Chanock SJ: **LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants.** *Bioinformatics* 2015, **31:**3555-3557.

50. Ma C, Blackwell T, Boehnke M, Scott LJ, GoT2D investigators: **Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants.** *Genet Epidemiol* 2013, **37:**539-550.

51. Bigdeli TB, Neale BM, Neale MC: **Statistical properties of single-marker tests for rare variants.** *Twin Res Hum Genet* 2014, **17:**143-150.

52. Gao X, Zhao S, Du Y, Yang Z, Tian Y, Zhao J, Yuan X, Santos BR, Wei D, Cui L, et al: **Data-driven subtypes of polycystic ovary syndrome and their association with clinical outcomes.** *Nat Med* 2025, **31:**4214-4224.

53. Moolhuijsen LME, Visser JA: **AMH in PCOS: Controlling the ovary, placenta, or brain?** *Curr Opin Endocr Metab Res* 2020, **12:**91-97.

54. Shi Y, Zhao H, Shi Y, Cao Y, Yang D, Li Z, Zhang B, Liang X, Li T, Chen J, et al: **Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome.** *Nat Genet* 2012, **44:**1020-1025.

55. Mbarek H, Steinberg S, Nyholt DR, Gordon SD, Miller MB, McRae AF, Hottenga JJ, Day FR, Willemsen G, de Geus EJ, et al: **Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility.** *Am J Hum Genet* 2016, **98:**898-908.

56. Bohaczuk SC, Cassin J, Slaiwa TI, Thackray VG, Mellon PL: **Distal enhancer potentiates activin- and GnRH-induced transcription of FSHB.** *Endocrinology* 2021, **162**.

57. Ruth KS, Day FR, Tyrrell J, Thompson DJ, Wood AR, Mahajan A, Beaumont RN, Wittemans L, Martin S, Busch AS, et al: **Using human genetics to understand the disease impacts of testosterone in men and women.** *Nat Med* 2020, **26:**252-258.

58. Ichikawa M, Konoshita T, Makino Y, Suzuki J, Ishizuka T, Nakamura H: **An association study of C9orf3, a novel component of the renin-angiotensin system, and hypertension in diabetes.** *Sci Rep* 2020, **10:**16111.

59. Pearson TA, Manolio TA: **How to interpret a genome-wide association study.** *JAMA* 2008, **299:**1335-1344.

60. Salinas YD, Wang Z, DeWan AT: **Statistical analysis of multiple phenotypes in genetic epidemiologic studies: From cross-phenotype associations to pleiotropy.** *Am J Epidemiol* 2018, **187:**855-863.

61. van der Ham K, Moolhuijsen LME, Brewer K, Sisk R, Dunaif A, Laven JSE, Louwers YV, Visser JA: **Clustering identifies subtypes with different phenotypic characteristics in women with polycystic ovary syndrome.** *J Clin Endocrinol Metab* 2024, **109:**3096-3107.

62. Burns KA, Stuckey AW, Wilson SG, Watts GF, Stuckey BGA: **Characterization of distinct polycystic ovary syndrome subtypes by cluster and principal component analyses.** *Front Endocrinol (Lausanne)* 2025, **16:**1572427.

63. Zhang Y, Movva VC, Williams MS, Lee MTM: **Polycystic ovary syndrome susceptibility loci inform disease etiological heterogeneity.** *J Clin Med* 2021, **10:**2688.

64. Stamou MI, Smith KT, Kim H, Balasubramanian R, Gray KJ, Udler MS: **Polycystic ovary syndrome physiologic pathways implicated through clustering of genetic loci.** *J Clin Endocrinol Metab* 2024, **109:**968-977.

65. Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M: **The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases.** *PLoS One* 2013, **8:**e76295.

66. Zhang X, Basile AO, Pendergrass SA, Ritchie MD: **Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico.** *BMC Bioinformatics* 2019, **20:**46.

67. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461:**747-753.

68. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Jr., Chatterjee N: **Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants.** *Proc Natl Acad Sci U S A* 2011, **108:**18026-18031.

69. Schoech AP, Jordan DM, Loh PR, Gazal S, O'Connor LJ, Balick DJ, Palamara PF, Finucane HK, Sunyaev SR, Price AL: **Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection.** *Nat Commun* 2019, **10:**790.

70. McEvoy BP, Powell JE, Goddard ME, Visscher PM: **Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs.** *Genome Res* 2011, **21:**821-829.

71. Shi H, Burch KS, Johnson R, Freund MK, Kichaev G, Mancuso N, Manuel AM, Dong N, Pasaniuc B: **Localizing components of shared transethnic genetic architecture of complex traits from gwas summary data.** *Am J Hum Genet* 2020, **106:**805-817.

72. Zhang J, Zhang S, Qiao J, Wang T, Zeng P: **Similarity and diversity of genetic architecture for complex traits between East Asian and European populations.** *BMC Genomics* 2023, **24:**314.

73. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Genomes P, Bustamante CD: **Demographic history and rare allele sharing among human populations.** *Proc Natl Acad Sci U S A* 2011, **108:**11983-11988.

74. Ntzani EE, Liberopoulos G, Manolio TA, Ioannidis JP: **Consistency of genome-wide associations across major ancestral groups.** *Hum Genet* 2012, **131:**1057-1071.

75. Bohaczuk SC, Thackray VG, Shen J, Skowronska-Krawczyk D, Mellon PL: **FSHB transcription is regulated by a novel 5' distal enhancer with a fertility-associated single nucleotide polymorphism.** *Endocrinology* 2021, **162**.

76. Mbarek H, Gordon SD, Duffy DL, Hubers N, Mortlock S, Beck JJ, Hottenga JJ, Pool R, Dolan CV, Actkins KV, et al: **Genome-wide association study meta-analysis of dizygotic twinning illuminates genetic regulation of female fecundity.** *Hum Reprod* 2024, **39:**240-257.

77. Laisk T, Kukuskina V, Palmer D, Laber S, Chen CY, Ferreira T, Rahmioglu N, Zondervan K, Becker C, Smoller JW, et al: **Large-scale meta-analysis highlights the hypothalamic-pituitary-gonadal axis in the genetic regulation of menstrual cycle length.** *Hum Mol Genet* 2018, **27:**4323-4332.

78. Diaz-Perales A, Quesada V, Sanchez LM, Ugalde AP, Suarez MF, Fueyo A, Lopez-Otin C: **Identification of human aminopeptidase O, a novel metalloprotease with structural similarity to aminopeptidase B and leukotriene A4 hydrolase.** *J Biol Chem* 2005, **280:**14310-14317.

79. Arefi S, Mottaghi S, Sharifi AM: **Studying the correlation of renin-angiotensin-system (RAS) components and insulin resistance in polycystic ovary syndrome (PCOs).** *Gynecol Endocrinol* 2013, **29:**470-473.

80. Zhao J, Zhang Y, Li W, Yao M, Liu C, Zhang Z, Wang C, Wang X, Meng K: **Research progress of the Fanconi anemia pathway and premature ovarian insufficiencydagger.** *Biol Reprod* 2023, **109:**570-585.

81. Keaton JM, Kamali Z, Xie T, Vaez A, Williams A, Goleva SB, Ani A, Evangelou E, Hellwege JN, Yengo L, et al: **Genome-wide analysis in over 1 million individuals of European ancestry yields improved polygenic risk scores for blood pressure traits.** *Nat Genet* 2024, **56:**778-791.

82. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL: **Leveraging polygenic functional enrichment to improve GWAS power.** *Am J Hum Genet* 2019, **104:**65-75.

83. Surendran P, Feofanova EV, Lahrouchi N, Ntalla I, Karthikeyan S, Cook J, Chen L, Mifsud B, Yao C, Kraja AT, et al: **Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals.** *Nat Genet* 2020, **52:**1314-1332.

84. Dowrick JM, Roy NC, Bayer S, Frampton CMA, Talley NJ, Gearry RB, Angeli-Gordon TR: **Unsupervised machine learning highlights the challenges of subtyping disorders of gut-brain interaction.** *Neurogastroenterol Motil* 2024, **36:**e14898.

85. Johansson A, Andreassen OA, Brunak S, Franks PW, Hedman H, Loos RJF, Meder B, Melen E, Wheelock CE, Jacobsson B: **Precision medicine in complex diseases-Molecular subgrouping for improved prediction and treatment stratification.** *J Intern Med* 2023, **294:**378-396.

86. Varghese JS, Carrillo-Larco RM, Narayan KV: **Achieving replicable subphenotypes of adult-onset diabetes.** *Lancet Diabetes Endocrinol* 2023, **11:**635-636.

87. Aoki Y, Hamren B, Clegg LE, Stahre C, Bhatt DL, Raz I, Scirica BM, Oscarsson J, Carlsson B: **Assessing reproducibility and utility of clustering of patients with type 2 diabetes and established CV disease (SAVOR -TIMI 53 trial).** *PLoS One* 2021, **16:**e0259372.

88. Wen Z, Hammoud MZ, Siegel CE, Laska EM, Abu-Amara D, Etkin A, Milad MR, Marmar CR: **Neuroimaging-based variability in subtyping biomarkers for psychiatric heterogeneity.** *Mol Psychiatry* 2025, **30:**1966-1975.

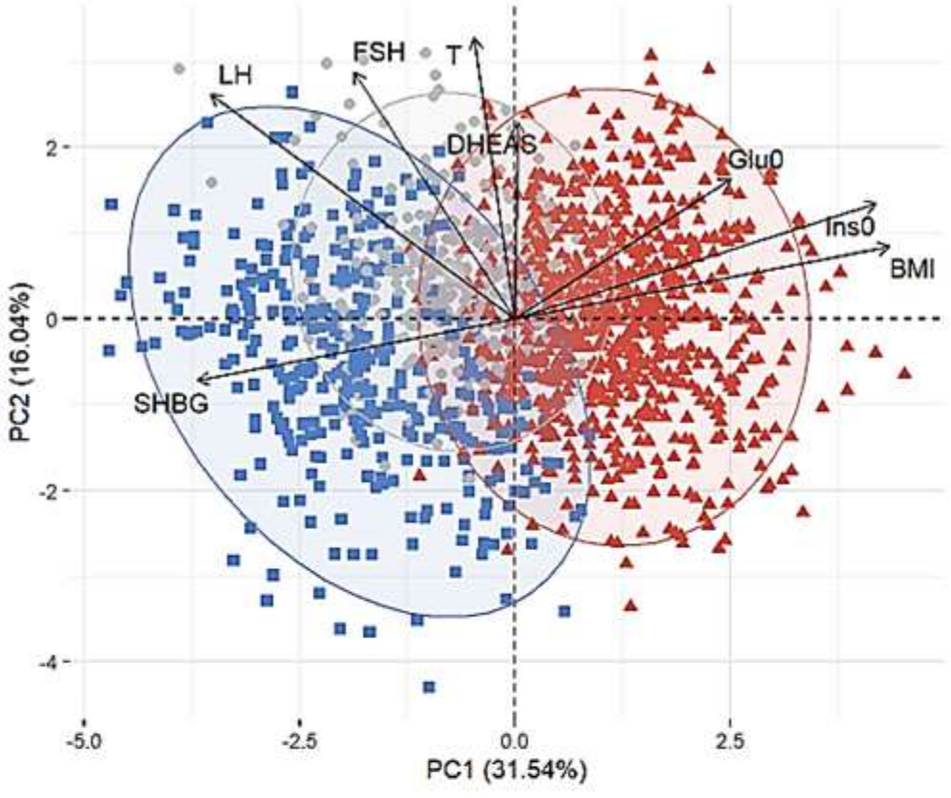
89. Zhao H, Xu Y, Xue B, Zhao S, Zhang M, Wu XK, Wang T, Wen Y, Li S, Zhang Q, et al: **Multi-ancestry genome-wide association analyses of polycystic ovary syndrome.** *Nat Genet* 2025, **57:**2669-2681.

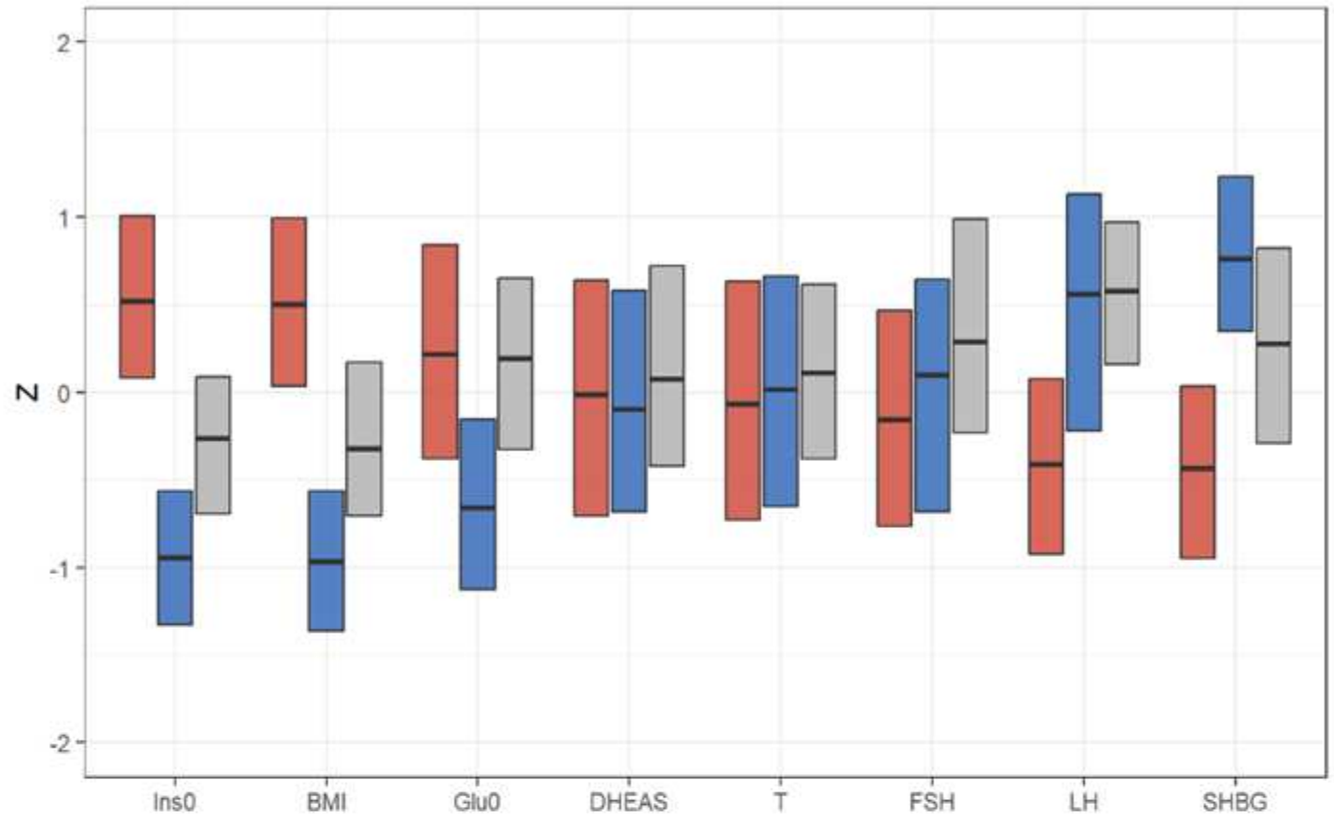
90. Sankaranarayanan L, Brewer KJ, Morrow S, Johnson GD, Barrera A, Venukuttan R, Sisk R, Dunaif A, Reddy TE: **Gene regulatory activity associated with polycystic ovary syndrome revealed DENND1A-dependent testosterone production.** *Nat Commun* 2025, **16:**7697.

91. Mukherjee S, Mez J, Trittschuh EH, Saykin AJ, Gibbons LE, Fardo DW, Wessels M, Bauman J, Moore M, Choi SE, et al: **Genetic data and cognitively defined late-onset Alzheimer's disease subgroups.** *Mol Psychiatry* 2020, **25:**2942-2951.

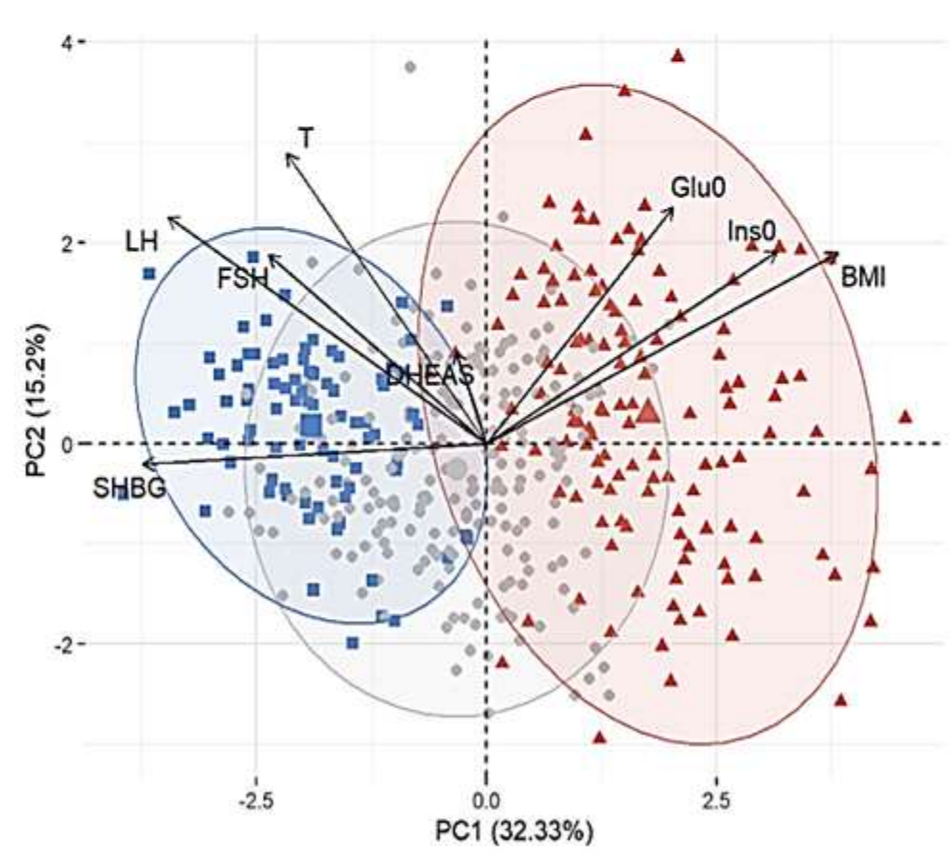
92. Haendel MA, Chute CG, Robinson PN: **Classification, ontology, and precision medicine.** *N Engl J Med* 2018, **379:**1452-1462.

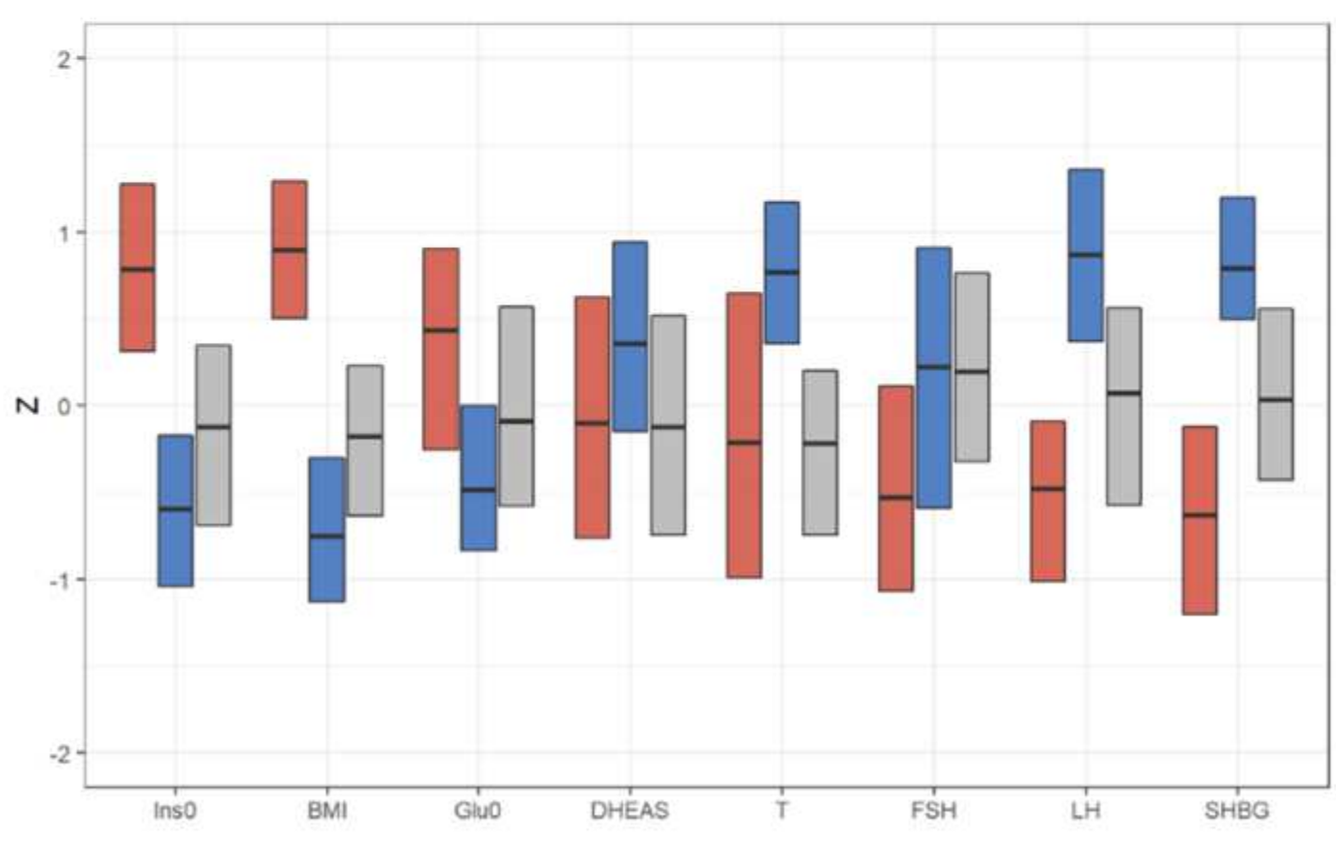
EUR





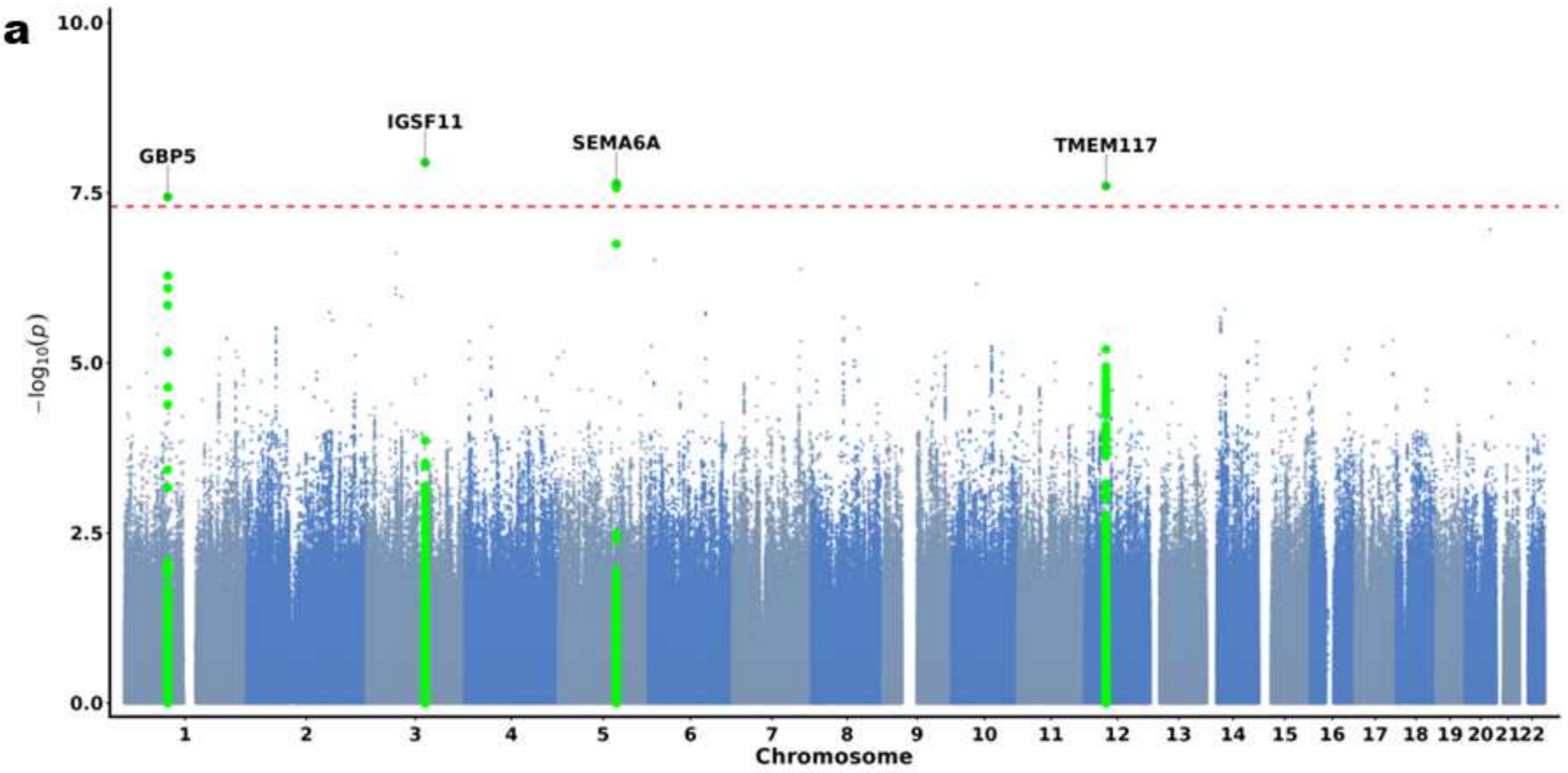
KOR



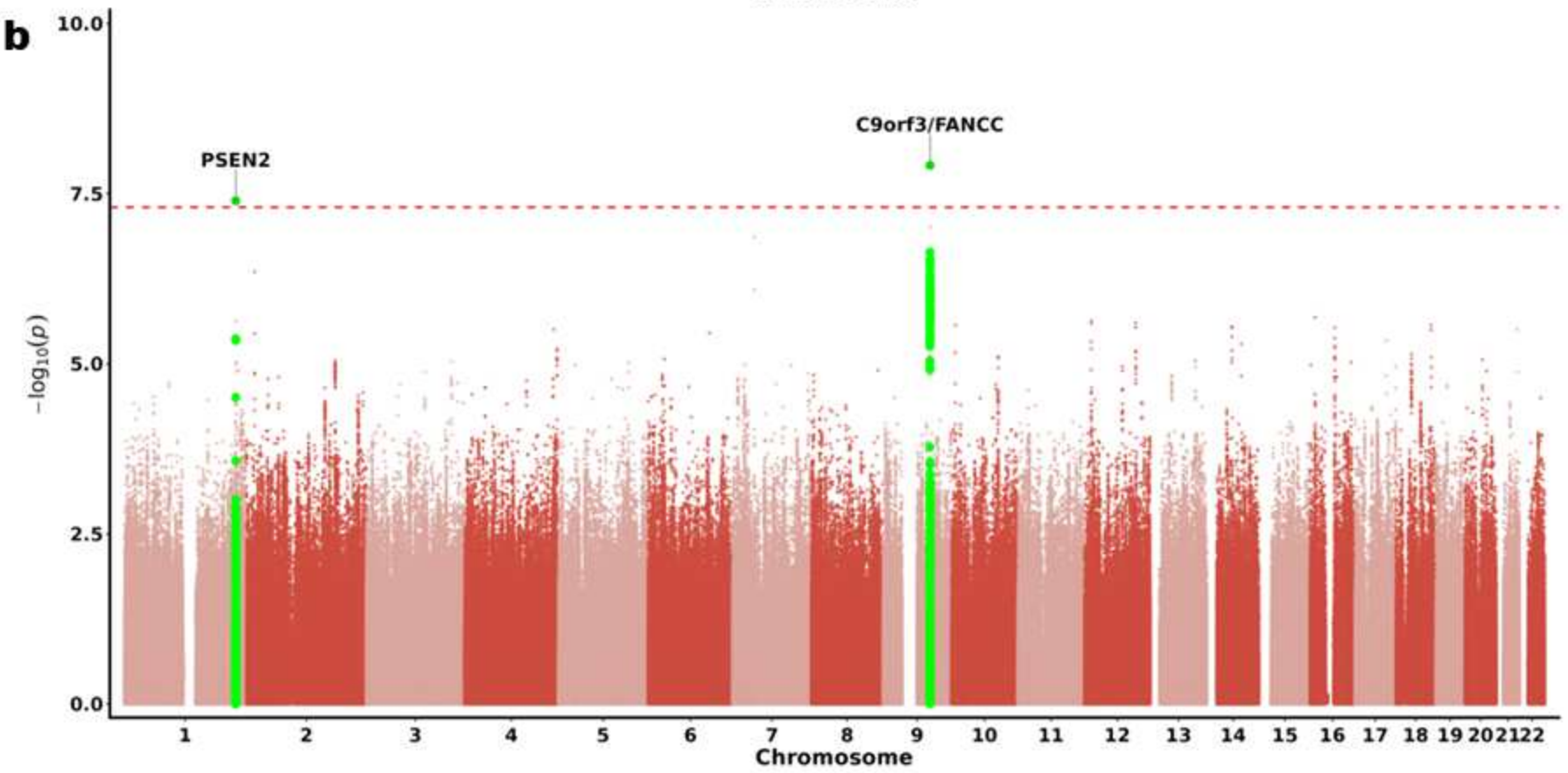


Subtype Reproductive Metabolic Background

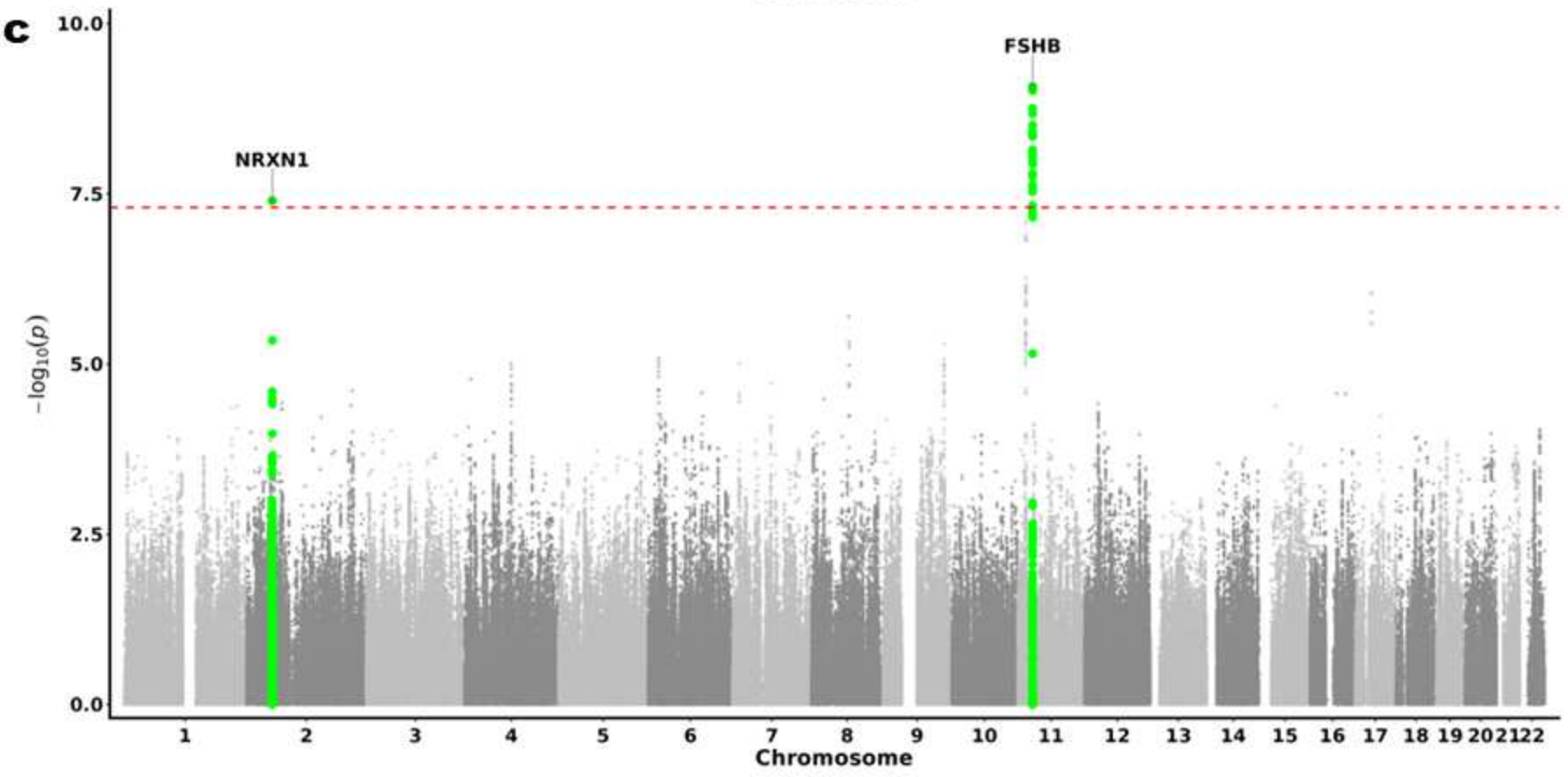
a

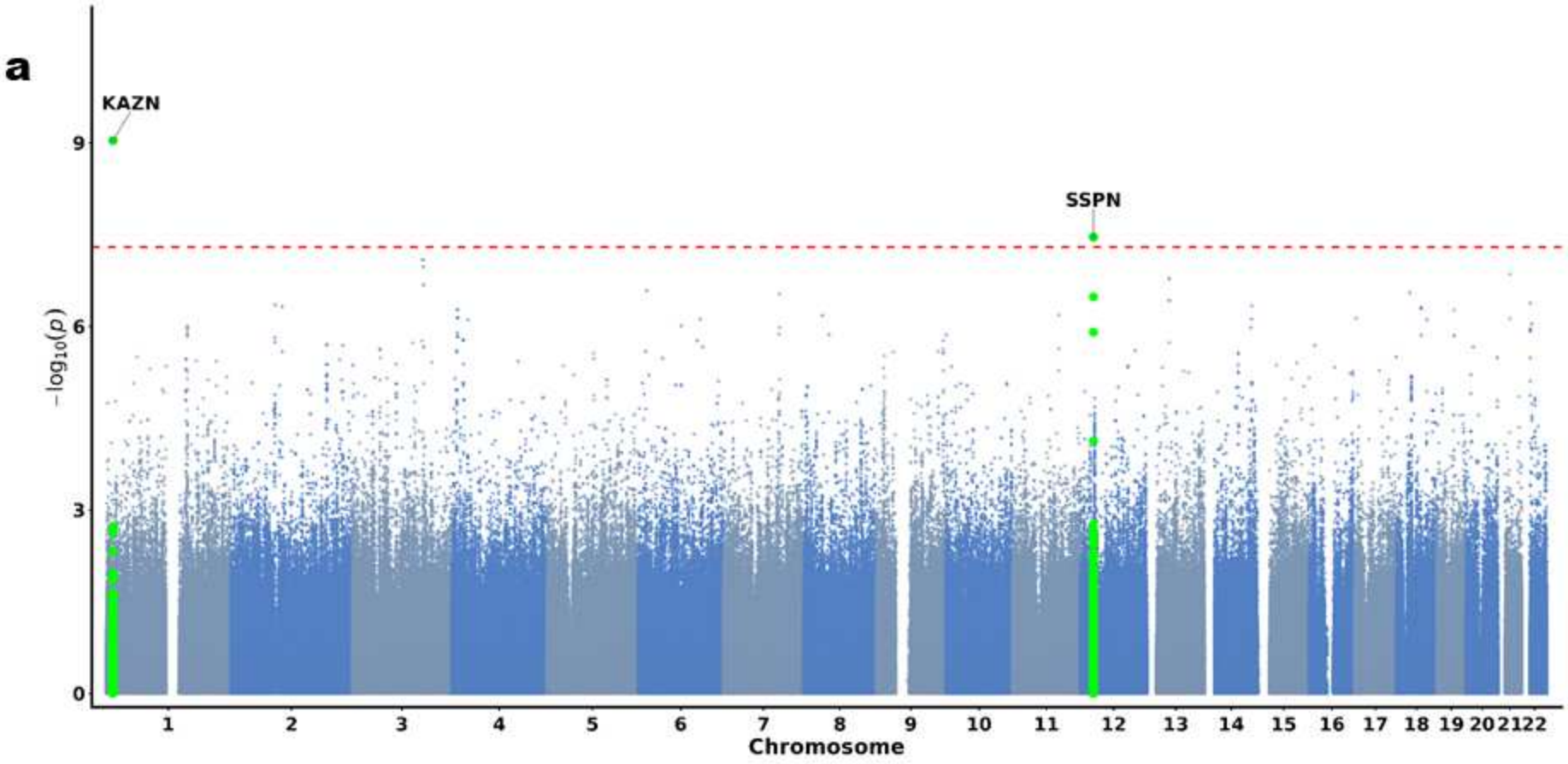

GBP5
IGSF11
SEMA6A
TMEM117
10.0
7.5
5.0
2.5
0.0
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

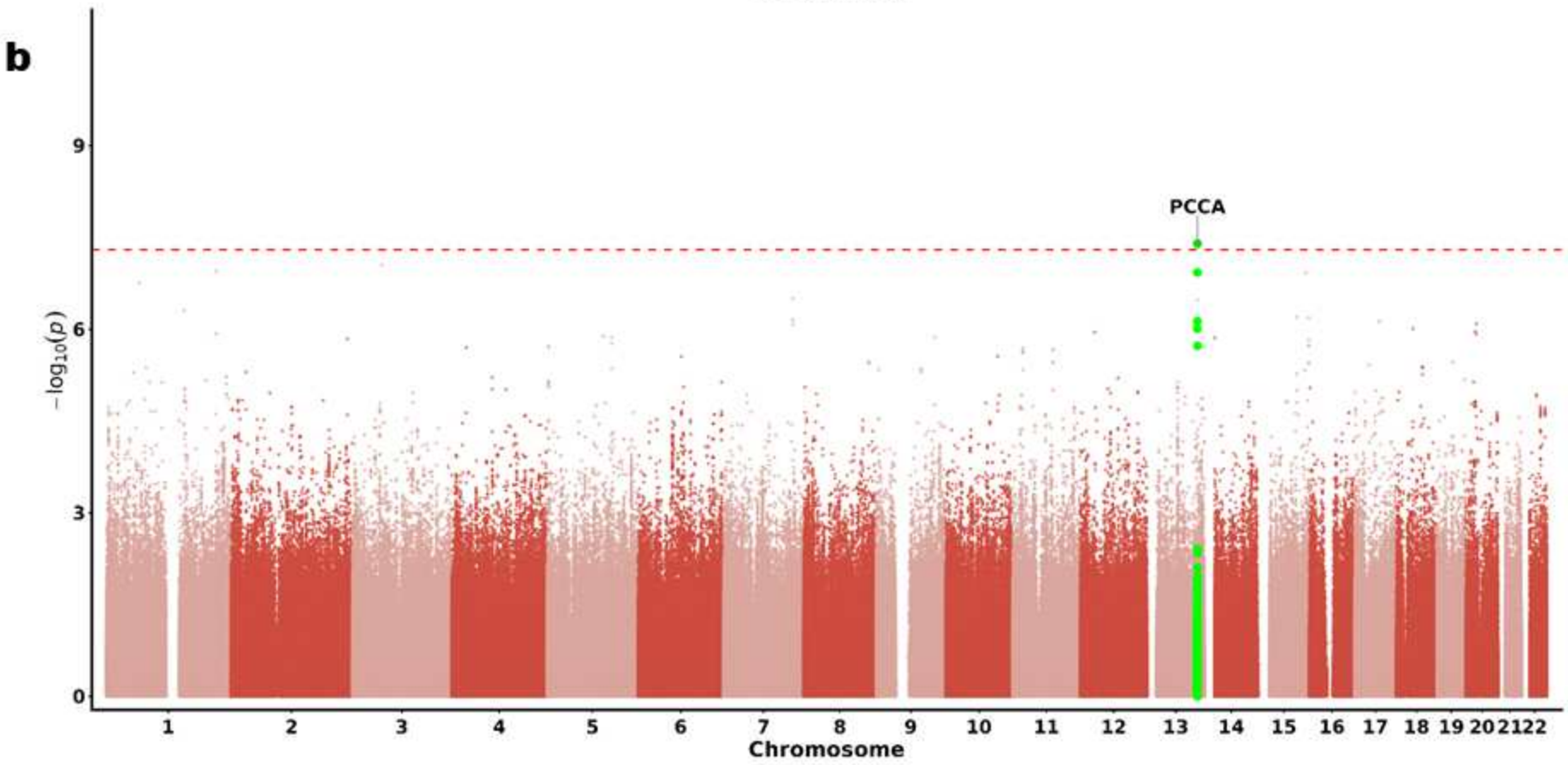
b

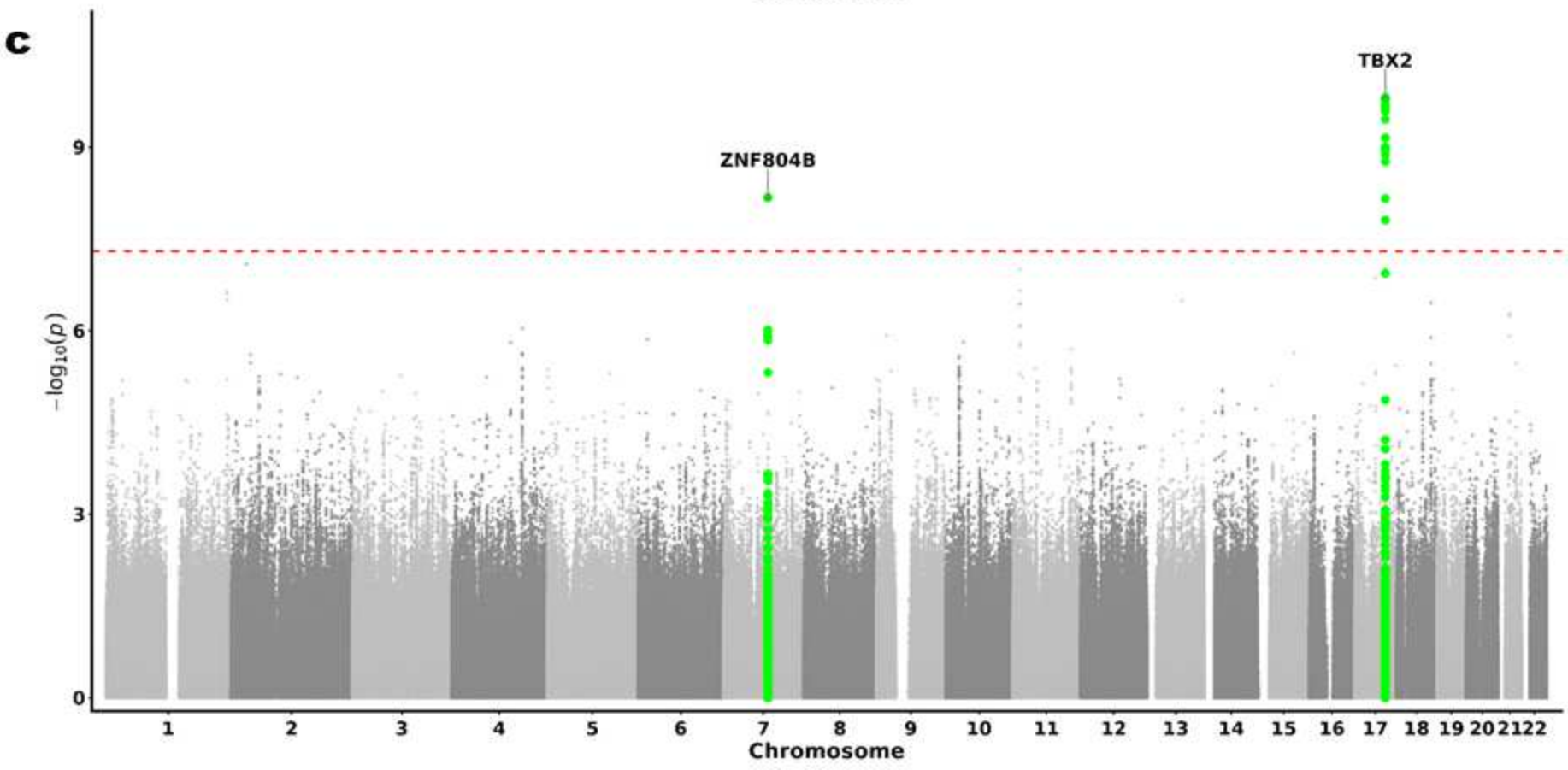

PSEN2
C9orf3/FANCC
10.0
7.5
5.0
2.5
0.0
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

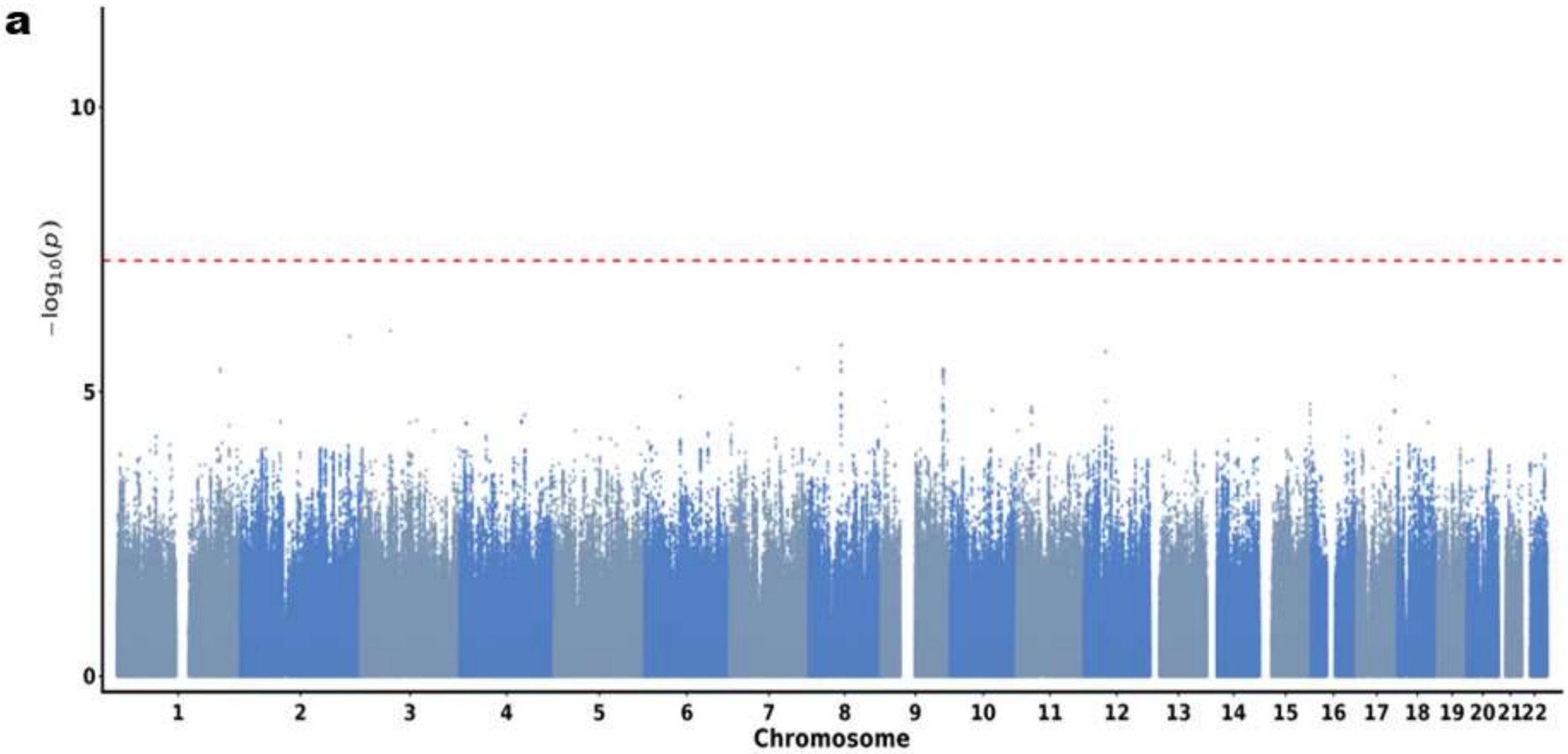
c

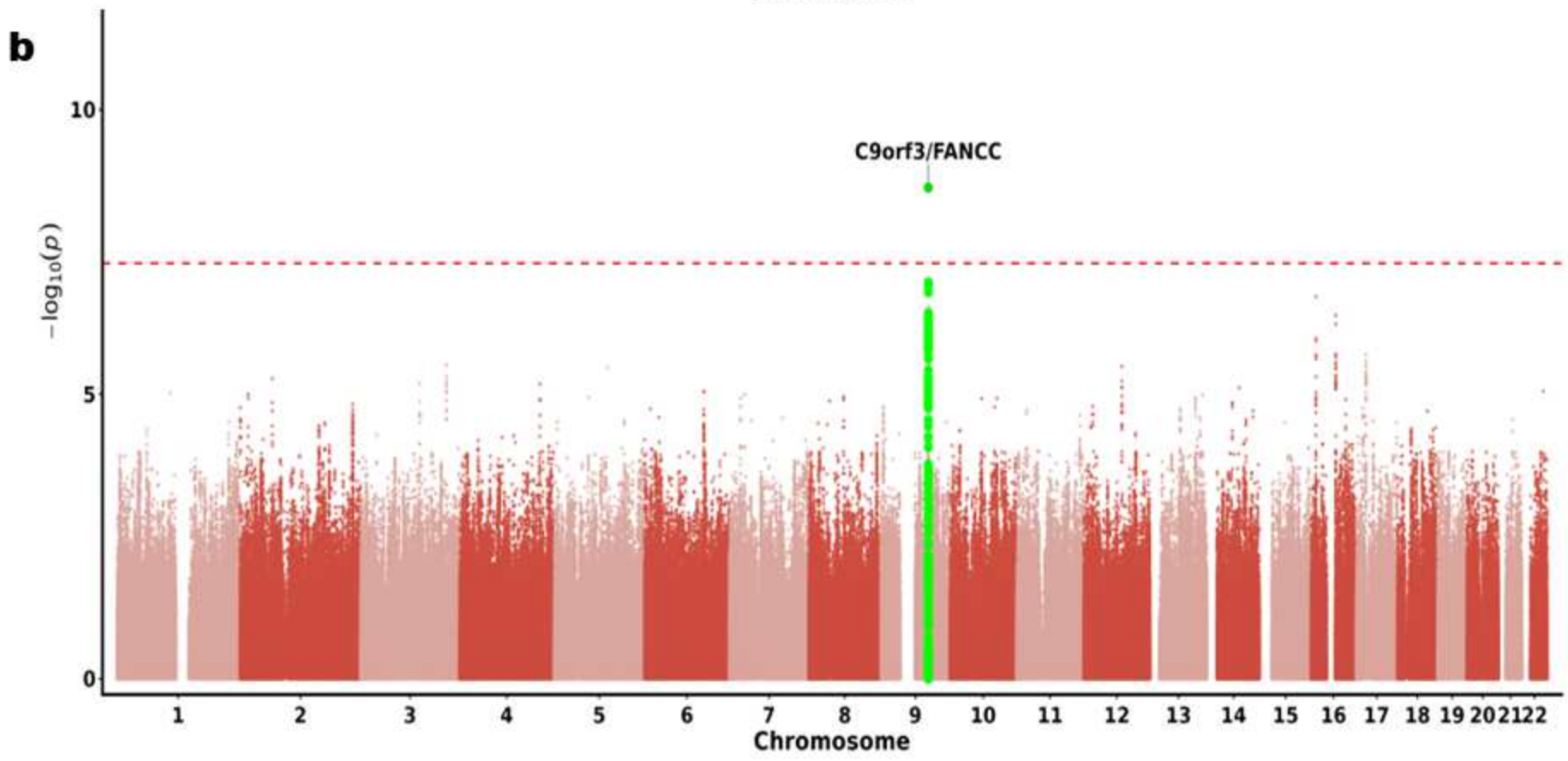

NRXN1
FSHB
10.0
7.5
5.0
2.5
0.0
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

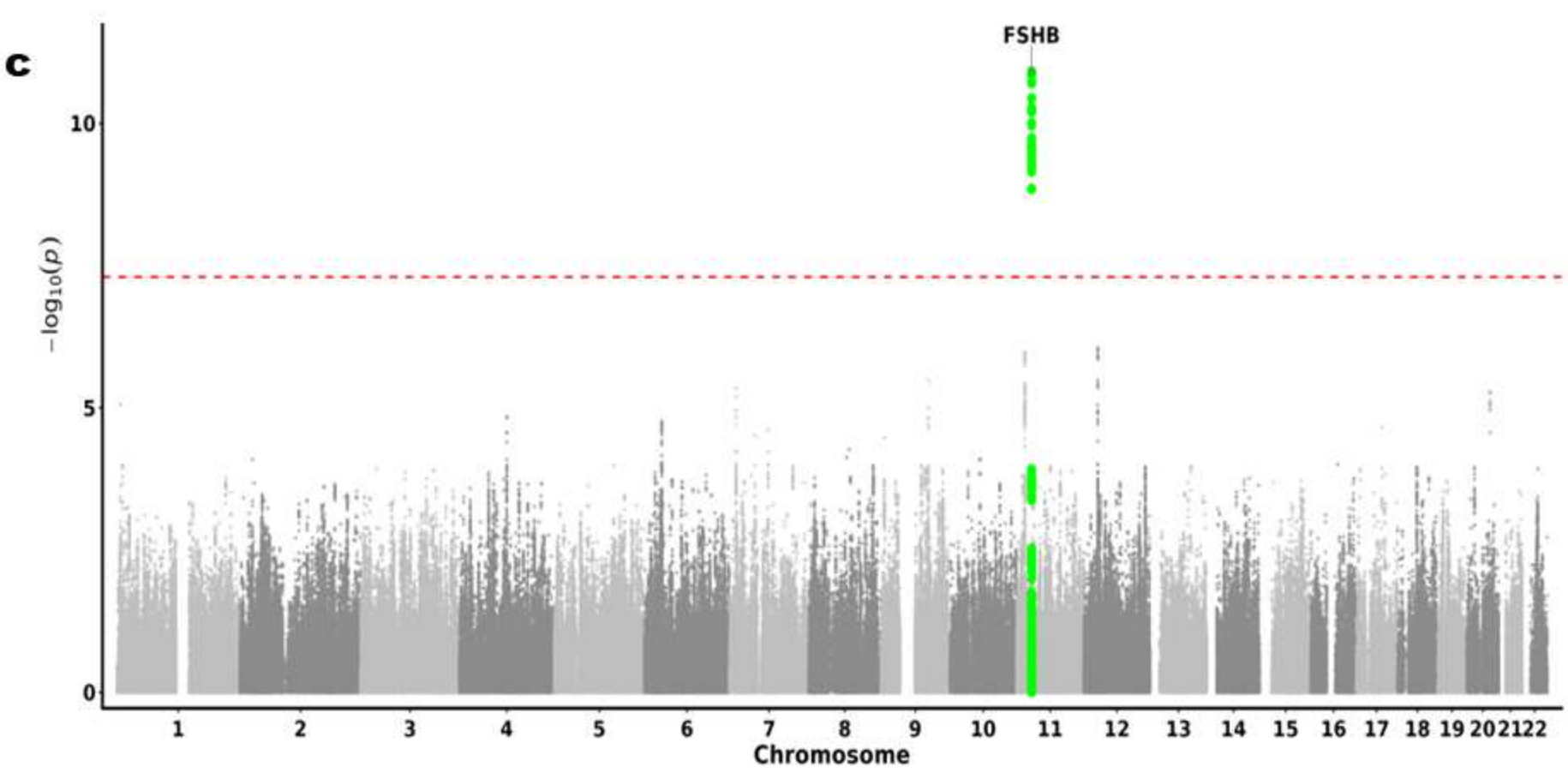

a
KAZN
SSPN
9
6
3
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

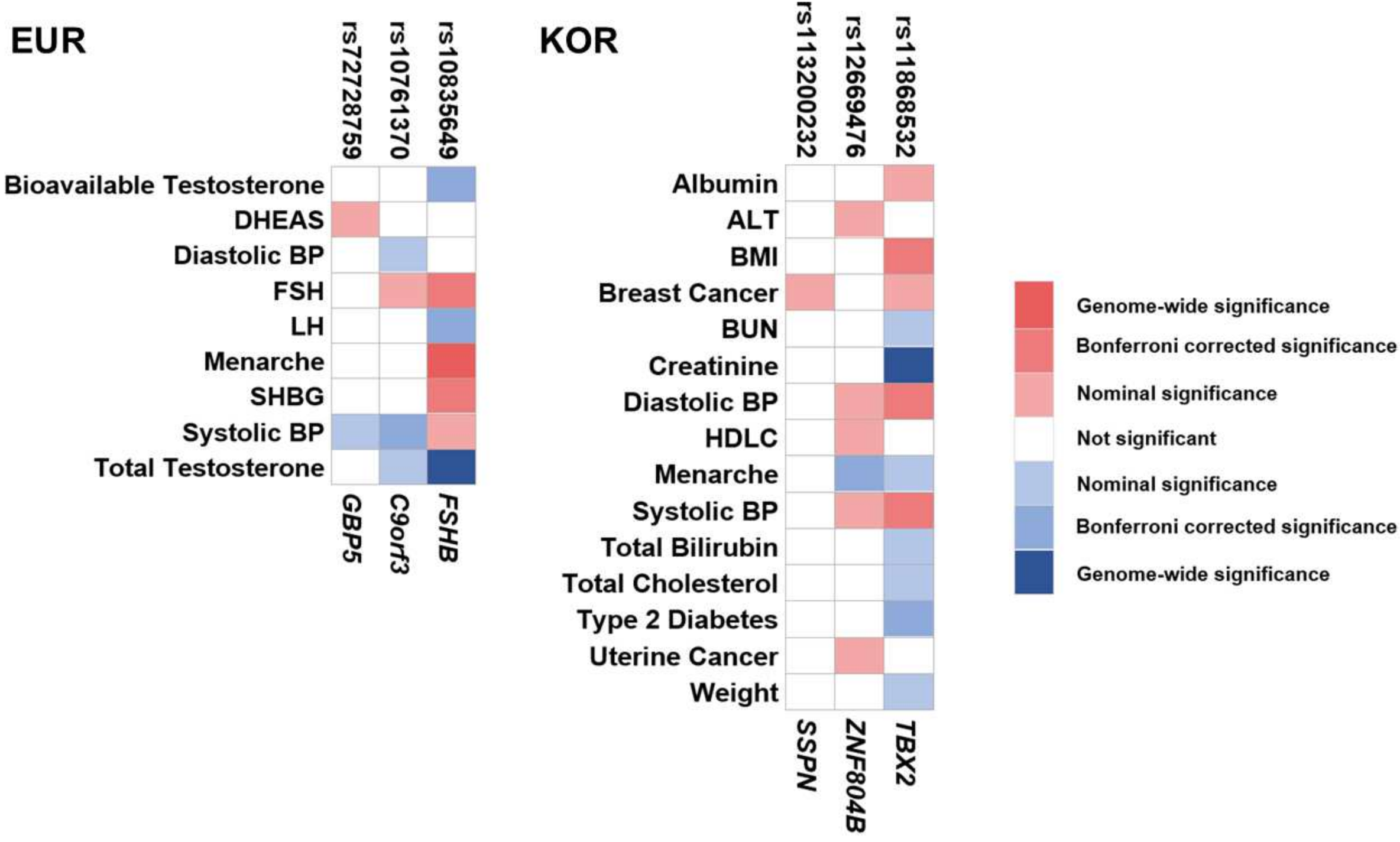

b
PCCA
9
6
3
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


c
TBX2
ZNF804B
9
6
3
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


a
10
5
0
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


b
C9orf3/FANCC
10
5
0
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


c
FSHB
10
5
0
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


EUR
rs72728759
rs10761370
rs10835649
Bioavailable Testosterone
DHEAS
Diastolic BP
FSH
LH
Menarche
SHBG
Systolic BP
Total Testosterone
GBP5
C9orf3
FSHB
KOR
rs113200232
rs12669476
rs11868532
Albumin
ALT
BMI
Breast Cancer
BUN
Creatinine
Diastolic BP
HDLC
Menarche
Systolic BP
Total Bilirubin
Total Cholesterol
Type 2 Diabetes
Uterine Cancer
Weight
SSPN
ZNF804B
TBX2
Genome-wide significance
Bonferroni corrected significance
Nominal significance
Not significant
Nominal significance
Bonferroni corrected significance
Genome-wide significance

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1SupplementalFiguresS1S4andTablesS1S5S7.docx
- Additionalfile2TableS6.xlsx