medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Understanding and Predicting Polycystic Ovary Syndrome through Shared Genetic Architecture with Testosterone, SHBG, and Inflammatory Markers

Lillian Kay Petersen[1,2,†], Garyk Brixi[1,†], Jun Li[4,5], Jie Hu[6], Zicheng Wang[1], Xikun Han[1], Anat Yaskolka Meir[1], Jaakko Tyrmi[7], Shruthi Mahalingaiah[1,8], Terhi Piltonen[9], and Liming Liang[1,3,*]

[†]These authors contributed equally
[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health
[2]Center for Nonlinear Studies, Los Alamos National Laboratory
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[4]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School
[5]Department of Nutrition, Harvard T.H. Chan School of Public Health
[6]Division of Women's Health, Department of Medicine, Brigham and Women's Hospital
[7]Faculty of Medicine and Health Technology, Tampere University
[8]Department of Environmental Health, Harvard T.H. Chan School of Public Health
[9]Department of Obstetrics and Gynaecology, PEDEGO Research Unit, Medical Research Center, Oulu University Hospital
[*]Corresponding author

February 23, 2024

Abstract

Polycystic ovary syndrome (PCOS) is a common hormonal disorder that affects one out of eight women and has high metabolic and psychological comorbidities. PCOS is thought to be associated with obesity, hormonal dysregulation, and systemic low-grade inflammation, but the underlying mechanisms remain unclear. Here we study the genetic relationship between PCOS and obesity, testosterone, sex hormone binding globulin (SHBG), and a wide-range of inflammatory markers. First, we created a large meta-analysis of PCOS (7,747 PCOS cases and 498,227 controls) and identified four novel genetic loci associated with PCOS. These novel loci have been previously associated with gene expression in multiple PCOS-relevant tissues including the thyroid and ovary. We then further incorporated GWASs for obesity (n=681,275), SHBG (n=190,366), testosterone (n=176,687), and 138 inflammatory biomarkers (average n=30,000). Using Mendelian randomization methods, we replicated genetic causal relationships from obesity and SHBG to PCOS. We identified significant genetic correlations between PCOS and eleven inflammatory biomarkers, including novel and strong correlations with death receptor 5 (LDSC rg = 0.54, FDR = 0.043), among others. Although no statistically significant causal relationship was observed between inflammatory markers and PCOS, 31 inflammatory biomarkers showed significant causal effects on SHBG or testosterone, supporting a potentially etiological role of chronic inflammation in influencing sex hormone levels. Finally, we show that combining the polygenic risk scores of PCOS and PCOS-related traits improves genetic prediction of PCOS cases in the UK Biobank and MGB Biobank, as compared to using only the risk score of PCOS. Together, these results support the theory that immune responses are altered in PCOS patients and that chronic inflammation may play a role in testosterone dysregulation.

1 Introduction

Polycystic ovary syndrome (PCOS) is a complex endocrine disorder that is estimated to affect between 5-20% of women of reproductive age [1]. PCOS is mainly diagnosed using the Rotterdam criteria, which requires the presence of two out of the following three symptoms: biochemical or clinical hyperandrogenism, irregular menstruation or anovulation, and polycystic ovarian morphology [2]. Women with PCOS report

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

lowered work ability and quality of life [3, 4, 5]. PCOS has been related to comorbidities including metabolic, reproductive and psychological disorders, obesity, diabetes, dyslipidemia, metabolic syndrome, obstructive sleep apnea, cardiovascular disease, subfertility, endometrial cancer, depression, anxiety, and eating disorders [6].

The pathogenesis of PCOS includes both genetic and environmental factors, but the specific mechanisms remain unclear [7]. PCOS is polygenic and highly heritable (heritability >70% based on twin studies) [8]. Genome-wide association studies (GWAS) have identified 22 genetic loci associated with PCOS, but the proportion of heritability explained remains low with a limited number of functional studies [9, 10, 11]. Hyperandrogenism and PCOS pathogenesis have been linked to factors such as weight gain, obesity, and insulin resistance; chronic low-grade inflammation [12]; and low sex hormone binding globulin (SHBG) levels [13]. Previous studies have found significant genetic correlations between PCOS, BMI, and waist-to-hip ratio (WHR) [14], and a potential causal role of BMI, type 2 diabetes, SHBG, and other traits in the development of PCOS based on previous Mendelian randomization (MR) studies [15]. PCOS patients often have elevated inflammation markers compared with age- and BMI-matched controls [16, 17, 18], and anti-inflammatory therapy can reverse PCOS-like traits in animal models [7]. This supports a potential role for altered immune response in PCOS.

Here we investigate the genetic relationship between PCOS, obesity, testosterone, and SHBG, and how this relationship is connected with and potentially mediated by inflammation. Figure 1 shows an overview of our study design.

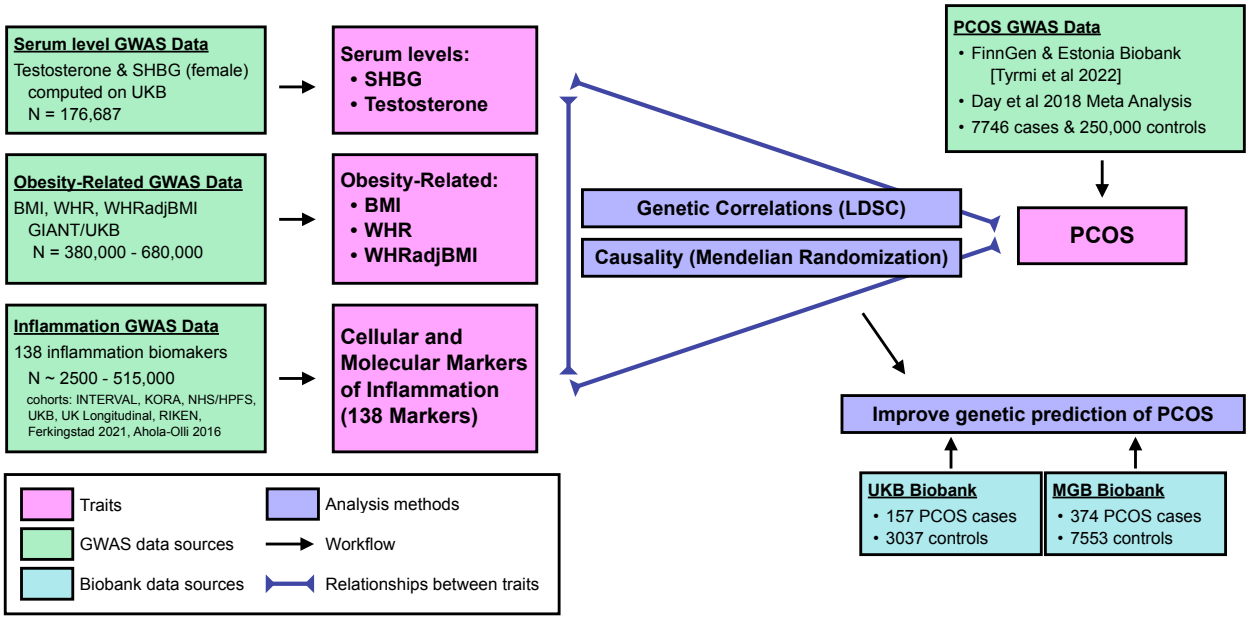

Figure 1: Flowchart depicting data sources and analysis strategies.

2 Results

2.1 PCOS Meta-analysis

We first combined Tyrmi et al. [9] and Day et al. [10] summary statistics to create a large meta-analysis of PCOS. The meta-analysis identified 26 genome-wide significant ($p < 5 \times 10^{-8}$) linkage disequilibrium (LD) independent loci with significant associations with PCOS. Four of thwese loci have not been previously reported to be associated with PCOS in the literature: rs61030588, rs11234902, rs56738967, rs78378222. Details about the lead variants are shown in table 1, and the full meta-analysis results are shown in figure 2.

All of the identified novel loci for PCOS were significant expression quantitative trait (eQTL) loci in multiple tissues [19]. Rs61030588 on chromosome 2 is within the gene *MSH6* and is a known eQTL in

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

CHR:BP Ref/Effect	RSID	Beta	Effect Allele Freq	p-value	Nearest Gene
Replicated PCOS Loci					
2: 43561780 G/A	rs7563201	-0.108	0.521	5.93×10^{-8}	*THADA*
2: 213387900 C/T	rs7564590	0.144	0.645	1.90×10^{-12}	*ERBB4*
5: 16836005 C/T	rs9312937	-0.125	0.457	1.90×10^{-12}	*MYO10*
8: 11621450 G/A	rs2740433	-0.140	0.344	1.91×10^{-11}	*GATA4*
9: 126637668 G/T	rs7028482	-0.304	0.065	1.49×10^{-14}	*DENND1A*
11: 30226356 C/T	rs11031005	-0.202	0.142	8.31×10^{-14}	*FSHB*
11: 113949232 C/T	rs1784692	0.203	0.138	3.69×10^{-12}	*ZBTB16*
12: 75978358 G/A	rs1148006	-0.117	0.705	4.70×10^{-8}	*KRR1*
22: 29098376 G/A	rs182075939	-0.523	0.046	1.93×10^{-16}	*CHEK2*
New PCOS Loci					
2: 47995854 G/A	rs61030588	-0.125	0.283	3.42×10^{-8}	*MSH6*
11: 86712340 G/A	rs11234902	-0.141	0.762	7.42×10^{-10}	*FZD4-DT*
16: 79740541 G/C	rs56738967	0.117	0.695	2.46×10^{-8}	*MAF*
17: 7571752 G/T	rs78378222	-0.405	0.021	4.47×10^{-8}	*TP53*

Table 1: Loci that reach genome-wide significance for PCOS in the meta-analysis and (top) are in LD with loci previously reported for PCOS; or (bottom) are not in LD with loci previously reported for PCOS.

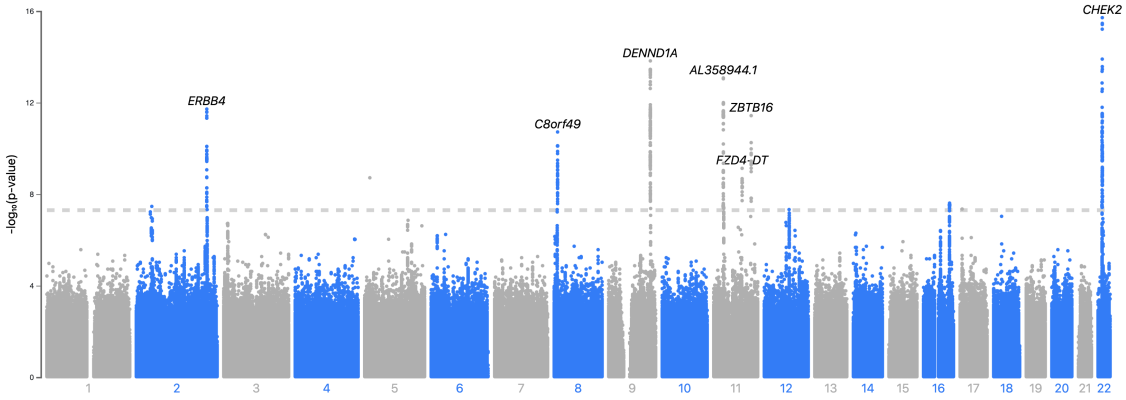

Figure 2: Manhattan plot of PCOS meta-analysis

multiple tissues including the thyroid and ovaries. *MSH6* is a gene involved in DNA repair, which has previously been implicated in menopause age and ovarian aging [20]. The novel PCOS locus rs11234902 on chromosome 11 is a significant thyroid tissue eQTL for *RP11-736K20.6*, an RNA gene. The novel locus on chromosome 16, rs56738967, is a significant eQTL for *MAFTRR* in a large number of tissues including the thyroid and ovary. *MAFTRR* is a lncRNA involved in gene regulation. The new locus on chromosome 17 is on an intronic region of *TP53*, a tumor suppressor gene. This variant has been reported to increase risk of uterine fibroids, gliomas, and lean mass, while other variants on *TP53* are inked to levels of SHBG and testosterone [21, 22, 23]. The rs78378222 locus is a significant eQTL for *TP53* in several tissues including adipose tissue. The full tables of eQTL hits for novel PCOS loci lead variants are included in the supplementary file 1.

2.2 Genetic Correlation

2.2.1 Genetic correlations between PCOS, obesity, testosterone, and SHBG

We curated GWAS summary statistics for body mass index (BMI) and female-specific GWAS for SHBG, testosterone, waist-to-hip ratio (WHR), and female-specific waist-hip ratio adjusted for BMI (WHRadjBMI), and computed the genetic correlations between these traits with PCOS using LDSC.

The genetic correlations between these PCOS-related traits are shown in table 3. PCOS has strong genetic correlations with SHBG (r_g=0.45), BMI (r_g=0.35), and WHR (r_g=0.35), and has weaker correlations with WHRadjBMI (r_g=0.17) and testosterone (r_g=0.16). SHBG is inversely correlated with all of obesity-related

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

traits especially WHR. We did not find significant correlations between testosterone and SHBG.

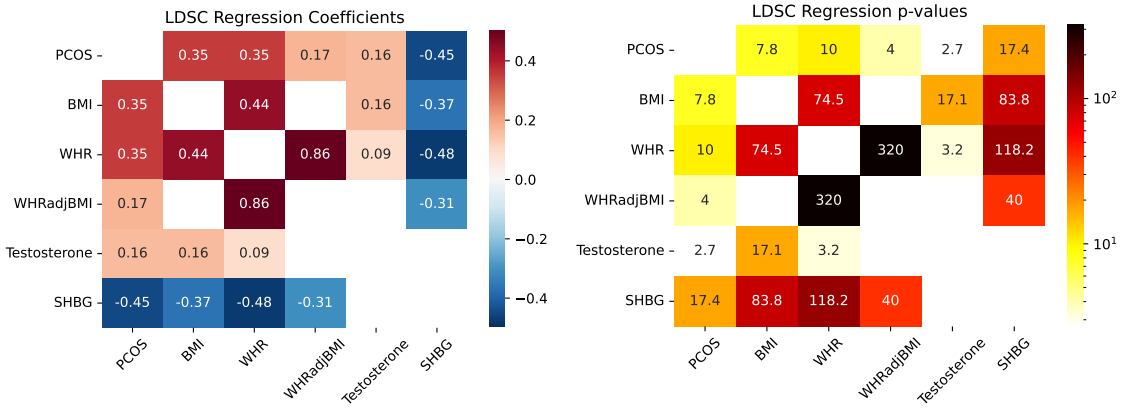

Figure 3: Genetic correlations and p-values (-log10), computed via LDSC, between the non-inflammatory traits in our study. Values that did not meet FDR < 0.05 are not shown.

2.2.2 Genetic correlations of inflammatory biomarkers with PCOS, obesity, testosterone, and SHBG

We created meta-analyses of 138 inflammatory biomarkers, including immune cell counts and biomarker serum levels. The sample sizes of these biomarkers ranged between 2,538 and 505,690 individuals, and SNP-based heritability ranged from 0 to 0.5. The full table describing these inflammatory biomarkers are shown in the supplementary file 2.

We first examined genetic correlations between inflammatory biomarkers and PCOS, SHBG and testosterone. For inflammatory biomarkers with significant genetic correlations with at least one of the three traits (FDR < 0.05), we further examined their genetic correlations with obesity traits (as shown in figure 4, with the values included in supplementary file 3).

PCOS showed significant genetic correlation with TRAILr2, IL2, leptin, IL1ra, HGF, CRP, adiponectin, as well as circulating counts of total white blood cells (WBC), lymphocytes (LymC), neutrophils (NeuC), and monocytes (MonC). Most of these inflammatory biomarkers that correlated with PCOS showed significant genetic correlation of similar magnitude with BMI, and WHR, or negative SHBG. The exception was TRAILr2, which is only related to BMI but not WHR.

SHBG is significantly genetically correlated with 36 inflammation markers, all of which were also significantly correlated to an obesity trait in the opposite directions. Testosterone only showed significant genetic correlations with three inflammatory markers, namely with leptin, CRP, and monocyte count.

Hierarchical clustering on the inflammatory LDSC correlations suggests that WHR, BMI, and low SHBG share a similar inflammatory profile, while PCOS has a inflammatory profile closer to WHRadjBMI and testosterone.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

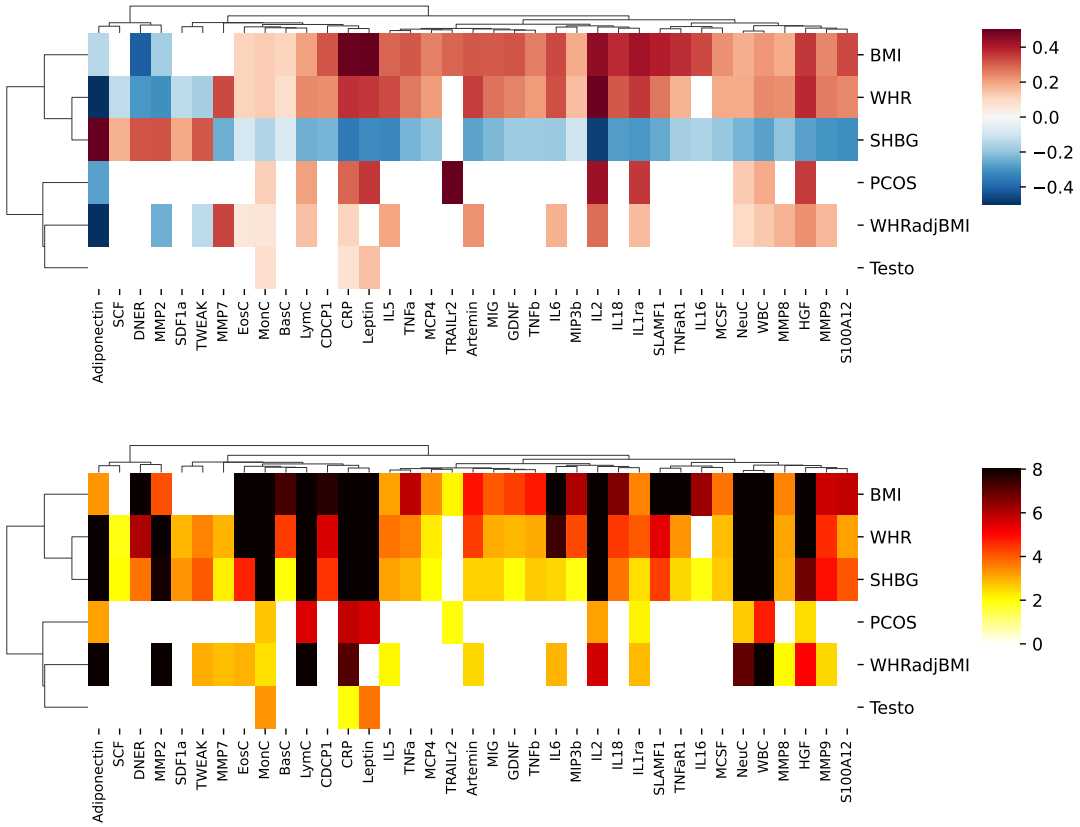

Figure 4: Genetic correlations (top) and -log10 p-value (bottom) computed via LDSC between the traits PCOS, BMI, WHR, WHRadjBMI, SHBG, testosterone and the inflammation markers. Correlations and p-values shown are FDR < 0.05. Rows and columns are clustered by genetic correlation using hierarchical clustering, and we reversed the correlation direction of SHBG before clustering.

2.2.3 Causal relationships between PCOS, obesity, testosterone, and SHBG

To investigate potential causal relationships, we conducted bi-directional MR analysis between PCOS, obesity traits, testosterone, and SHBG. We validated our MR results using several different methods, including CAUSE [24], mode-based estimation (MBE), MR-Egger, and the inverse-variance weighted (IVW) method. We did not test the causal relationship between PCOS and testosterone, as testosterone levels are often used to define PCOS. These MR results are shown in the supplementary file 4.

All MR methods suggest that higher BMI could lead to PCOS ($\text{effect}_{\text{CAUSE}} = 0.59$, $p_{\text{CAUSE}} = 6.4 \times 10^{-4}$). Most methods suggest that WHR has a causal effect on PCOS, and a few methods suggest that WHRadjBMI is causal. Most of the tests also suggest that low SHBG levels have a significant causal effect on PCOS with effect size about -0.25 ($p_{\text{CAUSE}} = 0.041$). PCOS did not show statistically significant causal effects on any of the tested traits (figure 5), which is consistent with previous MR studies [15]. All MR methods suggest that higher BMI, WHR, and WHRadjBMI may lower SHBG levels and that higher WHRadjBMI may decrease testosterone levels (see supplementary file 4).

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

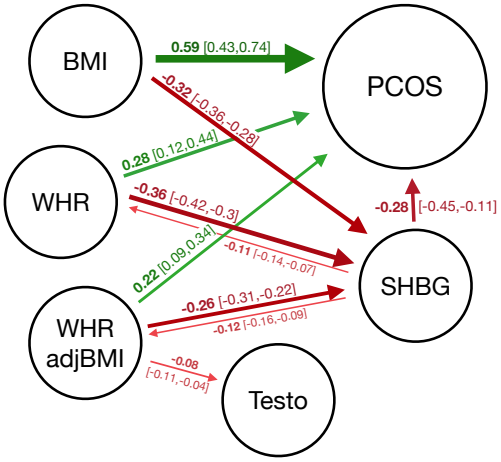

Figure 5: Mendelian randomization estimated causal effects and 95% credible intervals between BMI, WHR, WHRadjBMI, SHBG, Testosterone, and PCOS. Green arrows indicate positive causal effects and red arrows indicate negative causal effects. The thickness and shade of the arrows is proportional to effect size. All estimates are based on the CAUSE model and are significant ($p<0.05$) for the hypothesis that a causal model is a better fit than a sharing model.

2.2.4 Causal relationships between inflammatory biomarkers and PCOS, SHBG, and testosterone

Next we conducted bi-directional MR between PCOS, testosterone, SHBG and the panel of 138 inflammatory biomarkers. For inflammatory markers that are significantly related to any of these traits, we further conducted bi-directional MR between them with BMI and WHR to examine the role of obesity. Our main results were estimated with the MBE MR method, because it has less bias and lower type-I error rates than IVW and is significantly faster than CAUSE. We also examined results from other MR methods to gauge the robustness of the potential associations. Results of MR with inflammatory markers are detailed in supplementary file 5. We did not find statistically significant causality between PCOS and any inflammatory markers based on MR analysis, likely due to the low heritability of the PCOS GWAS.

Twenty-eight inflammatory biomarkers showed a significantly causal effect for SHBG. Of these 28 biomarkers, only CD36antg and CRP were found to decrease SHBG while the other 26 were found to increase SHBG. TWEAK has the strongest causal association with SHBG, an effect that was replicated in all of the MR methods ($\text{beta}_{\text{MBE}} = 0.46$, $p_{\text{MBE}} = 2.46 * 10^{-219}$, $\text{beta}_{\text{IVW}} = 0.260$, $p_{\text{IVW}} = 5.74 * 10^{-3}$, $\text{beta}_{\text{MR-Egger}} = 0.351$, $p_{\text{MR-Egger}} = 1.46 * 10^{-2}$).

Higher genetically-predicted testosterone was found to cause increased levels of CRP ($\text{beta}_{\text{MBE}} = 0.0926$, $p_{\text{MBE}} = 6.91 * 10^{-7}$) (see supplementary table 3). Four inflammatory biomarkers showed significant causal relations with testosterone ($\text{FDR}_{\text{MBE}} =< 0.01$): genetically-predicted TWEAK ($\text{beta}_{\text{MBE}} = 0.057$, $p_{\text{MBE}} = 3.7 * 10^{-8}$) and MMP9 ($\text{beta}_{\text{MBE}} = 0.043$, $p_{\text{MBE}} = 2.9 * 10^{-4}$) were found to increase testosterone, while genetically-predicted IL2Rb ($\text{beta}_{\text{MBE}} = -0.057$, $p_{\text{MBE}} = 7.44 * 10^{-5}$) and IP10 ($\text{beta}_{\text{MBE}} = -0.126$, $p_{\text{MBE}} = 1.33 * 10^{-5}$) were found to decrease testosterone. Interestingly, TWEAK is also the only biomarker which shows a causal effect for both testosterone and SHBG.

We next assessed whether inflammation could be partially mediating the causal effect from obesity to sex hormones. BMI showed a significant causal effect on 4 out of 28 biomarkers that had significant causality for SHBG, as visualized in figure 6a. MR suggests that higher BMI may elevate levels of MMP17, IL10Rb, and IL8, which may in turn increase SHBG levels. Higher BMI and testosterone could increase CRP levels, which may lower levels of SHBG.

Many of the inflammation markers that were found to increase SHBG were also found to decrease WHRadjBMI (figure 6b), suggesting that these markers could have a protective role. Of the 28 inflammatory biomarkers showing causal effect for SHBG, 15 are also significant for WHRadjBMI, 15 significant for WHR, and 1 (FASLG) for BMI. Details of the MR results are available in the supplementary file 5.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

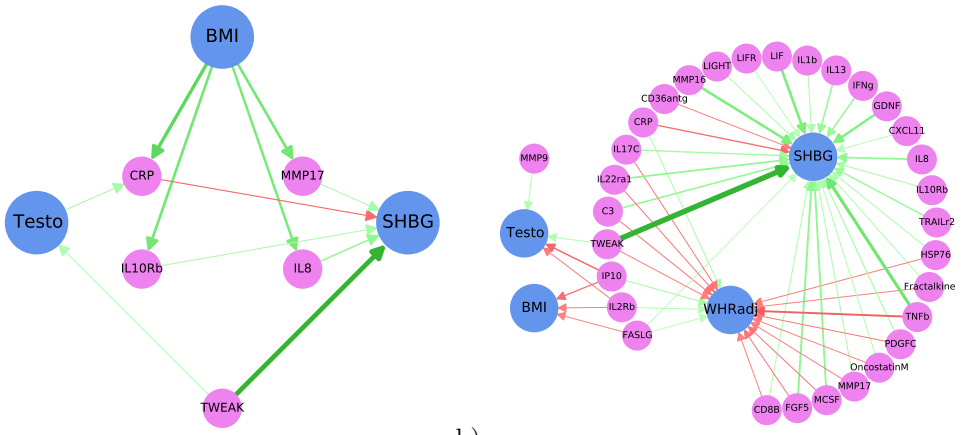

Figure 6: Mendelian randomization results between inflammatory biomarkers and PCOS-related traits. (a) Inflammatory biomarkers may mediate some of the causal relationship between obesity and hormones. (b) Inflammatory biomarkers which significantly regulate SHBG or Testosterone, and the causal relationship of these markers on BMI and WHR adjusted by BMI. Pink circles represent inflammatory markers, blue circles represent traits, and arrows show the direction of causality. Only MBE results are shown in this figure, and only effects with FDR < 0.01 are included. Green arrows indicate positive causality, red arrows indicate negative causality, and the width/shade of the arrow indicates strength. Full MR results can be found in the supplementary file 5.

2.3 Polygenic risk scores combining causal risk factors improve PCOS prediction

Finally, we created a model for predicting genetic risk of developing PCOS based on the combined polygenic risk scores (PRSs) of PCOS, BMI, WHR, WHRadjBMI, SHBG, and testosterone.

We created PRS from the PCOS, BMI, WHR, WHRadjBMI, SHBG, testosterone summary statistics and applied them to all women in the UK Biobank and Mass General Brigham (MGB) biobank, standardizing each PRS to mean 0 and variance 1. We then created lasso logistic regression models to predict PCOS cases using the PRSs, and compared the prediction to that based solely on PCOS PRS. We trained the logistic regressions and selected hyperparameters via a nested 10-fold cross validation to avoid overfitting.

Our new PRS-based model improved PCOS prediction in both the UK Biobank and the MGB Biobank on held-out data. In the UK Biobank, the area under the ROC curve (AUROC) improved from 0.59 when only using the PCOS PRS to 0.72 when combining the PRSs. In the MGB biobank, the AUROC improved from 0.59 to 0.61. UK Biobank and MGB Biobank model performance and coefficients are shown in figure 7.

The coefficients used to predict PCOS varied between the UK biobank and the MGB biobank. In the UK Biobank, BMI PRS had the largest effect size for predicting PCOS cases, followed the PCOS, SHBG, and the testosterone PRSs. In the MGB Biobank, the PCOS and BMI PRS accounted for most of the prediction power.

Further inclusion of including inflammation PRSs did not improve PCOS predictions in UK biobank or the MGB biobank.

3 Discussion

This study investigated genetic correlations and causal relationships between PCOS and obesity, testosterone, SHBG, and biomarkers indicative of a wide-range of inflammatory pathways. Our large GWAS meta-analysis of PCOS identified four novel loci. We demonstrated genetic correlations and potential causal relationships between obesity, testosterone, and SHBG with PCOS, and the potential role of chronic inflammation in these relationships. Interestingly, the genetic correlation between PCOS and testosterone is smaller compared with

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

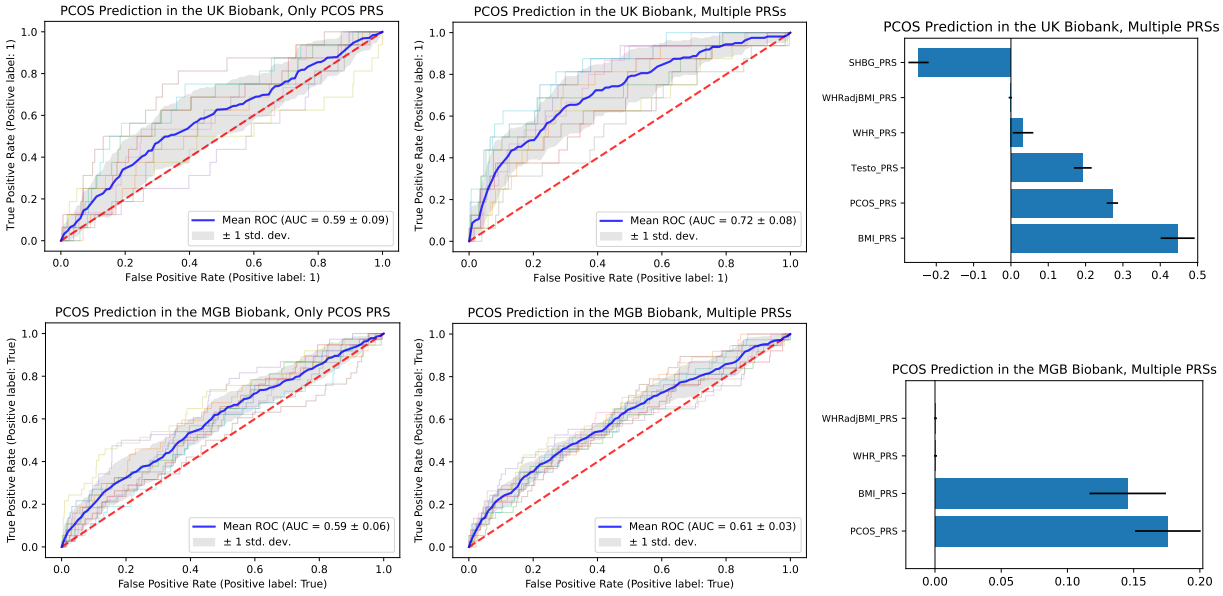

Figure 7: Genetic prediction of PCOS cases in the UK Biobank (top) and MGB biobank (bottom). A nested 10-fold cross validation lasso logistic model was trained on each biobank. The left plots show the receiver operating characteristic (ROC) curve for predicting PCOS cases when only using the PCOS polygenic risk score (PRS). The middle plots show the ROC curve when combining the PCOS PRS with the BMI, WHR, WHRadjBMI, SHBG, and testosterone PRSs, and the left plots show the corresponding model coefficients for each PRS.

those with SHBG, BMI, WHR, and WHRadjBMI, despite hyperandrogenism often being used to characterize PCOS. Finally, we showed that incorporating genetics of causal risk markers of PCOS, namely obesity, testosterone, and SHBG, improved genetic risk prediction of PCOS. This study provides evidence that PCOS shares genetic architecture with a range of inflammatory biomarkers. We found significant genetic correlations between PCOS and eleven inflammatory biomarkers, including biomarkers that have not been previously related to PCOS in the clinical literature, such as TRAILr2 (also known as death receptor 5). A previous study has found that TRAILr2 mediates testosterone-driven apoptosis of PCOS granulosa cells in culture [25], which supports the hypothesis that TRAILr2 is involved in the parthenogenesis of PCOS. While our study provides additional evidence that TRAILr2 may be related to PCOS, further research is needed.

Using Mendelian randomization, we found that obesity is likely a causal risk factor for PCOS, while SHBG may protect against PCOS, consistent with previous MR findings [26].

We found through MR that BMI, WHR, and WHRadjBMI all decrease SHBG, suggesting that the previously reported association between obesity and SHBG may be causal [27]. We further found that the relationship between WHR, WHRadjBMI and SHBG is bidirectional, with high SHBG in turn decreasing WHR and WHRadjBMI. This could indicate a positive feedback loop between central obesity and low SHBG, although verification is needed. Consistently, experiments in mice have previously reported that increasing SHBG downregulates de novo lipogenesis and reduces liver fat [28]. Using MR analysis, we found evidence that a broad array of inflammatory biomarkers may affect levels of SHBG, testosterone, and obesity. Of note, higher genetically-predicted TWEAK showed particularly strong causal effects on increasing SHBG and testosterone levels and lowering WHRadjBMI. We further found that testosterone may lead to higher CRP levels, consistent with previous findings that women with PCOS have higher CRP levels than BMI and age-matched controls [18]. These findings support the existing hypothesis that chronic inflammation may lead to dysregulated hormonal production in the ovaries, and also indicate that a more complex profile of inflammatory biomarkers may be involved in this process than previously considered.

The potential causal relationship between many inflammatory biomarkers and SHBG is surprising and requires further explanation. The majority of these inflammatory biomarkers appear to show protective

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

effect by showing a casual effect for increased SHBG and decreased WHRadjBMI and BMI. It should be noted that many of the directions from MR analysis are opposite to the directions of genetic correlations from LDSC. Further study is needed to validate our findings and elaborate on the mechanisms that underlie these results.

While the results of MR suggest causal links, there could also be mediators such as insulin resistance, diabetes, CVD, metabolic syndrome, and other conditions that are causal for testosterone, SHBG levels, and chronic inflammation. For this reason it is important to interpret the results as one piece of evidence and apply general caution as needed whenever dealing with Mendelian randomization studies.

Taking advantage of these related traits, we created a model to improve genetic prediction of PCOS by incorporating multiple PRS scores, and found that including information from the genetics of obesity, SHBG, and testosterone significantly improved PCOS prediction in two independent biobanks. This result indicates that the power of current PCOS GWAS is still limited for polygenic risk prediction and can be improved by incorporating PRSs from genetically-related traits. Since PCOS is hard to diagnose, up to 75% of cases remain unidentified [29], and using genetics to flag potential cases could help improve detection of high risk population and enable early intervention.

It is worth noting, though, that even the improved AUROC remains relatively low—0.72 in the UK Biobank and 0.61 in the MGB Biobank. A likely reason for this is that there are probably many people with undiagnosed or unreported PCOS in both biobanks. The population prevalence of PCOS is predicted to be between 5% and 15% in European populations [30], but its reported prevalence in the UK Biobank is approximately 0.01% and in the MGB biobank approximately 4.7%. PCOS is challenging to properly diagnose, requiring multiple clinical and laboratory assessments including a pelvic ultrasound. It is possible that many of the "false positives" in the model could be women who have PCOS but are undiagnosed. Another possibility is that these polygenic risk scores are not accurate enough to effectively separate women with PCOS from those without.

When we added inflammatory biomarkers into the model for genetic prediction of PCOS, they did not further improve PCOS prediction in either biobank. It is possible that any effect of these biomarkers on obesity, testosterone, and SHBG is already captured by PRSs of obesity, testosterone, and SHBG.

Our study has several limitations. The aforementioned under-diagnosis of PCOS cases not only affects the ability to create PRS scores within the biobanks, but also the power of PCOS GWAS due to false negatives within the biobanks. Furthermore, MR analysis has limitations and any violations to the assumptions may bias the results—MR can only suggest potential causal relationships that must be verified in clinical studies. Importantly, this study was limited to individuals of European ancestry due to the small case counts in biobanks; PCOS has high prevalence globally and thus future studies including diverse populations are needed [31, 32]. Since PCOS may present itself differently between populations, it is especially important that biobanks increase diversity to improve PCOS research for all.

In summary, our study identifies four novel genetic loci for PCOS and demonstrates shared genetic architecture and potential causal relationships between PCOS and obesity, SHBG, testosterone, and inflammation. Together, these results support theories that immune responses are altered in PCOS patients and that chronic inflammation plays a role in dysregulating testosterone and SHBG levels.

4 Methods

4.1 Data Sources

4.1.1 PCOS Summary Statistics Data

We obtained PCOS summary statistics from the PCOS GWAS in the FinnGen and Estonia Biobanks [9] and a recent cross-population PCOS meta-analysis in European populations [10]. In the FinnGen and Estonia biobanks there were a total of 3,609 cases and 229,788 controls. All cases were self-reported and all other women were considered controls. In Tyrmi et al. [9], GWASs were conducted on the FinnGen and Estonia cohort before they were combined in an inverse-variance-weighted meta-analysis. Both GWASs used population-specific imputation panels: the Sequencing Initiative Suomi V3 [33] for FinnGen and Mitt et al. [34] for EstBB. Associations were run using the SAIGE generalized mixed model [35], and included age,

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

genotype batches, and PCs 1-10 as covariates. Due to including rarer variant alleles, Tyrmi et al. [9] includes 22.8 million SNPs.

In Day et al. [10] there are 4,137 PCOS cases and 20,129 controls pooled together in a fixed-effect, inverse-weighted-variance meta-analysis from 6 cohorts (Rotterdam, British Birth Cohort, Estonian Genome Center of the University of Tartu (EGCUT), deCODE genetics, Chicago, and Boston). In total there are 8.8 million SNPs. The Estonian cohort used in Day et al. [10] has 157 cases and 2807 controls which overlap with cases and controls used in Tyrmi et al. [9]. This causes a 2% overlap between the two studies. Since we did not observe any inflation in LD regression intercept and the overlap is small, we assume this will not create disproportionate effects and continue with the analysis.

4.1.2 Obesity Summary Statistics Data

We obtained BMI summary statistics data from Yengo et al. [36], a meta-analysis of GIANT and UKB for a total N of 681,275 men and women. WHR and WHRadjBMI summary statistics are from a GIANT and UKB meta-analysis, conducted by Pulit et al. [37]. We used female-specific summary statistics, resulting in N=263,148 for WHR and N=262,759 for WHRadjBMI.

4.1.3 Testosterone and SHBG GWAS

SHBG and testosterone serum level GWAS were conducted using the UK Biobank females who identified as "White British" and matched ancestry based on principal components. Serum levels for SHBG were available for 190,366 women and serum levels for total testosterone were available for 176,687 women. GWAS was conducted using the BOLT-LMM algorithm, adjusting for the first 20 PCs, age, age^2, menopausal status, pre-menopausal oral contraceptive use, and postmenopausal hormone therapy use [38]. We replicated all genetic correlation analysis with pre-menopausal testosterone and SHBG and found the same results, so kept the analysis using the joined pre and post-menopausal serum levels in order to increase power.

4.1.4 PCOS in the UK Biobank

Samples in the UK Biobank were used to train and test the polygenic risk score model for improved genetic prediction of PCOS. We used UK Biobank release version 3 with participants limited to females who self-identified as "White British" and matched ancestry based on principal components. PCOS cases were defined by self report, by which there are 159 cases.

Controls were filtered based on Rotterdam phenotypes to try to minimize the number of false negatives. Controls were selected from females that did not report ICD codes indicating excess androgen or irregular menstruation. ICD codes used to indicate excess androgen and irregular menstruation are the same as in Zhang et al. [39]. To enable a more balanced ratio for classification in the logistic regression, nineteen controls were randomly matched by age to each case to match the lower estimated population prevalence of 5%.

4.1.5 PCOS in the MGB Biobank

Samples, genomic data, and health information were obtained from the Mass General Brigham Biobank, a biorepository of consented patients samples at Mass General Brigham (parent organization of Massachusetts General Hospital and Brigham and Women's Hospital). These samples were also used to train and test the polygenic risk score model for improved genetic prediction of PCOS. Participants were limited to females who self-identify as white.

PCOS cases were defined by ICD self report, through which there are 374 cases. The control criteria in the MGB biobank was the same as in the UK biobank. Women who reported ICD codes indicating excess androgen or irregular menstruation were removed from the analysis to minimize false negatives. Using this criteria, there were a total of 7,553 controls.

4.1.6 Inflammatory Biomarker Data and Summary Statistics

To characterize inflammation on both cellular and molecular level and from multiple inflammatory pathways, we curated GWAS for a total of 138 inflammatory biomarkers.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Blood immune cell types: GWAS for 6 biomarkers are curated for counts of white blood cells, neutrophils, lymphocytes, monocytes, eosinophils, basophils. We conducted GWAS for these cell types (inverse normal transformed) in the UK Biobank using BOLT-LMM, adjusting for 20 PCs, age, age2, sex, and study center. The GWAS summary statistics were further combined with published summary statistics from the Biobank Japan (BBJ) [40] using an inverse-variance weighted meta-analysis using METAL. **Lymphocyte subtypes:** GWAS for 6 lymphocyte subsets, including CD4+ T cells, CD8+ T cells, CD56+ natural killer (NK) cells, CD3+ T cells, CD19+ B cells, and the derived measure CD4:CD8 ratio, are obtained from Ferreira et al. [41].

Molecular biomarkers of inflammation: GWAS for 126 biomarkers, including CRP and biomarkers indicative of diverse inflammatory pathways, were curated by meta-analysis of published and newly conducted GWAS. For CRP, GWAS were conducted in the UK Biobank (similar methods with those for immune cell subtypes); the NHS/HPFS (using Rvtests [42], after inverse-normal transformation, adjusting for the first 10 PCs, age, sex, cohorts and sub-studies); and GWAS summary statistics from UKB and NHS/HPFS were further meta-analyzed with published GWAS summary statistics from the Biobank Japan. For other biomarkers, we conducted GWAS for adiponectin, leptin, ICAM1, IL6, TNFaR1, TNFaR2, and IL18 in the NHS/HPFS; we acquired other biomarkers GWAS summary data from several publicly available sources including the Ahola-Olli et al. [43], Dastani et al. [44], and Kilpeläinen et al. [45], and GWAS of proteomics including Suhre et al. [46], Sun et al. [47], and Ferkingstad et al. [48]. For GWAS of proteomics measured using aptamer based SOMAscan platform, some markers in the proteomics dataset had different aptamers for the same protein target; we chose the GWAS for the aptamer with more genome-wide significant signals. We conducted meta-analysis for the same circulating protein biomarkers using the METAL [49] with the inverse-variance-weighted method. The detailed information on the meta-analysis of these inflammatory biomarkers is included in the supplement.

4.2 Analyses

4.2.1 PCOS Meta-Analysis and Genome-Wide Significant Loci

We combined Tyrmi et al. [9] summary statistics and Day et al. [10] in METAL [49] using the inverse-variance-weighted method. The summary statistics were in genome build GRCh37 and analyzed in PLINK [50] to clump loci using the setting $p_1 = 5e-8$, $p_2 = 1e-5$, clump-kb = 1000, and $r2 = 0.01$. We compared clumps that reached genome-wide significance in the meta-analysis to PCOS-associated SNPs from previous studies in order to identify novel loci. We consulted GWAS catalog to check previous studies for significant loci [51] and used locuszoom to visualize the GWAS [52].

4.2.2 Genetic Correlations using LDSC

We ran LDSC [53] to find the genetic correlations. We first found genetic correlations between PCOS, BMI, WHR, WHRadjBMI, SHBG, and testosterone, and later we calculated genetic correlations between these traits and each of the 138 inflammation markers. We used a SNP list from HapMap3 [54], computed LD scores in European ancestry from 1000 Genomes [55], and limited SNPs to those with MAF>0.01.

During the inflammation analysis, p-values were corrected via false discovery rate (FDR), and only correlations with FDR$<$ 0.05 (138 inflammation markers $\times$ 6 traits = 828 tests) were included in the analysis.

4.2.3 Mendelian Randomization

We conducted Mendelian randomization to test causal relationships with PCOS and related metabolic, hormonal, and inflammatory traits. We tested for causality in both directions between PCOS, each obesity trait, each hormonal trait, and each inflammation marker. Since androgen excess can be used as part of diagnosing PCOS, testosterone and PCOS break some of the MR assumptions, and thus their causality was not tested.

For trait-to-trait MR analyses, we implemented several MR models. First we used the Causal Analysis Using Summary Effect estimates (CAUSE) model [24]. CAUSE accounts for correlated and uncorrelated horizontal pleiotropic effects and thereby avoids more false positives. To find significant SNPs that are not

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

in LD, we used PLINK [50] with parameters $p1 = 5e-8$, $p2 = 5e-8$, clump-kb = 1000, and $r2 = 0.01$. We also compared the CAUSE estimates to estimates calculated via the mode-based estimate (MBE) [56], MR-EGGER [57], and inverse variance weighting (IVW) methods. We calculated each of these tests via the Mendelian randomization R package [58].

For trait-to-inflammation or inflammation-to-trait MR analyses, we only used the Mendelian randomization R package [58]. For every test we mainly used mode-based estimate (MBE) method, which allows relaxation of the instrumental variable assumptions and has less bias and lower type-I error rates than other methods [56]. We also looked for consistency using the IVW, and MR-EGGER methods. The CAUSE method was not implemented, since it takes a too long of a time to run each test. We only report as significant associations with FDR < 0.01 (138 tests) to minimize the number of false positive correlations.

4.2.4 Polygenic Risk Score Model to Improve PCOS Prediction

Our goal was to improve PCOS prediction by combining the PCOS polygenic risk score (PRS) with genetic scores of related risk factors. First we included the PRSs of PCOS, obesity measurements, SHBG, and testosterone, and we later added 138 inflammatory PRSs to further improve the prediction. We compared these models to a model that only considers the PCOS PRS.

To calculate the PRSs, we used sBayesR [59] to create a list of variants and effect sizes for each trait (PCOS, Testosterone, SHBG, BMI, WHR, WHRadjBMI). We used decreasing p-values starting at 0.5, using the highest possible p-value where sBayesR converged. For the inflammation biomarkers, we used PRS-CS with the 1000 Genomes Phase 3 reference panel [55] and the PRS-CS auto setting to create a list of variants and effect sizes [59]. Then we used PLINK [50] to apply these SNP effects to create polygenic risk scores for every female of European descent in the UK Biobank and MGB Biobank.

We created lasso logistic regression models to predict PCOS cases from multiple PRSs in both the MGB biobank and UKB biobank. The logistic regressions were trained via a nested cross validation (CV), with a 10-fold outer CV and 5-fold inner CV. For the outer loop, the whole dataset was randomly split into 10 equal groups. Each group was used once as a holdout set, with the remaining 9 groups used as the training set; a randomly-split inner 5-fold cross validation within only the training set was used to tune the regularization parameters; the resulting regression model was used to predict the out-of-sample 10th group. The prediction performance was evaluated using the area under the receiver operating characteristic (ROC) curve. In evaluation, PCOS cases were weighted higher than controls (weights were inversely proportional to class frequencies) to account for class imbalances. Scikit-Learn was used to implement all models [60].

5 Acknowledgments

L.K.P. gratefully acknowledges the support of the U.S. Department of Energy (DOE) through the Los Alamos National Laboratory (LANL) LDRD Program and the Center for Nonlinear Studies for this work. All authors thank Jocelyn Neri for help implementing Rotterdam Criteria in the MGB biobank. This research was conducted using the UK Biobank Resource under Application #45052. The authors would like to thank all participants and staff from the Mass General Brigham Biobank and UK Biobank. We further thank the Channing Network Health Division and all participants and staff in the Nurses Health Study.

6 Author Contributions

L.K.P., G.B., and L.L. designed the research study. L.K.P and G.B. lead all analyses. J.L., X.H., L.P. and G.B. curated the meta analyses of inflammatory biomarkers. J.H. conducted the GWASs for SHBG and testosterone in the UKB. Z.H. helped create polygenic risk scores. T.P. and S.M. provided guidance on clinical PCOS phenotypes and diagnosis, A.Y.M. provided guidance on inflammatory marker interpretation, and J.T. provided guidance on GWAS hit interpretation. L.K.P. and G.B. wrote the paper, with input from all authors. All authors provided valuable feedback and helped shape the analysis and manuscript.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

7 Funding Source

The Nurses' Health Study, Nurses' Health Study II, and Health Professionals Follow-up Study are supported by NIH grants UM1CA186107, R01CA49449, U01CA176726, R01CA67262, and U01CA167552.

References

[1] Gurkan Bozdag, Sezcan Mumusoglu, Dila Zengin, Erdem Karabulut, and Bulent Okan Yildiz. The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. *Human reproduction*, 31(12):2841–2855, 2016.

[2] Helena J Teede, Chau Thien Tay, Joop Laven, Anuja Dokras, Lisa J Moran, Terhi T Piltonen, Michael F Costello, Jacky Boivin, Leanne M Redman, Jacqueline A Boyle, Robert J Norman, Aya Mousa, Anju E Joham, and International PCOS Network. Recommendations from the 2023 International Evidence-based Guideline for the Assessment and Management of Polycystic Ovary Syndrome. *Human Reproduction*, 38(9):1655–1679, September 2023. ISSN 0268-1161. doi: 10.1093/humrep/dead156. URL https://doi.org/10.1093/humrep/dead156.

[3] Salla Karjula, Laure Morin-Papunen, Stephen Franks, Juha Auvinen, Marjo-Riitta JÃ€rvelin, Juha S. Tapanainen, Jari Jokelainen, Jouko Miettunen, and Terhi T. Piltonen. Population-based Data at Ages 31 and 46 Show Decreased HRQoL and Life Satisfaction in Women with PCOS Symptoms. *The Journal of Clinical Endocrinology and Metabolism*, 105(6):1814–1826, June 2020. ISSN 1945-7197. doi: 10.1210/clinem/dgz256.

[4] Linda Kujanpää, Riikka K. Arffman, Paula Pesonen, Elisa Korhonen, Salla Karjula, Marjo-Riitta Järvelin, Stephen Franks, Juha S. Tapanainen, Laure Morin-Papunen, and Terhi T. Piltonen. Women with polycystic ovary syndrome are burdened with multimorbidity and medication use independent of body mass index at late fertile age: A population-based cohort study. *Acta Obstetricia Et Gynecologica Scandinavica*, 101(7):728–736, July 2022. ISSN 1600-0412. doi: 10.1111/aogs.14382.

[5] Linda Kujanpää, Riikka K Arffman, Eeva Vaaramo, Henna-Riikka Rossi, Jaana Laitinen, Laure Morin-Papunen, Juha Tapanainen, Leena Ala-Mursula, and Terhi T Piltonen. Women with polycystic ovary syndrome have poorer work ability and higher disability retirement rate at midlife: a Northern Finland Birth Cohort 1966 study. *European Journal of Endocrinology*, 187(3):479–488, July 2022. ISSN 0804-4643. doi: 10.1530/EJE-22-0027. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9422246/.

[6] Emily W Gilbert, Chau T Tay, Danielle S Hiam, Helena J Teede, and Lisa J Moran. Comorbidities and complications of polycystic ovary syndrome: An overview of systematic reviews. *Clinical endocrinology*, 89(6):683–699, 2018.

[7] Elisabet Stener-Victorin, Vasantha Padmanabhan, Kirsty A Walters, Rebecca E Campbell, Anna Benrick, Paolo Giacobini, Daniel A Dumesic, and David H Abbott. Animal Models to Understand the Etiology and Pathophysiology of Polycystic Ovary Syndrome. *Endocrine Reviews*, 41(4):bnaa010, August 2020. ISSN 0163-769X. doi: 10.1210/endrev/bnaa010. URL https://doi.org/10.1210/endrev/bnaa010.

[8] JM Vink, S Sadrzadeh, CB Lambalk, and DI Boomsma. Heritability of polycystic ovary syndrome in a Dutch twin-family study. *The Journal of Clinical Endocrinology & Metabolism*, 91(6):2100–2104, 2006.

[9] Jaakko S Tyrmi, Riikka K Arffman, Natalia Pujol-Gualdo, Venla Kurra, Laure Morin-Papunen, Eeva Sliz, FinnGen Consortium, Estonian Biobank Research Team, Terhi T Piltonen, Triin Laisk, Johannes Kettunen, and Hannele Laivuori. Leveraging Northern European population history: novel low-frequency variants for polycystic ovary syndrome. *Human Reproduction*, 37(2):352–365, 11 2021. ISSN 0268-1161. doi: 10.1093/humrep/deab250. URL https://doi.org/10.1093/humrep/deab250.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

[10] Felix Day, Tugce Karaderi, Michelle R. Jones, Cindy Meun, Chunyan He, Alex Drong, Peter Kraft, Nan Lin, Hongyan Huang, et al. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLOS Genetics*, 14(12):e1007813, December 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007813. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007813. Publisher: Public Library of Science.

[11] Hyejin Lee, Jee-Young Oh, Yeon-Ah Sung, Hyewon Chung, Hyung-Lae Kim, Gwang Sub Kim, Yoon Shin Cho, and Jin Taek Kim. Genome-wide association study identified new susceptibility loci for polycystic ovary syndrome. *Human Reproduction*, 30(3):723–731, 2015.

[12] Anju E Joham, Robert J Norman, Elisabet Stener-Victorin, Richard S Legro, Stephen Franks, Lisa J Moran, Jacqueline Boyle, and Helena J Teede. Polycystic ovary syndrome. *The Lancet Diabetes & Endocrinology*, 10(9):668–680, September 2022. ISSN 2213-8587. doi: 10.1016/S2213-8587(22)00163-2. URL https://www.sciencedirect.com/science/article/pii/S2213858722001632.

[13] Jing-ling Zhu, Zhuo Chen, Wen-jie Feng, Shuang-lian Long, and Zhong-Cheng Mo. Sex hormone-binding globulin and polycystic ovary syndrome. *Clinica Chimica Acta*, 499:142–148, December 2019. ISSN 0009-8981. doi: 10.1016/j.cca.2019.09.010. URL https://www.sciencedirect.com/science/article/pii/S000989811932039X.

[14] Qianwen Liu, Zhaozhong Zhu, Peter Kraft, Qiaolin Deng, Elisabet Stener-Victorin, and Xia Jiang. Genomic correlation, shared loci, and causal relationship between obesity and polycystic ovary syndrome: a large-scale genome-wide cross-trait analysis. *BMC medicine*, 20(1):1–13, 2022.

[15] Tiantian Zhu and Mark O Goodarzi. Causes and Consequences of Polycystic Ovary Syndrome: Insights From Mendelian Randomization. *The Journal of Clinical Endocrinology & Metabolism*, 107(3):e899–e911, 10 2021. ISSN 0021-972X. doi: 10.1210/clinem/dgab757. URL https://doi.org/10.1210/clinem/dgab757.

[16] E Rudnicka, M Kunicki, K Suchta, P Machura, M Grymowicz, and R Smolarczyk. Inflammatory markers in women with polycystic ovary syndrome. *BioMed Research International*, 2020, 2020.

[17] N Phelan, A O'Connor, T Kyaw Tun, N Correia, G Boran, HM Roche, and J Gibney. Leucocytosis in women with polycystic ovary syndrome (pcos) is incompletely explained by obesity and insulin resistance. *Clinical endocrinology*, 78(1):107–113, 2013.

[18] Raziye Keskin Kurt, Ayşe Güler Okyay, Ali Ulvi Hakverdi, Arif Gungoren, Kenan Serdar Dolapcioglu, Atilla Karateke, and Mustafa Ozcil Dogan. The effect of obesity on inflammatory markers in patients with pcos: a bmi-matched case–control study. *Archives of gynecology and obstetrics*, 290(2):315–319, 2014.

[19] GTEx Consortium Lead analysts: Aguet François 1 Brown Andrew A. 2 3 4 Castel Stephane E. 5 6 Davis Joe R. 7 8 He Yuan 9 Jo Brian 10 Mohammadi Pejman 5 6 Park YoSon 11 Parsana Princy 12 Segrè Ayellet V. 1 Strober Benjamin J. 9 Zappala Zachary 7 8, NIH program management: Addington Anjene 15 Guan Ping 16 Koester Susan 15 Little A. Roger 17 Lockhart Nicole C. 18 Moore Helen M. 16 Rao Abhi 16 Struewing Jeffery P. 19 Volpi Simona 19, Pathology: Sobin Leslie 30 Barcus Mary E. 30 Branton Philip A. 16, NIH Common Fund Nierras Concepcion R. 137, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[20] John RB Perry, Yi-Hsiang Hsu, Daniel I Chasman, Andrew D Johnson, Cathy Elks, Eva Albrecht, Irene L Andrulis, Jonathan Beesley, Gerald S Berenson, Sven Bergmann, et al. Dna mismatch repair gene msh6 implicated in determining age at natural menopause. *Human molecular genetics*, 23(9): 2490–2497, 2014.

[21] Yu-Fang Pei, Yao-Zhong Liu, Xiao-Lin Yang, Hong Zhang, Gui-Juan Feng, Xin-Tong Wei, and Lei Zhang. The genetic architecture of appendicular lean mass characterized by association analysis in the uk biobank study. *Communications biology*, 3(1):608, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

[22] Eeva Sliz, Jaakko S Tyrmi, Nilufer Rahmioglu, Krina T Zondervan, Christian M Becker, Outi Uimari, and Johannes Kettunen. Evidence of a causal effect of genetic tendency to gain muscle mass on uterine leiomyomata. *Nature Communications*, 14(1):542, 2023.

[23] Nasa Sinnott-Armstrong, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, Michael Wainberg, Hanna M Ollila, Tuomo Kiiskinen, et al. Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature genetics*, 53(2):185–194, 2021.

[24] Jean Morrison, Nicholas Knoblauch, Joseph H. Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747, July 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-0631-4. URL https://www.nature.com/articles/s41588-020-0631-4. Number: 7 Publisher: Nature Publishing Group.

[25] Jerilee MK Azhary, Miyuki Harada, Nozomi Takahashi, Emi Nose, Chisato Kunitomi, Hiroshi Koike, Tetsuya Hirata, Yasushi Hirota, Kaori Koga, Osamu Wada-Hiraike, et al. Endoplasmic reticulum stress activated by androgen enhances apoptosis of granulosa cells via induction of death receptor 5 in pcos. *Endocrinology*, 160(1):119–132, 2019.

[26] M A Brower, Y Hai, M R Jones, X Guo, Y D I Chen, J I Rotter, R M Krauss, R S Legro, R Azziz, and M O Goodarzi. Bidirectional Mendelian randomization to explore the causal relationships between body mass index and polycystic ovary syndrome. *Human Reproduction*, 34(1):127–136, January 2019. ISSN 0268-1161. doi: 10.1093/humrep/dey343. URL https://doi.org/10.1093/humrep/dey343.

[27] Lori A Cooper, Stephanie T Page, John K Amory, Bradley D Anawalt, and Alvin M Matsumoto. The association of obesity with sex hormone-binding globulin is stronger than the association with ageing–implications for the interpretation of total testosterone measurements. *Clinical endocrinology*, 83(6): 828–833, 2015.

[28] Cristina Saez-Lopez, Anna Barbosa-Desongles, Cristina Hernandez, Roger A. Dyer, Sheila M. Innis, et al. Sex Hormone-Binding Globulin Reduction in Metabolic Disorders May Play a Role in NAFLD Development. *Endocrinology*, 158(3):545–559, March 2017. ISSN 0013-7227. doi: 10.1210/en.2016-1668. URL https://doi.org/10.1210/en.2016-1668.

[29] Yoonjung Yoonie Joo, KyEra Actkins, Jennifer A Pacheco, Anna O Basile, Robert Carroll, David R Crosslin, Felix Day, Joshua C Denny, Digna R Velez Edwards, Hakon Hakonarson, John B Harley, Scott J Hebbring, Kevin Ho, Gail P Jarvik, Michelle Jones, Tugce Karaderi, Frank D Mentch, Cindy Meun, Bahram Namjou, Sarah Pendergrass, Marylyn D Ritchie, Ian B Stanaway, Margrit Urbanek, Theresa L Walunas, Maureen Smith, Rex L Chisholm, Abel N Kho, Lea Davis, M Geoffrey Hayes, and International PCOS Consortium. A Polygenic and Phenotypic Risk Prediction for Polycystic Ovary Syndrome Evaluated by Phenome-Wide Association Studies. *The Journal of Clinical Endocrinology & Metabolism*, 105(6):1918–1936, June 2020. ISSN 0021-972X. doi: 10.1210/clinem/dgz326. URL https://doi.org/10.1210/clinem/dgz326.

[30] Daria Lizneva, Larisa Suturina, Walidah Walker, Soumia Brakta, Larisa Gavrilova-Jordan, and Ricardo Azziz. Criteria, prevalence, and phenotypes of polycystic ovary syndrome. *Fertility and Sterility*, 106(1):6–15, July 2016. ISSN 0015-0282. doi: 10.1016/j.fertnstert.2016.05.003. URL https://www.sciencedirect.com/science/article/pii/S0015028216612323.

[31] Jennifer K. Hillman, Lauren N. C. Johnson, Meghana Limaye, Rebecca A. Feldman, Mary Sammel, and Anuja Dokras. Black women with polycystic ovary syndrome (PCOS) have increased risk for metabolic syndrome and cardiovascular disease compared with white women with PCOS. *Fertility and Sterility*, 101(2):530–535, February 2014. ISSN 0015-0282. doi: 10.1016/j.fertnstert.2013.10.055. URL https://www.sciencedirect.com/science/article/pii/S0015028213032536.

[32] Yue Zhao and Jie Qiao. Ethnic differences in the phenotypic expression of polycystic ovary syndrome. *Steroids*, 78(8):755–760, August 2013. ISSN 0039-128X. doi: 10.1016/j.steroids.2013.04.006. URL https://www.sciencedirect.com/science/article/pii/S0039128X13000883.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

[33] Mart Kals, Tiit Nikopensius, Kristi Läll, Kalle Pärn, Timo Tõnis Sikka, Jaana Suvisaari, Veikko Salomaa, Samuli Ripatti, Aarno Palotie, Andres Metspalu, Tõnu Esko, Priit Palta, and Reedik Mägi. Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses. *bioRxiv*, 2019. doi: 10.1101/579201. URL https://www.biorxiv.org/content/early/2019/04/04/579201.

[34] Mario Mitt, Mart Kals, Kalle Pärn, Stacey B. Gabriel, Eric S. Lander, Aarno Palotie, Samuli Ripatti, Andrew P. Morris, Andres Metspalu, Tõnu Esko, Reedik Mägi, and Priit Palta. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, 25(7):869–876, July 2017. ISSN 1476-5438. doi: 10.1038/ejhg.2017.51. URL https://www.nature.com/articles/ejhg201751. Number: 7 Publisher: Nature Publishing Group.

[35] Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, September 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0184-y. URL https://www.nature.com/articles/s41588-018-0184-y. Number: 9 Publisher: Nature Publishing Group.

[36] Loic Yengo, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, Joel Hirschhorn, Jian Yang, and Peter M Visscher. Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. *Human Molecular Genetics*, 27(20):3641–3649, October 2018. ISSN 0964-6906. doi: 10.1093/hmg/ddy271. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6488973/.

[37] Sara L Pulit, Charli Stoneman, Andrew P Morris, Andrew R Wood, Craig A Glastonbury, Jessica Tyrrell, LoÃ¯c Yengo, Teresa Ferreira, Eirini Marouli, Yingjie Ji, Jian Yang, Samuel Jones, Robin Beaumont, Damien C Croteau-Chonka, Thomas W Winkler, GIANT Consortium, Andrew T Hattersley, Ruth J F Loos, Joel N Hirschhorn, Peter M Visscher, Timothy M Frayling, Hanieh Yaghootkar, and Cecilia M Lindgren. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Human Molecular Genetics*, 28(1):166–174, January 2019. ISSN 0964-6906. doi: 10.1093/hmg/ddy327. URL https://doi.org/10.1093/hmg/ddy327.

[38] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3): 284–290, 2015.

[39] Yanfei Zhang, Kevin Ho, Jacob M. Keaton, Dustin N. Hartzel, Felix Day, Anne E. Justice, Navya S. Josyula, Sarah A. Pendergrass, Ky'Era Actkins, Lea K. Davis, Digna R. Velez Edwards, Brody Holohan, Andrea Ramirez, Ian B. Stanaway, David R. Crosslin, Gail P. Jarvik, Patrick Sleiman, Hakon Hakonarson, Marc S. Williams, and Ming Ta Michael Lee. A genome-wide association study of polycystic ovary syndrome identified from electronic health records. *American Journal of Obstetrics and Gynecology*, 223(4):559.e1–559.e21, October 2020. ISSN 0002-9378. doi: 10.1016/j.ajog.2020.04.004. URL https://www.sciencedirect.com/science/article/pii/S0002937820304282.

[40] Masahiro Kanai, Masato Akiyama, Atsushi Takahashi, Nana Matoba, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature genetics*, 50 (3):390–400, 2018.

[41] Manuel AR Ferreira, Massimo Mangino, Chanson J Brumme, Zhen Zhen Zhao, Sarah E Medland, Margaret J Wright, Dale R Nyholt, Scott Gordon, Megan Campbell, Brian P McEvoy, et al. Quantitative trait loci for cd4: Cd8 lymphocyte ratio are associated with risk of type 1 diabetes and hiv-1 immune control. *The American Journal of Human Genetics*, 86(1):88–92, 2010.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

[42] Xiaowei Zhan, Youna Hu, Bingshan Li, Goncalo R Abecasis, and Dajiang J Liu. Rvtests: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*, 32 (9):1423–1426, 2016.

[43] Ari V Ahola-Olli, Peter Würtz, Aki S Havulinna, Kristiina Aalto, Niina Pitkänen, Terho Lehtimäki, Mika Kähönen, Leo-Pekka Lyytikäinen, Emma Raitoharju, Ilkka Seppälä, et al. Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. *The American Journal of Human Genetics*, 100(1):40–50, 2017.

[44] Zari Dastani, Marie-France Hivert, Nicholas Timpson, John RB Perry, Xin Yuan, Robert A Scott, Peter Henneman, Iris M Heid, Jorge R Kizer, Leo-Pekka Lyytikäinen, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS genetics*, 8(3):e1002607, 2012.

[45] Tuomas O Kilpeläinen, Jayne F Martin Carli, Alicja A Skowronski, Qi Sun, Jennifer Kriebel, Mary F Feitosa, Åsa K Hedman, Alexander W Drong, James E Hayes, Jinghua Zhao, et al. Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nature communications*, 7(1):10494, 2016.

[46] Karsten Suhre, Matthias Arnold, Aditya Mukund Bhagwat, Richard J Cotton, Rudolf Engelke, Johannes Raffler, Hina Sarwath, Gaurav Thareja, Annika Wahl, Robert Kirk DeLisle, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature communications*, 8(1): 14357, 2017.

[47] Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, et al. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

[48] Egil Ferkingstad, Patrick Sulem, Bjarni A. Atlason, Gardar Sveinbjornsson, Magnus I. Magnusson, et al. Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics*, 53 (12):1712–1721, December 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00978-w. URL https://www.nature.com/articles/s41588-021-00978-w. Number: 12 Publisher: Nature Publishing Group.

[49] Cristen J. Willer, Yun Li, and GonÃ§alo R. Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, September 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq340. URL https://doi.org/10.1093/bioinformatics/btq340.

[50] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007. ISSN 0002-9297, 1537-6605. doi: 10.1086/519795. URL https://www.cell.com/ajhg/abstract/S0002-9297(07)61352-4. Publisher: Elsevier.

[51] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[52] Andrew P Boughton, Ryan P Welch, Matthew Flickinger, Peter VandeHaar, Daniel Taliun, GonÃ§alo R Abecasis, and Michael Boehnke. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics*, 37(18):3017–3018, March 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab186. URL https://doi.org/10.1093/bioinformatics/btab186. _eprint: https://academic.oup.com/bioinformatics/article-pdf/37/18/3017/40471331/btab186.pdf.

[53] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, March 2015. ISSN 1546-1718. doi: 10.1038/ng.3211. URL https://www.nature.com/articles/ng.3211. Number: 3 Publisher: Nature Publishing Group.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.17.23297115; this version posted February 23, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[54] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, September 2010. ISSN 1476-4687. doi: 10.1038/nature09298. URL https://www.nature.com/articles/nature09298. Number: 7311 Publisher: Nature Publishing Group.

[55] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, October 2015. ISSN 1476-4687. doi: 10.1038/nature15393. URL https://www.nature.com/articles/nature15393. Number: 7571 Publisher: Nature Publishing Group.

[56] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998, December 2017. ISSN 0300-5771. doi: 10.1093/ije/dyx102. URL https://doi.org/10.1093/ije/dyx102.

[57] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, April 2015. ISSN 0300-5771. doi: 10.1093/ije/dyv080. URL https://doi.org/10.1093/ije/dyv080.

[58] Olena O Yavorska and Stephen Burgess. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology*, 46(6):1734–1739, December 2017. ISSN 0300-5771. doi: 10.1093/ije/dyx034. URL https://doi.org/10.1093/ije/dyx034.

[59] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications*, 10(1):1–10, 2019.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.