Understanding and Predicting Polycystic Ovary Syndrome through Shared Genetics with Testosterone, SHBG, and Chronic Inflammation

Lillian Kay Petersen[1,2,*], Garyk Brixi[1,*], Jun Li[4,5], Jie Hu[1,6], 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]Center for Genomic Medicine and Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital and Harvard Medical School
[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

October 18, 2023

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 increased 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 chronic inflammation. First we create a large meta-analysis of PCOS (7,747 PCOS cases and 498,227 controls) and identify four novel genetic loci associated with PCOS. These novel loci have been previously associated with changes in gene expression in multiple tissues including the thyroid and ovary. We then analyze the PCOS meta analysis alongside GWASs for obesity (n=681,275), SHBG (n=190,366), testosterone (n=176,687), and meta-analyses of 130 inflammatory biomarkers (average n=30,000). We replicate potential causal relationships (via Mendelian randomization) from obesity and SHBG to PCOS and find extensive genetic correlations and causality between these traits and inflammatory biomarkers. We identify significant genetic correlations between PCOS and eight inflammatory biomarkers, including new relationships such as a strong correlation to death receptor 5 (LDSC rg = 0.54, FDR = 0.043). Although these results point to a shared genetic architecture between PCOS and inflammation, we did not find statistically significant causality between them; however, we did find 27 inflammatory biomarkers with significant causal effects on SHBG and four biomarkers with significant causal effects on testosterone. These results support the hypothesis that chronic inflammation can influence androgen and SHBG levels, though more research is needed to verify these results. Finally, we show that combining the polygenic risk scores of PCOS-related traits improves genetic prediction of PCOS cases in the UK Biobank and MGB Biobank compared to using only PCOS risk scores (0.72 AUC from 0.59 AUC and 0.61 AUC from 0.59 AUC in each biobank respectively).

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]. PCOS is a condition with high a high risk of comorbidities, and women with PCOS report lowered work ability and quality of life [3, 4, 5]. The most comorbidities of PCOS are metabolic, reproductive and psychological disorders, and include diseases such as obesity, diabetes, dyslipidemia, metabolic syndrome, obstructive sleep apnea, cardiovascular disease, subfertility, endometrial cancer, depression, anxiety, and eating disorders [6]. Recent clinical studies suggest distinct clusters of reproductive and metabolic PCOS, which has also been supported by genetic evidence [7].

The pathogenesis of PCOS is complicated, including both genetic and environmental factors, and the exact mechanisms remain unclear [8]. Hyperandrogenism and PCOS pathogenesis have been linked to factors such as weight gain, obesity, and insulin resistance; chronic low-grade inflammation [9]; and low sex hormone binding globulin (SHBG) levels [10]

PCOS is polygenic and is estimated to be highly heritable—twin studies have suggested that PCOS has heritability of over 70% [11]. 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 [12, 13, 14].

Previous studies have found significant genetic correlations between PCOS, BMI, and waist-to-hip ratio (WHR), suggesting a shared genetic architecture [15]. Previous Mendelian randomization studies of PCOS suggest a causal role of BMI, type 2 diabetes, SHBG, and other traits in the development of PCOS [16].

While elevated markers of chronic inflammation such as C-reactive protein (CRP), IL-6, and white blood cell count are associated with obesity, testosterone, and insulin resistance, PCOS patients often have elevated inflammation markers compared with age- and BMI-matched controls [17, 18, 19]. This supports the idea that inflammation and altered immune response is related to PCOS independently of related comorbidities [8]

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.

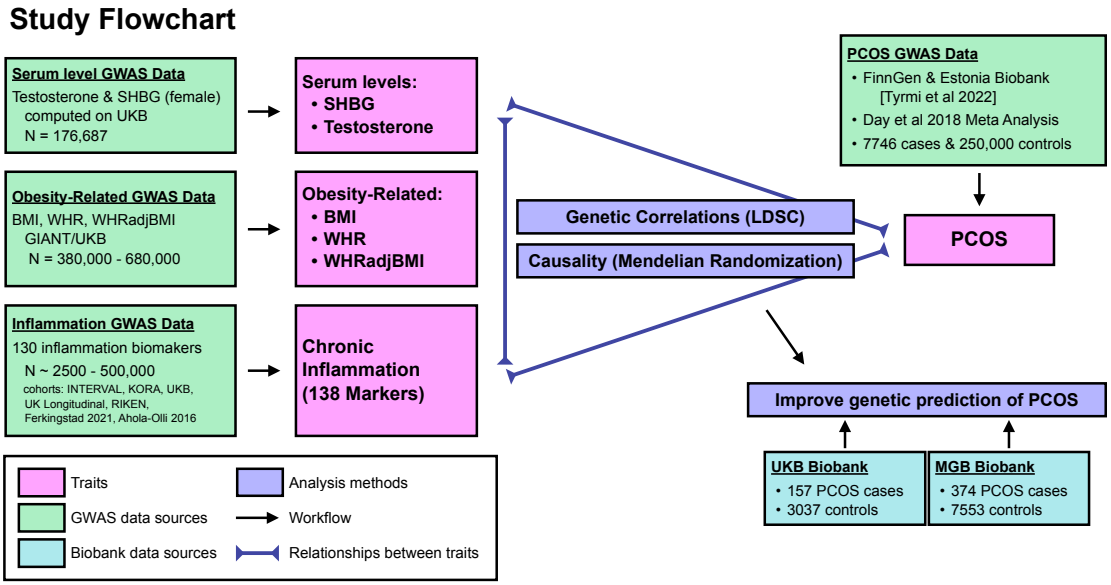

Figure 1: Flowchart overview of this study.

2 Results

2.1 PCOS Meta-analysis

We first combined Tyrmi et al. [12] and Day et al. [13] 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 these loci have not been previously reported to be associated with PCOS in the literature. Details about these loci and their most significant variants are shown in table 1, and the full meta-analysis results are shown in figure 2.

Replicated PCOS Loci

CHR:BP Ref/Effect	RSID	Beta	Effect Allele Freq	p-value	Nearest Gene
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

CHR:BP Ref/Effect	RSID	Beta	Effect Allele Freq	p-value	Nearest Gene
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: Top: loci that reach genome wide significance in the meta-analysis and are nearby loci previously reported for PCOS. Bottom: loci that reach genome wide significance in the meta-analysis and are not in LD with loci previously reported for PCOS.

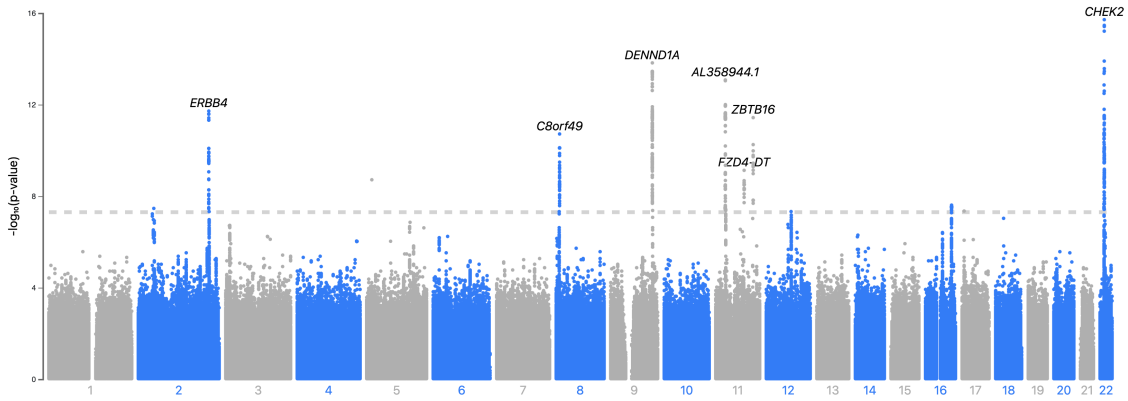

Figure 2: Manhattan plot of PCOS meta-analysis [20].

All of the identified novel loci for PCOS were significant expression quantitative trait (eQTL) loci in multiple tissues [21]. Rs61030588 on chromosome 2 is within the gene MSH6 andd is a known eQTL in 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 [22]. The novel PCOS locus rs11234902 on chromosome 11 is a significant eQTL of RP11-736K20.6, an RNA gene, in thyroid tissue. 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 [23, 24, 25]. 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 are included in the supplementary file 1.

2.2 Genetic Correlations

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

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

The genetic correlations between these PCOS-related traits are shown in table 3. PCOS has strong genetic correlation with SHBG (0.45), BMI (0.35), and WHR (0.35), and has a weaker correlations with WHRadjBMI (0.17) and testosterone (0.16). SHBG is correlated to all of the obesity-related traits with the strongest relationship with WHR. We did not find female levels of testosterone and SHBG to be significantly correlated.

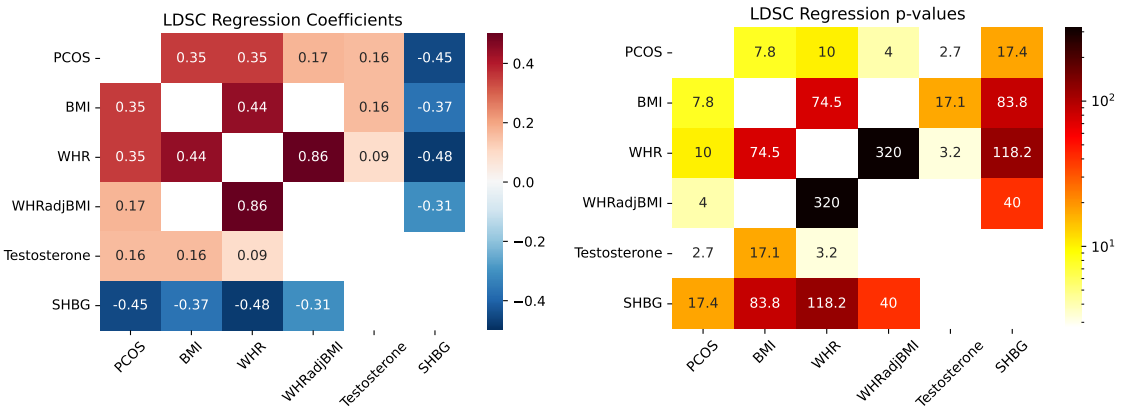

Figure 3: Genetic correlations and p-values (-log10), computed via LDSC, between the non-inflammatory traits in our study. Values that did not meet FDR < 0.05 are masked.

2.2.2 Genetic correlations between inflammatory biomarkers, PCOS, obesity, testosterone, and SHBG

We created meta-analyses of 130 inflammatory biomarkers, including immune cell counts and biomarker serum levels. Each inflammatory biomarker included between 2,538 and 505,690 individuals in the analysis, with LDSC-calculated heritability ranging 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, IL1ra, HGF, lymphocyte count (LymC), neutrophil count (NeuC), monocyte count (MonC), and white blood cell count (WBC). Most of the inflammatory biomarkers that correlated with PCOS have a similar magnitude of correlation with BMI, WHR, or negative SHBG. The exception was TRAILr2, which has a correlation to PCOS of 0.54, compared to the correlation of 0.3 with BMI, SHBG, or WHR.

The inflammatory markers that are significantly correlated with low SHBG are often also associated with BMI and WHR. SHBG is significantly correlated with 28 inflammation markers, all of which were also significantly correlated to an obesity trait, although in the opposite direction.

Testosterone had the fewest significant correlations with inflammation, with a significant correlation only with monocyte levels.

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.

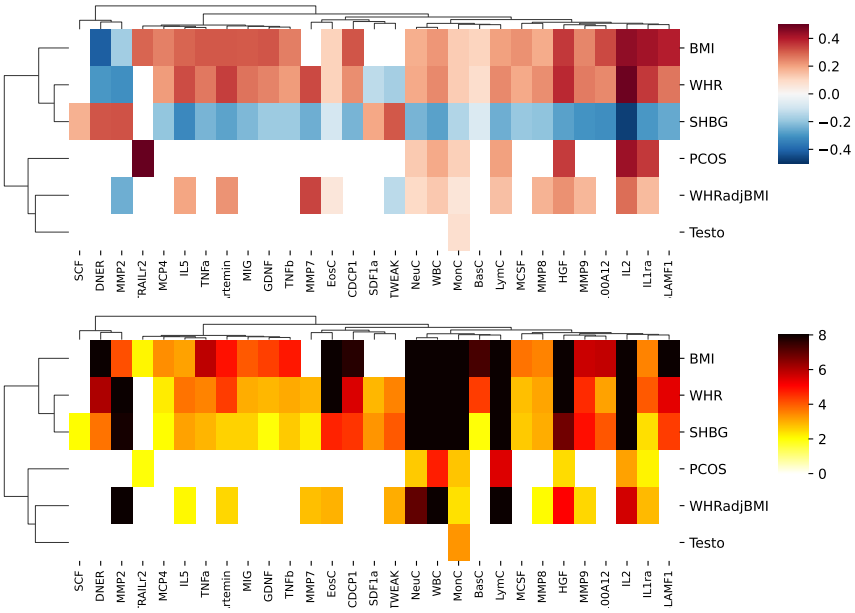

Figure 4: Genetic correlations (top) and -log10 p-value (bottom) computed via LDSC between the traits PCOS, BMI, WHR, WHRadjBMI, SHBG, testosterone and the inflammation markers. Correlations and p-values shown are FDR < 0.05. Rows and columns are clustered by genetic correlation coefficients using hierarchical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

2.3 Mendelian Randomization

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

To investigate the potential causal relationships, we conducted bi-directional Mendelian randomization (MR) between PCOS, obesity traits, testosterone, and SHBG. We replicated MR results using several different methods, including CAUSE [26], 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 BMI causes PCOS ($\text{effect}_{\text{CAUSE}} = 0.58$, $p_{\text{CAUSE}} = 3.7 \times 10^{-4}$). Most methods suggest that WHR has a causal effect on PCOS, and a few methods suggest that WHRadjBMI is causal. MBE and IVW suggest that SHBG has a causal effect on PCOS with effect size about -0.2 ($p_{\text{MBE}} = 0.02$, $p_{\text{IVW}} < 0.001$), but this result is borderline not significant according to CAUSE (p=0.057) and MR-EGGER (p=0.061).

PCOS was not found to have statistically significant causality for any of the other tested traits (figure 5) which is consistent with previous MR studies [16].

All MR methods suggest that BMI, WHR, and WHRadjBMI decrease SHBG and that WHRadjBMI decreases testosterone (see supplementary file 4).

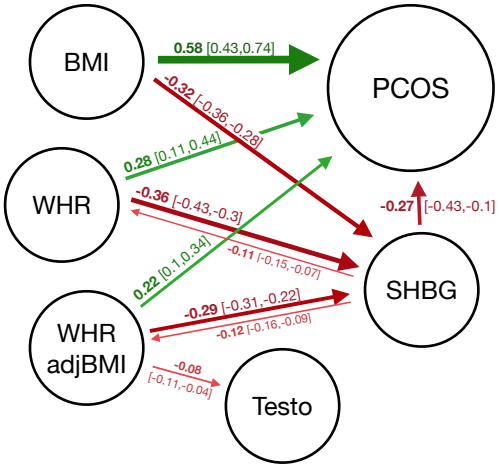

Figure 5: Mendelian randomization estimated causal effects and 95% confidence intervals between BMI, WHR, WHRadjBMI, SHBG, Testosterone, and PCOS. Green arrows indicate positive causality and red arrows indicate negative causality. The thickness and shade of the arrows is proportional to effect size. All estimates are based on the CAUSE model. All values are significant ($p<0.05$) except SHBG $\rightarrow$ PCOS ($p=0.057$), which is included because most MR methods found this relationship to be significant.

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

Next we conducted bi-directional MR between PCOS, testosterone, SHBG and the panel of 130 inflammatory biomarkers. We further conducted bi-directional MR between obesity traits and the inflammatory markers that are significantly related to one of the three primary traits to see how obesity may be involved in the relationship. Our primary results were estimated based on the MBE MR method, since 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 the PCOS and inflammatory markers using MR, likely due to the low heritability of the PCOS GWAS.

Four biomarkers showed significant causal relations with testosterone ($FDR_{MBE} =< 0.01$): genetically-predicted TWEAK ($estimate_{MBE} = 0.057$, $p_{MBE} = 3.7 * 10^{-8}$) and MMP9 ($estimate_{MBE} = 0.043$, $p_{MBE} = 2.9 * 10^{-4}$) were found to increase testosterone, while genetically-predicted IL2Rb ($estimate_{MBE} = -0.057$, $p_{MBE} = 7.44 * 10^{-5}$) and IP10 ($estimate_{MBE} = -0.126$, $p_{MBE} = 1.33 * 10^{-5}$) were found to decrease testosterone.

Twenty-seven inflammatory biomarkers were found to be significantly causal for SHBG. Of these 27 biomakers, only CD36antg was 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 ($estimate_{MBE} = 0.46$, $p_{MBE} = 2.46 * 10^{-219}$, $estimate_{IVW} = 0.260$, $p_{IVW} = 5.74 * 10^{-3}$, $estimate_{MR\text{-}Egger} = 0.351$, $p_{MR\text{-}Egger} = 1.46 * 10^{-2}$). Interestingly, TWEAK is also the only biomarker whose genetically-predicted level was found to be causal for both testosterone and SHBG.

We next assessed whether inflammation could be partially mediating the causal effect from obesity to sex hormones. Among the 27 significantly causal biomarkers for SHBG, three were found to be caused by BMI, as visualized in figure 6a. MR suggests that these markers are elevated by BMI, and in turn increase SHBG levels.

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 27 biomarkers causal for SHBG, 14 are also significant for WHRadjBMI, 14 significant for WHR, and 1 (FASLG) for BMI. Details of the MR results are available in the supplementary file 5.

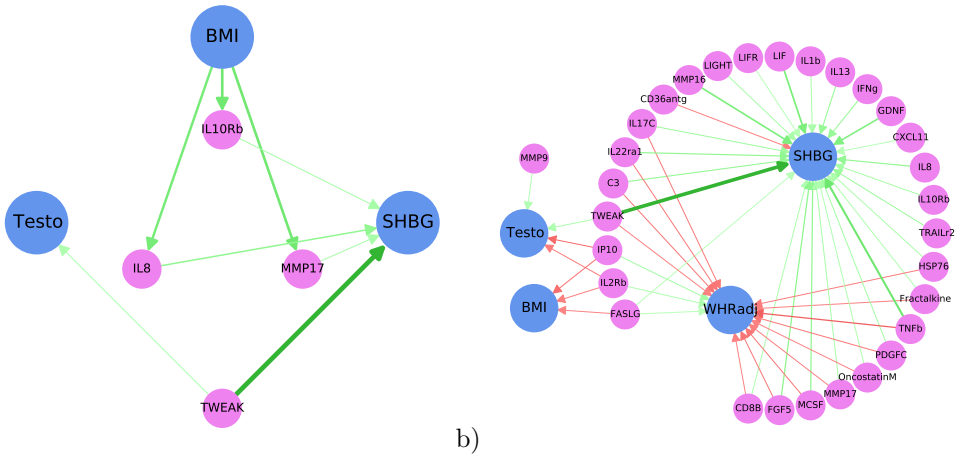

Figure 6: Mendelian randomization results between inflammatory biomarkers and PCOS-related traits. (a) Inflammatory biomarkers may mediate some of the causal relationship between obesity and hormones. (b) Inflammatory biomarkers which significantly regulate SHBG or Testosterone, and the causal relationship of these markers on BMI and WHR adjusted by BMI. Pink circles represent inflammation 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.

2.4 Combined polygenic risk scores 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 first calculated the PRSs of PCOS, BMI, WHR, WHRadjBMI, SHBG, and testosterone in women in the UK Biobank and Mass General Brigham (MGB) biobank. We then created lasso logistic regressions to predict PCOS cases based on all of these PRSs, and compared the prediction to that based just on the PCOS PRS. All PRSs were standardized to mean 0 and variance 1 before input into the model. 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.

We further explored whether including inflammation PRSs would improve PCOS predictions, but did not find any improvement in AUROC in either biobank.

3 Discussion

This study investigated genetic correlations and causal relationships between PCOS and obesity, testosterone, SHBG, and chronic inflammatory biomarkers. The large GWAS meta-analysis of PCOS identified four novel loci. We replicated and identified novel genetic correlations and potential causal relationships between PCOS, obesity, testosterone, SHBG, and inflammatory markers. Finally, genetic relationships between these traits were used to improve genetic risk assessment of PCOS.

Interestingly, we found that PCOS has a smaller genetic correlation with testosterone than it does with

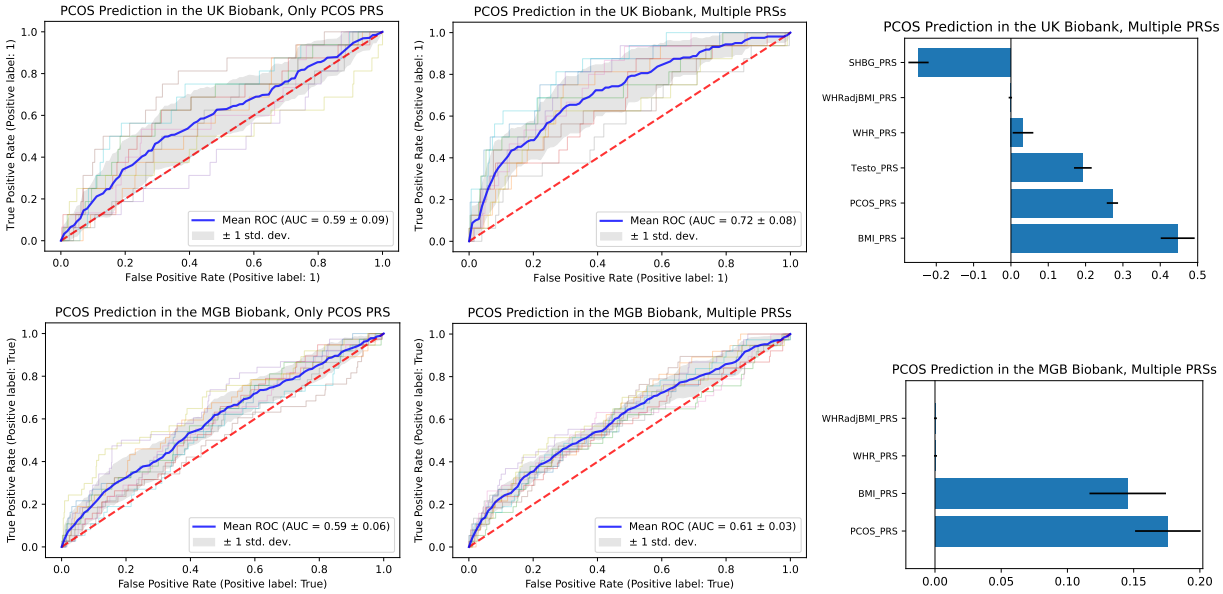

Figure 7: Genetic prediction of PCOS cases in the UK Biobank (top) and MGB biobank (bottom). A nested 10-fold cross validation lasso logistic model was trained on each biobank. The left plots show the receiver operating characteristic (ROC) curve for predicting PCOS cases when only using the PCOS polygenic 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.

SHBG, BMI, WHR, and WHRadjBMI, despite the common understanding of PCOS as a condition characterized by hyperandrogenism.

This study provides evidence that PCOS shares genetic architecture with a range of inflammatory biomarkers. We found significant genetic correlations between PCOS and eight 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). It is surprising, though, that we found TRAILr2 to both be positively correlated with PCOS and to increase SHBG, since high SHBG is thought to decrease PCOS risk [27]. A previous study has found that TRAILr2 mediates testosterone-driven apoptosis of PCOS granulosa cells in culture [28], 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 likely causes PCOS, while SHBG may protect against PCOS, consistent with previous MR findings [29].

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 [30]. 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 [31].

We also found evidence that SHBG, testosterone, and adipose tissue may be regulated by a broad array of inflammatory biomarkers. Among the strongest of these regulators is TWEAK, which we found to increase SHBG, increase testosterone, and decrease WHRadjBMI. These findings support the existing hypothesis that chronic inflammation can dysregulate hormonal production in the ovaries, but also implicate many more inflammatory biomarkers to 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 vast majority of regulating inflammatory biomarkers are found to be protective: most increase SHBG and decrease WHRadjBMI and BMI. Many of the directions of causality

don't agree with the direction of correlation found using LDSC. Further study is needed to explain the regulatory feedback mechanisms which underlie these results and confirm these findings.

While the results of MR suggest causal links, there could also be mediators such as insulin resistance, diabetes, CVD, metabolic syndrome, and other conditions which are causal for testosterone, SHBG levels, and 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 [32], 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 [33], 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. This could be due to under-diagnosis and noisy biobank data, or may suggest that PCOS-associated inflammation is already captured by the combined PRSs of PCOS, obesity, testosterone, and SHBG.

There are many limitations to this study that must be discussed. The aforementioned under-diagnosis of PCOS cases not only affects the ability to create PRS scores within the biobanks, but also can affects PCOS GWAS due to false negatives within the biobanks. Furthermore, Mendelian randomization can only suggest causal relationships which must be verified in clinical follow-up studies. Importantly, this study was limited to individuals of European ancestry, but PCOS has high prevalence around the world and requires more diverse studies [34, 35]. Since PCOS may present itself differently between populations, it is especially important that biobanks increase diversity to improve PCOS research for all.

Despite these limitations, this study represents a step forward in our understanding of PCOS and inflammation. It identifies novel variants, suggests inflammatory pathways may have causal relationships with hyperandrogenism, and demonstrates improved genetic prediction of PCOS in biobanks.

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 [12] and a recent cross-population PCOS meta-analysis in European populations [13]. 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. [12], 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 [36] for FinnGen and Mitt et al. [37] for EstBB. Associations were run using the SAIGE generalized mixed model [38], and included age, genotype batches, and PCs 1-10 as covariates. Due to including rarer variant alleles, Tyrmi et al. [12] includes 22.8 million SNPs.

In Day et al. [13] 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. [13] has 157 cases and 2807 controls which overlap with cases and controls used in Tyrmi et al. [12]. 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. [39], 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. [40]. 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 [41]. 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. [42]. 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 Inflammation 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 130 inflammatory biomarkers.

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) [43] 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. [44].

Molecular biomarkers of inflammation: GWAS for biomarkers indicative of diverse inflammatory pathways were curated by meta-analysis of published GWAS. We acquired summary data from several publicly available sources including the Ahola-Olli et al. [45], Dastani et al. [46], and Kilpeläinen et al. [47], and proteomics GWAS including Suhre et al. [48], Sun et al. [49], and Ferkingstad et al. [50]. 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 [51] with the inverse-variance-weighted method. The full list of inflammation biomarkers and their data sources is included in the supplementary file 2.

4.2 Analyses

4.2.1 PCOS Meta-Analysis and Genome-Wide Significant Loci

We combined Tyrmi et al. [12] summary statistics and Day et al. [13] in METAL [51] using the inverse-variance-weighted method. The summary statistics were in genome build GRCh37 and analyzed in PLINK [52] to clump loci using the setting $p_1 = 5 \times 10^{-8}$, $p_2 = 1 \times 10^{-5}$, 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 [53] and used locuszoom to visualize the GWAS [20].

4.2.2 Genetic Correlations using LDSC

We ran LDSC [54] 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 130 inflammation markers. We used a SNP list from HapMap3 [55], computed LD scores in European ancestry from 1000 Genomes [56], 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 (130 inflammation markers $\times$ 6 traits = 780 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 [26]. CAUSE accounts for correlated and uncorrelated horizontal pleiotropic effects and thereby avoids more false positives. To find significant SNPs that are not in LD, we used PLINK [52] with parameters $p1 = 5 \times 10^{-8}$, $p2 = 5 \times 10^{-8}$, clump-kb = 1000, and $r2 = 0.01$. We also compared the CAUSE estimates to estimates calculated via the mode-based estimate (MBE) [57], MR-EGGER [58], and inverse variance weighting (IVW) methods. We calculated each of these tests via the Mendelian randomization R package [59].

For trait-to-inflammation or inflammation-to-trait MR analyses, we only used the Mendelian randomization R package [59]. 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 [57]. 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 (130 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 130 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 [60] 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 inflammatory biomarkers, we used PRS-CS with the 1000 Genomes Phase 3 reference panel [56] and the PRS-CS auto setting to create a list of variants and effect sizes [60]. Then we used PLINK [52] 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 [61].

5 Acknowledgments

LKP 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. We would like to thank the participants and researchers from the UK Biobank and Mass General Brigham Biobank who contributed or collected data.

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, Marie L Misso, Michael F Costello, Anuja Dokras, Joop Laven, Lisa Moran, Terhi Piltonen, and Robert J Norman. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Human reproduction*, 33(9):1602–1618, 2018.

[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] Kujanpaa Linda, Riikka K. Arffman, Paula Pesonen, Elisa Korhonen, Salla Karjula, Marjo-Riitta Javelin, 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] Kujanpaa Linda, 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, September 2022. ISSN 1479-683X. doi: 10.1530/EJE-22-0027.

[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] Matthew Dapas and Andrea Dunaif. Deconstructing a Syndrome: Genomic Insights Into PCOS Causal Mechanisms and Classification. *Endocrine Reviews*, 43(6):927–965, December 2022. ISSN 0163-769X. doi: 10.1210/endrev/bnac001. URL https://doi.org/10.1210/endrev/bnac001.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Xianqin Qu and Richard Donnelly. Sex Hormone-Binding Globulin (SHBG) as an Early Biomarker and Therapeutic Target in Polycystic Ovary Syndrome. *International Journal of Molecular Sciences*, 21 (21):8191, January 2020. ISSN 1422-0067. doi: 10.3390/ijms21218191. URL https://www.mdpi.com/1422-0067/21/21/8191. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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/.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.