ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

COMMON AND RARE VARIANT ASSOCIATION APPROACHES IN TURKISH FAMILIES WITH POLYCYSTIC OVARIAN SYNDROME

ELİF ÖZ

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Dr. Onur Emre ONAT

ISTANBUL-2023

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

COMMON AND RARE VARIANT ASSOCIATION APPROACHES IN TURKISH FAMILIES WITH POLYCYSTIC OVARIAN SYNDROME

ELİF ÖZ

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Dr. Onur Emre ONAT

ISTANBUL-2023

Department:	Department of Biostatistics and Bioinformatics
Program:	Biostatistics and Bioinformatics Thesis Master Program
Thesis Title:	Common and Rare Variant Association Approaches in Turkish Families with Polycystic Ovarian Syndrome
Student's name and Surname:	Elif ÖZ
Date of Defence:	05/07/2023

This is to certify that I have examined this copy of master thesis. I have found that she/he prepared after fulfilling the specified requirements in the associated legislations before the final examining committee whose signatures are below.

Jury Member (Head of the Defense)	Prof. Dr. Osman Uğur SEZERMAN Acibadem University Institute of Health Sciences
Jury Member (Thesis Supervisor)	Dr. Onur Emre ONAT Bezmialem University Institute of Life Sciences
Jury Member	Prof. Dr. Ercan BASTU UMas Chan Medical School

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethicel rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

Date

"Elif Öz"

(Signature)

PREFACE AND ACKNOWLEDGEMENT

During the process of this thesis, I have learnt lots of things. I have gained lots of good memories and things that I wish will never become a memory.

I would first like to thank TUBITAK for their support for my Master's degree with the BIDEB 2210-E programme.

I would like to thank my supervisor, Dr. Onur Emre Onat who supported me way more than only a supervisor. His constructive feedbacks and his positive attitude during this project helped me too much, shed light on my way where I felt that I am stuck, and encouraged me to continue. He always found a way that eases my difficulties.

I would like to thank Prof. Uğur Sezerman, who supported me not only in this project but also thus far, throughout the period I was learning about not only computational biology but also many aspects of my academic self that I was developing these years.

I also want to acknowledge Merve Nur Köroğlu who eased my way so much during the project and always looked for a solution with me whenever I got stuck, and Nisa Esen who accompanied me through my way, and made it easier and more pleasant.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

AAFP	American Academy of Family Physicians
ACTH	Adrenocorticotropic Hormone
ADAT1	Adenosine Deaminase TRNA Specific 1
AK5	Adenylate Kinase 5
AMH	Anti Mullerian Hormone
AMHR2	Anti-Müllerian Hormone Receptor Type 2
AR	Androgen receptor
ARL14EP	ADP Ribosylation Factor Like GTPase 14 Effector Protein
ARMC3	Armadillo Repeat Containing 3
ASD	Autism Spectrum Disorder
ASRM	American Society for Reproductive Medicine
BMI	Body Mass Index
BMP4	Bone Morphogenetic Protein 4
BMP6	Bone Morphogenetic Protein 6
BMP9	Bone Morphogenetic Protein 9
BMP15	Bone Morphogenetic Protein 15
BQSR	Base Quality Score Recalibration
C19orf44	Chromosome 19 Open Reading Frame 44
CAPN10	Calpain 10
CASKIN2	CASK Interacting Protein 2
COL20A1	Collagen Type XX Alpha 1 Chain
CRYBG1	Crystallin Beta-Gamma Domain Containing 1
CWC27	CWC27 Spliceosome Associated Cyclophilin
CYP11a	Cytochrome P450 Family 11 Subfamily A Member 1
CYP17	Cytochrome P450 Family 17
CYP19	Cytochrome P450 Family 19
CYP21	Cytochrome P450 Family 21
DENND1A	DENN Domain Containing 1A
DHEA	Dehydroepiandrosterone
DHEAS	Dehydroepiandrosterone-sulfate

DHT	Dihydrotestosterone
DLG2	Discs Large MAGUK Scaffold Protein 2
DP	Depth
DUOX1	Dual Oxidase 1
EBG	Evidence-based Guideline
EHT	Efficient Hybrid Test
ERBB3	Erb-B2 Receptor Tyrosine Kinase 3
ERBB4	Erb-B2 Receptor Tyrosine Kinase 4
ESHRE	European Society for Human Reproduction and Embryology
EWAS	Exome-Wide Association Study
FANCC	FA Complementation Group C
FOX	Forkhead Box
FS	Fisher Scoring
FSH	Follicle Stimulating Hormone
FSHB	Follicle Stimulating Hormone Subunit Beta
FSHR	Follicle Stimulating Hormone Receptor
FTO	Fat Mass And Obesity Associated
GATA4	GATA Binding Protein 4
GATK	Genome Analysis Toolkit
GDF9	Growth Differentiation Factor 9
GLMM	Generalized Linear Mixed Model
GLUT4	Glucose Transporter Type 4
GMMAT	Generalized Mixel Model Association Test
GnRH	Gonadotrophin Releasing Hormone
GPCR	G-protein coupled receptor
GWAS	Genome-Wide Association Study
HPA	Hypothalamus Pituitary Adrenal
IGF1	Insulin-Like Growth Factor 1
INS	Insulin
INSR	Insulin Receptor
IRF1	Interferon Regulatory Factor 1
IRS1	Insulin Receptor Substrate Protein 1

IRS2	Insulin Receptor Substrate Protein 2
KRR1	KRR-R Motif-Containing Protein 1
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LH	Luteinizing Hormone
LHCGR	Luteinizing Hormone/Chorionic Gonadotropin Receptor
LIMA1	LIM Domain And Actin Binding 1
LMM	Linear Mixed Model
LoF	Loss of Function
LRT	Likelihood Ratio Test
MAF	Minor Allele Frequency
MAPRE1	Microtubule Associated Protein RP/EB Family Member 1
MQ	Mapping Quality
NEIL2	Nei Like DNA Glycosylase 2
NGS	Next Generation Sequencing
NIH	National Institute of Health
NMD	Nonsense-Mediated Decay
OGFOD3	2-Oxoglutarate And Iron Dependent Oxygenase Domain Containing 3
OR4K14	Olfactory Receptor Family 4 Subfamily K Member 14
PCA	Principal Component Analysis
PCDHGA11	Protocadherin Gamma Subfamily A, 11
PCO	Polycystic Ovary Syndrome 1
PCOS	Polycystic Ovary Syndrome
PKOS	Polikistik Over Sendromu
PLGRKT	Plasminogen Receptor With A C-Terminal Lysine
POM121L12	POM121 Transmembrane Nucleoporin Like 12
PRS	Polygenic Risk Score
PSENEN	Presenilin Enhancer, Gamma-Secretase Subunit
QD	Quality Normalized with Depth
QQ	Quantile-Quantile
QUAL	Quality
RAB5B	Ras-Related Protein Rab-5B

RAD50	RAD50 Double Strand Break Repair Protein
RNN	Recurrent Neural Network
SCN2A	Sodium Voltage-Gated Channel Alpha Subunit 2
SELENOP	Selenoprotein P
SHBG	Sex Hormone Binding Globulin
SKAT	Sequence Kernel Association Test
SKAT-O	Optimal Sequence Kernel Association Test
SLC22A18	Solute Carrier Family 22 Member 18
SLC25A47	Solute Carrier Family 25 Member 47
SMMAT	Set Mixed Model Association Tests
SNP	Single Nucleotide Polymorphism
SRD5A1	Steroid 5 Alpha-Reductase 1
SRD5A2	Steroid 5 Alpha-Reductase 2
STK11	Serine/Threonine Kinase 11
SULT2A1	Sulfotransferase 2A1
SVM	Support Vector Machine
TEX15	Testis Expressed 15, Meiosis And Synapsis Associated
THADA	Thyroid Adenoma-Associated
THAP8	THAP Domain Containing 8
TOX3	TOX High Mobility Group Box Family Member 3
TRP	Transient Receptor Potential
TRPM5	Transient Receptor Potential Cation Channel Subfamily M Member 5
VQSR	Variant Quality Score Recalibration
WDR26	WD Repeat Domain 26
WES	Whole Exome Sequencing
YAP1	Yes1 Associated Transcriptional Regulator
ZBTB16	Zinc Finger And BTB Domain Containing 16
ZFHX4	Zinc Finger Homeobox 4
ZNF534	Zinc Finger Protein 534

LIST OF FIGURES

LIST OF TABLES

ÖZET

Polikistik Over Sendromu Bulunan Türk Ailelerinde Yaygın ve Nadir Varyant İlişkilendirme Yaklaşımları

Polikistik Over Sendromu (PKOS), dünya çapında 5 kadından 1'ine kadar sık görülebilen ve ek metabolik komorbiditelere sahip, kısırlığın en yaygın nedeni olan karmaşık bir sendromdur. Durum hem genetik hem de çevresel faktörlerden etkilenir. Bu tür hastalıkların altında yatan mekanizmaları aydınlatabilmek için, NGS teknolojilerindeki gelişmeler sayesinde geliştirilen birçok istatistiksel girişim vardır. Söz konusu gendeki fonksiyon bozulmasının, ilgilenilen özelliğe katkısı hakkında yorum yapabilmek için; her bir gendeki patojenik varyant yükünü belirleyici, gen tabanlı varyant agregatif yaklaşımları geliştirilmiştir. Bu yaklaşımları Türk PKOS vakalarında uygulamak için, 190 aileden 203 PKOS hastası ve 52'si PKOS hastasının aile üyesi olan 815 kontrol bireyinin Tam Ekzom Dizilimi verileri, çeşitli gen tabanlı ilişkilendirme çalışma modellerinde kullanılmıştır. Amaç, potansiyel aday genleri belirlemek ve PCOS'un altında yatan moleküler mekanizmalar hakkında fikir edinmektir. İstatistiksel analiz sonuçları, varyantların aileler içinde ayrılması ve ilgili literatür araştırması, yumurtalık işleviyle veya insülinle ilişkili yolaklarda yer alan potansiyel aday genler hakkında değerli bilgiler sağlamıştır. İlginç bir şekilde, bu genlerin çoğu aynı zamanda otizmle de ilişkilidir ve bu, her iki koşulun altında yatan genetik mekanizmalarda potansiyel bir örtüşmeyi gösterir. Öte yandan, çalışmanın yaygın varyant analizi kısmı, WES verilerini analiz ederken, genom çapında bir temel verinin, ilişkili varyantların çoğunun genomun intergenik bölgelerinde yer aldığından dolayı uygun olmadığını ortaya koydu.

Anahtar Sözcükler: Polikistik over sendromu, Nadir varyant ilişkilendirme çalışmaları, Tüm ekzom dizileme, Ailesel veriler, Poligenik risk skoru

ABSTRACT

Common and Rare Variant Association Approaches in Turkish Families with Polycystic Ovarian Syndrome

Polycystic Ovary Syndrome (PCOS) is a complex condition that affects up to 1 in 5 women globally and is the most common cause of infertility, having additional metabolic comorbidities. The condition is affected by both genetic and environmental factors. To be able to shed light on these types of diseases, there are several statistical attempts developed in virtue of the enhancements in NGS technologies. Gene-based variant aggregative approaches are developed to take the pathogenic variant burden in every gene to be able to comment on the contribution of disruption of the gene in question to the trait of interest. To apply this approach in Turkish PCOS cases, Whole Exome Sequencing (WES) data of 203 PCOS patients from 190 families and 815 control individuals with 52 of those being family members of PCOS patients have been utilized for several gene-based association study models. The aim is to identify potential candidate genes and gain insights into the underlying molecular mechanisms of PCOS. The statistical analysis results, segregation of variants within families, and relevant literature search provided valuable insights into potential candidate genes which are involved either in ovarian function-related pathways or in insulin-related pathways associated with PCOS. Interestingly, most of these genes are also associated with autism, indicating a potential overlap in the genetic mechanisms underlying both conditions. On the other hand, common variant analysis part of the study revealed the fact that when analysing a WES data, a genome-wide base data is not appropriate, since the most of the associated variants are at the intergenic regions of the genome.

Keywords: Polycystic ovarian syndrome, Rare variant association studies, Whole exome sequencing, Familial data, Polygenic risk score

1 INTRODUCTION AND AIM

1.1 Polycystic Ovarian Syndrome

Being the most common endocrine disorder among its population, Polycystic ovary syndrome (PCOS) affects 4 to 20 percent of women of reproductive age globally (1). Consequently, it is also the most common cause of anovulatory infertility (2). Not only reproductive system is affected by PCOS, but metabolic and cardiovascular risk factors accompany it as well. Therefore, it is essential that PCOS be understood and treated to reduce the burden of associated comorbidities (2, 3). However, the aetiology of PCOS is still largely unknown albeit significant advancements in the understanding of its pathophysiology.

PCOS is a syndromic condition that not only consists of polycystic ovary morphology but is characterised also by hyperandrogenism and oligo- or anovulation. Diagnosis relies on the existence of at least 2 of the aforementioned 3 conditions, which is called Rotterdam criteria and currently the standard in the diagnosis of PCOS. Albeit diagnostic criteria had evolved starting from the initial National Institute of Health (NIH) consensus criteria which are followed by the European Society for Human Reproduction and Embryology/American Society for Reproductive Medicine (ESHRE/ASRM) consensus Rotterdam criteria and then the Androgen Excess and PCOS Society criteria, use of Rotterdam criteria is recommended by the Endocrine Society in 2013, the American Academy of Family Physicians (AAFP) Guidelines in 2016, and the International Evidence-based Guideline (EBG) for the diagnosis of PCOS 2018 (2, 4).

Since its metabolic nature, PCOS have several comorbidities such as obesity and abdominal adiposity. The reason why PCOS is such a complicated disease is that its comorbidities bidirectionally cause the maintenance of PCOS and therefore themselves. Obesity causes insulin resistance and hyperinsulinemia as high testosterone level in the serum does (5). Insulin has the ability to mimic the function of Luteinizing Hormone (LH) and therefore, this high level of insulin in the blood

indirectly increases the production of Gonadotrophin Releasing Hormone (GnRH) production. The higher level of GnRH causes excessive androgen production, making this feedback mechanism bidirectional. In addition, insulin has a negative effect on the level of Sex Hormone Binding Globulin (SHBG) which is a transport carrier protein responsible for binding estrogen and androgens and regulating their biological activity. High insulin level in the bloodstream also decreases the SHBG, increasing free testosterone level which is another mechanism to cause hyperandrogenism in PCOS. Hyperandrogenism causes hirsutism, alopecia, and acne which are commonly seen in PCOS patients (6).

It has been observed that women with PCOS have an increased risk of developing endometrial and ovarian cancers. Albeit the exact cause of this increased risk is not known, it is hypothesized that the hormonal imbalance seen in PCOS may be a contributing factor. PCOS patient women also have an increased risk of developing cardiovascular and cerebrovascular diseases, such as coronary artery disease and stroke, as well as venous thromboembolism. Additionally, the psychosocial impact of living with PCOS can lead to mood disorders such as depression and anxiety. On the other hand, the psychological impact of PCOS can vary greatly between individuals and can also be influenced by several different parameters (6).

Comorbidities of PCOS are affected by the lifestyle factors of patients. To illustrate, dietary habits and physical activity influence SHBG levels as well as body mass index (BMI) and obesity. In particular, it has been found that a low-carbohydrate diet can lead to increased SHBG levels in PCOS patients. Additionally, elevated SHBG levels have been associated with improved metabolic health markers such as lower fasting glucose and insulin levels. Inevitably, obese women with PCOS tend to have significantly lower SHBG concentrations than non-obese women.

It is a complex, heterogeneous disorder whose aetiology includes both genetic and environmental factors. Unlike Mendelian disorders, the genetic basis of complex disorders is far more intricate and difficult to discern. This is due to the fact that

these disorders are affected by a combination of both environmental and genetic factors, making it difficult to identify similarities between cases. As such, further research is needed in order to gain a better understanding of the underlying mechanisms of PCOS (6).

1.1.1 Genetics of PCOS

Being a complex disease affecting the metabolism and fertility of the patients, PCOS may be the result of mutations in any gene that affects ovaries either directly or indirectly. Genome-wide association studies (GWASs) and a GWAS meta-analysis had been applied to enlighten the aetiology of PCOS (7, 8, 9, 10, 11). These studies revealed the fact that the pathways responsible for PCOS pathogenesis involve androgen biosynthesis, gonadotrophin secretion and its function, ovarian ageing, and endocrine regulation.

The largest meta-GWAS of PCOS performed by the International PCOS consortium identified 14 genetic susceptibility loci (11). These loci contain PCOS candidate genes which are THADA, ERBB4, IRF1, RAD50, GATA4, NEIL2, PLGRKT, FANCC, DENND1A, ARL14EP, FSHB, YAP1, ZBTB16, ERBB3, RAB5B, KRR1, TOX3, MAPRE1.

THADA, ERRB3, ERBB4, YAP1, RAB5B, and MAPRE1 genes do not involve an ovary-specific role instead, they are related to apoptotic and proliferative functions of the cell. ERBB4 gene which is involved in mitogenesis and differentiation had been associated with different kinds of cancers. THADA and DENND1A are associated with insulin resistance and other metabolic abnormalities (12). In addition DENND1A protein is found in both the nucleus and cytoplasm of theca cells. IRF1 gene is a tumour suppressor which has also an activator role of genes involved in both innate and acquired immune responses. RAD50 and FANCC are also genes that are responsible for double-stranded DNA break repair. NEIL2 is responsible for base excision repair. Among these genes, only the FSHB gene is

involved in an ovary function-specific role. It encodes the beta subunit of follicle-stimulating hormone.

Genotype–phenotype correlation studies conducted in Han Chinese women with PCOS have demonstrated a significant association between variants in the THADA and DENND1A genes and endocrine and metabolic disturbances. Specifically, these studies have identified a positive correlation between PCOS-associated endocrine and metabolic disturbances, such as elevated levels of luteinizing hormone, testosterone, and insulin resistance, and the presence of certain THADA and DENND1A gene variants. These findings suggest that genetic variations in the THADA and DENND1A genes may be important determinants of the severity of PCOS-related disturbances (13).

In studies examining the European population, researchers have observed that a variant of the gene DENND1A is associated with an increased risk for androgen excess and anovulation. Additionally, a variant located near the FSHR gene has been associated with reduced levels of the follicle-stimulating hormone (FSH), while a variant near the RAB5B gene appears to be associated with impaired glucose metabolism (6).

1.1.1.1 Genes related to steroidogenesis and steroid hormone function

Cytochrome P450 protein family of enzymes are involved in a variety of metabolic processes including steroidogenesis. In particular, these proteins catalyze the oxidation of organic substances using molecular oxygen, allowing the organism to produce certain hormones and other molecules. The enzymes are found primarily in the endoplasmic reticulum, and mutations in the genes encoding these proteins can result in either increased or decreased aromatase activity. It had been shown that lean and obese PCOS patients show lower aromatase activity (14).

Being members of this family, CYP11a which catalyzes the conversion of cholesterol to pregnenolone, the rate-limiting step in the synthesis of the steroid

hormones, CYP17 which is an integral enzyme in the steroidogenic pathway that produces androgens and estrogens, CYP19 which catalyzes the last steps of estrogen biosynthesis, CYP21 which has a pivotal role in the synthesis of steroid hormones had been associated with PCOS up to date (14-22).

The AR (Androgen receptor) gene had also been proven to own an integral role in PCOS aetiology by both animal and human functional studies (23). The product of this gene acts as a transcription factor upon stimulation by steroid hormones.

The protein product of the SHBG gene binds to androgens, estrogens, and progestins in the bloodstream, reducing the levels of free hormones circulating in the bloodstream. Individuals with PCOS had been shown to have a lower level of SHBG and there are various single nucleotide polymorphisms (SNPs) that are associated with PCOS.

The SRD5A1 and SRD5A2 genes encode for two isoforms of the 5-alpha reductase enzyme, which converts testosterone to dihydrotestosterone (DHT) which is a more potent androgen. The SRD5A1 gene is expressed in the liver, skin, and brain, while the SRD5A2 gene is expressed in the prostate, seminal vesicles, and hair follicles. Both genes play an important role in the regulation of androgen levels in various tissues. Several studies have found that variants in the SRD5A1 and SRD5A2 genes are associated with an increased risk of PCOS (24-26).

One of the factors that contribute to the heterogeneity of PCOS is the genetic variation of SULT2A1. SULT2A1 is a gene that encodes for an enzyme called sulfotransferase 2A1. This enzyme is involved in the metabolism of dehydroepiandrosterone (DHEA) and its sulfate form, DHEAS. DHEA is a precursor to androgens, and its levels are often elevated in women with PCOS. DHEAS, on the other hand, is a storage form of DHEA that can be converted to DHEA as needed. Studies have shown that PCOS patients with different variants of the SULT2A1 gene exhibit variations in the DHEA/DHEAS ratio, which suggests that the SULT2A1 gene plays a role in the regulation of androgen levels in women with PCOS (24).

1.1.1.2 Genes related to folliculogenesis

Imbalances or functional degenerations in gonadotrophins which are the luteinizing hormone (LH) and the Follicular Stimulating Hormone (FSH) and their receptors which are Follicular Stimulating Hormone Receptor (FSHR) and Luteinizing Hormone/Chorionic Gonadotropin Receptor (LHCGR) have been associated with PCOS. Variations in LH whose protein product plays a critical role in the ovulation have been shown to be associated with an increased risk of PCOS. Deregulations in the LH gene expressioun have been shown to be associated with increased levels of LH in the bloodstream, hence disrupt the delicate hormonal balance that is necessary for normal ovulation and menstruation (22, 27).

The FSHR gene which is a G-protein coupled receptor (GPCR) has a pivotal role in the development and proper functioning of gonads. This gene encodes a transmembrane receptor protein found on the surface of ovarian follicle cells, which are responsible for producing and releasing estrogen. It is has an important role in FSH release from the pituitary gland. Recent research has shown that mutations in the FSHR gene may be associated with an increased risk of PCOS, by affecting insulin metabolism and causing insulin metabolism which is one of the most important comorbidities of PCOS. Additionally, studies have shown that women with PCOS often have lower levels of FSHR expression in their ovarian tissue, which may contribute to the abnormal follicular development and anovulation that both are characteristics of it (28).

The FSHB gene is responsible for encoding the beta subunit of the FSH which has a pivotal role in maturation of follicles during menstrual cycle. Since this gene directly affects the FSH expression and its function, variations in it cause disruptions of FSH function and therefore PCOS. Additionally, variations in the FSHB gene have been linked to other reproductive disorders, including primary ovarian insufficiency and male infertility (10).

The Anti-Mullerian Hormone (AMH) regulates transcriptions of some of the cytochrome P450 protein family of enzymes and has a negative effect on CYP17 transcription. Therefore an impairment in the AMH level subsequently causes an increase in the CYP17 because of its degenerated inhibition. This increase also causes an increase in the androgen level which is called hyperandrogenism as one of the main characteristics of PCOS (29).

AMHR2 gene encodes for the anti-Müllerian hormone receptor type 2, which plays an integral role in regulation of the maturation of follicles. The AMHR2 gene has been shown to be associated with PCOS. In women who suffer from PCOS, the AMHR2 gene is overexpressed, leading to excessive production of AMH, which causes a disruption in the follicular recruitment proves and in turn, leads to the development of ovarian cysts, irregular menstrual cycles, and infertility (30, 31).

There are several other genes that have been associated with PCOS due to their roles in follicular recruitment process, including BMP9, GDF9, BMP6, and BMP15. These genes are involved in the regulation of follicular maturation and ovulation steps due to the fact that their protein products are inhibitory transcription factors that control follicular growth keeping follicles from premature growth (32).

1.1.1.3 Genes related to insulin function

Hyperinsulinemia is one of the most important comorbidities of PCOS. The INS gene, which encodes insulin, is an integral gene in regulation of glucose metabolism. Women who suffer from PCOS, insulin resistance or type 2 diabetes is generally accompany PCOS. The high level of insulin that is caused by these conditions can stimulate the ovaries to produce more androgens than the normal amount, leading to one of the main characteristics of PCOS which is hyperandrogenism (33).

The INSR gene, which encodes the insulin receptor, is also associated to PCOS. Studies have shown that women with PCOS can have a decreased expression of INSR, which can in turn, contribute to insulin resistance. Insulin resistance, in other words decreased sensitivity to insulin, can lead to the overproduction of androgens with the aforementioned mechanisms (33). Insulin receptor substrate proteins 1 and 2 (IRS1 and IRS2) are activated upon phosphorylation by insulin receptor. Pathogenic variations in these genes have also been associated with insulin resistance and therefore PCOS. In addition, variations in these genes have been associated with obesity as well (34-36).

CAPN10 is also one of the PCOS-associated genes whose mechanisms are in insulin-related pathways, due to the role of CAPN10 in regulation of insulin secretion. Insulin resistance is important in PCOS, as mentioned earlier. Mutations in CAPN10 gene leads to impaired glucose metabolism and therefore insulin resistance. Several studies have reported a higher prevalence of CAPN10 gene mutations in women with PCOS compared to women without PCOS. These findings suggest that mutations in the CAPN10 gene can also contribute to the PCOS condition (37, 38).

1.1.1.4 Genes related to metabolism

PCO gene, also known as PCOS1, is one of the genes that are associated with PCOS aetiology with a metabolism-related mechanism. The PCO gene is found at one of the mostly associated regions of PCOS and is associated with PCOS in several studies (39).

The FTO gene has been widely studied and is known to be associated with obesity and type 2 diabetes. It is well-known that these conditions also contribute to aetiology of PCOS, as mentioned previously. The underlying mechanisms linking the obesity to PCOS are complex and still under investigation, yet it is well-known that obesity is one of the most important comorbidities of PCOS as highlighted previously. Therefore, albeit not directly, pathogenic mutations in FTO gene contributes to PCOS aetiology. (40, 41).

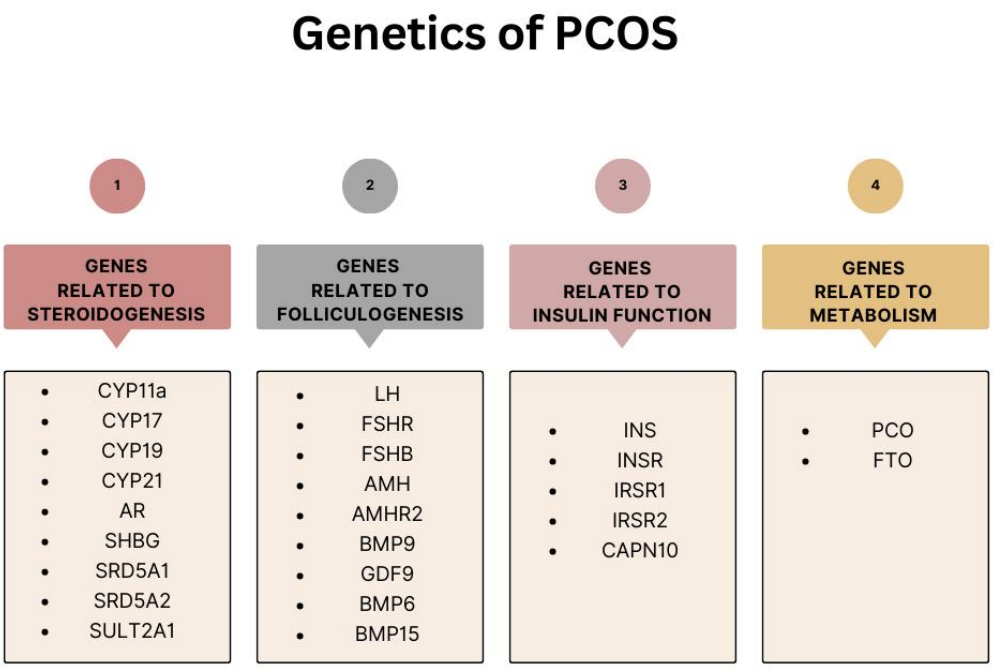

Figure 1. Genetics of PCOS

1.2 Rare Variant Association Studies

Reflections of technical and statistical enhancements in genomic science have contribute to the improvements in disease gene identification methods. Genome-wide association studies (GWASs) and exome-wide association studies (EWASs) are one of the mostly used methods that are used to detect rare variants that have a possibility to have a drastic effect on the function of the gene and therefore have a high impact on phenotypic traits. These studies make it possible to analyze a large number of variants by comparing the genotypes of patients and control individuals. As a result, these studies have provided new insights into the biological mechanisms underlying many phenotypes, identifying several variants that contribute to a number of traits and have elucidated the disrupted pathways causing disease.

However, this approach is unable to explain the mechanisms underlying multigenic and complex diseases. To be able to interpret the association of a trait with multiple markers there are different methods have been proposed. Chronologically, the first method claimed to achieve this is the Burden score test (42, 43). Burden-based methods are more powerful when most variants in a region are causal and the effects are in the same direction. On the other hand, they cannot account for the protective variants and their power decreases when there is a small number of associated variants. To overcome these handicaps, kernel-based methods which rely on a non-parametric approach have been proposed. Kernel methods are able to handle genetic markers in opposite directions since they take the genetic similarity between subject pairs into account. They are particularly useful in the analysis of genetic data, where the underlying relationships between genetic variants and disease risk can be highly complex and difficult to model using traditional statistical methods (44).

Albeit linear mixed models are commonly used in aggregative association analyses as well as kernel-based methods, they differ in the logic behind them and their applications. Linear mixed models are a class of regression models that incorporate fixed and random effects to account for variability in the data, while In contrast, kernel methods rely on a similarity function to map the data into a higher-dimensional space, where linear models can be used.

In this study, Burden Test, Sequence Kernel Association Test (SKAT), Optimal Sequence Kernel Association Test (SKAT-O), and Efficient Hybrid Test have been used utilizing the SKAT and the GMMAT R packages.

1.3 Polygenic Risk Score Analysis

Since complex diseases are often caused by multiple genetic factors and environmental influences it has been discovered that individual genetic variants have only small effects on the risk of developing such diseases. This realization has led to the development of a new technique called Polygenic Risk Score (PRS) analysis. In

PRS analysis, the small effect sizes of multiple SNPs are combined to create a single score that indicates an individual's risk of developing a particular trait or disease. The SNPs used in the analysis are derived from a completely separate GWAS that serves as the base data. The genotypes of the individuals being analyzed are referred to as the target data. While PRS is not yet widely used in clinical practice, it has the potential to be used for genetic screening to identify individuals at risk of developing certain diseases (45). In this study, the summary statistics data of the largest meta-GWAS of PCOS by the International PCOS Consortium (11) was used as base data.

A well-defined PRS can not only distinguish between individuals who are healthy and those who have a particular disease, but it can also enable disease sub-grouping. Since association studies are typically performed on a limited sample size, overfitting can occur in associations. Since in PRS calculation, there is a totally different target data used, the results of PRS analyses are more generalizable than association studies alone. Another important factor that PRS takes into account is linkage disequilibrium (LD). This refers to the tendency of certain genetic variants to be inherited together because they are located close to each other on a chromosome. PRS analysis accounts for LD when calculating risk scores, which is why it is a more reliable method than simply summing up the effect sizes of risk alleles. To be able to calculate LD accurately, relatives are needed to be excluded (46).

2 BACKGROUND

2.1 Healthy Menstrual Cycle

In a healthy menstrual cycle, GnRH, released from the brain, stimulates the pituitary gland for the delivery of gonadotrophins which are FSH and LH. Gonadotrophins reach the ovaries through the bloodstream and start folliculogenesis of 6-12 follicles. These maturing follicles start to release estrogen which causes the lining of the uterus to begin to build. One of those follicles dominates and the others dissolve. The dominating follicle fully matures and travels to the edge of the ovary and perforates it to be able to release the egg it carries. This stimulates one of the fallopian tubes to move through the ovary and capture the outshooting egg. After the egg is released, the punctured follicle collapses and turns into corpus luteum, causing progesterone release which in turn decreases GnRH production as a negative feedback mechanism. If the egg is fertilized by a sperm, the resulting zygote travels down to the uterus which is prepared for pregnancy and implants in its lining at the beginning of pregnancy. On the other hand, if the released egg is not fertilized in a couple of days, the corpus luteum starts to dissolve with the simultaneous decrease of the progesterone and estrogen levels in the bloodstream. This decrease causes the shedding of the uterine lining and eventually menses begins. The cycle then starts again with GnRH production from the brain (47).

2.1.1 Follicular recruitment

During folliculogenesis, AMH acts as a regulator in two ways. In initial recruitment, follicles are selected from their dormant primordial stage to develop. In cyclic recruitment, growing follicles are recruited to grow until their preovulatory stage, which is a response to FSH increase. The selection mechanism is that only large preantral and small antral follicles, which are sensitive enough to FSH can continue to grow (47).

In addition to AMH, oocytes also secrete other inhibitory transcription factors such as serine/threonine kinases such as STK11 and BMP4 and proapoptotic factors like forkhead box (FOX) to perform epithelial-mesenchymal interaction that controls follicular growth in a quiescence state. Moreover, follicular growth factors are involved in the regulation and control of ovarian folliculogenesis. These follicular growth factors, including BMP9, GDF9, BMP6, and BMP15, work in synchronicity and synergistically to regulate the growth and development of follicular cells (32).

It is important to note that the ovarian follicular recruitment process is an intricate one that is independent of gonadal hormones. Therefore, the regulation and control of this process are crucial for successful reproduction (47).

2.1.2 Steroid hormone synthesis

The process of ovulation in females involves a complex interplay of hormones and cellular mechanisms. One important aspect of this process is the role of LH in stimulating theca cells to produce androgens. Theca cells are located in the outer layer of the ovarian follicle and are responsible for producing androgens, which are a type of male sex hormone. When LH is released by the pituitary gland during the menstrual cycle, it binds to receptors on the surface of theca cells and triggers a cascade of biochemical reactions within the cells. This ultimately leads to the synthesis and release of androgens, primarily testosterone and androstenedione, into the bloodstream. These androgens then travel to nearby granulosa cells, which are located closer to the oocyte, where they are converted into estrogen through a process called aromatization (47).

The production of androgens by theca cells is an important step in the process of ovulation, as these hormones play a crucial role in follicular development and maturation. In addition, androgens are also important for maintaining a healthy reproductive system in females, as they help to regulate the menstrual cycle and promote the growth and maintenance of bone mass (47).

To regulate the levels of androgens and other hormones in the bloodstream, the liver produces a glycoprotein called SHBG. This protein is a highly specific glycoprotein that plays a crucial role in regulating the levels of sex hormones in the body. It binds to androgens, estrogens, and progestins in the bloodstream, reducing to maintain adequate levels of free hormones available to act on the body's cells. In addition to being produced by the liver, SHBG levels can also be influenced by other factors such as age, sex, and hormonal imbalances. For example, high levels of thyroxine hormone produced by the thyroid gland can stimulate the liver to produce more SHBG, which can lead to lower levels of free hormones in the bloodstream. Conversely, certain conditions such as obesity and insulin resistance can lead to lower levels of SHBG, which can result in higher levels of free hormones and hormonal imbalances (48).

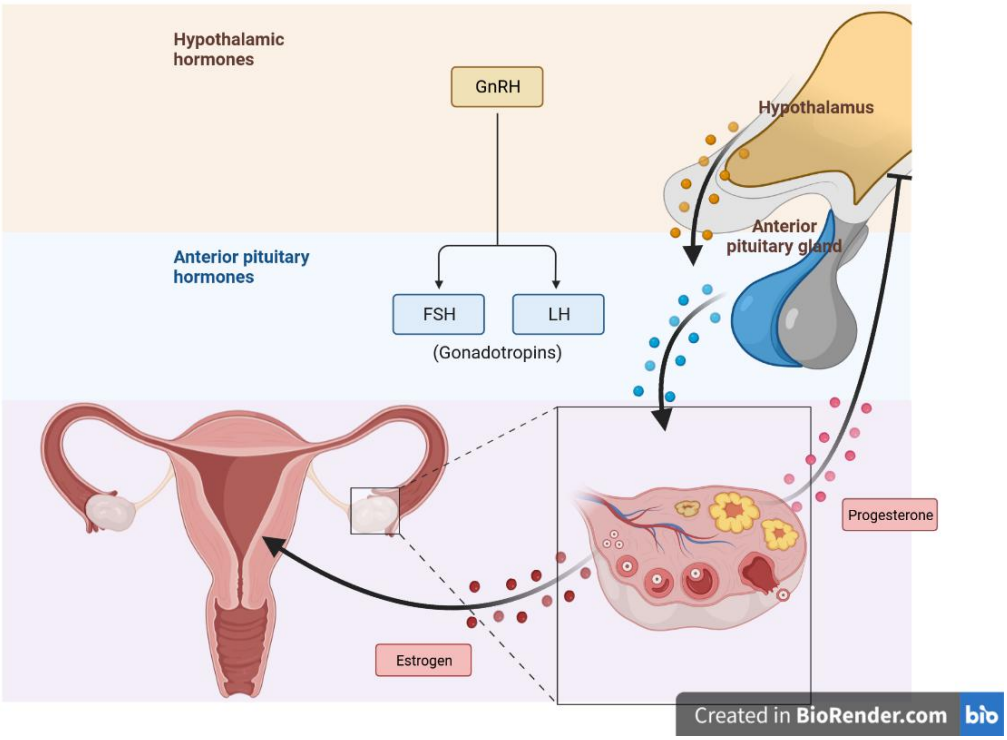

Figure 2. Normal Menstrual Cycle

2.2 Pathophysiology of Polycystic Ovary Syndrome

Due to the disruption in the mechanism which regulates ovulation through the synchronise interaction between LH, FSH, IGF1, and AMH, oligo- or anovulation, which are irregular and absent ovulation respectively, is common in PCOS. One of the main characteristics of PCOS, different from the normal menstrual cycle, is there is a higher release of GnRH into the bloodstream. Instead of a cyclic release which ascends and descends on a regular basis, GnRH release has a more frequent basis in PCOS condition which cause a higher level of LH, and therefore a disrupted LH:FSH ratio. The FSH level cannot suffice for the maturation of follicles, domination of one, and ovulation, albeit there is usually a normal amount of FSH release. One of the causes of this decreased sensitivity is the increased level of AMH which is constructed by growing follicles. In normal ovulation, it manages the growth of the follicle, production of estrogen and dominant follicle selection. As a negative regulator of folliculogenesis, AMH decreases the sensitivity of these growing follicles to FSH and inhibits the maturation of all follicles other than the dominant one. Therefore, an abnormal increase in the AMH is the reason why the normal FSH level cannot suffice for the maturation of follicles and cause all follicles to be stuck in their pre-antral and antral states which then turn into ovarian cysts as one of the main characteristics of PCOS (6).

If there is no ovulation, then there is no corpus luteum and therefore no progesterone release. Since progesterone gives negative feedback to GnRH release, the absence of progesterone leads to a higher level of GnRH. The follicles that cannot mature and ovulate form cysts, which are one of the characteristics of PCOS. On the other hand, the increased level of LH also causes theca cells to produce more androgens, which causes the other characteristic of PCOS, hyperandrogenism. Hyperandrogenism causes the common PCOS symptoms which are hirsutism, alopecia, and acne. Furthermore, this imbalance in the hormone levels causes menstrual irregularities, and infertility as well (6).

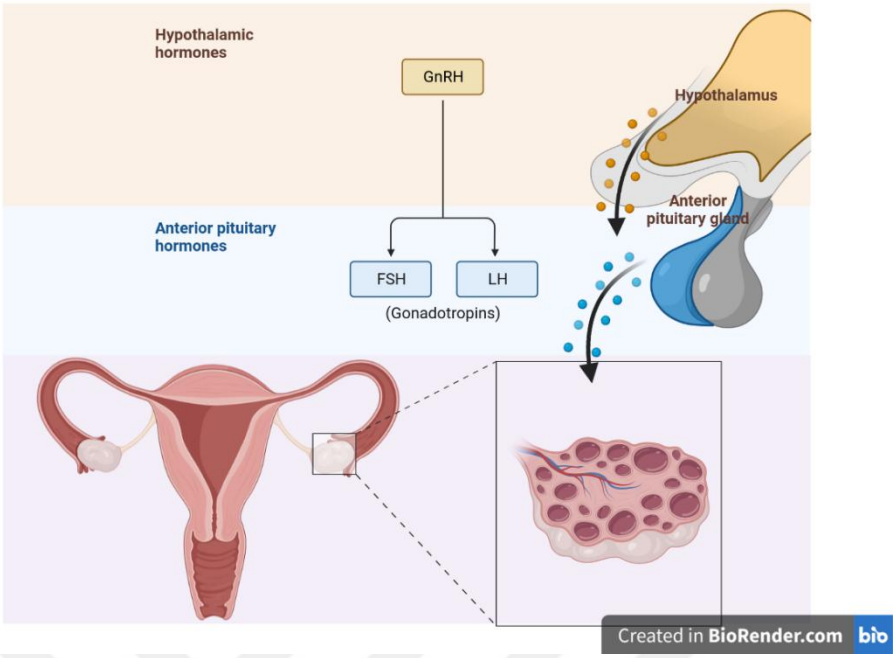

Figure 3. Menstrual Cycle in PCOS

Hyperinsulinemia, and as a result, insulin resistance is a common and important comorbidity of PCOS. Insulin can mimic the role of LH and upon stimulating ovary theca cells, cause more androgen release. On the other hand, the excess amount of androgen causes insulin insensitivity and therefore insulin resistance. Insulin resistance causes hyperinsulinemia and this high amount of insulin causes more androgen release as a bidirectional mechanism (49).

Hyperandrogenemia can also be caused by the dysregulation of responsible genes due to certain genetic abnormalities. To illustrate, low levels of SHBG have been associated with PCOS as well as its comorbidities, such as type 2 diabetes and cardiovascular disease. High levels of SHBG, on the other hand, have been linked to a decreased risk of developing certain types of cancers, including breast and prostate cancer. Researchers are increasingly exploring the potential therapeutic applications of SHBG in various medical conditions. For instance, the use of SHBG as a biomarker for the diagnosis and prognosis of certain diseases is currently being investigated. Moreover, the potential use of SHBG as a therapeutic agent for the treatment of metabolic disorders and certain cancers is also being explored (48, 49).

One of the androgens commonly associated with PCOS is DHEA and its sulfate form, DHEA-sulfate (DHEAS). DHEA and DHEAS are produced by the adrenal glands and are precursors to both estrogen and testosterone. Diagnosis of PCOS often involves measuring the levels of DHEA and DHEAS in the blood, along with other hormones such as FSH, LH, and testosterone. DHEA is sulfated to DHEAS, which is the potent inactive form, by the sulfotransferase 2A1 enzyme which is encoded by the SULT2A1 gene. DHEAS needs to be converted to DHEA to be converted to other androgens (50).

Recent research has shown that adrenal hyperandrogenism in PCOS patients is not solely dependent on the hypothalamus pituitary adrenal (HPA) axis, which is responsible for regulating hormone production by the adrenal glands. Instead, it appears that hyperresponsiveness toward androgen synthesis is the primary driver of excessive androgen production in these patients. Adrenal PCOS, which is characterized by the overproduction of androgens by the adrenal glands may be driven by several genetic causes, and two of the most common are cortisone reductase deficiency and 11 β hydroxysteroid dehydrogenase type 1 deficiency. In cortisone reductase deficiency, the enzyme responsible for converting cortisone to cortisol is impaired, leading to a buildup of cortisone in the body. This results in an increase in adrenocorticotropic hormone (ACTH) and subsequently, the adrenal glands produce more androgens. This excess androgen production contributes to the development of PCOS. Similarly, in 11 β hydroxysteroid dehydrogenase type 1 deficiency, cortisol production is impaired, leading to a buildup of cortisone in the body. This also elevates the level of ACTH, which in turn stimulates the adrenal glands to produce more androgens as a hallmark of PCOS (32).

2.2.1 PCOS types

2.2.1.1 Classic phenotype

Classic phenotype involves both Phenotype A and Phenotype B. These two phenotypes have hyperandrogenism and oligo- or anovulation in common. What differentiates Phenotype A from Phenotype B is that Phenotype A also involves polycystic ovarian morphology whereas in Phenotype B ovarian morphology is normal. Menstrual irregularities, hyperinsulinemia and insulin resistance are associated with this type of phenotype. In addition, PCOS patients with this type of phenotype are more prone to obesity. They also show a higher amount of AMH (51, 52).

2.2.1.2 Ovulatory phenotype

Phenotype C, which consists of hyperandrogenism, normal menstruation cycle, and polycystic ovarian morphology patients, constitutes the ovulatory phenotype. They have a slight increase in their androgen levels as well as their insulin levels (53).

2.2.1.3 Non-hyperandrogenic phenotype

The non-hyperandrogenic phenotype, so-called Phenotype D involves PCOS characteristics except hyperandrogenism. In this phenotype, patients show oligo- or anovulation and a polycystic morphology of ovaries, along with a normal androgen level. They have regular menstruation albeit having irregularities ever and anon (27, 54).

2.3 Rare Variant Association Studies

The application of advanced genetic analysis techniques, such as GWASs and EWASs, has advanced parallel to progress in next-generation sequencing technologies and enabled improved methods of identifying disease-causing genes.

These approaches enable the analysis of the association between the trait in question with a large number of variants by comparing genotypes of patients and control individuals. Such studies have facilitated the discovery of SNPs that are associated with phenotypic traits and have provided new insights into the biological mechanisms underlying them. To date, many variants have been identified that are linked to a variety of traits, thus revealing disrupted pathways causing disease.

2.3.1 Kernel-based methods

The classical genome-wide association approach is unable to capture the cumulative effect of several markers on the trait since it relies on the one SNP at a time principle. It assumes that every variant affects the phenotype individually and tests their effect independently of each other. Conversely, this situation does not apply to the real case scenario. In fact, there is a linkage disequilibrium between variants and they are segregating as haplotype blocks rather than individually. To address this issue and to be able to explain the contribution of each variant, various methods have been proposed. The earliest of these is the Burden score test (42, 43), which is more powerful when most variants in a region are causal and the effects are in the same direction. Nevertheless, this method has its limitations, such as its inability to account for protective variants and reduced power when there is a small number of associated variants. To overcome these, kernel-based methods have been developed. Kernel methods take the genetic similarity between subject pairs into account and are able to handle genetic markers in both directions. Since its linear nature, it has a flexible framework that makes it easy to account for non-genetic covariates (44). Consequently, they offer an improved capacity for interpreting the association of a trait with multiple markers.

Instead of the conventional regression approach of explicitly defining basis functions to capture the influence of all genetic markers, kernel methods provide an effective alternative by approximating the underlying smooth function with a kernel matrix. Kernel methods are well-suited for high-dimensional data and can be used to systematically identify nonlinear effects. Furthermore, they have the advantage of

being more computationally efficient than traditional regression methods as they do not require explicit basis functions to be defined. By constructing a kernel matrix, these methods can be used to accurately model the relationships between the genetic markers and the desired outcome (44).

2.3.1.1 SKAT with Liu method

When utilizing kernel-based methods under large sample sizes, moment-matching methods were initially proposed to approximate the null distribution for calculating p-values. This method involves calculating the first two moments of the score test statistic under the null hypothesis and using them to estimate the degrees of freedom and a factor to rescale the chi-square distribution. Liu et al. used the Satterthwaite approximation which is a type of moment-matching method (55).

Moment matching methods are used for parameter estimation. The main logic behind them is to match the observed moments of the sample distribution to the expected moments of the models' distribution, to be able to estimate the parameters of the model. This approach is preferred when the exact form of the target distribution is unknown or difficult to estimate. In addition, they are flexible and thus can be applied to a wide range of genetic association models. However, moment-matching methods have some limitations especially when dealing with small sample sizes or non-standard distributions (56).

The Satterthwaite approximation is a kind of parameter estimation algorithm that uses the moment-matching method. It is used to estimate the degrees of freedom in a t-test when the sample sizes are unequal. It is able to handle unequal sample sizes in genetic data arising from different group sizes. The approximation is based on a formula that takes into account the sample sizes and variances for each group. The resulting estimate is then used to calculate the t-value and associated p-value for the hypothesis test.

However, this method can be anti-conservative at the extreme tail and may lead to inaccurate results. On the other hand, by using higher moment-matching methods, more accurate p-values can be obtained, which can lead to better decision-making in statistical hypothesis testing (44).

2.3.1.2 SKAT with Davies method

Accurately computing the distribution of score test statistics can be a challenging task. To address this issue, Wu et al. (57) proposed a more precise asymptotic approximation of the distribution of score test statistics. The authors claimed that the distribution of the statistic is a linear combination of independent and identically distributed chi-squared random variables, and they applied the Davies method (58) to obtain an analytic solution for computing p-values.

The Davies method is used to analyze the distribution of a statistic and mostly used in genetic association studies. The logic behind it is to express the distribution of statistics as a linear combination of independent and identically distributed chi-squared random variables, which can then be used to calculate p-values and other measures of statistical significance. It is relatively simple to implement and can be easily adapted to different study designs and sample sizes, which make it proper for genetic studies. It is important to note that, this method is robust to violations of normality assumptions, given the non-normal distribution of many genetic traits. Finally, the Davies method has been shown to be more powerful than other commonly used methods in detecting genetic associations.

Albeit the Davies method is a powerful technique that provides accurate results and suitable for genetic studies, it requires the specification of a numerical accuracy parameter. If the parameter is not set appropriately, the results can be inaccurate, especially for very small p-values (44).

2.3.1.3 SKAT with Kuonen method

The method which Kuonen proposed is the saddlepoint method, which has the advantage of not requiring the accuracy parameter. In addition, it does not require the use of complex numerical integration techniques, which can be time-consuming and prone to errors. Instead, the saddlepoint method relies on the use of the saddlepoint approximation, which is a powerful technique that allows for the accurate calculation of p-values (44).

Saddlepoint approximation is a widely-used and highly effective method for approximating the probability density function of a test statistic. This technique is based on the idea which states that the logarithm of the moment-generating function of a statistic can be accurately approximated by a quadratic function around its maximum. This approximation allows for the estimation of the probability of observing an extreme value of the statistic. It is especially useful when the sample size is small or the distribution of the statistic is complex or unknown. However, it also is conservative for small sample sizes.

2.3.1.4 SKAT-O

SKAT-O is a statistical method that combines the Burden test and the SKAT to maximize the power of the analysis. Since the Burden test tests the association between a set of genetic variants and the phenotype of interest by collapsing all the variants into a single score and SKAT considers each variant individually, SKAT-O selects the best linear combination of them to maximize the power of the analysis. This is achieved by using a likelihood ratio test to compare the performance of different combinations of the two methods. The result is a more accurate and powerful test that can detect even small genetic effects that may be missed by other methods, conserving the computational efficiency of SKAT (59).

2.3.2 Likelihood ratio tests

In addition to score tests, other hypothesis testing procedures based on kernel methods have been proposed in past years. One such approach is the restricted likelihood ratio test approach, which was proposed by Zeng et al. in 2014 for aggregative rare variant testing (60). This approach is based on the idea of direct parameter estimation, which is advantageous because it requires fitting both the null (i.e., reduced) and alternative (i.e., full) models as a condition of hypothesis testing when using a likelihood ratio statistic, unlike score statistics, which only require parameter estimates for the null hypothesis model (44). The null hypothesis assumes there is no genetic association between the trait of interest and the genetic variant being tested. The alternative hypothesis, on the other hand, assumes that there is a genetic association between the trait of interest and the genetic variant. The test produces a p-value, which indicates the strength of evidence against the null hypothesis.

In genetic association studies, both score tests and likelihood ratio tests are commonly used to assess the contribution of genetic variants to phenotype. Score tests are based on regression models and use the score function to test the null hypothesis that the variant has no effect on the outcome of interest. On the other hand, likelihood ratio tests compare the likelihood of the data under the null hypothesis to the likelihood of the data under an alternative hypothesis that includes the genetic variant. Likelihood ratio tests can provide more precise estimates of the effect size of the genetic variant than score tests.

2.3.3 Linkage methods

Linkage methods are an important tool for genetic analysis in familial data, allowing researchers to utilize the segregation of variants across multiple individuals within a family. In the context of Mendelian disorders, parametric linkage methods are often used to identify genetic variants that are responsible for the disease phenotype. These methods rely on assumptions about the mode of inheritance and

the underlying genetic model, and thus may not be suitable for complex diseases with multiple genetic and environmental factors (61).

In contrast, non-parametric or model-free linkage methods, also known as allele-sharing methods, do not rely on these assumptions and are therefore better suited for the analysis of complex diseases. These methods assess the degree of similarity or sharing of genetic markers between affected individuals in a family, and compare this to what would be expected by chance. This allows for the identification of regions of the genome that are likely to contain disease-causing variants (62).

However, there are limitations to the use of linkage methods in WES data. WES provides a comprehensive view of the protein-coding regions of the genome, but there may be missing data due to technical limitations or incomplete coverage. This can reduce the statistical power of linkage analysis since missing data can lead to false negative results or reduced accuracy in estimating the degree of allele sharing.

2.3.4 Linear Mixed Models (LMMs)

Being an extension of traditional linear models, linear mixed models account for both fixed and random effects, which make them powerful when there are multiple factors affecting the trait of interest. They have the advantage of taking population structure and relatedness into consideration. These models are highly versatile and can be used for both continuous and binary traits (63-67).

However, LMMs have certain limitations. Most importantly, they fail to control the type 1 error rate, which makes them inappropriate for analyzing binary traits when population stratification leads to a violation of its constant-residual variance assumption. To address this issue, alternative methods for analyzing binary traits that take into account population structure and relatedness have been developed (68-71). These models can be particularly useful when dealing with complex traits that are influenced by multiple genetic and environmental factors.

Linear mixed models differentiate from likelihood ratio tests in their underlying assumptions. While likelihood ratio tests are used to compare the goodness of fit of different statistical models and can be used to test hypotheses about the effect of specific genetic variants on traits or diseases, linear mixed models are powerful tools for data containing familial relationships since they can account for the correlation between repeated measures and thus the relatedness between individuals. In addition, these models can also handle missing data.

2.3.5 Generalized Linear Mixed Models (GLMMs)

Linear mixed models are used to analyze data with continuous outcomes, while generalized linear mixed models are used to analyze data with categorical or discrete outcomes. Additionally, linear mixed models assume a normal distribution of the errors, while generalized linear mixed models do not make any assumptions about the distribution of the errors. Furthermore, linear mixed models can analyze one level of nested data, whereas generalized linear mixed models are able to analyze multiple levels of nested data (72).

2.3.5.1 Variant-Set Mixed Model Association Tests (SMMAT)

A set of variant set mixed model association tests (SMMATs) which both take familial relationships into account and control type 1 error rate for binary traits have been proposed by Chen et al. in 2019 (73). The SMMAT framework includes four tests: the burden test (SMMAT-B), SKAT (SMMAT-S), SKAT-O (SMMAT-O), and an efficient hybrid test (SMMAT-E) that combines the burden test and SKAT as a powerful alternative to SKAT-O. As a pivotal advancement, all of them use GLMM with only covariates, which need to be fit only once for all genetic variant sets in an analysis. This reduces the computational burden and makes the tests more efficient.

SMMAT-E combines both the burden test and an adjusted mixed model SKAT statistic, which is approximately asymptotically independent from the mixed model burden test statistic. It uses matrix projections to approximate the adjusted SKAT statistic from a global null model without any fixed effects for the variant set-specific genetic burden. This approach minimizes the computational cost as the global null model only needs to be fit once in a whole-genome analysis. The approximation is highly accurate, even in the presence of large genetic effects, making it an efficient and reliable tool for studying complex disease genetics.

All SMMATs use the same GLMM which assumes the phenotype follows a Bernoulli distribution and thus it uses logit as a link function. Initially, it constructs a null model that assumes there is no association between the variant set and the trait. The next step in the analysis involves the inclusion of the variant set as a predictor variable in the GLMM. The model is then compared to the null model using likelihood ratio tests to determine if there is a significant association between the variant set and the trait. If the variant set is found to be significantly associated with the trait, the GLMM is further refined by adding covariates to account for confounding factors. The final model is then used to estimate the effect size of the variant set on the trait and to identify any potential genetic markers that may be useful for predicting the trait.

When performing GWAS, it is important to take into account the presence of genetic variants that are not directly measured but are in linkage disequilibrium with the genotyped variants. This is where SMMAT come into play. SMMAT-B uses a Bayesian approach to estimate the effects of the genetic variants. It assumes that the effects of the variants are normally distributed and uses a burden score test to estimate the posterior distribution of the effects. SMMAT-S, on the other hand, is a frequentist approach that uses a score test to estimate the effects of genetic variants. It assumes that the effects of the variants are fixed and uses a variance component score-type test to test the null hypothesis. Finally, SMMAT-O is a hybrid of SMMAT-B and SMMAT-S. It uses a Bayesian approach to estimate the variance components of the genetic effects and a frequentist approach to estimate the effects

themselves. Weighting strategies are used to assign different weights to individual genetic variants, based on their contributions to the trait or disease under study. For example, some variants may have a stronger effect on the trait than others and hence would be assigned a higher weight. Several different weighting strategies can be used in the SMMAT framework. In this study, MAF-based weights were used.

2.4 Polygenic Risk Score Analysis

PRS have become an increasingly popular tool in precision medicine, as they allow for the prediction of an individual's risk of developing a complex disease based on their genetic profile. However, accurately calculating PRS can be a challenging task due to the high level of uncertainty and correlation between multiple variants. The main aims of PRS calculation are to adjust effect sizes to exclude variants that have high uncertainty and to take LD into account to estimate individual effect sizes by excluding the effect of correlation between multiple variants. To adjust effect sizes, various shrinkage strategies are currently applied, such as LASSO, elastic net, or Bayesian approaches. These strategies aim to reduce the impact of noise and uncertainty in the data, resulting in more accurate and reliable PRS calculations. Another important point of LD in PRS calculation is that, it requires exclusion of family members to prevent spurious LD. On the other hand, classic PRS calculation depends on p-value thresholding. This approach involves selecting SNPs based on their statistical significance, with only those that reach a certain threshold being included in the analysis. However, this approach can lead to the exclusion of potentially important SNPs that do not reach the threshold, resulting in an incomplete picture of an individual's genetic risk profile. For LD control, there are also different strategies. One approach is clumping, where only the SNP with the highest certainty (smallest p-value) is retained in a certain locus, while the others are excluded. Another approach is to account for LD between SNPs without clumping, which takes into account the correlation between SNPs and adjusts for it in the PRS calculation (74).

3 MATERIALS AND METHODS

3.1 Cohort Description

In this study, WES data of 203 PCOS patients from 190 families which was collected within the framework of the project titled "Genetics of Polycystic Ovarian Syndrome in Turkish Families" is used. Women between the ages of 18-40 years, with the presence of menstrual irregularity due to oligo- or anovulation, and with evidence of hyperandrogenism were preferentially sampled. Individuals were excluded from the study if they have a history of any pituitary disease, hypothalamic or hyperthalamic diseases, untreated thyroid diseases, and reproductive diseases except PCOS; if their hyperandrogenism and menstrual irregularity are caused by conditions such as congenital adrenal hyperplasia, androgen-secreting tumours, Cushing syndrome, and hyperprolactinemia; if they use danazol or androgenic progestins; and if onset of symptoms is in the third decade of their life or later. In addition, pregnant individuals were also excluded from the study.

There are 815 control individuals and 52 of those are family members of PCOS patients. Out of these control individuals, 261 are and 16 of the family member controls are male. This project was initiated in 2013 and was conducted in a joint effort between Bilkent University, Rockefeller University, and Hacettepe University. All patients and family members involved in this project were provided with information regarding the study and their written consent was obtained before participation.

Table 1. Cohort Gender Information

	Female	**Male**	**Total**
Case	203	-	203
Control	563	252	815
Total	766	252	1018

Table 2. Cohort Age and BMI Information

Variable	Control	PCOS
Age	38.97+-14.3 (7-95)	32.68+-10.32 (11-41)
BMI	45.3+-11 (14.5-40)	39.4+-12.51 (18.01-80.09)

3.2 Pre-processing

The WES data of cases and controls was aligned to the hg38 reference genome and processed according to the GATK best practice guidelines (75). The alignment process involves mapping reads to the reference genome and assigning them to precise locations so that they can be analysed. The GATK best practice guidelines provide a tool for every step of processing genomic data, ensuring that the quality of the output is high. These steps include duplication and removal of duplicate reads, indel realignment, base quality score recalibration (BQSR), variant calling and joint genotyping, and then variant quality score recalibration (VQSR). Variants are then filtered according to GATK Best Practices recommendations (76, 77). Annotations of variants are added to variants utilizing ensembl-vep (78), ANNOVAR (79), and SNP-EFF (80).

After the annotated vcf files have been obtained, graphs of variant quality, variant depth, mean depth of individual, variant missingness, mean missingness of individual, heterozygosity, and minor allele frequency (MAF) were obtained utilizing vcftools (81), and tidyverse R package [version 2.0.0]. The filtering criteria were decided due to information gained from the graphs. Individuals having a heterozygosity of 3 standard deviations below the average and a mean depth below 20 have been excluded from the study. Variant filtering criteria were decided as 30 for variant quality, and 8 for depth. For the other parameters, GATK Best Practices Guidelines were followed.

Table 3. Filtering Criteria

In house MAF	< 0.01
HRC MAF	< 0.001
1000G MAF	< 0.001
ESP6500_AA MAF	< 0.001
ESP6500_EA MAF	< 0.001
FS	< 200
QD	> 2
MQ	> 20
QUAL	> 30
ReadPosRankSum	> -20
DP	> 8

To be able to observe if there is an outlier in the population and examine the structure of the cohort, Principal Component Analysis (PCA) was conducted upon merging the cohort data with that of the 1000 Genomes Project. To be able to merge these 2 datasets properly, different chromosome files of the 1000 Genomes dataset were concatenated and the resulting file was filtered due to genotyping rate of 0.8, individual missingness of 0.2, and MAF of 0.05. Then, common SNPs between the 1000 Genomes dataset and our cohort have been determined and uncorresponding SNPs were excluded. Finally, the merged file is pruned and the Principal Components were calculated with PLINK (Version 1.9) (82). Principal Components 1 and 2 were visualized by utilizing R and also Principal Components 3, 4, and 5 were examined in a 3D plot with Scatterplot3d package of R, since it had been shown that 2D PCA plots can be inefficient to differentiate populations (83).

3.3 Variant Filtering

Initially, synonymous variants were filtered with MAF thresholds of 10^{-3}, $5*10^{-4}$, 10^{-4}, $5*10^{-5}$, 10^{-5}, and $5*10^{-6}$. To optimize the MAF threshold, data produced by every 6 thresholds were analyzed via binary functions of SKAT-O and Burden test, utilizing the SKAT R package (Version 2.2.5) (59, 84-86). Then, QQ plots of p-values were obtained via the gaston R package (Version 1.5.9) (87).

Secondly, to compare the two approaches, the same data were also analysed via SMMAT-S, SMMAT-O, SMMAT-B, and SMMAT-E, utilizing the GMMAT R package (Version 1.4.0) (88). Then, QQ plots of p-values were obtained via the gaston R package (Version 1.5.9) (87).

Out of the tested 6 MAF thresholds, a MAF of 10^{-4} was selected to continue. 13 different filtering criteria with different parameters and thresholds were settled (Figure 4) .

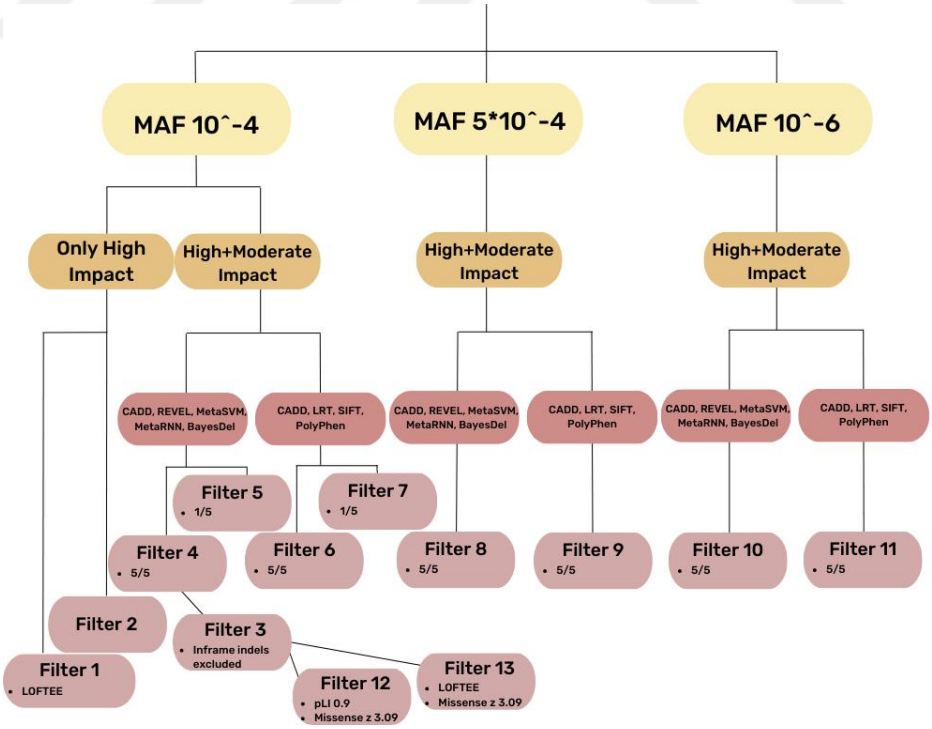

Figure 4. Filtering Criteria that Models were Built

These filtering criteria include predictions from several in silico pathogenicity prediction tools. These tools include LOFTEE, CADD, REVEL, MetaSVM, MetaRNN, LRT, SIFT, and PolyPhen (Version 2). It is important to note that even the tool with the highest accuracy has its Achilles heel, which means protein families in which the tool's predictions are unreliable. It had been proven that choosing optimal tools and pathogenicity thresholds depends on the protein family of interest and the structure of the data. In addition, even the tools that have the highest accuracy and sensitivity give solely a prediction about the pathogenicity of the variant, if there is no functional evidence (89).

The LOFTEE (Loss-Of-Function Transcript Effect Estimator) is a statistical method that is used to identify variants that are likely to cause loss-of-function (LoF) in a given transcript. The method utilizes a combination of functional annotations and variant-level metrics to estimate the likelihood that a given variant will cause loss-of-function in a particular transcript. The LOFTEE method uses a variety of functional annotations, including protein domain annotations, splice site annotations, and conservation scores, to identify variants that are likely to cause loss of function. In addition, the method uses variant-level metrics, such as the number of reads supporting the variant and the quality of the reads, to estimate the likelihood that a given variant will cause loss-of-function in a particular transcript and differentiates these LoF variants into 2 categories which are high confidence and low confidence (90).

The CADD (Combined Annotation-Dependent Depletion) tool is a sophisticated method that utilizes multiple annotations to assess the pathogenicity of genetic variants. This tool is based on the concept that deleterious genetic variants tend to be rare in the population, but are also conserved across species and are often located in functional regions of the genome. CADD uses various annotations, including DNA sequence conservation, protein structure, and functional genomics data, to score genetic variants based on their potential to cause disease. The tool assigns a single score to each variant, which can be used to rank variants according to their predicted pathogenicity. It has been extensively validated and outperforms other popular

prediction tools in terms of accuracy and sensitivity. It has been used to identify disease-causing variants in numerous genetic disorders, including cancer, neurological disorders, and developmental disorders (91).

The REVEL is a powerful tool that is based on a machine learning algorithm that has been trained on a large dataset of known pathogenic and benign variants. This algorithm uses a combination of features, including evolutionary conservation, functional annotation, and population frequency, to predict whether a variant is likely to be pathogenic or benign. It has been shown to have high accuracy in predicting the pathogenicity of variants across a wide range of genetic disorders. It has been particularly useful in the field of cancer genomics, where it has been used to identify novel driver mutations and predict the response of tumours to targeted therapies (92).

MetaSVM is one of the most powerful pathogenicity prediction tools. It uses support vector machine (SVM) which is a supervised machine learning algorithm that is commonly used in classification problems. It works by finding the optimal hyperplane that separates the data points of different classes, thus allowing for the accurate classification of new, unseen data. In the context of pathogenicity prediction, MetaSVM uses SVM to classify genetic variants as either pathogenic or benign based on their features (e.g. amino acid changes, conservation scores, etc.). The algorithm is trained on a large dataset of known pathogenic and benign variants, allowing it to learn patterns and make accurate predictions. One advantage of using SVM as the statistical logic for MetaSVM is its ability to handle high-dimensional data. This is particularly useful in genomics, where a large number of features (e.g. thousands of genetic variants) may be considered in pathogenicity prediction. SVM is also known for its ability to handle non-linear relationships between features, which can be important in predicting pathogenicity (93).

The MetaRNN is founded on a solid statistical framework that leverages machine learning algorithms to predict the pathogenicity of genetic variants. It employs a recurrent neural network (RNN) architecture that is well-suited for processing sequential data, such as genetic sequences. This architecture allows the

algorithm to take into account the temporal dependencies and context of the genetic information, which is critical for accurately predicting pathogenicity. Furthermore, the MetaRNN tool also incorporates a range of features that enhance its predictive power. These include sequence conservation scores, functional annotations, and phylogenetic information, among others. By integrating these features into the machine learning model, the tool can provide more comprehensive and accurate predictions of pathogenicity (94).

The likelihood ratio test (LRT) tool is based on the likelihood function, which compares the null hypothesis which claims that the variant is benign and the alternative hypothesis which suggests that the variant is pathogenic. It is used to evaluate the evidence in support of a variant being pathogenic. The likelihood ratio is calculated by dividing the likelihood of the alternative model by the likelihood of the null model. A high likelihood ratio indicates strong evidence in favour of the alternative hypothesis, while a low likelihood ratio supports the null hypothesis. Albeit the LRT is a powerful tool for evaluating the pathogenicity of genetic variants, it is important to note that it is not infallible. The accuracy of the LRT depends on the quality of the data used to construct the statistical models, as well as the assumptions made about the underlying biology. Additionally, the LRT is just one piece of evidence that should be considered when evaluating the pathogenicity of a variant. Other factors, such as functional assays and clinical data, should also be taken into account (95).

The SIFT (Sorting Intolerant From Tolerant) is a powerful computational tool that is widely used. At its core, the logic behind SIFT is based on the observation that functionally important amino acid residues in proteins tend to be conserved across different species, while non-functional residues can vary more widely. SIFT uses this principle to predict whether specific amino acid substitutions in a protein sequence are likely to be deleterious or benign. To do this, the tool first compares the amino acid of interest to a set of related sequences from other species. It then calculates a score based on how conserved the amino acid is across these sequences. If the score is below a certain threshold, SIFT predicts that the substitution is likely

to be deleterious. The statistical model behind SIFT has been extensively validated and has proved to be highly accurate in predicting the impact of amino acid substitutions on protein function. In addition to its accuracy, SIFT is also notable for its speed and scalability. It can analyze large datasets of protein sequences in a matter of minutes, making it a particularly useful tool for large-scale genomics projects (96).

Polyphen is a widely used pathogenicity prediction tool that uses the training set of known pathogenic and benign variants to build its prediction model. Polyphen uses The accuracy of Polyphen's predictions has been extensively validated through experimental studies and is widely accepted as a reliable tool for predicting the pathogenicity of genetic variants (97).

3.4 Rare Variant Association Analysis of Turkish PCOS Cohort

All of the produced data were analysed via SMMAT-S, SMMAT-O, SMMAT-B, and SMMAT-E, utilizing the GMMAT R package (Version 1.4.0) (88). Then, QQ plots of p-values of all analyses were obtained via the gaston R package (Version 1.5.9) (87). Then, the segregation of variants within families was examined. Finally, the most important candidate genes were selected and a relevant literature search was conducted. The functions of genes were checked from GeneCards (98).

3.5 Imputation

To be able to fill empty spaces caused by the missingness of the WES data with respect to haplotype blocks, a two-step imputation was conducted. Initially, the cohort data were imputed with Turkish Genome Panel by utilizing Minimac4 (99). To achieve this, the Turkish Genome Panel and the 1000 Genomes phase 3 data that were found in vcf format have been converted into m3vcf format with the help of Minimac3 (100). Subsequently, the resulting m3vcf files were converted into msav files. Then, the cohort data were phased with the Eagle (Version 2.4.1). The phased vcf was then imputed first with the Turkish Genome panel and then with the 1000 Genomes dataset.

3.6 Polygenic Risk Score (PRS) Calculation

After the imputation step, family members of PCOS patients and control individuals were excluded to prevent spurious LD. Then, the resulting file was converted into PLINK format, which allows for efficient handling of large-scale genetic data and compatibility with various analysis software. The next step involved the calculation of PRS, which are used to predict an individual's risk of developing a particular disease based on their genetic makeup. This was done using classical p-value thresholding and clumping method which involves setting a significance level for genetic variants and only including those that pass the threshold in the risk score calculation. As a p-value threshold, 0.5 was chosen. The clumping method, on the other hand, involves grouping variants that are in linkage disequilibrium with each other and selecting the most significant variant from each group for the risk score calculation. The analysis then repeated with both mostly prioritized genes in our association analyses and top genes of GWAS.

4 RESULTS

4.1 Quality Statistics

Graphs of variant quality, variant depth, mean depth of individual, variant missingness, mean missingness of individual, heterozygosity, and minor allele frequency (MAF) were obtained (Figure 5, 6).

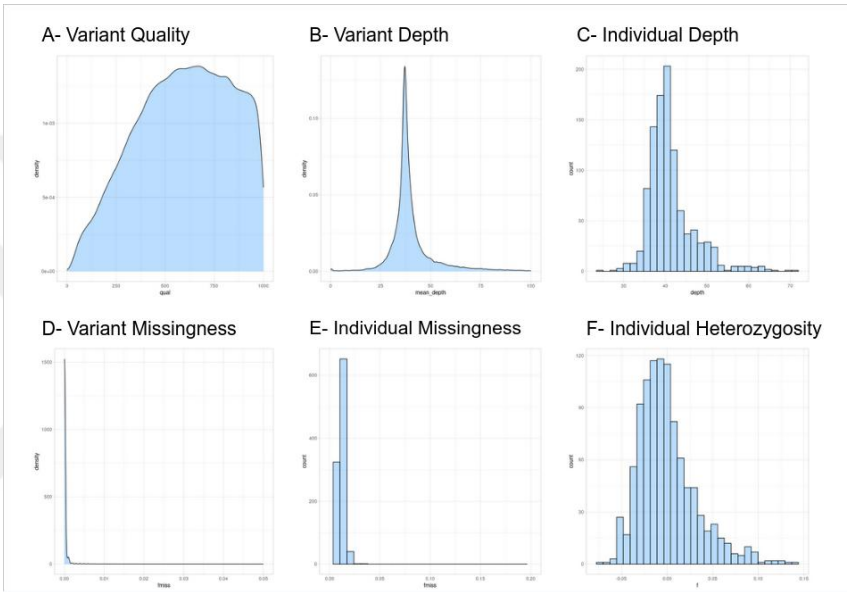

Figure 5. Quality Parameter Graphs of Cohort Data

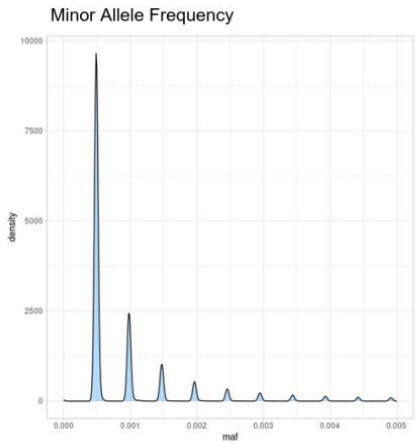

Figure 6. Minor Allele Frequency Graph of Cohort

4.2 Principal Component Analysis

To be able to observe if there is an outlier in the population and examine the structure of the cohort, Principal Component Analysis (PCA) was conducted (Figure 7, 8).

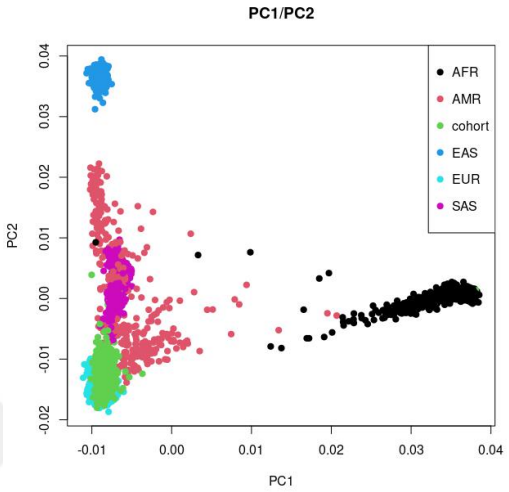

Figure 7. 2D-PCA Result of Cohort Data that is Merged with 1kG Data

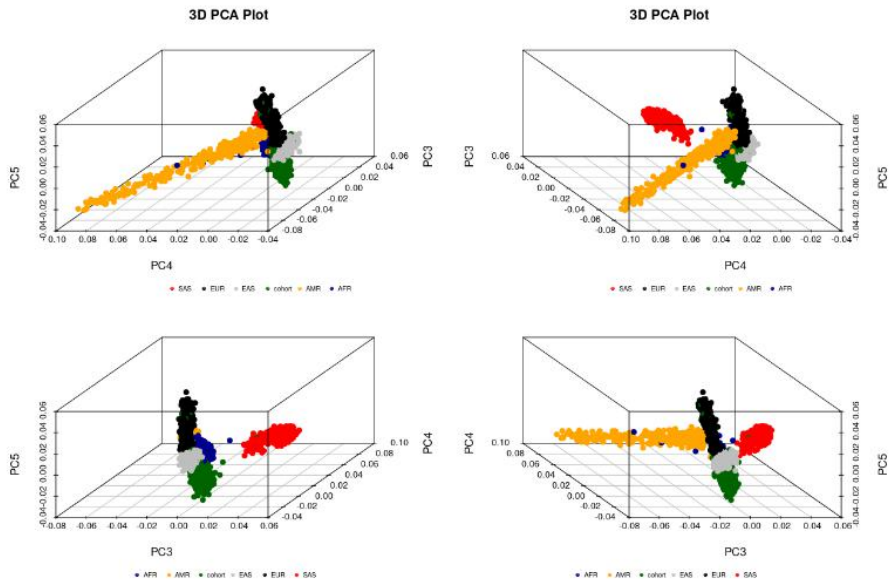

Figure 8. 3D-PCA Result of Cohort Data that is Merged with 1kG Data

4.3 MAF Threshold Optimization with SKAT Package

To optimize the MAF threshold, data produced by every 6 thresholds were analyzed via binary functions of SKAT-O and Burden test (Figure 9, 10).

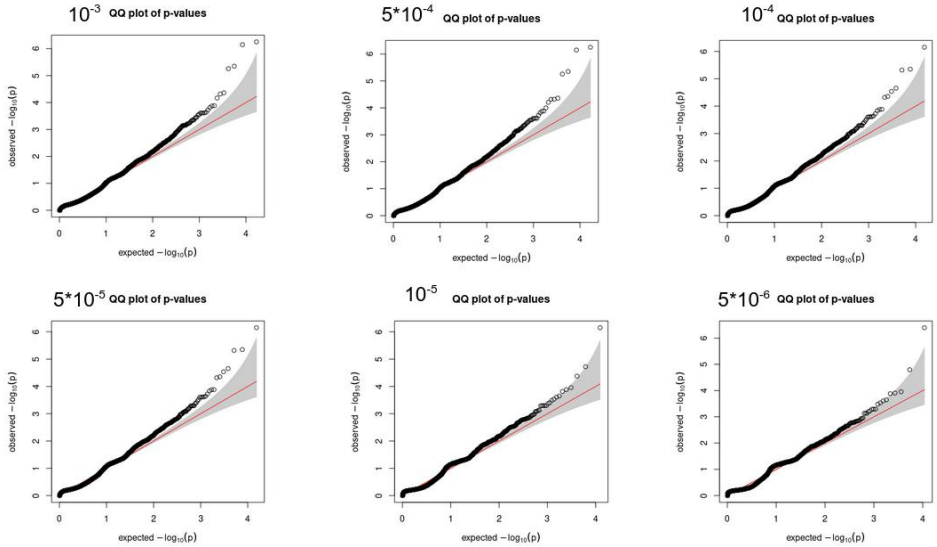

Figure 9. QQ Plots of p-values of SKAT-O with Synonymous Variants Filtered with Different in-house MAF Thresholds

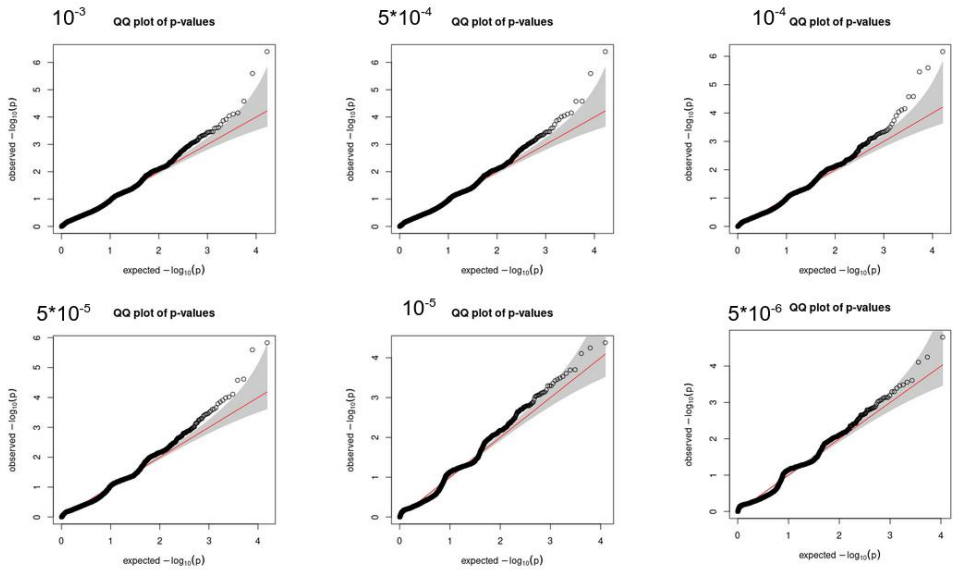

Figure 10. QQ Plots of p-values of Burden Test with Synonymous Variants Filtered with Different in-house MAF Thresholds

4.4 MAF Threshold Optimization with GMMAT Package

To compare the classical kernel methods with SMMAT, the synonymous variants were also analysed via SMMAT-S, SMMAT-O, SMMAT-B, and SMMAT-E (Figure 11-16).

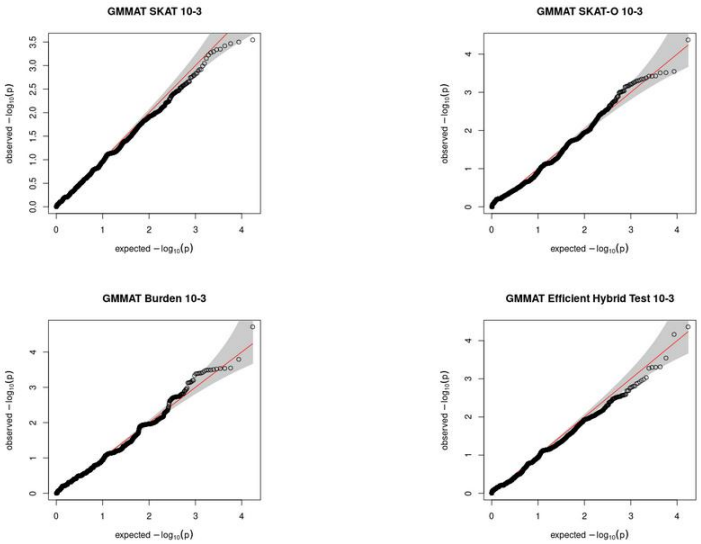

Figure 11. QQ Plots of p-values of SMMAT Models with Synonymous Variants Filtered with in-house MAF Threshold of 10^{-3}

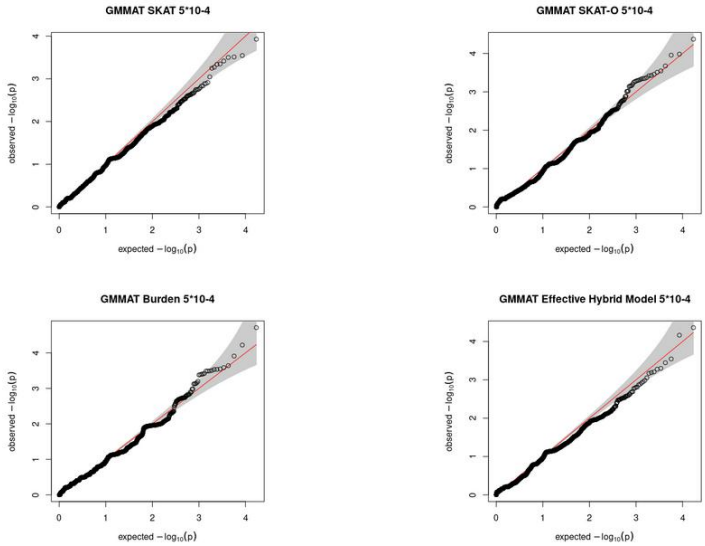

Figure 12. QQ Plots of p-values of SMMAT Models with Synonymous Variants Filtered with in-house MAF Threshold of $5*10^{-4}$

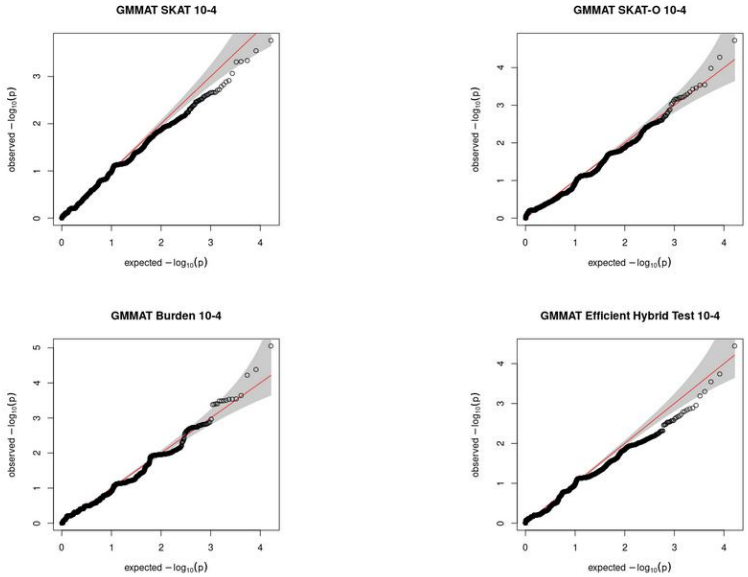

Figure 13. QQ Plots of p-values of SMMAT Models with Synonymous Variants Filtered with in-house MAF Threshold of 10^{-4}

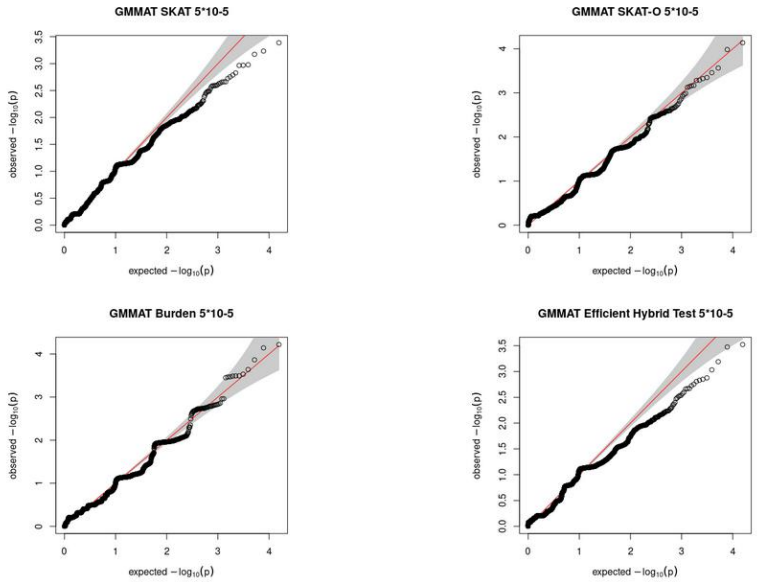

Figure 14. QQ Plots of p-values of SMMAT Models with Synonymous Variants Filtered with in-house MAF Threshold of $5*10^{-5}$

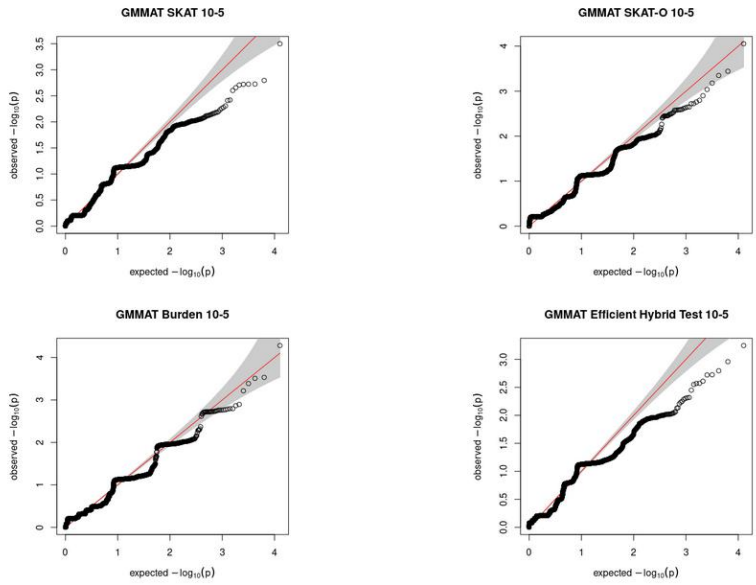

Figure 15. QQ Plots of p-values of SMMAT Models with Synonymous Variants Filtered with in-house MAF Threshold of 10^{-5}

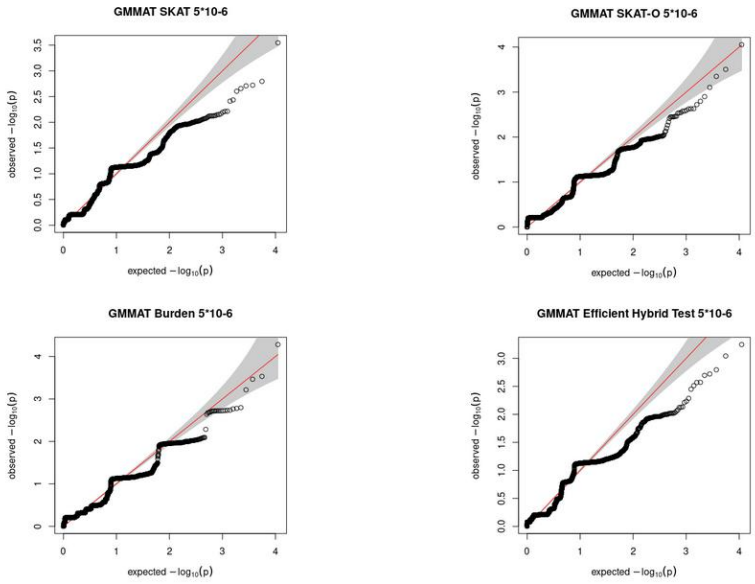

Figure 16. QQ Plots of p-values of SMMAT Models with Synonymous Variants Filtered with in-house MAF Threshold of $5*10^{-6}$

4.5 SMMAT Results

4.5.1 Filter 1

Data that had been filtered via Filter 1, which involves high confidence LoF variants were analysed with all 4 methods. QQ plots are given at Figure 17. The top associated genes are given at Table 4, and segregations of variants in these genes are given at Table 5 and Table 6.

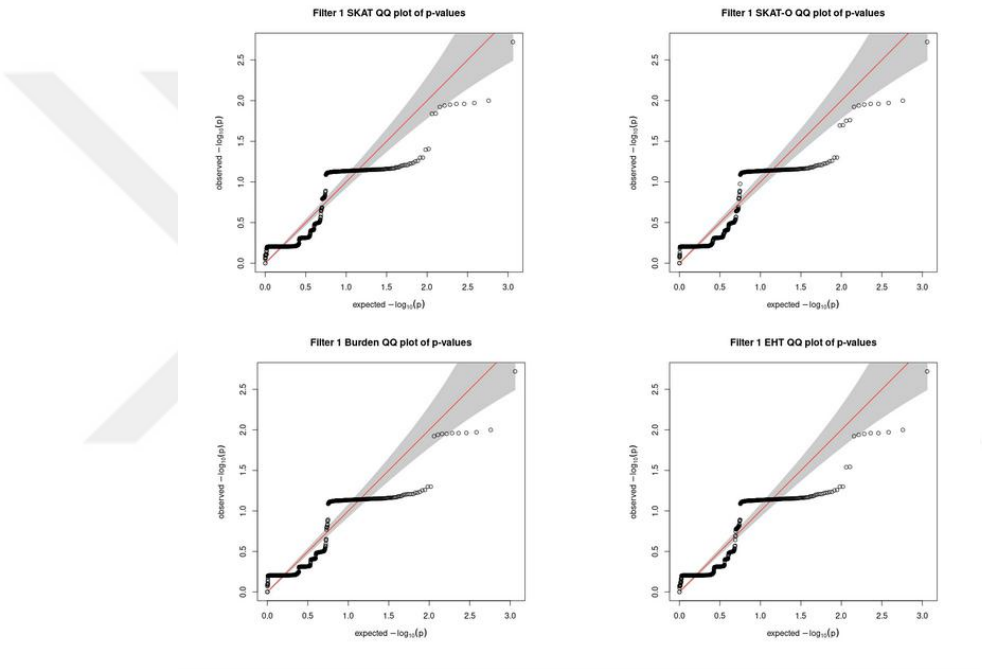

Figure 17. QQ Plots of p-values of SMMAT Models with Filter 1

Table 4. Top 2 p-values of SMMAT Models with Filter 1

group	B.pval	group	S.pval	group	O.pval	group	E.pval
DLG2	0.001901099	DLG2	0.001901099	DLG2	0.001901099	DLG2	0.001901099
ZNF534	0.010015977	ZNF534	0.010015977	ZNF534	0.010015977	ZNF534	0.010015977

Table 5. Segregation of Variants in DLG2

Variant	Family ID	Individual ID	Zygosity	PCOS
c.742-1G>T	PCOS-035	14-097	0/1	+
		14-098	0/1	+
		14-099	0/1	+

Table 6. Segregation of Variants in ZNF534

Variant	Family ID	Individual ID	Zygosity	PCOS
c.311-2A>G	OB-0340	17-323	0/1	+
		17-324	0/1	+

4.5.2 Filter 2

Data that had been filtered via Filter 2, which involves only LoF variants were analysed with all 4 methods. QQ plots are given at Figure 18. The top associated genes are given at Table 7, and segregations of variants in these genes are given at Table 8 and Table 9.

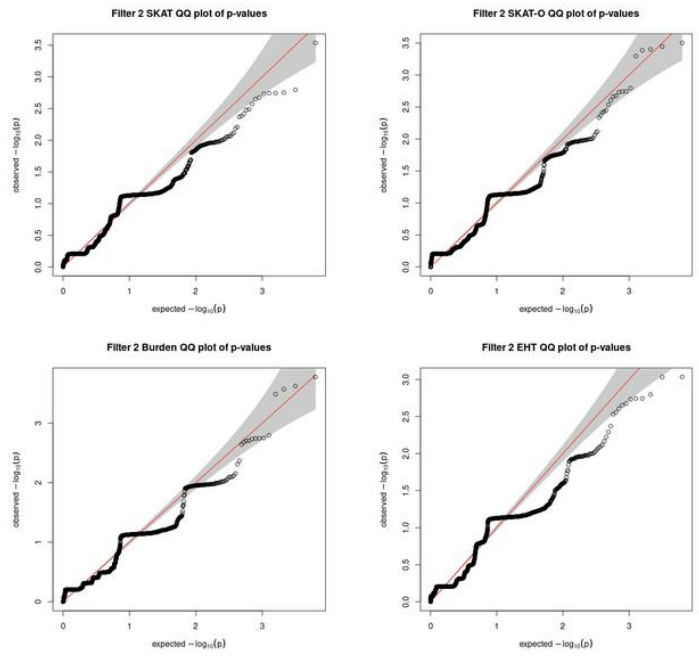

Figure 18. QQ Plots of p-values of SMMAT Models with Filter 2

Table 7. Top 2 p-values of SMMAT Models with Filter 2

group	B.pval	group	S.pval	group	O.pval	group	E.pval
COL20A1	0.000167944	PCDHGA11	0.000290625	PCDHGA11	0.000314317	COL20A1	0.000925419
C19orf44	0.000238777			COL20A1	0.000359342	PCDHGA11	0.00093057

Table 8. Segregation of Variants in COL20A1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.756delG:p.N255Tfs*8	OB-0026	13-168	0/1	M
	OB-0037	13-422	0/1	-
c.2209delT:p.S737Pfs*3	PCOS-012	14-057	0/1	+
		14-058	0/1	+
		14-061	0/1	+
		14-065	0/1	-
c.756delG:p.N255Tfs*8	OB-0146	14-079	0/1	+
		14-180	0/1	M
	PCOS-092	16-208	0/1	+
	PCOS-106	16-429	0/1	+
c.C1609T:p.Q537X	PCOS-129	16-453	0/1	+
c.C1618T:p.R540X	PCOS-130	16-454	0/1	+
c.C753G:p.Y251X	PCOS-131	16-455	0/1	+
c.3482delG:p.G1162Afs*40	OB-0311	17-290	0/1	-
c.756delG:p.N255Tfs*8	OB-0330	17-313	0/1	+
c.1023delC:p.L342Cfs*44	OB-0521	18-062	0/1	-
c.2358+1G>A	OB-0909	18-497	0/1	M

Table 9. Segregation of Variants in PCDHGA11

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1002delG:p.M334Ifs*2	OB-0154	15-283	0/1	+
	OB-0160	16-553	0/1	+
	OB-0348	17-339	0/1	+
	PCOS-097	16-213	0/1	+
c.1563delC:p.F521Lfs*9	OB-0493	18-033	0/1	-
c.1928dupT:p.V644Sfs*40	OB-0177	16-144	0/1	-
	OB-0217	15-442	0/1	M
	OB-0885	18-469	0/1	-
c.C1206G:p.Y402X	OB-0369	17-362	0/1	-
c.C1371A:p.Y457X	OB-0555	18-104	0/1	-

4.5.3 Filter 3

Data that had been filtered via Filter 3, which involves variants which are predicted to be pathogenic by all of CADD, REVEL, MetaSVM, MetaRNN, and BayesDel; and those which are not non-frameshift indels were analysed with all 4 methods. QQ plots are given at Figure 19. The top associated genes are given at Table 10, and segregations of variants in these genes are given at Table 11 and 12.

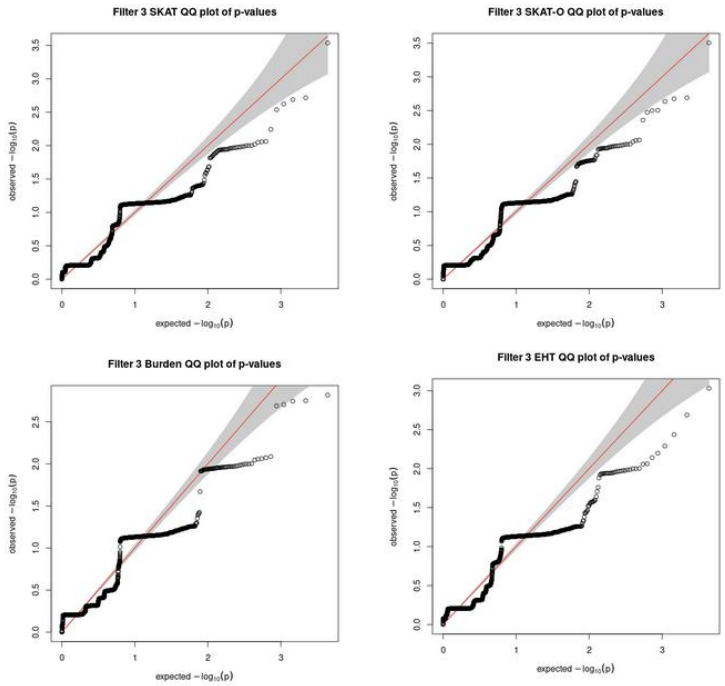

Figure 19. QQ Plots of p-values of SMMAT Models with Filter 3

Table 10. Top 2 p-values of SMMAT Models with Filter 3

group	B.pval	group	S.pval	group	O.pval	group	E.pval
PCDHGA11	0.001972835	PCDHGA11	0.000290625	PCDHGA11	0.000314317	PCDHGA11	0.00093057
		CRYBG1	0.001926974	CRYBG1	0.002314522		

Table 11. Segregation of Variants in PCDHGA11

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1002delG:p.M334Ifs*2	OB-0154	15-283	0/1	+
	OB-0160	16-553	0/1	+
	OB-0348	17-339	0/1	+
	PCOS-097	16-213	0/1	+
c.1563delC:p.F521Lfs*9	OB-0493	18-033	0/1	-

Table 12. Segregation of Variants in CRYBG1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.590_591insAGAGCTGGGC:p.A203Gfs*11	OB-0301	17-280	0/1	-
c.4646_4647insAT:p.G1551Mfs*28	PCOS-011	14-078	0/1	+
		14-208	0/1	+
	PCOS-035	14-098	0/1	+

4.5.4 Filter 4

Data that had been filtered via Filter 4, which involves variants which are predicted to be pathogenic by all of CADD, REVEL, MetaSVM, MetaRNN, and BayesDel were analysed with all 4 methods. QQ plots are given at Figure 20. The top associated genes are given at Table 13, and segregations of variants in these genes are given at Table 14-18.

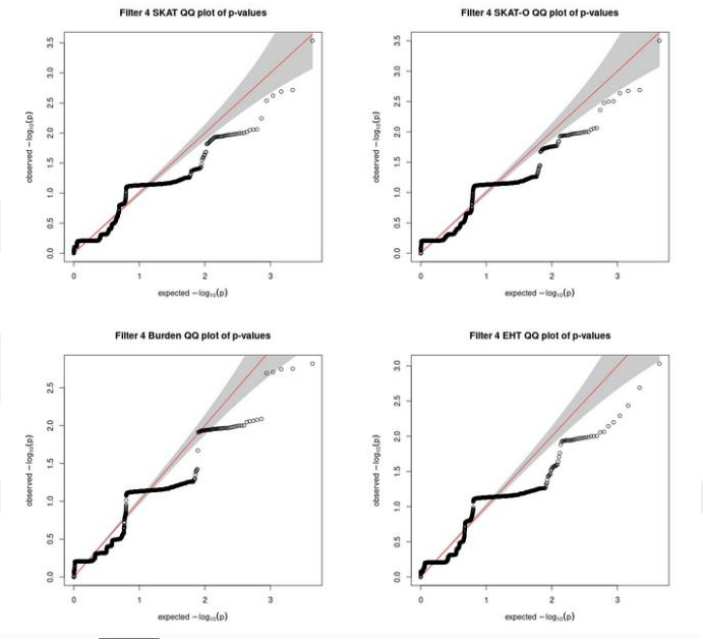

Figure 20. QQ Plots of p-values of SMMAT Models with Filter 4

Table 13. Top 2 p-values of SMMAT Models with Filter 4

group	B.pval	group	S.pval	group	O.pval	group	E.pval
C19orf44	0.00152099	PCDHGA11	0.000290625	PCDHGA11	0.000314317	PCDHGA11	0.00093057
TEX15	0.00177192 5	CRYBG1	0.00192697 4	CWC27	0.00204473 3	CWC27	0.00204 4733

Table 14. Segregation of Variants in PCDHGA11

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1002delG:p.M334Ifs*2	OB-0154	15-283	0/1	+
	OB-0160	16-553	0/1	+
	OB-0348	17-339	0/1	+
	PCOS-097	16-213	0/1	+
c.1563delC:p.F521Lfs*9	OB-0493	18-033	0/1	-

Table 15. Segregation of Variants in C19orf44

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1707_1709del:p.Q571del	OB-0795	18-361	0/1	+
c.220dupC:p.R75Qfs*6	PCOS-004	14-067 14-068	0/1 0/1	+ +

Table 16. Segregation of Variants in CRYBG1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.590_591insAGAGCTGGGC:p.A203Gfs*11	OB-0301	17-280	0/1	-
c.4646_4647insAT:p.G1551Mfs*28	PCOS-011 PCOS-035	14-078 14-208 14-098	0/1 0/1 0/1	+ + +

Table 17. Segregation of Variants in CWC27

Variant	Family ID	Individual ID	Zygosity	PCOS
c.917_919del:p.K308del	OB-0782 PCOS-028 PCOS-116	18-344 16-450 16-440	0/1 0/1 0/1	+ + +

Table 18. Segregation of Variants in TEX15

Variant	Family ID	Individual ID	Zygosity	PCOS
c.4219_4221del:p.E1407del	OB-0223 OB-0508	17-142 18-049	0/1 0/1	M +
c.2734_2748del:p.I912_E916del	OB-0511 OB-0910	18-052 18-498	0/1 0/1	+ +
c.5208delC:p.L1737Wfs*9	PCOS-093	16-209	0/1	+

4.5.5 Filter 5

Data that had been filtered via Filter 5, which involves variants which are predicted to be pathogenic by at least one of CADD, REVEL, MetaSVM, MetaRNN, and BayesDel were analysed with all 4 methods. QQ plots are given at Figure 21. The top associated genes are given at Table 19, and segregations of variants in these genes are given at Table 20-23.

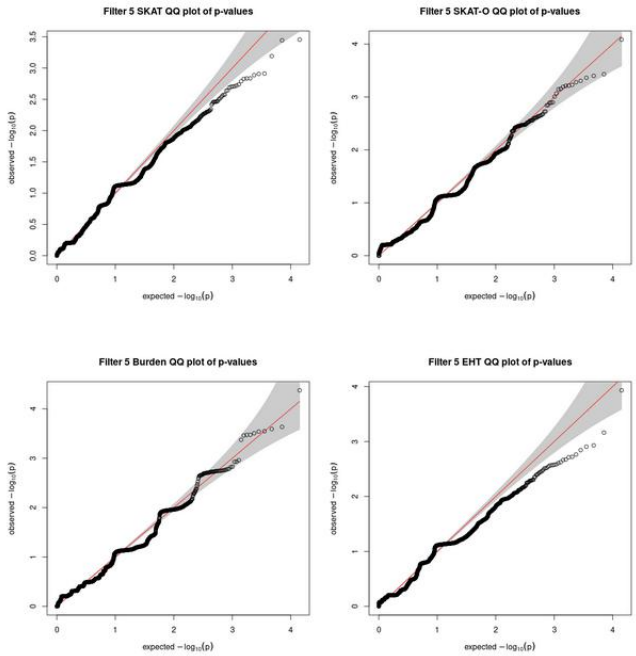

Figure 21. QQ Plots of p-values of SMMAT Models with Filter 5

Table 19. Top 2 p-values of SMMAT Models with Filter 5

group	B.pval	group	S.pval	group	O.pval	group	E.pval
COL20A1	4.22E-05	TRPM5	0.00035122	COL20A1	8.24E-05	COL20A1	0.000117151
AK5	0.000231714	PCDHGA11	0.000360555				

Table 20. Segregation of Variants in COL20A1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.C3572T:p.P1191L	OB-0080	14-132	0/1	+
c.G564T:p.W188C	OB-0208	16-327	0/1	-
	OB-0874	18-460	0/1	-
	OB-0946	18-552	0/1	-
	PCOS-061	16-177	0/1	+
c.3482delG:p.G1162Afs*40	OB-0311	17-290	0/1	-
c.G2728A:p.E910K	OB-0361	17-353	0/1	+
c.C2494G:p.P832A	OB-0801	18-367	0/1	+
c.G1321C:p.E441Q	OB-0867	18-442	0/1	M
c.2358+1G>A	OB-0909	18-497	0/1	M
c.2209delT:p.S737Pfs*3	PCOS-012	14-057	0/1	+
		14-058	0/1	+
		14-061	0/1	+
		14-065	0/1	-
c.763delA:p.N255Tfs*8	PCOS-106	16-429	0/1	+
c.C1609T:p.Q537X	PCOS-129	16-453	0/1	+

Table 21. Segregation of Variants in TRPM5

Variant	Family ID	Individual ID	Zygosity	PCOS
c.T2227G:p.F743V	LEAN-001	13-446	0/1	+
		13-447	0/1	+
c.825_841del:p.L278Tfs*45	OB-0135	14-203	0/1	-
		14-205	0/1	-
		14-206	0/1	-
c.G482T:p.R161L	OB-0256	17-196	0/1	-
c.G3178A:p.E1060K	OB-0406	17-413	0/1	M
	OB-0428	17-447	0/1	-
	OB-0760	18-322	0/1	+
	OB-0779	18-341	0/1	M
c.825_841del:p.L278Tfs*45	OB-0874	18-460	0/1	-
c.A1618G:p.M540V	OB-0945	18-551	0/1	+
c.T2227G:p.F743V	PCOS-136	16-460	0/1	+
c.T2227G:p.F743V	LEAN-001	13-446	0/1	+
		13-447	0/1	+

Table 22. Segregation of Variants in AK5

Variant	Family ID	Individual ID	Zygosity	PCOS
c.C821T:p.A274V	OB-0209	16-348	0/1	+
c.T1208C:p.M403T	PCOS-055	16-170	0/1	+
c.A1310G:p.Y437C	PCOS-143	16-467	0/1	+
c.G1504A:p.V502M	PCOS-147	16-471	0/1	+

Table 23. Segregation of Variants in PCDHGA11

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1002delG:p.M334Ifs*2	OB-0154	15-283	0/1	+
	OB-0160	16-553	0/1	+
	OB-0348	17-339	0/1	+
	PCOS-097	16-213	0/1	+
c.1563delC:p.F521Lfs*9	OB-0493	18-033	0/1	-
c.A104G:p.Y35C	OB-0558	18-107	0/1	-
c.A1442C:p.D481A	OB-0569	18-119	0/1	-
c.A2071G:p.T691A	OB-0808	18-374	0/1	M
c.A2543G:p.H848R	OB-0178	16-219	0/1	M
c.A637T:p.T213S	OB-0170	14-193	0/1	-
c.A758G:p.E253G	OB-0769	18-331	0/1	-
	PCOS-096	16-212	0/1	+
c.C1206G:p.Y402X	OB-0369	17-362	0/1	-
c.C1262G:p.T421R	OB-0971	18-619	0/1	M
c.C1371A:p.Y457X	OB-0555	18-104	0/1	-
c.C1376A:p.A459D	OB-0465	17-495	0/1	M
c.C1693T:p.P565S	OB-0124	16-271	0/1	M
c.C1371A:p.Y457X	OB-0555	18-104	0/1	-
c.C1376A:p.A459D	OB-0465	17-495	0/1	M
c.C395T:p.P132L	OB-0558	18-107	0/1	-
c.C514T:p.L172F	OB-0006	13-414	0/1	-
c.C908A:p.T303K	OB-0107	14-010	0/1	-
c.G1018C:p.V340L	OB-0253	17-185	0/1	-
		17-192	0/1	-
		17-193	0/1	-

Table 23. Segregation of Variants in PCDHGA11 (cont.)

c.G2000T:p.S667I	OB-0378	17-376	0/1	M
c.G2536C:p.D846H	OB-0396	17-400	0/1	-
c.G2567A:p.R856K	OB-0750 PCOS-026	18-308 14-037	0/1 0/1	- +
c.G2570C:p.G857A	PCOS-153	16-477	0/1	+
c.G2617A:p.G873S	OB-0929	18-517	0/1	-
c.G2668A:p.V890M	ETM-032	13-110	0/1	-
c.T1468C:p.S490P	OB-0991	18-890	0/1	-
c.T1856C:p.V619A	OB-0080	14-130 14-134	0/1 0/1	- M
c.T2534C:p.L845P	OB-0026	13-166	0/1	M
c.T539G:p.L180R	OB-0170	14-193	0/1	-

4.5.6 Filter 6

Data that had been filtered via Filter 6, which involves variants which are predicted to be pathogenic by all of CADD, PolyPhen, SIFT, and LRT were analysed with all 4 methods. QQ plots are given at Figure 22. The top associated genes are given at Table 24, and segregations of variants in these genes are given at Table 25 and Table 26.

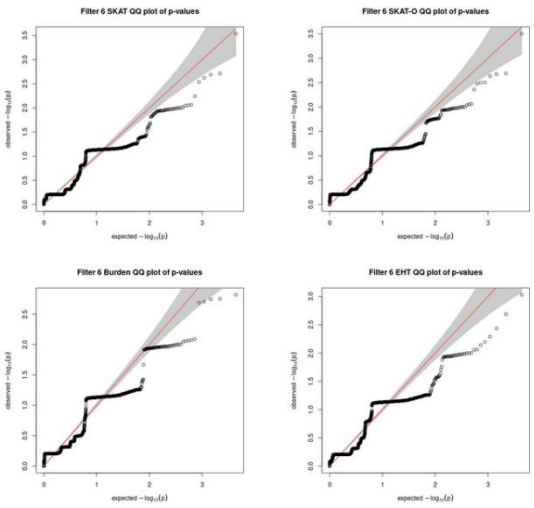

Figure 22. QQ Plots of p-values of SMMAT Models with Filter 6

Table 24. Top 2 p-values of SMMAT Models with Filter 6

group	B.pval	group	S.pval	group	O.pval	group	E.pval
C19orf44	0.00152099	PCDHGA11	0.000290625	PCDHGA11	0.000314317	PCDHGA11	0.00093057
TEX15	0.001771925	CRYBG1	0.001926974	CWC27	0.002044733	CWC27	0.002044733

Table 25. Segregation of Variants in PCDHGA11

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1002delG:p.M334Ifs*2	OB-0154 OB-0160 OB-0348 PCOS-097	15-283 16-553 17-339 16-213	0/1 0/1 0/1 0/1	+ + + +
c.1563delC:p.F521Lfs*9	OB-0493	18-033	0/1	-

Table 26. Segregation of Variants in CRYBG1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.590_591insAGAGCTGGGC:p.A203Gfs*11	OB-0301	17-280	0/1	-
c.4646_4647insAT:p.G1551Mfs*28	PCOS-011 PCOS-035	14-078 14-208 14-098	0/1 0/1 0/1	+ + +

4.5.7 Filter 7

Data that had been filtered via Filter 7, which involves variants which are predicted to be pathogenic by at least one of of CADD, PolyPhen, SIFT, and LRT were analysed with all 4 methods. QQ plots are given at Figure 23. The top associated genes are given at Table 27, and segregations of variants in these genes are given at Table 28-34.

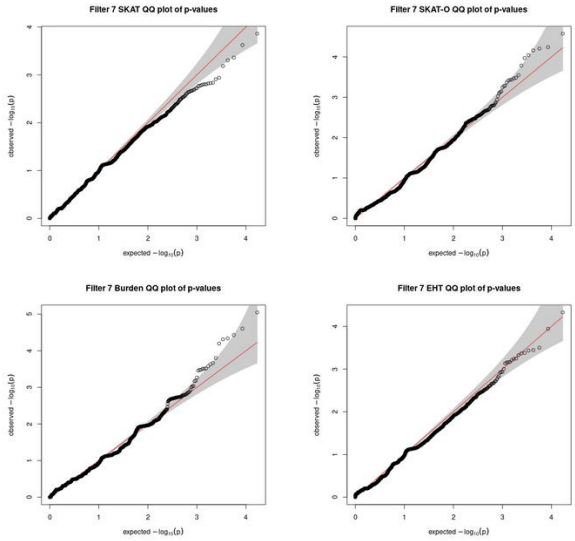

Figure 23. QQ Plots of p-values of SMMAT Models with Filter 7

Table 27. Most important p-values of SMMAT Models with Filter 7

group	B.pval	group	S.pval	group	O.pval	group	E.pval
THAP8	9.00E-06	ST8SIA6	0.000137825	THAP8	2.66E-05	ARMC3	4.71E-05
ARMC3	2.48E-05			ARMC3	5.74E-05		
OR4K14	3.73E-05			OR4K14	6.22E-05		
ADAT1	4.60E-05			POM121L12	6.89E-05		
POM121L12	4.88E-05			ADAT1	9.09E-05		
SELENOP	6.40E-05						

Table 28. Segregation of Variants in THAP8

Variant	Family ID	Individual ID	Zygosity	PCOS
c.612dupT:p.D205*	PCOS-012	14-058	0/1	+
c.G514A:p.G172R	PCOS-133	16-457	0/1	+
c.C220T:p.R74C	PCOS-137	16-461	0/1	+
c.C78A:p.F26L	PCOS-001	16-171	0/1	+
c.A62G:p.N21S	PCOS-090	16-206	0/1	+
c.C22T:p.P8S	OB-0807	18-373	0/1	+

Table 29. Segregation of Variants in ARMC3

Variant	Family ID	Individual ID	Zygosity	PCOS
c.670_675del:p.S224_R225del	OB-0080 OB-0449	14-132 14-134 17-478	0/1 0/1 0/1	+ M -
c.C1801T:p.R601X	OB-0403 PCOS-106	17-410 16-429	0/1 0/1	+ +
c.C56A:p.P19Q	OB-0455	17-485	0/1	-
c.C1003A:p.L335I	OB-0495	18-035	0/1	-
c.C56A:p.P19Q	OB-0553	18-102	0/1	-
c.C1210T:p.R404X	OB-0742	18-300	0/1	+
c.T1827G:p.I609M	OB-1058	18-1037	0/1	+
c.A1240G:p.K414E	PCOS-012	14-061 14-066	0/1 0/1	+ +
c.T752C:p.I251T	PCOS-067	16-183	0/1	+
c.C349G:p.L117V	PCOS-099	16-215	0/1	+

Table 30. Segregation of Variants in OR4K14

Variant	Family ID	Individual ID	Zygosity	PCOS
c.C343A:p.L115I	OB-0365	17-357	0/1	+
c.G686A:p.R229H	OB-0417	17-433	0/1	+
c.G917A:p.R306Q	PCOS-004	14-067 14-068	0/1 0/1	+ +
c.A649G:p.I217V	PCOS-109	16-433	0/1	+

Table 31. Segregation of Variants in ADAT1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.C1007T:p.T336I	LEAN-001	13-446	0/1	+
		13-447	0/1	+
		13-153	0/1	-
	OB-0066			
c.G630C:p.M210I	OB-0214	16-368	0/1	+
c.1267_1269del:p.K423del	OB-0247	17-175	0/1	-
	OB-0319	17-302	0/1	-
c.T112C:p.W38R	OB-0345	17-330	0/1	+
c.G10A:p.A4T	OB-0381	17-384	0/1	-
c.1267_1269del:p.K423del	OB-0975	18-860	0/1	-
c.G13A:p.D5N	PCOS-012	14-057	0/1	+
		14-058	0/1	+
		17-165	0/1	-
	OB-0240			
c.G608A:p.R203H	PCOS-086	16-202	0/1	+
c.C652T:p.R218X	PCOS-108	16-432	0/1	+

Table 32. Segregation of Variants in POM121L12

Variant	Family ID	Individual ID	Zygosity	PCOS
c.170dupC:p.L58Pfs*88	OB-0330	17-313	0/1	+
c.A292G:p.R98G	PCOS-084	16-200	0/1	+
c.C208T:p.R70C	LEAN-001	13-446	0/1	+
	OB-0401	17-408	0/1	-
	OB-0976	18-862	0/1	-
	PCOS-085	16-201	0/1	+
c.G329C:p.R110P	PCOS-129	16-453	0/1	+
c.G488T:p.R163L	PCOS-151	16-475	0/1	+
c.G52A:p.A18T	PCOS-135	16-459	0/1	+
c.G82C:p.A28P	OB-0771	18-333	0/1	-

Table 33. Segregation of Variants in SELENOP

Variant	Family ID	Individual ID	Zygosity	PCOS
c.G416+1A	OB-0133	16-130	0/1	+
c.T898C:p.Ter300Argext*?	OB-0949	18-560	0/1	+
c.G422A:p.Gly141Asp	PCOS-035	14-097	0/1	+
		14-098	0/1	+
c.C610T:p.His204Tyr	PCOS-087	16-203	0/1	+

Table 34. Segregation of Variants in ST8SIA6

Variant	Family ID	Individual ID	Zygosity	PCOS
c.753dupA:p.A252Sfs*40	OB-0022	13-472	0/1	M
		13-477	0/1	-
		13-478	0/1	-
		13-479	0/1	-
c.A1045C:p.T349P	OB-0063	13-389	0/1	M
	OB-0118	14-147	0/1	M
		14-148	0/1	-
		15-122	0/1	M
	OB-0167	14-190	0/1	-

Table 34. Segregation of Variants in ST8SIA6 (cont.)

c.A280T:p.K94X	OB-0214	16-368	0/1	+
c.C763G:p.L255V	PCOS-059	16-175	0/1	+
c.G885T:p.K295N	OB-0319	17-302	0/1	-
c.T161C:p.L54P	OB-0109	16-405 16-408	0/1 0/1	- M
c.T307C:p.Y103H	OB-0445	17-474	0/1	-
c.T592C:p.S198P	OB-0119	16-126	0/1	-
c.T890A:p.L297Q	PCOS-006 PCOS-028&126 PCOS-116 PCOS-123 PCOS-135	14-056 16-450 16-440 16-447 16-459	0/1 0/1 0/1 0/1 0/1	+ + + + +
c.T929C:p.F310S	OB-0222	17-141	0/1	M

4.5.8 Filter 8

Data that had been filtered via Filter 8, which involves variants which are predicted to be pathogenic by all of CADD, REVEL, MetaSVM, MetaRNN, and BayesDel and have a MAF smaller than $5*10^{-4}$ were analysed with all 4 methods. QQ plots are given at Figure 24. The top associated genes are given at Table 35, and segregations of variants in these genes are given at Table 36-39.

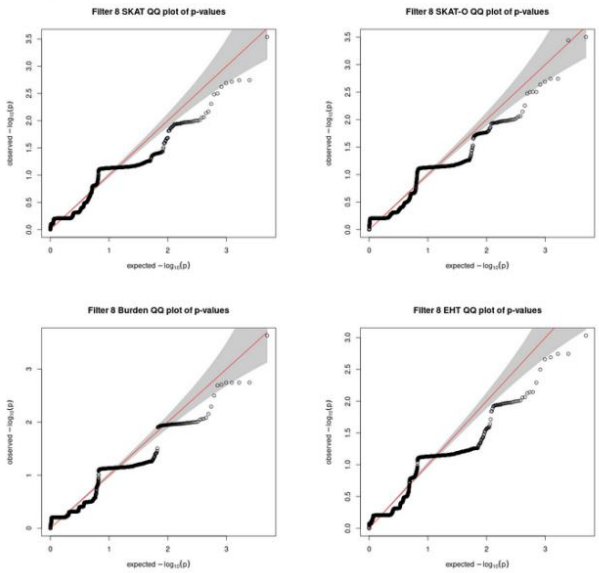

Figure 24. QQ Plots of p-values of SMMAT Models with Filter 8

Table 35. Top 2 p-values of SMMAT Models with Filter 8

group	B.pval	group	S.pval	group	O.pval	group	E.pval
C19orf44	0.000234816	PCDHGA11	0.000290625	PCDHGA11	0.000314317	PCDHGA11	0.00093057
OGFOD3	0.001800507	PDHA2	0.001802189	C19orf44	0.00036369	PDHA2	0.001802189

Table 36. Segregation of Variants in PCDHGA11

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1002delG:p.M334Ifs*2	OB-0154	15-283	0/1	+
	OB-0160	16-553	0/1	+
	OB-0348	17-339	0/1	+
	PCOS-097	16-213	0/1	+
c.1563delC:p.F521Lfs*9	OB-0493	18-033	0/1	-
c.1928dupT:p.V644Sfs*40	OB-0177	16-144	0/1	-
	OB-0217	15-442	0/1	M
	OB-0885	18-469	0/1	-

Table 37. Segregation of Variants in C19orf44

Variant	Family ID	Individual ID	Zygosity	PCOS
c.548_549del:p.S185Pfs*2	OB-0511	18-052	0/1	+
c.1707_1709del:p.Q571del	OB-0795	18-361	0/1	+
c.220dupC:p.R75Qfs*6	PCOS-004	14-067	0/1	+
		14-068	0/1	+

Table 38. Segregation of Variants in OGFOD3

Variant	Family ID	Individual ID	Zygosity	PCOS
c.518_520del:p.F173del	OB-0784	18-346	0/1	+
c.838_840del:p.F280del	OB-0945	18-551	0/1	+

Table 39. Segregation of Variants in PSENEN

Variant	Family ID	Individual ID	Zygosity	PCOS
c.-58-1G>A	PCOS-093	16-209	0/1	+
	PCOS-149	16-473	0/1	+
	PCOS-151	16-475	0/1	+

4.5.9 Filter 9

Data that had been filtered via Filter 9, which involves variants which are predicted to be pathogenic by all of CADD, PolyPhen, SIFT, and LRT and have a MAF smaller than $5*10^{-4}$ were analysed with all 4 methods. QQ plots are given at Figure 25. The top associated genes are given at Table 40, and segregations of variants in these genes are given at Table 41 and Table 42.

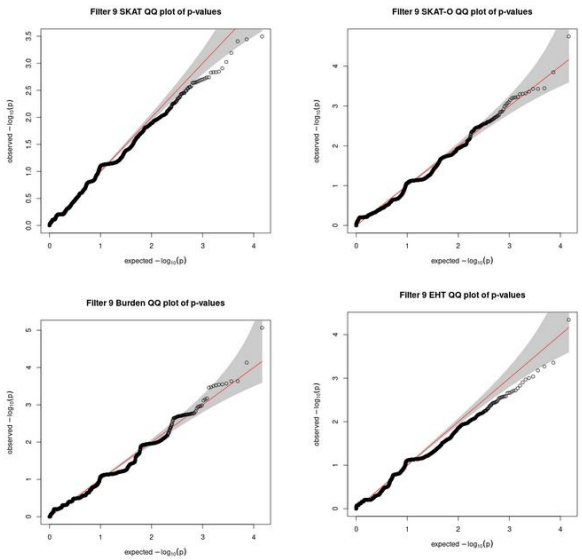

Figure 25. QQ Plots of p-values of SMMAT Models with Filter 9

Table 40. Top 2 p-values of SMMAT Models with Filter 9

group	B.pval	group	S.pval	group	O.pval	group	E.pval
COL20A1	8.61E-06	TRPM5	0.000320239	COL20A1	1.80E-05	COL20A1	4.55E-05
UTP6	7.37E-05	PCDHGA11	0.000360555				

Table 41. Segregation of Variants in COL20A1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.C3572T:p.P1191L	OB-0080	14-132	0/1	+
c.G564T:p.W188C	OB-0208	16-327	0/1	-
	OB-0874	18-460	0/1	-
	OB-0946	18-552	0/1	-
	PCOS-061	16-177	0/1	+
c.3482delG:p.G1162Afs*40	OB-0311	17-290	0/1	-
c.G2728A:p.E910K	OB-0361	17-353	0/1	+
c.C2494G:p.P832A	OB-0801	18-367	0/1	+
c.G1321C:p.E441Q	OB-0867	18-442	0/1	M
c.2358+1G>A	OB-0909	18-497	0/1	M
c.2209delT:p.S737Pfs*3	PCOS-012	14-057	0/1	+
		14-058	0/1	+
		14-061	0/1	+
		14-065	0/1	-
c.763delA:p.N255Tfs*8	PCOS-106	16-429	0/1	+
c.C1609T:p.Q537X	PCOS-129	16-453	0/1	+
c.C1618T:p.R540X	PCOS-130	16-454	0/1	+
c.C753G:p.Y251X	PCOS-131	16-455	0/1	+

Table 41. Segregation of Variants in COL20A1 (cont.)

c.756delG:p.N255Tfs*8	OB-0026	13-168	0/1	M
	OB-0037	13-422	0/1	-
	OB-0146	14-079	0/1	+
		14-180	0/1	M
	OB-0330	17-313	0/1	+
	PCOS-092	16-208	0/1	+

Table 42. Segregation of Variants in PCDHGA11

Variant	Family ID	Individual ID	Zygosity	PCOS
c.825_841del:p.L278Tfs*45	OB-0135	14-203	0/1	-
		14-205	0/1	-
		14-206	0/1	-
	OB-0874	18-460	0/1	-
c.A1618G:p.M540V	OB-0945	18-551	0/1	+
c.C2311T:p.L771F	OB-0779	18-341	0/1	M
c.G2194A:p.V732M	OB-0428	17-447	0/1	-
c.G3178A:p.E1060K	OB-0406	17-413	0/1	M
c.G482T:p.R161L	OB-0256	17-196	0/1	-
c.T2227G:p.F743V	LEAN-001	13-446	0/1	+
		13-447	0/1	+
	OB-0760	18-322	0/1	+
	PCOS-136	16-460	0/1	+

4.5.10 Filter 10

Data that had been filtered via Filter 10, which involves variants which are predicted to be pathogenic by all of CADD, REVEL, MetaSVM, MetaRNN, and BayesDel and have a MAF smaller than 10^{-6} were analysed with all 4 methods. QQ plots are given at Figure 26. The top associated genes are given at Table 43, and segregations of variants in these genes are given at Table 44-46.

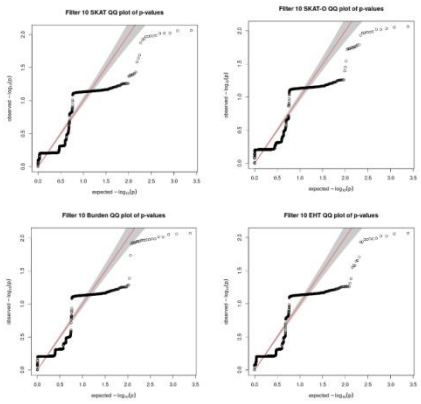

Figure 26. QQ Plots of p-values of SMMAT Models with Filter 10

Table 43. Top 2 p-values of SMMAT Models with Filter 10

group	B.pval	group	S.pval	group	O.pval	group	E.pval
COL20A1	0.00843015	SLC22A18	0.008662653	SLC22A18	0.008662653	SLC22A18	0.008662653
SLC22A18	0.008662653	SLC25A47	0.008776714	SLC25A47	0.008776714	SLC25A47	0.008776714
SLC25A47	0.008776714			COL20A1	0.009174735		

Table 44. Segregation of Variants in COL20A1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.3482delG:p.G1162Afs*40	OB-0311	17-290	0/1	-
c.2209delT:p.S737Pfs*3	PCOS-012	14-057 14-058 14-061 14-065	0/1 0/1 0/1 0/1	+ + + -
c.763delA:p.N255Tfs*8	PCOS-106	16-429	0/1	+

Table 45. Segregation of Variants in SLC22A18

Variant	Family ID	Individual ID	Zygosity	PCOS
c.783_784insGCGTA:p.Y263*	OB-0204 OB-0941	16-505 18-547	0/1 0/1	+ +

Table 46. Segregation of Variants in SLC25A47

Variant	Family ID	Individual ID	Zygosity	PCOS
c.391_392insAGCAGC:p.Q133_R134insQQ	OB-0508 OB-0511	18-049 18-052	0/1 0/1	+ +

4.5.11 Filter 11

Data that had been filtered via Filter 11, which involves variants which are predicted to be pathogenic by all of CADD, PolyPhen, SIFT, and LRT and have a MAF smaller than 10^{-6} were analysed with all 4 methods. QQ plots are given at Figure 27. The top associated genes are given at Table 47, and segregations of variants in these genes are given at Table 48-50.

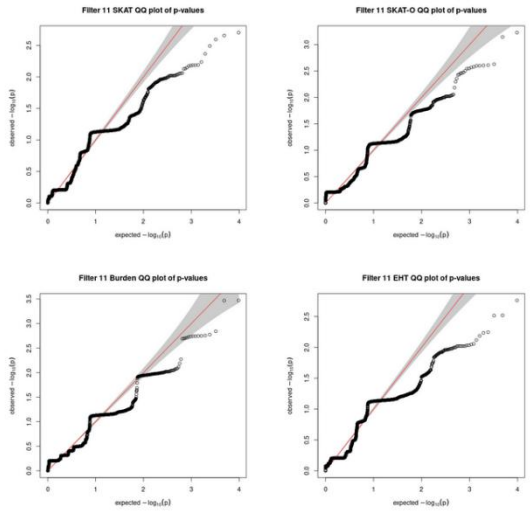

Figure 27. QQ Plots of p-values of SMMAT Models with Filter 11

Table 47. Top 2 p-values of SMMAT Models with Filter 11

group	B.pval	group	S.pval	group	O.pval	group	E.pval
ZFHX4	0.000337827	DUOX1	0.001981105	ARHGAP27	0.000591317	LIMA1	0.001735585
ARHGAP27	0.000340987	DLG2	0.002204948	ZFHX4	0.000715809	ZFHX4	0.003037801

Table 48. Segregation of Variants in DLG2

Variant	Family ID	Individual ID	Zygosity	PCOS
c.742-1G>T	PCOS-035	14-097 14-098 14-099	0/1 0/1 0/1	+ + +
c.G1001C:p.G334A	OB-0093	13-355	0/1	M

Table 49. Segregation of Variants in LIMA1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.A809C:p.D270A	LEAN-001	13-446 13-447	0/1 0/1	+ +
c.A670G:p.K224E	OB-0143 OB-0463	15-117 17-493 17-497	0/1 0/1 0/1	- - -
c.T1270A:p.S424T	OB-1004	18-922	0/1	+
c.A809C:p.D270A	PCOS-082	16-198	0/1	+

Table 50. Segregation of Variants in DUOX1

Variant	Family ID	Individual ID	Zygosity	PCOS
c.A3946G:p.T1316A	OB-0133	16-130 16-131	0/1 0/1	+ +
c.C551T:p.S184F	OB-0313	17-293	0/1	-
c.C3496T:p.L1166F	OB-0538	18-087	0/1	M
c.A4192G:p.I1398V	OB-0577	18-131	0/1	-
c.G3665C:p.W1222S	PCOS-035	14-097 14-098	0/1 0/1	+ +
c.G4412T:p.C1471F	PCOS-136	16-460	0/1	+

4.5.12 Filter 12

Data that had been filtered via Filter 12, which involves cariants which are predicted to be pathogenic by all of CADD, REVEL, MetaSVM, MetaRNN, and BayesDel; and those are not non-frameshift indels; and filtered with a pLI of 0.9, missense z-score of 3.09 were analysed with all 4 methods. QQ plots are given at Figure 28. The top associated genes are given at Table 51, and segregations of variants in these genes are given at Table 52 and Table 53.

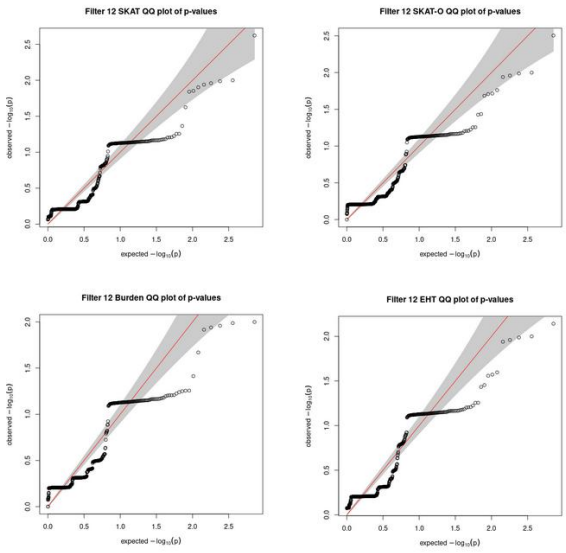

Figure 28. QQ Plots of p-values of SMMAT Models with Filter 12

Table 51. Top 2 p-values of SMMAT Models with Filter 12

group	B.pval	group	S.pval	group	O.pval	group	E.pval
CASKIN2	0.010015977	WDR26	0.002386931	WDR26	0.003131521	WDR26	0.007218914
SCN2A	0.010300007	CASKIN2	0.010015977	CASKIN2	0.010015977	CASKIN2	0.010015977

Table 52. Segregation of Variants in CASKIN2

Variant	Family ID	Individual ID	Zygosity	PCOS
c.965_973del:p.V322_S324del	OB-0340	17-323 17-324	0/1 0/1	+ +

Table 53. Segregation of Variants in WDR26

Variant	Family ID	Individual ID	Zygosity	PCOS
c.G222A:p.L74L	OB-0728	18-285	0/1	-
c.G214T:p.V72L	OB-0749	18-307	0/1	M
c.36_38del:p.G25del	PCOS-083 PCOS-090 PCOS-138	16-199 16-206 16-462	0/1 0/1 0/1	+ + +

4.5.13 Filter 13

Data that had been filtered via Filter 13, which involves cariants which are predicted to be pathogenic by all of CADD, REVEL, MetaSVM, MetaRNN, and BayesDel; and those are not non-frameshift indels; and filtered with LOFTEE and missense z-score of 3.09 were analysed with all 4 methods. QQ plots are given at Figure 29. The top associated genes are given at Table 54, and segregations of variants in these genes are given at Table 56.

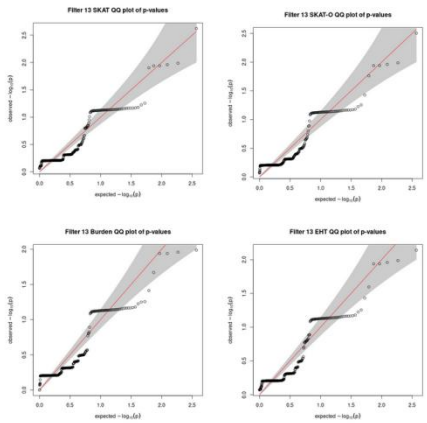

Figure 29. QQ Plots of p-values of SMMAT Models with Filter 13

Table 54. Top 2 p-values of SMMAT Models with Filter 13

group	B.pval	group	S.pval	group	O.pval	group	E.pval
SCN2A	0.010300007	WDR26	0.002386931	WDR26	0.003131521	WDR26	0.007218914
		SCN2A	0.010300007	SCN2A	0.010300007	SCN2A	0.010300007

Table 55. Segregation of Variants in SCN2A

Variant	Family ID	Individual ID	Zygosity	PCOS
c.1994_1996del:p.A666del	PCOS-091	16-207	0/1	+
	PCOS-146	16-470	0/1	+

Table 56. Segregation of Variants in WDR26

Variant	Family ID	Individual ID	Zygosity	PCOS
c.G214T:p.V72L	OB-0749	18-307	0/1	M
c.36_38del:p.G25del	PCOS-083	16-199	0/1	+
	PCOS-090	16-206	0/1	+
	PCOS-138	16-462	0/1	+

4.6 Polygenic Risk Score Analysis

PRS analysis was done with a p-value threshold of 0.5 (Figure 30). The analysis then repeated with both mostly prioritized genes in our association analyses (Figure 31) and top genes of GWAS (Figure 32).

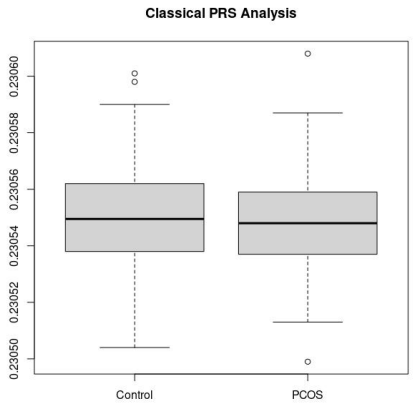

Figure 30. Boxplot of Classical PRS Results

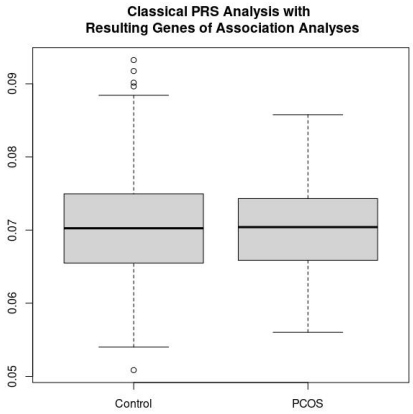

Figure 31. Boxplot of Classical PRS Results with Resulting Genes of Association Analyses

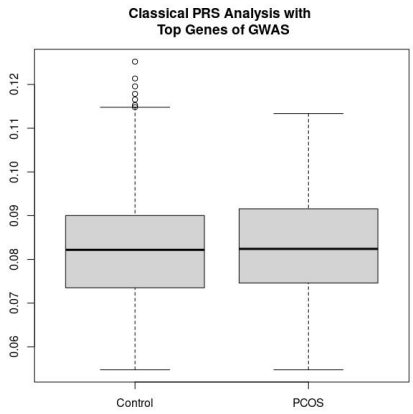

Figure 32. Boxplot of Classical PRS Results with Top Genes of PCOS

5 DISCUSSION

After conducting a comprehensive analysis of the quality graphs, it is evident that our dataset possesses high-quality standards. The missingness rate, which is an indicator of the number of missing values in the dataset, is remarkably low. This low rate of missingness and the high genotyping rate suggests that the dataset is complete and reliable, making it an excellent resource for further analysis and modelling. Furthermore, the quality and depth of the data are also noteworthy. Quality refers to the accuracy and consistency of the data, while depth relates to the amount of information available in the dataset. The high quality of the data implies that it is free from errors and inconsistencies, and the depth of the data suggests that it contains a significant amount of information that can be utilized for various purposes.

When analyzing genetic variation in a population, the MAF is a key parameter to consider. MAF refers to the frequency of the less common allele at a given genetic locus and is an important indicator of genetic diversity. One widely used resource for studying human genetic variation is the 1000 Genomes Project, which aims to provide a comprehensive map of genetic variation across different populations. By analyzing data from thousands of individuals, this project has generated estimates of MAF for a large number of genetic variants. It is important to note that MAF can vary widely depending on the population being studied, as genetic variation is influenced by factors such as geography, history, and cultural practices. Therefore, it is important to consider the context in which MAF values are being interpreted and to be aware of potential sources of bias or confounding. In the 1000 Genomes project, it is found that individuals from all studied populations, rare variants which have a frequency less than 0.005 constitute the majority of the variants (101). This observation suggests that despite the historic bottlenecks, populations have experienced significant growth and diversification in recent times. It had been shown in the same study that forces such as natural selection and genetic drift have shaped the frequency distribution of variants in ways that deviate from the assumptions of a constant population size model. The majority of variants observed in our study were

also found to be rarer than 0.001. This finding demonstrates that historic diversification and genetic drift are also valid for the Turkish population.

Upon examining the plots of PCA, it is evident that there are no outliers present in the cohort. This is a positive sign as outliers can significantly impact the accuracy of the results and lead to erroneous conclusions. Additionally, the PCA plots indicate that the cohort is relatively close to the European population, which is not surprising given that the cohort is composed of Turkish individuals. Overall, the absence of outliers suggests that the data have a strong potential to give reliable results.

Upon analyzing the results of the SKAT and GMMAT packages, it becomes evident that the inclusion of familial correlations plays a crucial role in enhancing the model. The comparison of QQ plots generated from both packages shows that the SMMAT algorithm is more effective when familial correlations are taken into account. This finding is essential as it emphasizes the importance of incorporating familial correlations in genetic studies to improve the accuracy and reliability of the results obtained. The inclusion of familial correlations in the model helps to reduce the noise in the data and improves the power of the analysis. The SMMAT algorithm provides a reliable and efficient approach to account for these correlations and hence the study continued with the SMMAT algorithm.

Filter 1, Filter 2, Filter 3, Filter 4, Filter 6, Filter 8, Filter 10, Filter 11, Filter 12, and Filter 13, have stringent filtering criteria, which results in a low number of remaining variants. While this may seem like a disadvantage, it is important to note that these filters are designed to identify only the highest-quality variants, which are more likely to be biologically relevant. Despite the lower number of remaining variants, these models still have the potential to capture significant genes that are involved in PCOS. This is because the strict filtering criteria help to reduce the number of false positive results, which can be a common issue in genetic studies. Furthermore, the segregation of variants in top genes in these models has been examined comprehensively to identify the most relevant genes. It is also worth noting that the resulting p-values, while not highly significant, may still be biologically relevant. P-values are used to determine the statistical significance of an

observation, but they do not necessarily reflect the biological importance of the result. Therefore, it is possible that even though the p-values are not highly significant, the identified variants may still be relevant to the development of PCOS.

On the other hand, Filter 5, Filter 7, and Filter 9 are the most powerful models. Their resulting p-values are more significant, indicating a high degree of confidence in their results, and QQ plots do not have any significant deviation.

All 4 algorithms prioritized the DLG2 gene in Filter 1 criteria. In this model, there is only 1 marker in the DLG2 gene which is the c.742-1G>T splice acceptor variant. This variant is not seen in any individual in the GnomAD database, despite the high coverage of the region. This is a null variant because the DLG2 gene is loss of function intolerant, gnomAD Loss-of-Function Observed/Expected = 0.212. This variant is seen in the PCOS-035 family in which all samples have PCOS and they all carry the heterozygous variant.

The DLG2 gene which is known to be involved in the regulation of cell growth and division, plays an essential role in regulating synaptic plasticity, which is vital for learning and memory. Previous studies have linked variations in this gene to the development of neurodevelopmental disorders, such as schizophrenia and autism spectrum disorders. It is noteworthy because recent studies have shown that there is a significant correlation between PCOS and Autism Spectrum Disorder (ASD) among both women and their children. It had been shown by various studies that have large sample sizes and high-quality data that women with PCOS have an increased probability of giving birth to a child with ASD. Furthermore, research has also found that women with PCOS are more likely to be diagnosed with ASD themselves which implies that there may be a shared genetic factor that contributes to both PCOS and ASD (102). Additionally, DLG2 has been implicated in the regulation of insulin secretion and may contribute to the pathogenesis of insulin-resistant diabetes which is an important comorbidity of PCOS (103).

In Filter 1 criteria, the second prioritized gene is ZNF534 which also has only one splice acceptor variant c.311-2A>G in the OB-0340 family. This family include 2 samples who both are PCOS patients and carry mutation heterozygous.

In Filter 2 criteria, while SKAT and SKAT-O prioritize the PCDHGA11 gene, Burden and EHT prioritize the COL20A1 gene. However, in Filter 3, the PCDHGA11 gene emerges as the top gene in all models, and in all models except for Burden with Filter 4, Filter 6, and Filter 8. The frameshift c.1002delG:p.M334Ifs*2 mutation is the most important variant in the PCDHGA11 gene, and it is predicted to cause NMD (nonsense-mediated decay) due to the gene's loss of function intolerant nature (gnomAD Loss-of-Function Observed/Expected = 0.606). The exon containing the mutation affects 16 functional domains, and the truncated region contains 7 pathogenic variants. Interestingly, the variant has been found in 4 individuals from 4 different families, and there are no other family members with the same variant. COL20A1 is also the top gene in all models except SKAT in Filter 5, and Filter 9 and in the Burden model with Filter 10.

A very recent study has shown that the PCDHGA11 gene is deregulated in PCOS patients (104). This gene encodes for protocadherin gamma subfamily A 11 protein, which is involved in cell-cell adhesion and communication. While the link between PCDHGA11 and PCOS is yet to be established, recent studies have linked protocadherins to female reproductive disorders.

According to the results obtained from the Filter 3 model, the CRYBG1 gene is ranked as the second most significant gene. The most crucial variant of this gene in this model is c.4646_4647insAT:p.G1551Mfs*28, which has been found in two different families (PCOS-011 and PCOS-035) affected by PCOS. This mutation is a frameshift variant that leads to a premature stop codon and the production of a truncated protein. However, there is another PCOS patient in PCOS-035 that does not carry the mutation.

The protein encoded by CRYBG1 is a crucial actin cytoskeleton-binding protein that is a valuable tumour suppressor gene. Its involvement in cancer prevention is believed to be due to its role in regulating cell division and proliferation (105). On the other hand, recent studies have also linked CRYBG1 to the regulation of fat-cell differentiation and abdominal adiposity which is one of the main characteristics of PCOS. In particular, the protein product of this gene has been found to be involved in the early stages of adipocyte differentiation, where it plays a critical role in promoting the accumulation of lipids and fat in adipocytes (106).

In Filter 4, the Burden test prioritizes the C19orf44 gene. In addition to the C19orf44 gene, the other three genes that are prioritized by all four tests in Filter 4 are CRYBG1, CWC27, and TEX15.

In particular, SKAT has identified the C19orf44 gene as a top priority in both Filter 6 and Filter 8. Additionally, SKAT has prioritized the CRYBG1 gene in Filter 6. In Filter 8, SKAT has also prioritized the OGFOD3 and PSENEN genes. PSENEN gene has been shown as an upstream regulator in ovarian development and function that is involved in PCOS (107). Interestingly, this gene is found to be one of the most important biomarkers of autism (108). Its prioritization highlights again the potential intersection between PCOS and autism.

Of the 4 tests conducted in Filter 10, only the Burden test did not prioritize the SLC22A18 gene. Notably, the gene was found to contain only one marker, frameshift c.783_784insGCGTA:p.Y263*, which was identified in two individuals from two different families. These individuals carried the variant heterozygously. One of the families does not have any other sample however there are 5 PCOS-female and 3 male individuals that do not carry the mutation in OB-0204. In addition to the SLC22A18 gene, the SLC25A47 gene was also prioritized in the same model. These genes are solute carrier family proteins that are known to play a role in mitochondrial function. SLC22A18 gene product is one of the proteins that are known as orphan transporters which have no known ligand (109).

The mostly prioritized variants in Filter 11 are DLG2, ZFHX4, LIMA1, and DUOX1. ZFHX4 is a transcription factor that is involved in the regulation of gene expression. Variants in this gene have been associated with a range of conditions, including fasting blood glucose measurement and type 2 diabetes (110). On the other hand, this gene is known to be one of the androgen-regulated genes (111). LIMA1 gene product has been shown to interact with actin filaments by cross-linking and stabilizing them, thereby promoting the formation of stress fibers. Stress fibers are actin bundles that are essential for many cellular processes, including cell migration, cytokinesis, and cell adhesion, as well as insulin-stimulated glucose transport in muscle and adipose tissue. Specifically, the presence of stress fibers has been shown to be necessary for the proper translocation of the glucose transporter GLUT4 to the plasma membrane, which facilitates glucose uptake into cells. In this context, LIMA1's ability to stabilize actin filaments and promote stress fiber formation may be crucial for the insulin-stimulated transport of GLUT4 (112). In the context of PCOS, the LIMA1 gene may play a role in the development of insulin resistance which is one of the main comorbidities of PCOS.

The mostly prioritized variants in Filter 12 are CASKIN2 and WDR26, while the mostly prioritized variants in Filter 13 are SCN2A and WDR26. All of them are associated with neurological disorders including autism (113-115). CASKIN2 is a protein that is involved in the regulation of synaptic transmission and neuronal development. It has been found to be essential for the proper localization and function of ion channels and receptors in the brain. SCN2A is another gene that has been extensively studied due to its role in encoding the alpha subunit of the voltage-gated sodium channel. This gene is crucial for the proper functioning of neurons, and its dysregulation can lead to severe neurological consequences. On the other hand, WDR26 is a gene that encodes a protein involved in various cellular processes, including the cell cycle, DNA repair, and apoptosis. It has been found to be essential for the proper functioning of the immune system and the development of various types of cancer (92).

SKAT has identified the TRPM5 gene as a top-priority candidate gene for PCOS in both Filter 5 and Filter 9. TRPM5 is a member of the transient receptor potential (TRP) family of ion channels and is known to play a vital role in regulating insulin secretion in pancreatic beta-cells (116). Given the established link between insulin resistance and PCOS, the identification of TRPM5 as a candidate gene for PCOS is noteworthy. Furthermore, in Filter 5, SKAT also prioritized the AK5 and PCDHGA11 genes. AK5 encodes for adenylate kinase 5, an enzyme that catalyzes the transfer of phosphate groups between nucleotides. Previous studies have shown that it is involved in the regulation of insulin secretion and glucose metabolism (117).

In Filter 7, mostly prioritized genes are THAP8, ARMC3, OR4K14, ADAT1, POM121L12, and SELENOP. THAP8 is a transcription factor that has been linked to the regulation of gene expression in several biological processes, including cell proliferation and differentiation. ARMC3 is a protein-coding gene that has been suggested to play a role in ciliary function, which is essential for several physiological processes, including sensory perception and embryonic development (118). OR4K14 is an olfactory receptor gene that has been implicated in the sense of smell. ADAT1 is an enzyme that catalyzes the deamination of adenosine to inosine, and it has been suggested to play a role in RNA editing. POM121L12 is a gene that encodes a protein involved in the nuclear pore complex assembly, which is essential for the transport of molecules between the nucleus and the cytoplasm. Finally, SELENOP is a gene that encodes a protein that has been implicated in the regulation of thyroid hormone metabolism and antioxidant defence mechanisms.

When investigating PRS analysis results, it have been shown that classical PRS approach is not sufficient to discriminate between affected and control individuals when base data is from a genome-wide study whereas target data is WES. The most significant variants in GWAS are in the intergenic regions which are not covered in WES data. There are some attempts to assign effect sizes of most significant variants of GWAS to included variants of WES that is found at the same haplotype block. However, currently these approaches are very computationally inefficient that, in our sample size the PRS calculation would take more than 2 years to compute.

6 CONCLUSION

WES data of 203 PCOS patients from 190 families and 815 control individuals with 52 of those being family members of PCOS patients have been utilized for several gene-based association study models. The aim was to identify potential candidate genes and gain insights into the underlying molecular mechanisms of PCOS.

Comparing these models, it has been shown that including family correlations enhances the model. The statistical analysis results, segregation of variants within families, and relevant literature search provided valuable insights into potential candidate genes associated with PCOS. It was found that DLG2, PCDHGA11, CRYBG1, PSENEN, SLC22A18, SLC25A47, LIMA1, CASKIN2, SCN2A, and WDR26 are important candidate genes for PCOS. These genes are involved either in ovarian function-related pathways or in insulin-related pathways, indicating their potential role in PCOS.

Interestingly, most of these genes are also associated with autism. This correlation is significant as studies have found that women with PCOS are more likely to either have a child with autism or have autism themselves. This finding suggests a potential overlap in the genetic mechanisms underlying both conditions.

The analysis of common variants in the study revealed that when analysing data from a WES study, utilizing a genome-wide base data set may not be suitable. This is due to the fact that the majority of the associated variants are found to be located in the intergenic regions of the genome. Further research should be conducted to better understand the functional significance of such variants and their potential implications for disease pathogenesis.

7 REFERENCES

1. Deswal R, Narwal V, Dang A, Pundir CS. The Prevalence of Polycystic Ovary Syndrome: A Brief Systematic Review. J Hum Reprod Sci. 2020;13(4):261–71.

2. Joham AE, Norman RJ, Stener-Victorin E, Legro RS, Franks S, Moran LJ, et al. Polycystic ovary syndrome. The Lancet Diabetes & Endocrinology. 2022 Sep 1;10(9):668–80.

3. Singh S, Pal N, Shubham S, Sarma DK, Verma V, Marotta F, et al. Polycystic Ovary Syndrome: Etiology, Current Management, and Future Therapeutics. Journal of Clinical Medicine. 2023 Jan;12(4):1454.

4. Goh JE, Farrukh MJ, Keshavarzi F, Yap CS, Saleem Z, Salman M, et al. Assessment of prevalence, knowledge of polycystic ovary syndrome and health-related practices among women in klang valley: A cross-sectional survey. Frontiers in Endocrinology [Internet]. 2022 [cited 2023 Apr 29];13. Available from: https://www.frontiersin.org/articles/10.3389/fendo.2022.985588

5. Ottarsdottir K, Nilsson AG, Hellgren M, Lindblad U, Daka B. The association between serum testosterone and insulin resistance: a longitudinal study. Endocr Connect. 2018 Dec 4;7(12):1491–500.

6. Azziz R, Carmina E, Chen Z, Dunaif A, Laven JSE, Legro RS, et al. Polycystic ovary syndrome. Nat Rev Dis Primers. 2016 Aug 11;2(1):1–18.

7. Chen ZJ, Zhao H, He L, Shi Y, Qin Y, Shi Y, et al. Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. Nat Genet. 2011 Jan;43(1):55–9.

8. Shi Y, Zhao H, Shi Y, Cao Y, Yang D, Li Z, et al. Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. Nat Genet. 2012 Sep;44(9):1020–5.

9. Day FR, Hinds DA, Tung JY, Stolk L, Styrkarsdottir U, Saxena R, et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. Nat Commun. 2015 Sep 29;6:8464.

10. Hayes MG, Urbanek M, Ehrmann DA, Armstrong LL, Lee JY, Sisk R, et al. Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. Nat Commun. 2015 Aug 18;6:7502.

11. Day F, Karaderi T, Jones MR, Meun C, He C, Drong A, et al. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. PLoS Genet. 2018 Dec 19;14(12):e1007813.

12. Tian Y, Li J, Su S, Cao Y, Wang Z, Zhao S, et al. PCOS-GWAS Susceptibility Variants in THADA, INSR, TOX3, and DENND1A Are Associated With Metabolic Syndrome or Insulin Resistance in Women With PCOS. Frontiers in Endocrinology [Internet]. 2020 [cited 2023 May 12];11. Available from: https://www.frontiersin.org/articles/10.3389/fendo.2020.00274

13. Cui L, Zhao H, Zhang B, Qu Z, Liu J, Liang X, et al. Genotype–phenotype correlations of PCOS susceptibility SNPs identified by GWAS in a large cohort of Han Chinese women. Human Reproduction. 2013 Feb 1;28(2):538–44.

14. Takayama K, Suzuki T, Bulun SE, Sasano H, Yilmaz B, Sebastian S. Organization of the Human Aromatase P450 (CYP19) Gene. Semin Reprod Med. 2004 Jan;22(1):5–9.
15. Rosenfield RL, Barnes RB, Cara JF, Lucky AW. Dysregulation of cytochrome P450c17α as the cause of polycystic ovarian syndrome**Supported in part by grants HD-06308 and Rr-00055 from the United States Public Health Service, Bethesda, Maryland. Fertility and Sterility. 1990 May 1;53(5):785–91.
16. Carey AH, Waterworth D, Patel K, White D, Little J, Novelli P, et al. Polycystic ovaries and premature male pattern baldness are associated with one allele of the steroid metabolism gene CYP17. Human Molecular Genetics. 1994 Oct 1;3(10):1873–6.
17. Wickenheisser JK, Quinn PG, Nelson VL, Legro RS, Strauss JF III, McAllister. JM. Differential Activity of the Cytochrome P450 17α-Hydroxylase and Steroidogenic Acute Regulatory Protein Gene Promoters in Normal and Polycystic Ovary Syndrome Theca Cells1. The Journal of Clinical Endocrinology & Metabolism. 2000 Jun 1;85(6):2304–11.
18. Witchel SF, Aston CE. The role of heterozygosity for CYP21 in the polycystic ovary syndrome. J Pediatr Endocrinol Metab. 2000 Jan 1;13 Suppl 5:1315–7.
19. Gharani N, Waterworth DM, Batty S, White D, Gilling-Smith C, Conway GS, et al. Association of the Steroid Synthesis Gene Cyp11a with Polycystic Ovary Syndrome and Hyperandrogenism. Human Molecular Genetics. 1997 Mar 1;6(3):397–402.
20. Diamanti-Kandarakis E, Bartzis MI, Bergiele AT, Tsianateli TC, Kouli CR. Microsatellite polymorphism (tttta)n at −528 base pairs of gene CYP11α influences hyperandrogenemia in patients with polycystic ovary syndrome. Fertility and Sterility. 2000 Apr 1;73(4):735–41.
21. Wang Y, Wu X, Cao Y, Yi L, Chen J. A microsatellite polymorphism (tttta)n in the promoter of the CYP11a gene in Chinese women with polycystic ovary syndrome. Fertility and Sterility. 2006 Jul 1;86(1):223–6.
22. Franks S, Stark J, Hardy K. Follicle dynamics and anovulation in polycystic ovary syndrome. Human Reproduction Update. 2008 Sep 1;14(5):539.
23. Rodriguez Paris V, Bertoldo MJ. The Mechanism of Androgen Actions in PCOS Etiology. Medical Sciences. 2019 Sep;7(9):89.
24. Goodarzi MO, Shah NA, Antoine HJ, Pall M, Guo X, Azziz R. Variants in the 5alpha-reductase type 1 and type 2 genes are associated with polycystic ovary syndrome and the severity of hirsutism in affected women. J Clin Endocrinol Metab. 2006 Oct;91(10):4085–91.
25. Graupp M, Wehr E, Schweighofer N, Pieber TR, Obermayer-Pietsch B. Association of genetic variants in the two isoforms of 5α-reductase, SRD5A1 and SRD5A2, in lean patients with polycystic ovary syndrome. Eur J Obstet Gynecol Reprod Biol. 2011 Aug;157(2):175–9.
26. Caglayan AO, Dundar M, Tanriverdi F, Baysal NA, Unluhizarci K, Ozkul Y, et al. Idiopathic hirsutism: local and peripheral expression of aromatase (CYP19A) and 5α-reductase genes (SRD5A1 and SRD5A2). Fertility and Sterility. 2011 Aug 1;96(2):479–82.

27. Dewailly D, Catteau-Jonard S, Reyss AC, Leroy M, Pigny P. Oligoanovulation with Polycystic Ovaries But Not Overt Hyperandrogenism. The Journal of Clinical Endocrinology & Metabolism. 2006 Oct 1;91(10):3922–7.
28. Laven JSE. Follicle Stimulating Hormone Receptor (FSHR) Polymorphisms and Polycystic Ovary Syndrome (PCOS). Front Endocrinol (Lausanne). 2019 Feb 12;10:23.
29. Ran Y, Yi Q, Li C. The Relationship of Anti-Mullerian Hormone in Polycystic Ovary Syndrome Patients with Different Subgroups. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy. 2021;14:1419.
30. Georgopoulos NA, Karagiannidou E, Koika V, Roupas ND, Armeni A, Marioli D, et al. Increased Frequency of the Anti-Müllerian-Inhibiting Hormone Receptor 2 (AMHR2) 482 A>G Polymorphism in Women With Polycystic Ovary Syndrome: Relationship to Luteinizing Hormone Levels. J Clin Endocrinol Metab. 2013 Nov;98(11):E1866–70.
31. Gorsic LK, Dapas M, Legro RS, Hayes MG, Urbanek M. Functional Genetic Variation in the Anti-Müllerian Hormone Pathway in Women With Polycystic Ovary Syndrome. J Clin Endocrinol Metab. 2019 Feb 20;104(7):2855–74.
32. Nautiyal H, Imam SS, Alshehri S, Ghoneim MM, Afzal M, Alzarea SI, et al. Polycystic Ovarian Syndrome: A Complex Disease with a Genetics Approach. Biomedicines. 2022 Mar;10(3):540.
33. Shaaban Z, Khoradmehr A, Amiri-Yekta A, Nowzari F, Jafarzadeh Shirazi MR, Tamadon A. Pathophysiologic Mechanisms of Insulin Secretion and Signaling-Related Genes in Etiology of Polycystic Ovary Syndrome. Genet Res (Camb). 2021 Dec 6;2021:7781823.
34. Dilek S, Ertunc D, Tok EC, Erdal EM, Aktas A. Association of Gly972Arg variant of insulin receptor substrate-1 with metabolic features in women with polycystic ovary syndrome. Fertility and Sterility. 2005 Aug 1;84(2):407–12.
35. Ioannidis A, Ikonomi E, Dimou NL, Douma L, Bagos PG. Polymorphisms of the insulin receptor and the insulin receptor substrates genes in polycystic ovary syndrome: A Mendelian randomization meta-analysis. Molecular Genetics and Metabolism. 2010 Feb 1;99(2):174–83.
36. Ruan Y, Ma J, Xie X. Association of IRS-1 and IRS-2 genes polymorphisms with polycystic ovary syndrome: a meta-analysis. Endocrine Journal. 2012;59(7):601–9.
37. Gonzalez A, Abril E, Roca A, Aragón MJ, Figueroa MJ, Velarde P, et al. Comment: CAPN10 alleles are associated with polycystic ovary syndrome. J Clin Endocrinol Metab. 2002 Aug;87(8):3971–6.
38. Dasgupta S, Sirisha PVS, Neelaveni K, Anuradha K, Reddy BM. Association of CAPN10 SNPs and Haplotypes with Polycystic Ovary Syndrome among South Indian Women. PLoS One. 2012 Feb 23;7(2):e32192.
39. Khan MJ, Ullah A, Basit S. Genetic Basis of Polycystic Ovary Syndrome (PCOS): Current Perspectives. Appl Clin Genet. 2019 Dec 24;12:249–60.
40. Liu AL, Xie HJ, Xie HY, Liu J, Yin J, Hu JS, et al. Association between fat mass and obesity associated (FTO) gene rs9939609 A/T polymorphism and polycystic ovary syndrome: a systematic review and meta-analysis. BMC Med Genet. 2017 Aug 21;18:89.

41. Song DK, Lee H, Oh JY, Hong YS, Sung YA. FTO Gene Variants Are Associated with PCOS Susceptibility and Hyperandrogenemia in Young Korean Women. Diabetes Metab J. 2014 Aug;38(4):302–10.
42. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007 Feb 3;615(1–2):28–56.
43. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009 Feb;5(2):e1000384.
44. Larson NB, Chen J, Schaid DJ. A Review of Kernel Methods for Genetic Association Studies. Genet Epidemiol. 2019 Mar;43(2):122–36.
45. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Jun;27(2):e1608.
46. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020 Sep;15(9):2759–72.
47. Mihm M, Gangooly S, Muttukrishna S. The normal menstrual cycle in women. Animal Reproduction Science. 2011 Apr 1;124(3):229–36.
48. Xing C, Zhang J, Zhao H, He B. Effect of Sex Hormone-Binding Globulin on Polycystic Ovary Syndrome: Mechanisms, Manifestations, Genetics, and Treatment. Int J Womens Health. 2022 Feb 2;14:91–105.
49. Li Y, Fang L, Yan Y, Wang Z, Wu Z, Jia Q, et al. Association between human SHBG gene polymorphisms and risk of PCOS: a meta-analysis. Reprod Biomed Online. 2021 Jan;42(1):227–36.
50. Goodarzi MO, Carmina E, Azziz R. DHEA, DHEAS and PCOS. The Journal of Steroid Biochemistry and Molecular Biology. 2015 Jan 1;145:213–25.
51. Goverde AJ, van Koert AJB, Eijkemans MJ, Knauff EAH, Westerveld HE, Fauser BCJM, et al. Indicators for metabolic disturbances in anovulatory women with polycystic ovary syndrome diagnosed according to the Rotterdam consensus criteria. Human Reproduction. 2009 Mar 1;24(3):710–7.
52. Guastella E, Longo RA, Carmina E. Clinical and endocrine characteristics of the main polycystic ovary syndrome phenotypes. Fertil Steril. 2010 Nov;94(6):2197–201.
53. Lizneva D, Suturina L, Walker W, Brakta S, Gavrilova-Jordan L, Azziz R. Criteria, prevalence, and phenotypes of polycystic ovary syndrome. Fertility and Sterility. 2016 Jul 1;106(1):6–15.
54. Moghetti P, Tosi F, Bonin C, Di Sarra D, Fiers T, Kaufman JM, et al. Divergences in Insulin Resistance Between the Different Phenotypes of the Polycystic Ovary Syndrome. The Journal of Clinical Endocrinology & Metabolism. 2013 Apr 1;98(4):E628–37.
55. Liu D, Lin X, Ghosh D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics. 2007 Dec;63(4):1079–88.
56. Astolfi A, Scarciotti G, Simard J, Faedo N, Ringwood JV. Model Reduction by Moment Matching: Beyond Linearity A Review of the Last 10 Years. In: 2020 59th IEEE Conference on Decision and Control (CDC). 2020. p. 1–16.

57. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. The American Journal of Human Genetics. 2011 Jul 15;89(1):82–93.
58. Davies RB. The Distribution of a Linear Combination of $\chi 2$ Random Variables. Journal of the Royal Statistical Society Series C: Applied Statistics. 1980 Nov 1;29(3):323–33.
59. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am J Hum Genet. 2012 Aug 10;91(2):224–37.
60. Zeng P, Zhao Y, Liu J, Liu L, Zhang L, Wang T, et al. Likelihood Ratio Tests in Rare Variant Detection for Continuous Phenotypes. Annals of Human Genetics. 2014;78(5):320–32.
61. Bailey-Wilson JE. Parametric and Nonparametric Linkage Analysis. In: Encyclopedia of Life Sciences [Internet]. John Wiley & Sons, Ltd; 2018 [cited 2023 May 17]. p. 1–7. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0005403.pub2
62. Zhao L, He Z, Zhang D, Wang GT, Renton AE, Vardarajan BN, et al. A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late-Onset Alzheimer Disease via WGS Data. The American Journal of Human Genetics. 2019 Oct 3;105(4):822–35.
63. Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, et al. SNP Set Association Analysis for Familial Data. Genetic Epidemiology. 2012;36(8):797–810.
64. Jiang Y, Chiu CY, Yan Q, Chen W, Gorin MB, Conley YP, et al. Gene-Based Association Testing of Dichotomous Traits With Generalized Functional Linear Mixed Models Using Extended Pedigrees: Applications to Age-Related Macular Degeneration. J Am Stat Assoc. 2021;116(534):531–45.
65. Wang X, Lee S, Zhu X, Redline S, Lin X. GEE-Based SNP Set Association Test for Continuous and Discrete Traits in Family-Based Association Studies. Genetic Epidemiology. 2013;37(8):778–86.
66. Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, et al. Adjusted Sequence Kernel Association Test for Rare Variants Controlling for Cryptic and Family Relatedness. Genetic Epidemiology. 2013;37(4):366–76.
67. Yan Q, Tiwari HK, Yi N, Gao G, Zhang K, Lin WY, et al. A Sequence Kernel Association Test for Dichotomous Traits in Family Samples under a Generalized Linear Mixed Model. Hum Hered. 2015;79(2):60–8.
68. Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. Genet Epidemiol. 2013 Jul;37(5):10.1002/gepi.21727.
69. Saad M, Wijsman EM. Combining Family- and Population-Based Imputation Data for Association Analysis of Rare and Common Variants in Large Pedigrees. Genetic Epidemiology. 2014;38(7):579–90.
70. Choi S, Lee S, Cichon S, Nöthen MM, Lange C, Park T, et al. FARVAT: a family-based rare variant association test. Bioinformatics. 2014 Nov 15;30(22):3197–205.

71. Yan Q, Fang Z, Chen W. KMgene: a unified R package for gene-based association analysis for complex traits. Bioinformatics. 2018 Jun 15;34(12):2144–6.
72. Chen H, Meigs JB, Dupuis J. Sequence Kernel Association Test for Quantitative Traits in Family Samples. Genetic Epidemiology. 2013;37(2):196–204.
73. Chen H, Huffman JE, Brody JA, Wang C, Lee S, Li Z, et al. Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. Am J Hum Genet. 2019 Feb 7;104(2):260–74.
74. Ni G, Zeng J, Revez JA, Wang Y, Zheng Z, Ge T, et al. A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. Biol Psychiatry. 2021 Nov 1;90(9):611–20.
75. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.
76. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491–8.
77. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11.10.1-11.10.33.
78. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Jun 6;17(1):122.
79. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.
80. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012 Apr 1;6(2):80–92.
81. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.
82. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
83. Why most Principal Component Analyses (PCA) in population genetic studies are wrong | bioRxiv [Internet]. [cited 2023 Jun 8]. Available from: https://www.biorxiv.org/content/10.1101/2021.04.11.439381v4.abstract
84. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012 Sep;13(4):762–75.
85. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011 Jul 15;89(1):82–93.

86. Lee S, Fuchsberger C, Kim S, Scott L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. Biostatistics. 2016 Jan;17(1):1–15.
87. 46th European Mathematical Genetics Meeting (EMGM) 2018, Cagliari, Italy, April 18-20, 2018: Abstracts. Human Heredity. 2018 Apr 18;83(1):1–29.
88. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet. 2016 Apr 7;98(4):653–66.
89. Zaucha J, Heinzinger M, Tarnovskaya S, Rost B, Frishman D. Family-specific analysis of variant pathogenicity prediction tools. NAR Genom Bioinform. 2020 Jun;2(2):lqaa014.
90. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.
91. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 Jan 8;47(D1):D886–94.
92. Huffman N, Palmieri D, Coppola V. The CTLH Complex in Cancer Cell Plasticity. Journal of Oncology. 2019 Nov 30;2019:1–13.
93. Kim S, Jhong JH, Lee J, Koo JY. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017 Jan 26;10:2.
94. Li C, Zhi D, Wang K, Liu X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Medicine. 2022 Oct 8;14(1):115.
95. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009 Sep;19(9):1553–61.
96. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016 Jan;11(1):1–9.
97. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;0 7:Unit7.20.
98. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Current Protocols in Bioinformatics. 2016;54(1):1.30.1-1.30.33.
99. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2015 Mar 1;31(5):782–4.
100. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.
101. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov;491(7422):56–65.
102. Katsigianni M, Karageorgiou V, Lambrinoudaki I, Siristatidis C. Maternal polycystic ovarian syndrome in autism spectrum disorder: a systematic review and meta-analysis. Mol Psychiatry. 2019 Dec;24(12):1787–97.

103. Yang CH, Mangiafico SP, Waibel M, Loudovaris T, Loh K, Thomas HE, et al. E2f8 and Dlg2 genes have independent effects on impaired insulin secretion associated with hyperglycaemia. Diabetologia. 2020 Jul 1;63(7):1333–48.
104. Zanjirband M, Baharlooie M, Safaeinejad Z, Nasr-Esfahani MH. Transcriptomic screening to identify hub genes and drug signatures for PCOS based on RNA-Seq data in granulosa cells. Computers in Biology and Medicine. 2023 Mar 1;154:106601.
105. Haffner MC, Esopi DM, Chaux A, Gürel M, Ghosh S, Vaghasia AM, et al. AIM1 is an actin-binding protein that suppresses cell migration and micrometastatic dissemination. Nat Commun. 2017 Jul 26;8:142.
106. Shin J, Syme C, Wang D, Richer L, Pike GB, Gaudet D, et al. Novel Genetic Locus of Visceral Fat and Systemic Inflammation. J Clin Endocrinol Metab. 2019 Apr 3;104(9):3735–42.
107. Azumah R, Hummitzsch K, Hartanti MD, St. John JC, Anderson RA, Rodgers RJ. Analysis of Upstream Regulators, Networks, and Pathways Associated With the Expression Patterns of Polycystic Ovary Syndrome Candidate Genes During Fetal Ovary Development. Front Genet. 2022 Feb 7;12:762177.
108. Latkowski T, Osowski S. Gene selection in autism – Comparative study. Neurocomputing. 2017 Aug 9;250:37–44.
109. Yee SW, Stecula A, Chien HC, Zou L, Feofanova EV, van Borselen M, et al. Unraveling the functional role of the orphan solute carrier, SLC22A24 in the transport of steroid conjugates through metabolomic and genome-wide association studies. PLoS Genet. 2019 Sep 25;15(9):e1008208.
110. Mansour A, Mousa M, Abdelmannan D, Tay G, Hassoun A, Alsafar H. Microvascular and macrovascular complications of type 2 diabetes mellitus: Exome wide association analyses. Front Endocrinol (Lausanne). 2023 Mar 23;14:1143067.
111. Jin HJ, Kim J, Yu J. Androgen receptor genomic regulation. Transl Androl Urol. 2013 Sep;2(3):158–77.
112. Kheterpal I, Scherp P, Kelley L, Wang Z, Johnson W, Ribnicky D, et al. Bioactives from Artemisia dracunculus L. Enhance Insulin Sensitivity via Modulation of Skeletal Muscle Protein Phosphorylation. Nutrition. 2014;30(0 0):S43–51.
113. Smirnova E, Kwan JJ, Siu R, Gao X, Zoidl G, Demeler B, et al. A new mode of SAM domain mediated oligomerization observed in the CASKIN2 neuronal scaffolding protein. Cell Communication and Signaling. 2016 Aug 22;14(1):17.
114. Kruth KA, Grisolano TM, Ahern CA, Williams AJ. SCN2A channelopathies in the autism spectrum of neuropsychiatric disorders: a role for pluripotent stem cells? Mol Autism. 2020 Apr 7;11:23.
115. Skraban CM, Grand KL, Deardorff MA. WDR26-Related Intellectual Disability. In: Adam MP, Mirzaa GM, Pagon RA, Wallace SE, Bean LJ, Gripp KW, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2023 Jun 5]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK540448/

116. Brixel LR, Monteilh-Zoller MK, Ingenbrandt CS, Fleig A, Penner R, Enklaar T, et al. TRPM5 regulates glucose-stimulated insulin secretion. Pflugers Arch. 2010 Jun;460(1):69–76.
117. Stanojevic V, Habener JF, Holz GG, Leech CA. Cytosolic adenylate kinases regulate K-ATP channel activity in human β-cells. Biochemical and biophysical research communications. 2008 Apr 4;368(3):614.
118. Huang Y, Jiang Z, Gao X, Luo P, Jiang X. ARMC Subfamily: Structures, Functions, Evolutions, Interactions, and Diseases. Frontiers in Molecular Biosciences [Internet]. 2021 [cited 2023 Jun 4];8. Available from: https://www.frontiersin.org/articles/10.3389/fmolb.2021.791597

8 CURRICULUM VITAE