ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

COMMON AND RARE VARIANT ASSOCIATION APPROACHES IN TURKISH FAMILIES WITH OBESITY

FATMA NİSA ESEN

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Asst. Prof. ONUR EMRE ONAT

ISTANBUL-2024

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

COMMON AND RARE VARIANT ASSOCIATION APPROACHES IN TURKISH FAMILIES WITH OBESITY

FATMA NİSA ESEN

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Asst. Prof. ONUR EMRE ONAT

ISTANBUL-2024

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.

12/06/2024
Fatma Nisa Esen

PREFACE AND ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Asst. Prof. Onur Emre Onat, for giving me the opportunity to participate in this valuable project. His profound advice and endless patience have guided me and made academic research truly enjoyable. His consistently positive and solution-oriented attitude helped me during tough times, and his insights were invaluable whenever I felt stuck.

My gratitude extends to Prof. Uğur Sezerman, who has supported me throughout this project and has taught me countless lessons in the field of bioinformatics, continually challenging me to grow as a scientist. His continuous support helped me through the difficult times and allowed me to continue my research.

I would like to offer my special thanks to Assoc. Prof. Ahmet Yeşilyurt for his unwavering support and belief in me. His invaluable mentorship has shaped my career and illuminated my path in countless ways. His guidance has allowed me to gain new perspectives and empowered me to think outside the box. I am extremely grateful for the opportunity to work with such a brilliant mind and will always be in his debt.

I would like to thank Prof. Tayfun Özçelik, whose years of work and dedication made this project possible. I also want to acknowledge my dearest friends and colleagues: Merve Nur Köroğlu, who supported me throughout this project; Elif Öz, who accompanied me every step of the way and helped me shape my research methods; and Faruk Üstünel, who tirelessly worked with me for countless hours, offering his deep insights into computational biology and helping me solve every problem I encountered.

Finally, I would like to thank to my parents, my teammates and friends Hanifenur Mancılar, Hilal Yıldız Er, Damlanur Şakar, Dr. Eyyüp Üçtepe, Dr. Sait Tümer, Ebru Bayhan Aydın, Semih Karabulut, and Songül Harşıt. Without their tremendous understanding and encouragement in the past years, it would be impossible for me to complete my study.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

ACTH	Adrenocorticotropin
ADRB1	Adrenoceptor Beta 1
ADRB2	Adrenoceptor Beta 2
ADRB3	Adrenoceptor Beta 3
AFF4	ALF Transcription Elongation Factor 4
AGRP	Agouti-Related Protein
ALMS1	ALMS1 centrosome and basal body associated protein
ARH	Arcuate Nucleus Of Hypothalamus
ARID5B	AT-Rich Interaction Domain 5B
ARL6	ADP Ribosylation Factor Like GTPase 6
BBIP1	BBsome Interacting Protein 1
BBS	Bardet-Biedel Syndrome
BBS1	Bardet-Biedel Syndrome 1
BBS10	Bardet-Biedel Syndrome 10
BBS12	Bardet-Biedel Syndrome 12
BBS2	Bardet-Biedel Syndrome 2
BBS4	Bardet-Biedel Syndrome 4
BBS5	Bardet-Biedel Syndrome 5
BBS7	Bardet-Biedel Syndrome 7
BBS9	Bardet-Biedel Syndrome 9
BMI	Body Mass Index
BQSR	Base Quality Score Recalibration
BWA-MEM	Burrows-Wheeler Aligner-Maximal Exact Match
CAST	Cohort Allelic Sums Test
CEP164	Centrosomal Protein 164
CEP290	Centrosomal Protein 290
CFAP418	Cilia And Flagella Associated Protein 418
CHD	Coronary Heart Disease
CMC	Combined Multivaritate and Collapsing Method
CREBBP	CREB binding protein

CVD	Cardiovascular Disease
DBSNP	Database Of Single Nucleotide Polymorphisms
DD	Developmental Delay
EHMT1	Euchromatic Histone Lysine Methyltransferase 1
EHT	Efficient Hybrid Test
FTO	FTO Alpha-Ketoglutarate Dependent Dioxygenase
GABA	Gamma- Aminobutyric Acid
GATK	Genome Analysis Toolkit
GBD	Global Burden of Disease
GI	Gastrointestinal
GLMM	Generalized Linear Mixed Model
GNAS	Gs Alpha Subunit
GoF	Gain-Of-Function
GPCR	G Protein-Coupled Receptor
GQ	Genotype Quality
GRM	Genetic Relationship Matrix
GWAS	Genome-Wide Association Study
ID	Intellectual Disability
IFT172	Intraflagellar Transport 172
IFT27	Intraflagellar Transport 27
IFT74	Intraflagellar Transport 74
IRX3	Iroquois Homeobox 3
IRX5	Iroquois Homeobox 5
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
LEP	Leptin
LEPR	Leptin Receptor
LMM	Linear Mixed Model
LoF	Loss-Of-Function
LOFTEE	Loss-Of-Function Transcript Effect Estimator
LZTFL1	Leucine Zipper Transcription Factor Like 1
MAF	Minor Allele Frequency

MAGEL2	MAGE Family Member L2
MC1R	Melanocortin 1 Receptor
MC2R	Melanocortin 2 Receptor
MC3R	Melanocortin 3 Receptor
MC4R	Melanocortin 4 Receptor
MC5R	Melanocortin 5 Receptor
MCR	Melanocortin Receptor
MGI	Mouse Genome Informatics
MHC	Major Histocompatibility Complex
MKKS	MKKS Centrosomal Shuttling Protein
MKRN3	Makorin Ring Finger Protein 3
MKS1	MKS Transition Zone Complex Subunit 1
MQ	Mapping Quality
MSH	Melanocyte Stimulating Hormones
NDN	Necdin
NGS	Next Generation Sequencing
NPAP1	Nuclear Pore Associated Protein 1
NPY	Neuropeptide Y
NTRK2	Neurotrophic Receptor Tyrosine Kinase 2
OR	Odds Ratio
PC1/3	Prohormone Convertase 1/3
PCA	Principle Component Analysis
PCSK1	Proprotein Convertase Subtilisin/Kexin Type 1
POMC	Proopiomelanocortin
PRS	Polygenic Risk Score
PTPN22	Protein Tyrosine Phosphatase Non-Receptor Type 22
PWCR	Prader-Willi Critical Region
PWS	Prader-Willi Syndrome
RAB23	RAS-Associated Protein Rab23
RPGRIP1L	RPGR-Interacting Protein 1-Like
RVAS	Rare Variant Association Study
SCAPER	S-phase Cyclin A Associated Protein In The ER

SCLT1	Sodium Channel And Clathrin Linker 1
SDCCAG8	SHH Signaling And Ciliogenesis Regulator
SEMA3A	Semaphorin 3A
SEMA3G	Semaphorin 3G
SKAT	Sequence Kernel Association Test
SKAT-O	Optimal Sequence Kernel Association Test
SLC6A14	Solute Carrier Family 6 Member 14
SMMAT	Set Mixed Model Association Test
SMMAT-B	Set Mixed Model Association Test-Burden
SMMAT-E	Set Mixed Model Association Test-Efficient Hybrid Test
SMMAT-O	Set Mixed Model Association Test-SKAT-O
SMMAT-S	Set Mixed Model Association Test-SKAT
SNP	Single Nucleotide Polymorphism
SNRPN	Small Nuclear Ribonucleoprotein Polypeptide N
STAT3	Signal Transducer And Activator Of Transcription 3
TRIM32	Tripartite Motif Containing 32
TTC8	Tetratricopeptide Repeat Domain 8
UCP2	Uncoupling Protein 2
UCP3	Uncoupling Protein 3
UK	United Kingdom
USD	United States Dollars
UTR	Untranslated Region
VCF	Variant Call Format
VEP	Variant Effect Predictor
VPS13B	Vacuolar Protein Sorting 13 Homolog B
VQSR	Variant Quality Score Recalibration
VTA	Ventral Tegmental Area
WDPCP	WD Repeat Containing Planar Cell Polarity Effector
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organization
WOF	World Obesity Federation

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Common And Rare Variant Association Approaches In Turkish Families With Obesity

Obesity, defined as a body-mass index (BMI) greater than or equal to 30, is a multifactorial complex disorder which is currently amongst the major global health issues. Advancements in sequencing technologies have enabled the study of complex disorders influenced by both environmental and genetic factors, leading to the development of various statistical strategies to elucidate the underlying biological mechanisms. To apply these approaches to Turkish patients, we conducted common and rare variant association analyses using various statistical models on whole-exome sequencing (WES) data from 929 obesity patients from 782 families and 2,924 controls. Multiple gene-based aggregative burden and kernel-based methods, empowered by family data, were utilized for rare variant association analyses. Subsequently, pathway enrichment analysis was conducted with the most significantly associated genes identified from the rare variant analyses. Our primary objective was to identify candidate genes and pathways that could provide insights into the molecular mechanisms underlying obesity. Pathway enrichment analysis of the top candidate genes revealed enrichment in the taste transduction, regulation of insulin secretion, and serotonin signaling pathways, all of which have been previously implicated in mechanisms contributing to obesity. However, our common variant analysis suggested that using genome-wide association study (GWAS) data as the base for WES analysis may not be suitable, as significant variants are often located in deep intronic and intergenic regions not covered by WES data.

Keywords: Obesity, Whole exome sequencing, Rare variant association, Relatedness, Polygenic risk score.

ÖZET

Obeziteli Türk Ailelerde Yaygın ve Nadir Varyant İlişkilendirme Yaklaşımları

Vücut kitle indeksinin (BMI) 30'dan büyük veya ona eşit olması olarak tanımlanan obezite, günümüzün önemli küresel sağlık sorunları arasında yer alan çok faktörlü kompleks bir hastalıktır. Dizileme teknolojilerindeki ilerlemeler, hem çevresel hem de genetik faktörlerin etkilediği karmaşık bozuklukların incelenmesini mümkün kılmış ve bu biyolojik mekanizmaları aydınlatmak için çeşitli istatistiksel stratejilerin geliştirilmesine olanak sağlamıştır. Bu yaklaşımları Türk obezite olgularında uygulamak amacıyla, 782 aileden 929 obezite hastası ve 2924 kontrolün tüm ekzom dizileme (WES) verileri üzerinde çeşitli istatistiksel modeller kullanarak yaygın ve nadir varyant ilişkilendirme analizleri gerçekleştirdik. Aile verileriyle güçlendirilmiş çoklu gen temelli toplam yük ve kernel (çekirdek) temelli yöntemler nadir varyant ilişkilendirme analizleri için kullanıldı. Daha sonra, nadir varyant analizlerinden elde edilen en anlamlı şekilde ilişkili genler ile yolak zenginleştirme analizi yapıldı. Birincil amacımız, obezitenin altında yatan moleküler mekanizmalar hakkında bilgi sağlayabilecek aday genleri ve yolları belirlemekti. En iyi aday genlerin yolak zenginleştirme analizi, tat transdüksiyonu, insülin sekresyonunun düzenlenmesi ve serotonin sinyal yollarında zenginleşme olduğunu ortaya koydu; bunların tümü, obeziteye katkıda bulunan mekanizmalarla daha önce ilişkilendirilmiştir. Bununla birlikte, yaygın varyant analizimiz, WES analizine temel olarak genom çapında ilişkilendirme çalışması (GWAS) verilerininin kullanılmasının uygun olmayabileceğini, çünkü anlamlı varyantların genellikle WES verileriyle kapsanmayan derin intronik ve intergenik bölgelerde bulunduğunu öne sürdü.

Anahtar Sözcükler: Obezite, Tüm ekzom dizileme, Nadir variant ilişkilendirme, Akrabalık, Poligenik risk skoru.

1 INTRODUCTION AND AIM

1.1 Obesity

Obesity is a multifactorial complex disorder characterized by the accumulation of excess body fat, which adversely impacts health. It is defined as a body-mass index (BMI) greater than or equal to 30, which is calculated by dividing an individual's weight by the square of their height (1). There has been a substantial rise in the worldwide occurrence of obesity an currently, it's amongst the major global health issues (2). According to the World Health Organization (WHO), 16% of adults (18 years old and over), and 8% of children and adolescents (5-19 years old) of the population were reported to have obesity as of 2022. This indicates that between 1990 and 2022, the global prevalence of obesity increased by more than two-fold. (3). In a 2023 report of World Obesity Atlas, obesity is estimated affect nearly 2 billion adults, children and adolescents by 2035 (4).

Obesity is a significant contributor to various life-threatening disorders, making it a critical public health issue. The condition is linked to a series of critical disorders, including cardiovascular disease (CVD), type 2 diabetes, osteoarthritis, sleep apnea, chronic kidney disease, and several types of cancer, including colon, rectum, esophagus, and liver (5). The Global Burden of Disease (GBD) research has provided alarming statistics on the impact of obesity, revealing that increased Body Mass Index (BMI) values were primarily responsible for 4 million deaths in 2015 alone. Cardiovascular disease emerged as the primary cause of mortality related to high BMI, accounting for 2.7 million deaths, which is approximately two-thirds of the total deaths attributed to obesity. Second leading cause of death in comorbid conditions of obesity was type 2 diabetes, contributing to nearly 1 million deaths. Furthermore, it is predicted that an elevated body fat percentage contributes to around 4-9% of all cancer cases (6). By 2019, total number of deaths caused by higher-than-optimal BMI was estimated to be 5 million by GBD (7). Obesity also imposes a significant psychosocial burden and with individuals often having body image, self-esteem, and life quality issues (8). Multiple studies have shown that around 20% to 60% of people with obesity

suffer from a psychiatric illness including depression, anxiety, and eating disorders (8–13). Childhood and adolescent obesity have shown to affect school performance and is often associated with anxiety, depression, attention deficit hyperactivity disorder. It substantially affects quality of life, compounded by stigma, discrimination and bullying (14).

The economic impacts of the obesity and its comorbidities on the individuals, families, and nations are also crucial. Besides the significant burden on health care systems, obesity also causes reduced productivity in the workplace, permanent impairment, and death leading to inevitable decrease in economic growth (15). GBD estimates that the global costs of obesity will reach 3 trillion USD per year by 2030 and more than 18 trillion USD by 2060 (16).

Weight gain, and ultimately obesity, happens when the intake of energy surpasses the expenditure of energy over a given duration. This results in the buildup of surplus body fat in adipose tissue, ultimately leading to a rise in BMI (17).

Significant increase in obesity prevalence is strongly associated with environmental factors such as high caloric foods and low physical activity (sedentary lifestyle). However, studies have shown that obesity is a complex multifactorial disorder, which indicates that the development and severity of obesity is associated not only with environmental and psychosocial influences but also with genetic factors. This suggests that some individuals have increased susceptibility to obesity due to DNA variations whereas others have protection against it (18,19).

To understand the aetiology and complex nature of obesity, all the environmental and genetic factors that affect food intake, nutrient turnover, lipid metabolism, fat storage in adipose and non-adipose tissues, thermogenesis have to be considered. These processes are regulated by central nervous and endocrine systems in body. This would indicate that genes and genomic regions that regulate these processes are associated with the control of the body weight and composition (17,18).

1.1.1 Genetics of obesity

In the past 2 decades, genetic factors in obesity have been widely studied. First, family studies have shown the role of genetic factors in obesity, indicating that the risk of childhood obesity increases with a positive family history of obesity. Studies on twins have shown that the rate of obesity is between 70% and 90% among identical twins, while it ranges from 35% to 45% in non-identical twins. This demonstrates that genetic variations are essential in the development of BMI and obesity. Moreover, genetic factors that regulate and influence body fat percentage, wait circumference, physical activity level, eating habits and energy expenditure also affects BMI and development of obesity (20,21).

It has been a challenge to enlighten the genetics of obesity due to complex nature of the metabolic and neurological pathways as well as genetics of multifactorial disorders. With the development of technologies such as genotyping arrays and next generation sequencing (NGS) and data analysis methods such as genome-wide association studies (GWAS) and rare variant association testing, discovery of the genetic associations of obesity have rapidly increased (19). Recent research have identified more than 500 genes and 120 chromosomal regions that are linked to obesity in the human genome (22,23). Obesity could be divided into two groups as syndromic and non-syndromic, which are linked to different genes. And non-syndromic obesity could be further categorized as monogenic and polygenic obesity. Syndromic obesity may result from a pleiotropic Mendelian condition, where multiple, seemingly unrelated organ systems are affected by a single gene mutation; or result from chromosomal rearrangements. Monogenic obesity is the result of inheriting mutations in a single gene, while polygenic obesity involves several variations in many genes and genomic loci (22–24).

1.1.1.1 Syndromic obesity

Syndromic obesity is a term used to describe obesity that is accompanied by other physical characteristics such as developmental delay (DD), cognitive impairment,

congenital malformations, organ-specific abnormalities or dysmorphic features. Syndromic obesity is rare, shows high variability in phenotypes, and often follows monogenic inheritance. Syndromic obesity can occur from single gene mutations as well as chromosomal rearrangements (25,26). A recent comprehensive analysis revealed the presence of 79 distinct obesity syndromes, with obesity being identified as a primary characteristic in 55 of these syndromes, while in the remaining 24 syndromes, the occurrence of obesity was greater than that in the general population (27). Among these, best-known genetic syndromes are Prader-Willi syndrome and Bardet-Biedel syndrome.

Prader-Willi syndrome (PWS) is a neurodevelopmental disorder caused by errors in genomic imprinting (epigenetics) with deletion of paternally expressed genes located on chromosome 15. In most cases, PWS results from de novo deletion of paternal 15q11-q13 region, also known as Prader-Willi Critical Region (PWCR). Maternal uniparental disomy (UPD) is seen about 20%-30% of PWS cases. PWS is an example of syndromic obesity resulted from chromosomal rearrangements is characterized by intellectual disability, hypotonia, short stature, dysmorphic facial features, hormonal deficiencies in addition to obesity (28).

Bardet-Biedel syndrome (BBS) is an autosomal recessive multisystem ciliopathy characterized by intellectual disability, progressive rod-cone dystrophy, polydactyly, hyperphagia with severe obesity, type 2 diabetes, hypogonadism, and renal abnormalities. Ciliopathies are genetic disorders result from mutations in genes encoding proteins that localize to the cilium–centrosome complex, causing abnormal formation or function of cilia. By 2024, biallelic pathogenic variants in 25 genes have been associated with BBS. These are BBIP1, BBS1, BBS2, BBS4, BBS5, BBS7, BBS9, TTC8, ARL6, BBS10, BBS12, MKKS, CFAP418, CEP164, CEP290, IFT27, IFT74, IFT172, LZTFL1, MKS1, SCAPER, SCLT1, SDCCAG8, TRIM32, and WDPCP (29). Despite the molecular mechanism of obesity in BBS is not fully understood, impairments in the leptin-melanocortin pathway have been suggested to play a role. (30,31).

Other forms of syndromic obesity encompasses disorders such as 5p13 microdeletion syndrome, 16p11.2 deletion, Albright hereditary osteodystrophy (AHO) associated with GNAS mutation, Alstrom syndrome associated with ALMS1 mutation, CHOPS syndrome caused by AFF4 mutation, Carpenter syndrome caused by RAB23 mutation, Cohen syndrome-VPS13B mutation, Rubinstein-Taybi syndrome resulting from CREBBP mutation, Obesity, hyperphagia, and developmental delay syndrome-NTRK2 mutation (28).

1.1.1.2 Non-syndromic obesity

Non-syndromic forms of human obesity could be originated from monogenic or polygenic inheritance. Monogenic obesity is an uncommon, severe form of obesity that occurs early in life and is caused by a mutation in a single gene in various genes that has a Mendelian inheritance pattern. Polygenic obesity, usually referred to as common obesity, is linked to multiple common genetic variations in many genes and genetic loci.

1.1.1.2.1 Monogenic obesity

Monogenic obesity results from deleterious mutations in single genes and follows Mendelian inheritance patterns, which means it is inherited in autosomal dominant or autosomal recessive manner. Monogenic obesity generally presents with early age of onset and severe clinical findings. Single gene variants in monogenic obesity have large effects and show high penetrance. Majority of obesity cases in population are polygenic obesity, whereas monogenic obesity is rare (28,32).

Molecular and biological mechanisms associated with obesity were primarily identified from monogenic obesity. Primary candidate genes and pathways for obesity were discovered initially in mice studies and these were *ob (obese)* and *db (diabetes)* genes (33,34). Subsequently, it was demonstrated that the ob gene is responsible for encoding leptin, an adipose tissue hormone. It has been established that lack of leptin leads to significant obesity in mice (35). Following closely, the *db* gene in mice was

demonstrated to code for leptin receptor (LEPR), which receives and transmits signals from leptin hormone (36). Knockout *ob* and knockout *db* mice showed severe excessive eating (hyperphagia) and obesity (35,36). In Agouti "lethal yellow" mice, it was revealed that the obesity caused by ectopic expression of agouti protein, which is the antagonist of melanocortin 1 and 4 receptors (MC1R and MC4R), revealing the melanocortin pathway and its association with weight regulation (37).

Early studies to identify obesity genes in human were performed in families, in particularly consanguineous pedigrees, with multiple members with severe childhood obesity and unaffected members. Investigation for candidate genes based on mouse studies was performed with Sanger sequencing. Leptin (LEP) (38) and leptin receptor (LEPR) (39) genes were the first monogenic obesity in human identified. Shortly after this finding, scientists identified the specific genes responsible for monogenic obesity in the melanocortin pathway, which are PCSK1(40), MC4R (41), and POMC (42). The discoveries lead to the deeper understanding of monogenic obesity, mainly leptin–melanocortin pathway and its association with eating habits and body weight.

The leptin gene, LEP, is located on chromosome 7 and it encodes 146 amino acid long peptide hormone leptin. Leptin hormone is produced by white adipose tissue and it mainly regulates satiety and energy expenditure in body. Leptin acts through binding its on receptor, LepR, on the cell surface. The leptin receptors are located on neuronal, cardiac, hepatic, intestinal and pancreatic tissues. Leptin's satiety regulation is through the activation of LepR in hypothalamic neurons. Activation of LepR triggers the activation of STAT3, a transcription factor that promotes the production of anorexigenic peptides. These peptides inhibit food intake and increase energy expenditure. The level of circulating leptin in the blood correlates with the level of leptin in adipose tissue. This relationship acts as a feedback mechanism to maintain a stable internal environment by controlling energy reserves and food consumption. When biallelic pathogenic mutations present in LEP or LEPR gene, this pathway becomes dysregulated, leading to failure in food intake reduction despite the presence of excess adiposity and elevated levels of circulating leptin(43,44). Discovery of leptin signaling pathway created significant interest in clinical application for treating

obesity. However, administering exogenous leptin has shown to be not effective in most patients with obesity when most of them exhibited notably increased plasma leptin levels but reduced responsiveness to leptin (45,46).

The proopiomelanocortin gene (POMC) is located on chromosome 2 and it encodes POMC protein, which is a precursor polypeptide hormone synthesized mainly in the hypothalamus. By a series of post-translational modifications and enzymatic cleavage of POMC precursors, several polypeptides are produced. These include endorphins such as β-endorphin, corticotropins such as adrenocorticotropin (ACTH), and melanotropins such as alpha- beta- and gamma- melanocyte-stimulating hormones (MSH). POMC-derived peptides exert their actions primarily though the two distinct group of receptors, melanocortin receptors (MCRs) and endorphin/opioid receptors (47). Pathogenic mutations in POMC gene that affect the formation of MSHs, which are anorexigenic peptides playing a crucial role in controlling appetite and energy usage, result in obesity (48).

Melanocortin receptor 4 gene (MC4R) is located on chromosome 18, which encodes the G protein-coupled receptor (GPCR) involved in the hypothalamic leptin-melanocortin signaling pathway. MC4R is mainly produced in hypothalamus and serves as a crucial component of a neuronal circuit that is targeted by leptin and involved in maintaining energy balance. MC4R is activated by POMC-derived alpha-beta- and gamma-MSH and ACTH. Both autosomal dominant and autosomal recessive pathogenic loss-of-function (LoF) mutations were reported in patients with monogenic obesity. Heterozygous mutations in the MC4R gene are seen in approximately 1% to 6% of individuals with obesity, particularly those who develop obesity at an early age or during childhood. These mutations can have different levels of impact and expression, resulting in a range of obesity severity and related consequences. Homozygous or compound heterozygous variants however are reported less frequently but present with more severe phenotype (49). On the other hand, gain-of-function (GoF) variants in MC4R have been reported to be protective against obesity and negatively associated with the phenotype (50).

The proprotein convertase subtilisin/kexin type 1 gene (PCSK1) is located on chromosome 5 and it encodes prohormone convertase 1/3 (PC1/3), which is involved in the processing of a various neuropeptides and prohormones in endocrine tissues. The activating cleavage of certain peptide hormone precursors, such as proopiomelanocortin, proinsulin, proglucagon, and proghrelin, which regulates control of appetite, glucose homeostasis, energy regulation, depends on PC1/3 activity (51). Rare biallelic pathogenic variants in PCSK1 have been associated with monogenic obesity with multiple endocrine abnormalities (52).

The majority of mutations causing monogenic obesity have been discovered in individuals who have severe and early-onset obesity, typically before the age of 10. Monogenic obesity often follows an autosomal recessive inheritance pattern, however autosomal dominant inheritance can also occur. Consanguineous populations have higher rates of monogenic obesity as they harbor greater rates of homozygous pathogenic mutations. For instance, in a consanguineous Pakistani population, mutations in LEP, LEPR and MC4R accounts to nearly 30% of cases with severe childhood obesity (53).

Advancements in high-throughput sequencing technologies and discovery of next generation sequencing (NGS) has given an opportunity to extensive candidate gene screening, leading to the discovery of new monogenic obesity genes and pathways. This includes class 3 semphorins (SEMA3A-G), which plays a role in hypotalamic melanocortin circuit development including POMC (54).

1.1.1.2.2 Polygenic obesity

Discovery of single gene variants associated with obesity has paved the way to enlighten the basis of biology and pathophysiology of obesity. However, only a small portion of the cases harbor single gene variants in monogenic obesity genes. This has led to research to focus on polygenic inheritance in obesity (22).

Polygenic obesity, also called common obesity, constitutes approximately 60% of inherited obesity. The variants associated with polygenic inheritance are often common variants, have small effects on their own, but their combined effect is significant and result in certain phenotypes. Furthermore, variants across the genes and genomic regions associated with a polygenic trait differ from one individual to another. Thus, studying polygenic obesity is more challenging in comparison to monogenic obesity. Large-scale genome-wide association studies (GWAS) have led to the identification of new genes and pathways related with polygenic obesity. These studies analyze the association between millions of common genetic variants, without any prior notion about their potential function (22,28). To date, large-scale GWAS have identified over 300 genetic loci for obesity associated traits (55).

One of the most important GWAS discovery for obesity and body weight was FTO locus. FTO is mainly expressed in in adipose tissues and the skeletal muscles in humans, although its most prominent expression occurs in the hypothalamic arcuate nucleus, which plays a significant role in energy balance metabolism (56). This suggest the potential significant effects of FTO in energy balance and appetite regulation. GWASs have demonstrated that the FTO is the most important candidate obesity gene in children as well as adults. Multiple common and non-protein coding single nucleotide polymorphisms (SNPs) in FTO locus exhibited a remarkably strong correlation with an increase in body weight (56,57).

FTO SNPs and their association with BMI were first linked in European individuals with diabetes. Research showed that 16% of people with the homozygous FTO risk allele rs9939609 gained over 3 kg of weight, and their risk of obesity increased by 1.67 times, compared to those without the risk allele (56). The importance of the FTO locus has been verified by numerous GWAS studies on obesity-related phenotypes in individuals of European origin. Multiple GWAS studies on obesity and associated phenotypes in European populations revealed new obesity-associated FTO SNPs in the intron 1 including rs9930506, rs1421085, rs8050136, rs1121980, further confirming the importance of the locus (58–61). Studies in multiple different Asian populations showed that the A allele of rs9939609 was associated with obesity (62).

The FTO risk allele rs17817449 is also significantly correlated with weight in Chinese people (63). Global studies carried in South American (64) and Middle Eastern (65) populations revealed that the FTO risk alleles have the most significant association with BMI in these populations as well. Another study carried in adult Pakistani females with obesity showed the link between rs9939609 variant and increased levels of levels of fasting blood glucose and plasma insulin (66).

People who carry FTO risk alleles show a preference to consuming meals that are high in energy and experience decreased feelings of fullness, leading to overeating and the development of obesity (57). The biological mechanism of association of FTO locus and BMI is complex and is only getting resolved recently. Research suggests that FTO locus may regulate the expression of other genes, such as RPGRIP1L, IRX3 and IRX5. These genes directly and indirectly take part in biological pathways such as appetite regulation, thermogenesis, browning of adipocytes (67,68). For example, it is suggested that obesity associated FTO variant rs1421085 may disrupt ARID5B repressor binding leading to a decrease in the expression of IRX3 and IRX5 during initial phases of adipocyte differentiation. This mechanism could result in a cell-autonomous transition from thermogenesis and browning of white adipocytes to lipid accumulation, increased fat stores, and weight gain (69).

Despite the significant correlation with BMI, FTO alleles still have small effect on obesity susceptibility. A large-scale study conducted by The Genetic Investigation of Anthropometric Traits (GIANT) Consortium found that each FTO risk allele raises the chance of obesity by 1.20 to 1.32 times, and increases BMI by 0.37 kg/m2. Other variants in polygenic inheritance have relatively smaller effect on BMI, ranging from 0.08 to 0.3 kg/m^2 (70). Overall, the association between FTO SNPs with increased high energy food intake, increased appetite and reduced satiety is well-established (70,71). On the other hand, there was no observed correlation between FTO SNPs and physical activity levels (72,73).

Being a monogenic obesity gene, multiple GWAS showed MC4R gene also plays a significant role in polygenic obesity. Common polymorphisms in MC4R, such as

rs52820871 (p.I251L), rs2229616 (p.V103I) and many more have been associated with polygenic obesity susceptibility across different populations. As expected, effects of these polygenic variants on BMI and obesity-associated traits are much smaller compared to MC4R variants that cause monogenic obesity (74,75). A large-scale study in over 500,000 individuals in UK Biobank dataset confirmed that individuals with more polygenic risk variants in MC4R have higher BMI compared to individuals with low polygenic MC4R score (76).

Subsequent GWAS studies have discovered novel genes and genomic regions linked to adiposity characteristics, BMI, and waist-to-hip ratio. Multiple variants in beta-adrenergic receptors ADRB1, ADRB2 and ADRB3, uncoupling receptors UCP2 and UCP3, and SLC6A14 gene have been associated with polygenic obesity. ADRB1 and ADRB3 genes encode beta-adrenergic receptors that plays a role in energy utilization and lipolysis. ADRB1 variants p.Arg389Gly and p.Gly389Arg have been associated with long-term weight gain and childhood and adult-onset overweight (77,78). ADRB2 variant p. Gln27Glu has been linked to increase in BMI, elevated leptin and triglyceride levels (79,80). Common ADRB3 variant p.Trp64Arg is linked to increase in the white adipose tissue and a decreased lipolytic activity (81). Uncoupling receptors UCP2 and UCP3 are highly expressed in brown adipose tissue and take part in thermogenesis. UCP2 variants p.Ala55Val, a fourty-five base long indel in the 3'UTR, and c.2866G>A variant in the promoter region associated have been associated with increase in BMI and obesity (82–84). Finally, SLC6A14 gene modulates the availability of tryptophan for the serotonin synthesis, which affects energy balance and appetite control. It is mainly expressed in hypothalamic regions which control feeding behavior. SLC6A14 polymorphisms rs2071877 and rs2011162 have been strongly linked to polygenic obesity (85,86). Even with the groundbreaking discoveries in polygenic obesity genetics, the existing knowledge of genetic variations can only account for a small portion of the variation in BMI. Further research is needed to uncover obesity associated genetic variants and biological pathways that may play a role in body weight.

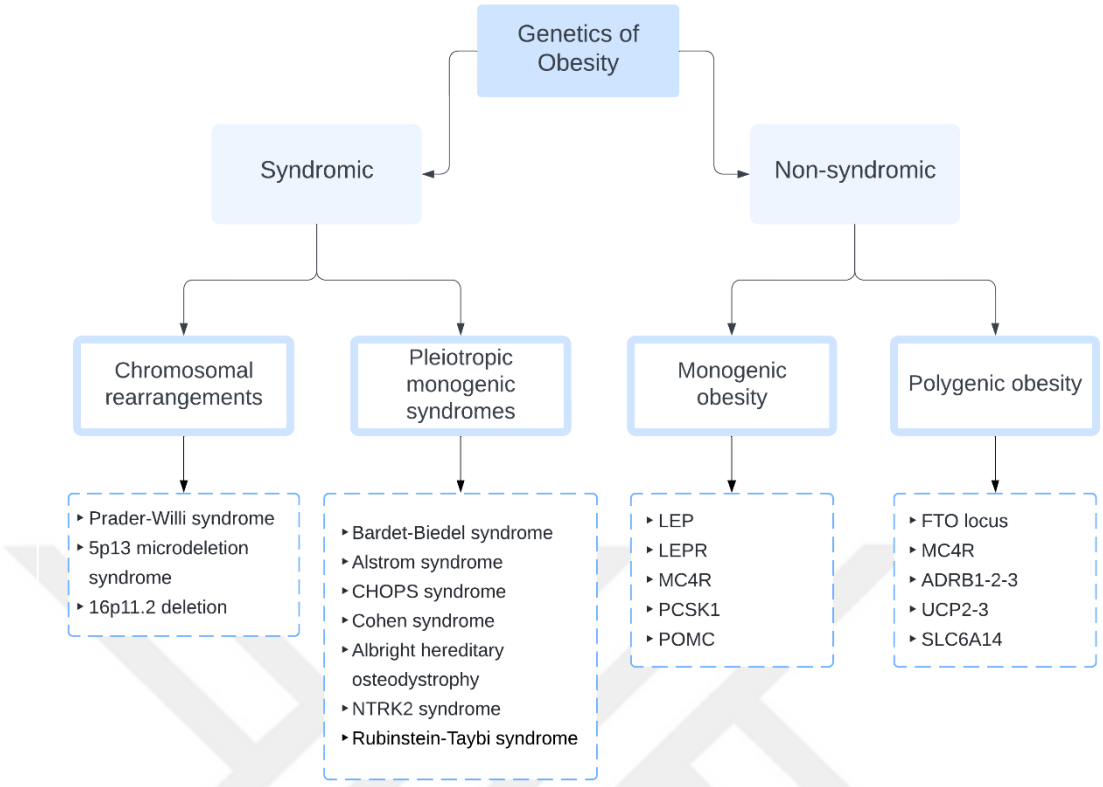

Figure 1. Genetics of Obesity

1.2 Polygenic Risk Score Analysis

Mendelian phenotypes are defined as phenotypical traits that follows Mendel's laws of inheritance. Many dichotomous traits as Mendel evaluated in his pea plant experiments follow this model. However, continuous (quantitative) traits, such as height, body weight, skin pigmentation are complex, meaning they do not present with "single gene mutation", they are inherited through multiple genetic variants across the genome, and sometimes affected by the environmental factors.

Discovery of majority of complex traits and common disorders not following Mendelian inheritance patterns has prompted the development of genome-wide association studies (GWASs). More than 20 years, GWASs have been essential in uncovering the genetic mechanisms of common complex disorders and traits. These traits exhibit a high degree of polygenicity and are linked to several common variations spread throughout the genome, often numbering in the hundreds or even thousands

(87). With the successful application of GWASs many significant genomic loci-trait associations, such as FTO and obesity (58) have been discovered.

GWASs analyze a large number of common genetic variations (usually defined as variants with minor allele frequency ≥1%) without a prior hypothesis about their potential function. The goal is to discover the variants that show a statistically significant association with a given trait or disease. The outcome of GWASs is blocks of SNPs, also known as risk loci, that exhibit a statistically significant association with the trait being studied. To able calculate an individual risk for a complex trait using GWAS results, polygenic risk score (PRS) analysis method was developed (88).

PRS is defined as a single value estimate of an individual's genetic risk for a phenotype, calculated via by combining and measuring the impact of many common variation in genome. Each variant used in PRS have a small effect on its own for a given trait or condition, but the additive effect of the hundreds, thousands of these variants result in a certain phenotype. A PRS is typically formed from summing the variants weighted by their effect sizes (odds-ratio (OR) for binary traits or beta coefficient for continuous traits) derived from a GWAS, also known as reference panel/data. The result is a single score for each individual's genetic risk for a trait or phenotype, with higher scores indicating higher risk. An important factor that effects the PRS calculation is linkage disequilibrium (LD). LD refers to the phenomenon of genetic variations located closely tend to be inherited together. PRS analysis take LD into consideration when calculating risk scores, which results in more accurate prediction than simply summing up the effect sizes of risk alleles. It is necessary to exclude relative individuals in order to address LD accurately (89,90).

In the past decade, hundreds of publications have mentioned PRS, with notable developments in the construction and assessment of PRS and many new applications that have been suggested. However, due to the complexity and challenges in accurate modeling and calculation, PRS has only recently been used in clinical applications in a few disorders. Currently, PRS is used in clinical setups for coronary heart disease

(CHD) and breast cancer, and the data strongly suggest that PRS calculation for these disorders improve patient outcomes (91).

The biggest current limitation of PRS is its limited capacity to accurately apply to a wide range of ancestral backgrounds and study populations. Accurate PRS calculation requires ancestry-matched genotype-level reference data, but the majority of the GWASs were conducted on participants of European ancestries. Using a reference panel constructed from a GWAS study based on one population for a different ethnic group may result in attenuated predictive accuracy of PRS calculation due to unpredictable bias caused by differential genetic drift (92,93). In this study, the summary statistics data of a recent obesity GWAS study conducted in UK Biobank dataset which has 7973 European ancestry cases with severe obesity and 142553 European ancestry controls (94) was used as reference data.

1.3 Rare Variant Association Analysis

GWASs have successfully identified thousands of associations between common variants and complex traits and disorders. However, much of the heritability of these traits and phenotypes is still unknown. Because the low-frequency and rare variants are not investigated in regular GWASs, their impact in contribution to complex traits are not evaluated in these assays. Therefore, novel techniques have been devised in rare variant association studies (RVASs) to evaluate the importance and impact of rare variants on complex traits (95).

With the paradigm-shifting drop in sequencing price, whole exome and whole genome sequencing (WES and WGS) have been widely used in studies of population genetics, as well as rare and complex diseases and traits. Due to extensive population and disease sequencing research, the current number of genetic variations in dbSNP exceeds 60 million. These studies have revealed that most human genetic variations are rare, further indicating the importance of RVASs for complex traits (96,97). In RVASs, burden tests are the first and most common methods that are used. Burden test mainly aims to identify genes that are enriched of rare deleterious variants in

phenotypically similar individuals. Burden-based tests include aggregating individuals who carry rare variations and comparing their average phenotype or disease prevalence with that of those who do not carry these variants. One major limitation of burden-based methods is the strong assumption of all tested rare variants are causal and they influence the phenotype in the same direction. To overcome this issue, kernel-based variance component tests, such as SKAT (Sequence Kernel Association Test) and combination of both burden and kernel-based methods, such as SKAT-O (Optimal Sequence Kernel Association Test) have been developed (96).

Kernel-based variance component tests, such as SKAT are non-burden tests that aggregate individual variant-score test statistics with weights when effects of SNPs are linearly modelled. This enables kernel-based methods to accurately handle genetic variants that have effects in different directions. Kernel methods like SKAT are especially powerful when a genomic region consists both deleterious and protective variants or many noncausal variants are expected to be present in the region. Different kernel methods to calculate p-values can be used depending on the sample and study characteristics. Despite providing significant power and making only a few assumptions of the effects of rare variants, SKAT also has limitations. SKAT is less powerful and prone to produce type I errors for small case-control studies. Additionally, if the majority of the tested rare variants in a region are causal and influence the phenotype in the same direction, SKAT becomes less powerful than burden tests (98).

To be able to handle the both scenarios that a genomic region might have many causal variants, or it may have many noncausal variants, combination of burden test and kernel-based tests have been developed. The main purpose of these tests, such as SKAT-O, is to take advantage of the strengths of both tests from which they originate (98).

Association analyses frequently employ linear mixed models. Linear mixed models are an expansion of linear models that incorporate both fixed and random factors. They are especially powerful when there is non-independence in data, such as

relatedness in a genetic dataset. To be able to handle the relatedness in a regression model, multiple new methods that use a linear mixed model to test for both quantitative and binary traits have been developed. These methods include a random variable to account for the correlation between individuals within the same pedigree, extending the kernel-based methods for handling pedigree data (99).

In this study, Burden test, SKAT, SKAT-O and Efficient Hybrid Test (EHT) are used to perform rare variant association testing in WES data.

1.4 Pathway Enrichment Analysis

Pathway enrichment analysis (PEA), also known as functional enrichment analysis, is a method used to identify biological pathways that are significantly overrepresented among a group of genes compared to what would be expected by chance. This method helps identify specific pathways that are enriched in a gene set by employing statistical techniques and leveraging data from biological processes and pathway databases available online (100).

With the widespread use of high-throughput technologies, the term “big data” was introduced to biological studies. Multiomics studies have become a general practice in biology and these studies often generate large and complex datasets. Analyzing and interpreting this big data presents a major challenge for omics studies. Functional evaluation of resulting gene lists, which can range from tens to hundreds of genes, is complex and error-prone when done manually without statistical approaches. This challenge led researchers to develop new computational methods such as PEA. PEA has emerged as a fundamental bioinformatics method for omics studies by helping the interpretation of large gene lists through the identification of statistically enriched pathways. The introduction of PEA and similar computational methods has transformed the field of bioinformatics. These tools make interpreting data more accurate and reliable, while also speeding up the discovery process. Researchers can now manage larger datasets more efficiently and uncover new relationships and patterns that were hidden before. This is especially important for personalized

medicine, as understanding individual genetic differences can lead to customized treatments and better patient outcomes.

At the very basis of PEA, enriched pathways in a gene list are tested using different statistical approaches considering the number of genes in the list annotated in a certain pathway and their relative ranking. There are multiple tools and methods for PEA with different statistical approaches. (101,102).

In this study, Enrichr which is a web-based tool developed by Chen et al. (103) is used to perform PEA based on rare variant association analyses results.

2 BACKGROUND

2.1 Pathophysiology of Obesity

Obesity fundamentally develops from an imbalance between the calorie intake and energy expenditure. However, the pathophysiology of obesity is associated with a complex interplay of factors which are genetic, environmental, and socioeconomic. Although the environmental and socioeconomic factors that influence behavior and cannot be addressed solely at the molecular level, the identification of key genes, molecules, and pathways involved in obesity unveils fundamental physiological mechanisms that can be addressed at the molecular level. Beyond the simplistic “calories in minus calories out” perspective on obesity, complex physiological factors involved in energy metabolism, neurological system, immune system, and gut microbiome play a pivotal role in the disorder.

The perspective that obesity simply occurs when caloric intake surpasses energy expenditure has led many to view obesity as a result of personal failings such as laziness, sloth, or lack of willpower. Nevertheless, mounting evidence suggests that the development of obesity is far more complex than a simple imbalance of calories. Essentially, humans possess an "evolutionary physiology" that is inclined to preserve body fat for survival. This critical evolutionary factor has fostered predisposition to excessive weight gain in today’s era of abundant and readily available calories. Hence, understanding the development of obesity primarily revolves around comprehending the energy homeostasis system, which fundamentally regulates energy intake and expenditure (104).

Research that investigates the adaptive responses of humans and animals to changes in body weight yields valuable insights into the behavior of the energy balance system. Caloric restriction leads to weight loss and causes an increase in appetite drive while decreasing energy expenditure. These responses exhibit resistance to additional weight loss and promote the restoration of lost weight. They can endure for several years, as long as body-fat levels have not returned to their original state. Conversely,

research involving individuals of normal weight show the response for weight gain induced by forced overfeeding is an increase in energy expenditure and reduced hunger. Once the overfeeding stopped, decrease in appetite, and increase in energy expenditure was observed (105). This observation indicates that the dysfunction of energy homeostasis is necessary to development of obesity. Ongoing research indicates that the energy homeostasis system is highly complex, consisting of both central and peripheral pathways that have complex interactions with each other.

2.1.1 Leptin-melanocortin pathway

As previously mentioned, molecular mechanisms associated with obesity were primarily identified from in mice studies. The *ob* gene was shown to encode leptin, a peptide hormone released from adipose tissue (35) and the *db* gene in mice was shown to encode leptin receptor, which receives and transmits signals from leptin hormone (36). Knockout *ob* and knockout *db* mice showed severe excessive eating (hyperphagia), hyperglycemia, hyperinsulinemia, and obesity (35,36).

Following these groundbreaking discoveries, research on human and mouse genetics has revealed several neural pathways in the brain that have a central function in regulating mammalian appetite and metabolism. The most well-defined mechanism is the hypothalamic leptin-melanocortin signaling pathway, which genetic disruptions primarily account for most monogenic severe obesity disorders. Moreover, more subtle polymorphic variations in this pathway influence the population's distribution of body weight.

Leptin, encoded by human LEP gene, is a peptide hormone primarily synthesized by white adipose tissue. The amount of leptin in the bloodstream is directly correlated with the amount of adipose tissue. Leptin carries out its actions by binding to the leptin receptor, which is encoded by human LEPR gene, on the surface of cells. Leptin receptors can be found on various types of tissues, including neuronal, hepatic, pancreatic, cardiac, and perivascular intestinal. The primary action site of leptin is brain, specifically the brainstem and hypothalamus. This was discovered as a result of

multiple studies that showed intravenous injection of leptin led to its accumulation in specific brain regions (106). Moreover, has been demonstrated that intravenous administration of leptin stimulates certain groups of neurons pre-dominantly located in brainstem and hypothalamic nuclei which are known to play a role in controlling energy balance (107). Adipose-derived leptin crosses the blood-brain barrier and accesses the hypothalamic arcuate nucleus (ARH), which is the central regulator of energy homeostasis. Here, leptin acts on ARH by inhibiting and stimulating the components on melanocortin system (108).

The melanocortin system is a critical component of ARH that regulates feeding behavior and energy homeostasis. It consists of three main components: the pro-peptide proopiomelanocortin (POMC), which is processed by prohormone convertases into biologically active moieties such as α-melanocyte stimulating hormone (MSH), β-MSH, γ-MSH, and adrenocorticotrophin (ACTH); five G protein-coupled melanocortin receptors (MCRs, MC1R-MC5R) which mediate the actions of these peptides; and finally endogenous antagonists of these receptors, known as agouti and agouti-related protein (AGRP). ARH neurons are named based on expressing one of these 3 components; POMC neurons, AGRP neurons, MCR neurons (109).

AGRP neurons secrete orexigenic neuropeptides AGRP and neuropeptide Y (NPY), along with the neurotransmitter γ-aminobutyric acid (GABA). AGRP is a naturally occurring substance that acts as a reverse agonist for melanocortin receptors, inhibiting their activity. POMC neurons synthesize Pomc, which undergoes post-translational modifications to generate several melanocortin ligands, namely α-, β-, or γ-MSH. These ligands act as endogenous agonists of MCR. There are five subtypes of MCR, namely MC1R, MC2R, MC3R, MC4R, and MC5R. Among these, MC3R and MC4R are mainly found in the brain. The binding of these receptors, especially the MC4R, with the natural melanocortin ligands plays a crucial role in the regulation of energy balance. (110). In the “the classic leptin-melanocortin model” leptin takes action by stimulating POMC neurons and inhibiting AGRP neurons in ARH. This results in activation of the MC3R and MC4R, causing the reduce in food intake and decrease in appetite; suggesting that leptin directly controls satiety on POMC and

AGRP neurons. In the classic model, elevated levels of leptin would stimulate POMC neurons and suppress AGRP neurons, thereby fostering a feeling of satiety. Conversely, reduced levels of leptin would stimulate AGRP neurons and suppress POMC neurons, so stimulating food consumption. When activated, the primary AGRP and POMC neurons affect secondary neurons, which are located out of ARH, by synthesizing NPY, AGRP, or POMC-derived α-MSH (108,110). These peptides would subsequently trigger broad impacts on energy homeostasis through their action on second-order neurons. However, recent studies propose that leptin directly acts on AGRP neurons to control appetite, but it may not control food intake by directly acting on POMC neurons, as the classic model suggests (111).

Several studies have shown that the POMC neurons are specifically affected by leptin in regulating glucose metabolism, rather than influencing food consumption (112–114). Nevertheless, this does not diminish the significance of POMC in controlling energy balance. In fact, POMC plays a crucial role in regulating energy balance. Recent studies have conclusively shown that LEPR expressing POMC neurons are not directly involved in the central regulation of feeding, and other POMC neurons that do not express LEPR exist. These LEPR- POMC neurons could be involved in energy homeostasis. Studies using single-cell sequencing have revealed the presence of numerous different transcriptional patterns in POMC neurons, with just a small percentage expressing Lepr. The majority of POMC neurons likely play a primary role in energy homeostasis (115,116). Deletion of POMC cells in mice resulted increased food intake and body weight in mice (117). Furthermore, loss of function mutations in human POMC result in severe obesity. This further suggest the critical role of POMC in regulating energy balance, and it can be concluded that LEPR-POMC neurons are likely to be involved in this process (111).

2.1.2 Neurotransmitters

Several neurotransmitters in the reward pathways are involved in energy homeostasis through regulating food intake and/or energy expenditure. Two of major neurotransmitters are serotonin and dopamine.

Serotonin is synthesized in the raphe nuclei located in the brainstem and has a role in the controlling food intake by sending signals to various brain regions involved in homeostatic and reward regulatory systems. These two main neurological pathways regulate eating behavior. Hedonic or reward pathways are primarily located in corticolimbic areas and homeostatic pathways are mainly located in hypothalamus and brainstem as mentioned before (118). The hedonic system primarily promotes food-seeking behavior by signaling pleasure or reward after food consumption whereas homeostatic system process metabolic feedback for energy intake and expenditure. Serotonin signaling in central nervous system is proposed to reduce the motivational and reward-related food consumption. Conversely, peripherally acting serotonin promotes energy absorption and storage (119). In homeostatic pathway, reduced serotonin signaling in the hypothalamus plays a role in obesity by disrupting negative feedback from ingested energy on food intake, leading to excessive eating. In the simplistic model, serotonin mediates food intake in ARH by activating α-MSH which is produced by POMC neurons, and inhibiting NPY and AGRP neurons. Overall, higher levels of serotonin signaling are associated with reduced food intake, while lower levels of serotonin signaling lead to hyperphagia and weight gain (118–120).

Dopamine is synthesized in the ventral tegmental area (VTA) and substantia nigra in brain and plays a major role in reward signaling in food intake. Dopamine signaling mainly mediates food consumption in hedonic pathways. Food that is enjoyable activates dopaminergic neurons in VTA that connects to nucleus accumbens, amygdala and prefrontal cortex. This brain regions together are engaged in reward-related behavior. Studies have shown that reduced dopamine signaling promote overeating of palatable food to compensate for reduced reward signaling in brain (120).

2.1.3 Gastrointestinal hormones

Gastrointestinal (GI) hormones mainly regulate food digestion and motility of gastrointestinal tract. Another important function of these hormones secreted by endocrine cells in the gut, stomach and pancreas is to signal appetite and regulate food intake regulation, thereby contributing to energy homeostasis in body. Several GI

hormones are secreted in pre-prandial (before eating) or postprandial (after eating) periods to promote satiety or hunger. Differences in circulating levels of these hormones have been observed between obese individuals and those with normal weight. This suggest that dysregulation of GI hormones may contribute to pathogenesis of obesity. Two orexigenic (hunger) hormones, ghrelin and motilin; and three anorexigenic (satiety) hormones, cholecystokinin, glucagon-like-peptide-1, peptide-YY are involved in the process (121).

Ghrelin is a 28 amino acid long orexigenic peptide hormone which is synthesized in stomach. Ghrelin is primarily secreted during preprandial state to stimulate hunger and influence food intake. This is achieved by ghrelin binding to growth hormone secretagogue receptors on AGRP/NPY neurons, which promotes food consumption. Ghrelin also stimulates VTA and hippocampus, indicating that it also plays a role in hedonic pathways and reward signaling. Fasting ghrelin levels are lower in obese patients, and obesity is associated with reduced postprandial ghrelin suppression. Another orexigenic hormone motilin is a 22 amino acid long orexigenic peptide hormone which is synthesized in small intestine. Main role of motilin is promoting GI motility; however, recent studies have shown that it is one of the key signals of hunger. In obesity, higher preprandial plasma motilin levels with less fluctuations have been observed (110,121).

Anorexigenic hormones of GI include cholecystokinin, glucagon-like-peptide-1, and peptide-YY. These hormones are mainly secreted in different parts of intestine and colon in response to food intake. Postprandial secretion of the anorexigenic hormones promote satiety. In obesity, delayed and/or reduced levels of these hormones after food consumption have been observed. However, it is not yet established whether altered levels of these hormones contribute to cause of obesity or are a consequence of it (122).

2.1.4 Gut microbiome

Gut microbiota have been defined as a separate metabolic organ in the body involved in many biological processes including metabolic, immunologic, and

endocrine processes. It is one of the largest components of human body, weighing around 1 to 2 kilograms and contains over 100 trillion microbial cells. This system mainly digests the food that cannot be digested by human cells. The gut microbiota produces many substances that regulate physiological activities including fatty acids, vitamins, anti-inflammatory and analgesic substances, as well as neurotoxins and immunotoxins. These substances enter the human blood and circulate around the body, thus get involved in many biological processes. A healthy gut microbiota is essential for proper metabolic and immunological activity. Alterations and imbalances of gut microbiota have been observed in individuals with cardiovascular diseases, diabetes and obesity. This observation has led studies to focus on gut microbiota and its association with critical physiological processes such as energy homeostasis in human body (123).

It has been demonstrated that gut microbiota communicates with brain through many pathways including endocrine, neural and immune system. Gut microbiota influences central nervous system, and central nervous system affects the composition of gut microbiota through the gut-brain axis (124). Initial studies in "germ-free" mice showed that this specific strain of mice lacking gut microbiome are leaner than control group mice despite consuming more calories. Fecal transplantation from control mice to germ-free mice resulted in high insulin resistance, increased body fat, indicating the association between gut microbiome with energy regulation and body fat rates (125). Another study showed obesity might be associated with two dominant bacterial phyla: Firmicutes and Bacteriodetes. In obese mice, proportional increase in Firmicutes and nearly 50% decrease of Bacteriodetes were observed (126).

Similar results were obtained in human obesity studies, where the abundance of Firmicutes in obese children's gut was increased whereas proportion of Bacteriodetes was decreased (127). However the evidence that altered gut microbiota is causal for obesity is not well established yet. Further studies are needed to reveal the causal-consequential relationship between gut microbiome and obesity (123).

2.2 Polygenic Risk Score Analysis

GWASs has become one of the focus areas to enlighten the pathophysiology of complex disorders. As majority of common disorders such as CVD, obesity, type 2 diabetes are complex, the interest in calculating individual risks for a trait or a disorder with PRS have significantly increased. However, due to the complexity and challenges in accurate modeling and calculation, PRS has only recently been used in clinical applications in a few disorders.

The aim of PRS is calculating a single value estimate representing an individual's susceptibility to a specific trait or disorder. This calculation is based sum of each variant's value weighted by effect size estimates derived from GWAS summary statistic data. To avoid the overfitting, which occurs when a model is over-optimized to sample data and performs poorly in independent datasets, there are three important tasks that needs to be handled: Using ancestry-matched genotype-level reference data, adjusting effect size estimates, and accounting for LD. To adjust the effect sizes, two main approaches have been used. First method is classical PRS approach, which is p-value thresholding. In p-value thresholding, only SNPs with a p-value below a certain threshold are included in the calculation. SNPs that fail to pass this threshold is excluded from the calculation. Because there isn't a certain optimal threshold value, PRS is usually calculated with different thresholds and the prediction is optimized accordingly. Second method is using different statistical methods, such as LASSO, Bayesian approaches, ridge regression to perform shrinkage. Each method performs shrinkage with different parameters and approaches, but the main goal is to include the SNPs that are associated with the trait with accurate estimate sizes. For accounting LD, there are also different strategies. In the most common approach, called clumping, SNP with the smallest p value at a locus is retained and others are excluded. The goal is to ensure retained SNPs are mostly independent of each other, so overfitting doesn't occur. A different strategy is LD modelling, which includes all SNPs without clumping and adjusts based on LD between the SNPs (90).

2.3 Rare Variant Association Analysis

GWASs commonly investigate common variants and their association with complex phenotypes, however a large missing heritability for these traits is observed. Studies have shown that the majority of human genetic variations are rare, indicating the significance of evaluating these variants. With RVASs, rare variants that are not investigated in GWASs and their associations with certain traits are assessed. There are multiple different statistical approaches used in RVASs, including burden tests, kernel-based methods such as SKAT, combined methods such as SKAT-O, linear mixed models and generalized linear mixed models. Each of these tests comes with its own advantages and disadvantages. In cases where there is no prior information about the proportion of causative variants in a region, employing multiple methods and optimizing the results can aid in building a robust disease model (96).

2.3.1 Burden tests

Burden tests are widely used in RVASs. The methodology for burden tests is based on evaluating the combined impact of several rare variations within a gene or genomic loci, rather than examining each variant separately. Burden test combines the information of multiple rare variants in a gene or region into a single score, then test the association of this score with the trait by comparing phenotypically similar individuals. In the simplest form, this test asks the question whether individuals who carry rare variants in a gene exhibit similar physical traits or not. This question is assessed by combining the carriers of rare variants and comparing their phenotype or disease status with non-carriers (95).

One of the most commonly used burden-based tests is cohort allelic sums test (CAST). This test is simply based on the difference between the number of rare variants observed in cases and controls. It is applied to assess whether this difference is statistically significant or not. CAST can be implemented with Fisher's Exact Test or chi-square test (128). Another burden-based method is combined multivariate and collapsing method (CMC). CMC is a unified method which uses allele frequencies to

divide variants into subgroups. Within each group, variants are collapsed into a marker variable. Then, a multivariate test, usually Hotelling's T2 test is applied to assess the groups of marker variables (129).

Burden-based tests are very powerful when the majority of variants are causative, and they affect the phenotype in the same direction. This power becomes a significant limitation when there are variants in a gene affecting the trait in different directions or only a small portion of variants are causative (96).

2.3.2 Kernel-based methods

Burden based tests lose power in the presence of variants affecting the phenotype in different directions or a small portion of variants are causative as mentioned above To overcome this issue, kernel-based variance component tests have been developed. Instead of simply aggregating variants, kernel-based tests aggregate individual variant-score test statistics with weights when effects of SNPs are linearly modelled. This allows accurate handling variants that have effects in different directions. Kernel methods are particularly powerful when a genomic region contain both deleterious and protective variants or many noncausal variants are expected to be present. Kernel methods transform data with nonlinear effects into a higher-dimensional space, enabling the modeling of complex relationships between data points. Another important feature of these tests is their flexibility to accommodate non-genetic covariates due to their linear foundation. Moreover, kernel-based methods are computationally efficient, making them well-suited for application in large-scale genetic studies (96).

2.3.2.1 SKAT

SKAT is a widely used kernel-based variance componence tests in RVASs as mentioned previously. SKAT uses multiple regression approaches to test the association between variants in a gene and the trait. Traits can be binary, as case-control, or continuous. As proposed by Wu et al. in 2011 (98), SKAT can be used both

for common and rare variants. However, it is commonly used in RVASs. Using a regression-based framework such as SKAT allows to calculate a p-value for each gene or region and adjust the phenotype according to covariates. SKAT essentially directly regress the phenotype on variants and covariates, enabling different variants to have effects in different directions or have no affect at all. Another important aspect of this method is that it does not require selection thresholds which eliminates the possibility of excluding variants that are associated with the phenotype. Different kernel methods for calculating p-values in SKAT can be used depending on the sample and study characteristics. These include moment-matching methods, Davies method, and Kuonen method (96).

In 2007, Liu et al. proposed the use of variance component score test for pathway-level analysis in microarray expression data (130). This method was then modified by Kwee et al. for aggregative genetic variant testing (131). This method, called moment-matching, aims to approximate the null distribution to accurately calculate p values in kernel tests with large sample size. Moment-matching methods involves calculating the first two moments of the score test statistic Q value under the null hypothesis. These moments are then utilized to estimate the degrees of freedom and a factor to rescale the chi-square distribution. The base of moment-matching is aligning the observed moments of sample distribution to expected moments of a model's distribution in order to estimate the model's parameters. The method is mainly used when the type of sample distribution is unknown or challenging to estimate. Both Liu et al. and Kwee et al. used the Satterthwaite approximation, which is a parameter estimation approach to estimate the effective degrees of freedom in a two-sample t-test. Moment-matching methods are less powerful in small sample sets or in the presence of non-standard distribution (132).

To achieve a more accurate approximation of the distribution of score test statistics, Wu et al. (133) have proposed to apply Davies method (134). This method is an algorithm that can calculate the distribution function of a linear combination of independent and identically distributed chi-squared random variables. Davies method aims an analytic solution to compute p-values, involving inversion of the characteristic

function of a linear combination of chi-squared random variables. This method is widely used in genetic association studies due to its adaptability to various study designs as well as its superior performance compared to other methods. However, the challenge of Davies method is the necessity to specify a numerical accuracy parameter. Poor specification of this parameter can lead to inaccurate results, particularly for very small p-values. Another limitation for this method is that it can be conservative for small sample sets (132).

Chen et al. have introduced to use the saddlepoint method, which is proposed by Kuonen (135), in genetic association studies. Kuonen's saddlepoint method offers a highly accurate estimation for the probability density function or probability mass function of a distribution, based on the moment generating function. Unlike Davies method, the saddlepoint approach does not require a specification of accuracy parameter. This advantage enables the saddlepoint method to handle extremely low p-values effectively (136). Similar to the Davies method, Kuonen's method can also be conservative for small sample sets (132).

Although SKAT offers substantial power and relies on only few assumptions regarding the impact of rare variants, it has several limits. SKAT exhibits reduced statistical power and increased tendency to produce type I errors when applied to small case-control studies. Furthermore, if the majority of the tested rare variants examined in a certain region are determined to be causal and have an impact on the phenotype in the same direction, the power of SKAT decreases compared to burden tests (98).

2.3.2.2 SKAT-O

Both burden test and SKAT have their own advantages and limitations when used in genetic association studies. To be able to handle the both scenarios that a genomic region might have many causal variants, or it may have many noncausal variants, SKAT-O, a combination of burden test and SKAT have been developed. The main purpose of SKAT-O is to maximize the power of the test by taking advantage of the strengths of both tests from which it originates. Furthermore, SKAT-O holds proper

type I error rates for association studies with a small sample size, exhibits higher power across a broad range of settings, and demonstrates greater robustness compared to both SKAT and burden tests (98).

SKAT-O dynamically adjusts its behavior based on the comparative power of the burden test and SKAT. It functions like the burden test when the burden test is more powerful, and functions like SKAT when SKAT offers greater statistical power than the burden test (98). This adaptive behavior is enabled by defining a parameter ρ, which allows for capturing signals with different degrees of correlation. When ρ=1, the test functions as burden test, assuming all variants have the effect in same direction. When ρ=0, the test functions as SKAT, assuming variants may have effects in opposite directions.

SKAT-O offers flexibility to adjust phenotype by covariates maintains computational efficiency similar to SKAT. However, SKAT-O p-values may be less reliable compared to SKAT p-values, and the performance of SKAT-O implementations is influenced by factors such as the number and correlation of genetic variants (132).

2.3.3 Linear mixed models

Linear mixed models are extensively used in association analyses due to their exceptional flexibility and straightforward interpretability. LMMs are an extension of simple linear models incorporating both fixed and random effects. They are particularly effective in scenarios where data exhibit non-independence due to factors such as relatedness or population stratification. The flexible nature of LMMs allows for effective testing of both binary and continuous traits. To address non-independence within regression models, several methods have been developed that utilize linear mixed models to test for both quantitative and binary traits. These methods incorporate a random variable to consider the association between related individuals, thereby extending the kernel-based methods for handling pedigree data (99).

2.3.4 Generalized linear mixed models

Generalized linear mixed model algorithms offers flexible and powerful models for correlated data analysis of GWASs. Unlike linear mixed models (LMMs) which assume a normal distribution of the response variable, GLMMs can handle non-normal response variables, making them more accurate and powerful for diverse data structures. Moreover, GLMMs can be utilized to test the associations for both binary and continuous trait and phenotypes (137).

The ability of GLMMs to accommodate different types of data is particularly important in the context of large association studies, where the traits being studied often do not follow a normal distribution. For example, many phenotypic traits in humans, such as the presence or absence of a disease (a binary trait) or the measurement of blood pressure (a continuous trait), require flexible modeling approaches. GLMMs are able to handle these varying data types, making them a valuable tool for researchers seeking to understand the genetic underpinnings of complex traits.

One of the key advantages of using GLMMs in variant association testing is their ability to account for both fixed and random effects. This is crucial because genetic data often involve hierarchical structures, such as relatedness between individuals or varying population structures. GLMMs can model these structures effectively, providing more accurate estimates of genetic associations. This leads to more reliable identification of candidate genes and genetic variants that are associated with specific traits.

Additionally, GLMMs offer improved statistical power compared to traditional LMMs. By accommodating non-normal response variables, GLMMs reduce the risk of model misspecification, which can lead to biased estimates and reduced power to detect true genetic associations. This makes GLMMs particularly useful in studies with complex data structures or when dealing with traits that exhibit significant deviations from normality (137).

2.3.4.1 Variant-set mixed model association tests

Population stratification and relatedness are significant sources of confounding variables in genetic association studies. LMMs incorporating a genetic relationship matrix (GRM) effectively address relatedness and population stratification. However, they may struggle to accurately analyze genetic associations of binary traits due to challenges in specifying appropriate mean-variance relationships. For this reason, Chen et al. have proposed a series of variant-set mixed model association tests (SMMATs) for binary and continuous traits, using GLMM framework. SMMATs can be applied to large-scale WES and WGS studies. They account for familial relationships and are designed to effectively control the type 1 error rate, particularly for binary traits. Integrating relatedness into the statistical models especially important to enhance the test results by distinguishing segregating and non-segregating variants, which leads to more robust results. Family data plays a critical role in all genome studies, whether it is diagnostic testing for a rare disorder or an association study that asses a disease or a phenotype, the first step after testing probands is evaluation of segregation of the variant in affected and unaffected family members. This segregation information is critical for confirming the biological relevance of candidate genes and variants. Accounting for family relationships with both affected and unaffected individuals within the same pedigree enables the development of more reliable statistical models, which are essential for identifying and validating candidate genes and variants with biological relevance.

The SMMAT framework comprises four tests: the burden test (SMMAT-B), SKAT (SMMAT-S), SKAT-O (SMMAT-O), and an efficient hybrid test (SMMAT-E) that integrates the burden test and SKAT, offering a potential alternative to SKAT-O. All four SMMATs share a common reduced model under the null hypothesis, GLMM with only covariates, which is fitted once for all genetic variant sets in an analysis. This approach allows these tests to be constructed using shared single-variant scores and their covariance matrices, thereby enhancing computational efficiency. The GLMM used in SMMATs assumes a Bernoulli distribution for the phenotype, utilizing a logit as link function. Initially, a null model is constructed assuming no association

between the variant set and the trait. Then the variant set is incorporated as a predictor variable in the GLMM. The model is then compared to the null model to assess if there is a significant association between the variant set and the trait. If a significant association is detected, the GLMM is refined by adding covariates such as GRM to address confounding factors. The final model is used to estimate the effect size of the variant set on the trait and identify potential genetic markers for trait prediction.

Each SMMAT have its own advantages to address different confounding factors and data structure. SMMAT-B assumes a normally distributed variant effects and uses burden testing to assess the association between genetic markers and the trait. SMMAT-S, which is a SKAT framework, takes a kernel-based approach to test the null hypothesis and estimate the effects of variants. SMMAT-S has no prior assumption of variants in a gene effecting the trait in the same direction. SMMAT-O witch is the linear combination of SMMAT-B and SMMAT-S combines both burden and SKAT approaches to assess the association between genetic variants and the trait. And finally, SMMAT-E, which is an alternative joint test for SMMAT-O, combines burden test and adjusted SKAT which are asymptotically independent from each other. The method use matrix projections to estimate the modified SKAT statistic, taking into account a global null model that does not include any fixed effects for the specific genetic load of the variant set. SMMAT-E is especially powerful in the presence of large-scale datasets and overpowers the classical SKAT-O approach by minimizing the computational time (138).

2.4 Pathway Enrichment Analysis

Pathway enrichment analysis (PEA) is a bioinformatics method employed to detect biological pathways that exhibit a much higher representation within a set of genes. This approach utilizes statistical methods and uses data from online biological process and pathway databases to find distinct pathways that are enriched in a gene list. Omics studies generate large and complex datasets that pose challenges for analysis and interpretation. These studies often yield tens to hundreds of candidate genes requiring evaluation in terms of functional and biological annotations. Manual

evaluation of such gene lists is laborious, time-consuming, and prone to errors. Therefore, methods like PEA have been developed to simplify this process and provide insights into the functional implications of genes (100,101).

Multiomics studies generate large and complex datasets that are challenging to analyze. Manually evaluating gene lists, which can include tens to hundreds of genes, is complex and prone to errors. To address this, researchers developed computational methods such as PEA, which has become essential for interpreting large gene lists by identifying statistically enriched pathways. PEA was initially developed to evaluate RNA expression data, which involves analyzing hundreds to thousands of genes with varying expression profiles. Since then, PEA has been applied to other studies, including GWAS, metabolic profiling, and single-cell sequencing. PEA uncovers patterns in experimental results, providing insight into the underlying molecular basis of disorders (139).

Several approaches that use different statistical testing methods for PEA have been developed. PEA methods can depend on ranking. Ranked tests take input of gene lists as a ranked list whereas non-ranked evaluates each gene equally without any ranking.

Based on the null hypothesis, there are two main approaches for PEA. These are competitive and self-contained. Competitive approach calculates p-values by comparing whether the given gene list is significantly enriched in pathways compared to remaining genes of the entire genome. Self-contained approach acts on the assumption of that genes in a specified list have the same level of association with the phenotype. It directly tests the association between the given gene set and pathways without any comparison with the genes in the rest of the genome. Competitive methods successfully show enrichment of genes in a particular pathways whereas self-contained methods are more effective in discovering new pathways. Both methods have their own advantages and disadvantages. For this reason, hybrids models that are empowered by both methods are used in various algorithms.

Another category of PEA is based on mathematical model that is used to calculate p-values. PEA tests typically either use permutation or exact test, such as Fisher's exact test, to compute p-values. Exact tests, like Fisher's exact test, provide a direct approach to calculating p-values. These tests compute the exact p-values based on the observed data without relying on approximations or resampling methods. This makes exact tests highly efficient and precise, especially when dealing with small sample sizes or sparse data. Their ability to deliver exact p-values ensures a high level of accuracy, which is essential for confirming the significance of gene-pathway associations in genetic studies. On the other hand, permutation tests offer a flexible and customizable approach to p-value calculation. Permutation tests involve resampling the data multiple times to create a distribution of the test statistic under the null hypothesis. By comparing the observed test statistic to this distribution, empirical p-values can be estimated. This method is especially useful when dealing with complex or non-standard data structures, as it does not assume a specific distribution of the test statistic. The flexibility of permutation tests allows researchers to tailor the analysis to the specific characteristics of their data, enhancing the reliability of the results.

Both exact tests and permutation tests have been modified and incorporated into various PEA algorithms to suit different research needs. For instance, hybrid models that combine elements of both exact and permutation tests have been developed to leverage the strengths of each approach. These hybrid models can provide more accurate and robust p-value estimates by balancing the precision of exact tests with the flexibility of permutation tests.

To avoid significantly enhanced p-values due to repeated testing, multiple-testing correction methods are used in PEA algorithms. Multiple-testing correction allows to systematically reduce p-values for more accurate calculation. The most common correction methods for PEA are Benjamini–Hochberg false discovery rate (FDR) and Bonferroni correction. The FDR approach manages the anticipated ratio of false positive findings within the identified important outcomes, making it particularly useful in scenarios when dealing with large datasets and multiple comparisons. This

method is less stringent than the Bonferroni correction, which adjusts p-values by dividing the desired significance level by the number of tests performed. While the Bonferroni correction reduces the likelihood of type I errors, it can be overly conservative, especially in high-dimensional data, leading to a higher rate of type II errors (100,101).

Incorporating these correction methods into PEA algorithms enhances the reliability of the results by mitigating the risk of false discoveries. The choice of correction method can depend on the specific context of the study and the balance between sensitivity and specificity that is desired. Employing the combination of these methods is often preferred to ensure a comprehensive and robust analysis (100, 101).

In this study, Enrichr which is a web-based tool that uses Fisher's exact test and p-value adjustment developed by Chen et al. (103) is used to perform PEA based on rare variant association analyses results.

3 MATERIALS AND METHODS

3.1 Study Cohort

WES data of 929 obesity patients from 782 families which was collected within the framework of the project titled "The Genetics of Obesity in Turkish Families: Identification of Casual Gene Mutations in. Obesity." started in 2013 with the cooperation of Bilkent University and Rockefeller University was used as the target cohort. The probands were collected from Hacettepe University Hospital and Keçiören Training and Research Hospital. This study was approved by the institution of ethics committees of Bilkent University (No: 2017_12_11_01), Hacettepe University (No: GO 13/88-18) and Rockefeller University (No: JFN-0791). All patients and family members participating in this project were informed about the study, and their written consent was obtained prior to their involvement.

Probands with childhood obesity (age 2-18 years) consist of 58 individuals with a BMI greater than or equal to the 95th percentile. Probands with adult obesity consist of 871individuals (age over 18) with a BMI greater than or equal to 30. The majority of the obesity cohort comprises probands with severe obesity (BMI greater than or equal to 40 for adults, BMI ≥140% of the 95th percentile for children), accounting for approximately 86% of all probands (n=801). This emphasis on severe obesity helps minimize potential environmental factors contributing to the disorder.

WES data of 2924 individuals in total, 76 people without obesity (BMI less than 30 for adults, less than 85th percentile for children) which are the family members of obesity cohort and 2848 people with neurological disorders which was obtained through the same collaboration, were used as the control group. The study was conducted as a case-control study rather than including BMI as a continuous trait due to missing BMI information for the 2848 control individuals from the neurological disorders cohort. Variant allele frequencies of control group (2848 people with neurological disorders) was obtained from Turkish Genome Panel published by Kars et al. in 2021 (140).

Table 1. Obesity Cohort Age, Gender and BMI Information

Probands	Average Age	Average BMI	Gender	
			Female	Male
Children	14.71 (7-18)	34.90 (95-99th percentile)	28	30
Adults	40.42 (19-87)	48.60 (30.0-95.9)	649	222
Total			677	252

3.2 Data Pre-processing and Quality Control

WES data of cases and controls were aligned to homo sapiens reference genome assembly GRCh38/hg38. The alignment was performed utilizing Burrows-Wheeler Aligner-Maximal Exact Match (BWA-MEM) toolkit. Genome Analysis Toolkit (GATK) was used for duplicate read removal, insertion-deletion realignment, base quality score recalibration (BQSR), variant calling and joint genotyping, and variant quality score recalibration (VQSR). VQSR and variant quality filtering were performed according to GATK Best Practices recommendations (141).

Variant level and individual level quality statistics from variant calling format (VCF) files were obtained using vcftools (Version 0.1.13) (142). Variant level quality metrics include quality, depth, missingness; individual level quality metrics include mean depth, mean missingness and heterozygosity. Histogram graphs for each quality parameter were obtained using ggplot2 R package (Version 3.4.4) (143). Variant filtering criteria were decided as 20 for mapping quality (MQ), and 8 for depth, 20 for genotype quality (GQ). For the other parameters, GATK Best Practices Guidelines were followed (141).

Principle component analysis (PCA) was performed to analyze ethnicity-based population structure. Identifying individuals with divergent ancestry can be accomplished by integrating the genotypes of the study population with genotypes from a reference dataset containing individuals of known ethnicities. For this study, cohort data of 1000 Genomes Project Phase 3 was used as reference dataset. Step by step workflow provided by plinkQC R package (version 0.3.4) was followed. Pre-

processing of study cohort and reference cohort datasets upon merging and calculating principle components were performed utilizing PLINK (version 1.9) (144). Pre-processing steps include filtering A-T and G-C SNPs out, pruning, checking possible mismatches and allele flips and removing, and finally marching the datasets. Then, principal components were calculated by PLINK. PCA plot for principal components 1 and 2 compared individuals was obtained utilizing plinkQC R package.

Variant annotation of VCF files was performed with Ensembl Variant Effect Predictor (VEP) (version 111) (145) and ANNOVAR (146). dbNSFP database (version 4.7) was used to obtain computational pathogenicity scores and population allele frequencies from various databases such as gnomAD (version 4.0), EXAC, ALFA, 1000 Genomes (147).

Despite multiple quality control steps during data preprocessing, insertion and deletion variants with multiallelic sites that pass the quality filters can still effect the results. These variants are particularly common in low-complexity genomic regions, such as repetitive and homopolymer sequences. Managing these variants is challenging in both data processing and annotation. Excluding all multiallelic frameshift and non-frameshift indels can lead to the loss of true positive variants. Accurate annotation of these variants is also difficult due to discrepancies in variant calls between databases. Therefore, multiallelic variants were not excluded in the pre-analysis step and were manually evaluated after obtaining SMMAT results.

3.3 Variant Filtering

Variant filtering was performed in several different ways to perform rare variant association analyses. SNPSift (Version 5.2c) (148) and vembrane (Version 1.0.4) (149) tools were utilized to filter VCF files. To optimize MAF threshold, synonymous variants were first filtered to by MAF threshold of 10^{-2} in 1000 Genomes dataset and in-house MAF. Filtered synonymous variants file then filtered with MAF thresholds of, $5*10^{-4}$, 10^{-4}, $5*10^{-5}$, 10^{-5}, 10^{-6} and $5*10^{-6}$ in gnomAD exomes (Version 4.0), gnomAD genomes (Version 4.0), ExAC and ALFA databases. SMMAT-B, SMMAT-

S, SMMAT-O, and SMMAT-E tests were performed utilizing GMMAT R Package (Version 1.4.2) (150). QQ plots of p- values were obtained using the gaston R package (Version 1.6.0) (151).

3.3.1 Filtering models

Following MAF threshold optimization using synonymous variants test results, a MAF of 10^{-5} was chosen. Seven different filtering criteria were established, each utilizing various parameters and thresholds (Figure 2).

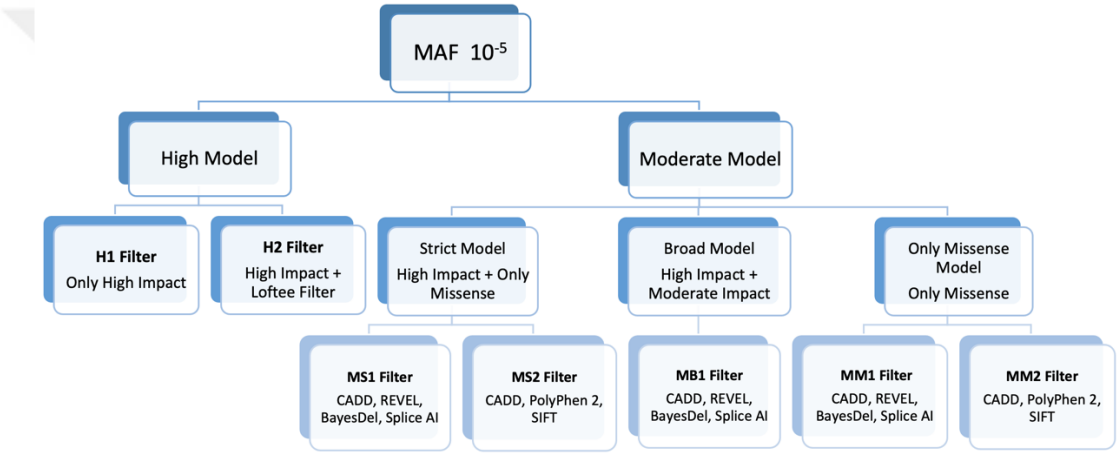

Figure 2. Filtering Models

Variant filtering is one of the most crucial steps in rare variant association studies. The final model will be built on filtered variants in VCF, hence filtering may can go two directions: losing true disease-causing variants or including majority of benign variants that would lead to no genetic burden in a gene that would distinguish disease and control group. For this reason, variants that have low potential of being disease causing, such as synonymous variants that are not located in the last five base pairs in exons, which may potentially impact splicing, intronic variants outside of canonical splicing region (+-2) that again have low probability of being deleterious, untranslated region (UTR) variants were not included in the filters. Two group of variant impacts which are high and moderate, were obtained from Ensembl VEP annotation.

High impact variants have the highest probability of causing the disease. These variants expect to cause loss of protein function by either causing the production of truncated protein, which means a protein shorter than its original sequence, and loss of important domains or triggering nonsense-mediated decay (NMD) which is a mechanism that leads to decaying of mRNA due to existence of premature termination codon on mature mRNA. In case of NMD, the gene is unable to produce its final product, which is usually a protein, and completely loses its function. Consequently, high impact variants are the most likely to be deleterious. Missense variants, on the other hand, do not result in protein truncation. However, depending on the specific amino acid change and its location, these variants can have significant deleterious effects. They can disrupt the function and composition of proteins, potentially leading to altered protein activity. Similarly, small in-frame insertions and deletions typically do not trigger nonsense-mediated decay (NMD), but like missense variants, they can impact protein function based on the affected amino acids' location and significance.

In high model, only high impact effect variants, which are transcript ablation, splice acceptor, splice donor, stop gain, frameshift, stop lost, start lost, transcript amplification, feature elongation, feature truncation were filtered. In the H1 (High-1) filter, these variants are analyzed. In H2 (High-2) filter, high impact variants were filtered and LOFTEE filter was applied.

In the MS1, MS2, MB1, MM1 and MM2 filters of moderate model, two different combinations of six in-silico variant pathogenicity prediction tools with two different threshold setups were applied. In-silico prediction tools utilized include CADD (Version1.7), REVEL, BayesDel (Version 1.0), Splice AI, Polyphen 2, and SIFT. Tools and their associated thresholds in this study were primarily selected based on Clinical Genome Resource (ClinGen) recommendations, although threshold adjustments were made to avoid overfiltering given that ClinGen's recommendations were initially intended for diagnostic variant evaluation (152,153).

In moderate strict (MS) models, high impact variants and missense variants were selected. For MS1(Moderate Strict-1) filter, these variants were filtered with

pathogenicity prediction threshold scores of CADD > 25, REVEL > 0.5, BayesDel > 0.0 and Splice AI > 0.2. For MS2 (Moderate Strict-2) filter, pathogenicity prediction threshold scores of CADD > 25, PolyPhen 2 (HVAR) rank score > 0.8, SIFT converted rankscore > 0.8 were utilized.

In moderate broad (MB) model, high impact variants and moderate impact variants, which are missense, in-frame insertion and in-frame deletion were filtered. For MB1 (Moderate Broad-1) filter, these variants were filtered with pathogenicity prediction threshold scores of CADD > 20, REVEL > 0.2, BayesDel > -0.25 and Splice AI > 0.1.

In only missense (MM) models, only missense variants were filtered. For MM1 (Moderate Missense-1) filter, these variants were filtered with pathogenicity prediction threshold scores of CADD > 25, REVEL > 0.5, BayesDel > 0.0 and Splice AI > 0.2. For MM2 (Moderate Missense-2) filter, pathogenicity prediction threshold scores of CADD > 25, PolyPhen 2 (HVAR) rank score > 0.8, SIFT converted rankscore > 0.8 were utilized.

3.3.2 Pathogenicity prediction tools

LOFTEE (Loss-Of-Function Transcript Effect Estimator) is a statistical approach developed to predict variants likely to cause loss of function (LoF) in proteins. It filters out variants that may escape nonsense-mediated decay, terminal truncation variants, and rescued splice variants. LOFTEE differentiates high-certainty loss-of-function (LoF) genetic variations from annotation errors and detects potential splice variants that occur outside the critical splice location.

CADD (Combined Annotation-Dependent Depletion) is a scoring system that ranks both single nucleotide variants (SNVs) and insertion-deletions (INDELs) by considering multiple genetic variation annotations. These annotations include surrounding sequence context, gene model, evolutionary constraint, functional predictions, and epigenetic characteristics. CADD employs a machine learning model

to integrate these annotations into a unified score. The obtained scores are converted into PHRED-like rank scores using the overall distribution of scores for around 9 billion possible SNVs, improving their compatibility and interpretability (154).

REVEL (Rare Exome Variant Ensemble Learner), provides a combined and adjusted pathogenicity score for missense variants based on individual tools. REVEL was trained using pathogenic and benign missense variations (MAF <1%) that are known, but ignoring the ones that were previously used to train its constituent tools. Independent testing demonstrated that REVEL exhibits superior overall performance compared to individual tools. REVEL is widely used in genetic diagnosis for assessing missense variant pathogenicity due to its demonstrated high accuracy and reliability (155).

BayesDel is a pathogenicity prediction meta score similar to REVEL. It provides a deleteriousness score for both coding and non-coding variants, including SNVs and INDELs. It uses naive Bayesian approach that handles missing values naturally, eliminating the need for imputation which can introduce bias and limit its utility. This method supports diverse data formats including extended pedigrees, case–controls, and accommodates features like incomplete penetrance and variable expressivity. BayesDel also provides an option (addAF) to incorporate allele frequency as a covariate to adjust the final score (156). In this study, BayesDel scores calculated with MAF were utilized.

Splice AI is a deep learning network designed for variants located in splice sites and those that may impact splicing. It provides a prediction for both SNVs and INDELs that may affect splicing. By analyzing the primary nucleotide sequence, Splice AI predicts splicing patterns and identifies noncoding mutations that disrupt the normal exon-intron structure, leading to significant consequences on the resulting protein. To evaluate the tool's performance, RNA sequencing was utilized to compare the tool's predictions with experimental data. The results demonstrated the tool's high accuracy in predicting variant effects. Overall, Splice AI is highly recommended for assessing

potential splice-altering variants due to its robust performance and accurate predictions (157).

PolyPhen-2 (Polymorphism Phenotyping version 2) is a pathogenicity prediction model for amino acid substitutions resulting from SNVs. It considers structural and comparative evolutionary annotations and builds a score which estimates the probability of the missense mutation being damaging or benign. The prediction method of PolyPhen-2 utilizes machine-learning classification. Polyphen-2 provides 2 sets of scores based on different training sets: HVAR (Human Variation) and HDIV (Human Diversity). HVAR trained model was built to distinguish mutations with drastic effects for diagnostic purposes, while HDIV model was built to evaluate rare variants that are deleterious but have milder effects (158). In this study, HVAR scores were used.

SIFT (Sorting Intolerant From Tolerant) is a prediction algorithm for amino acid changes as a consequence of nonsynonymous SNVs. This tool operates on the assumption that critical positions within a protein sequence have been conserved through evolution, implying that substitutions these sites could potentially affect the function of the protein. By utilizing sequence homology, SIFT is able to forecast the consequences of several potential changes (159).

3.4 Rare Variant Association Analysis

Rare variant association analysis was conducted using VCF files from eight different filtering models with SMMAT-B, SMMAT-S, SMMAT-O, and SMMAT-E tests, employing the GMMAT R package (Version 1.4.2) (150). QQ plots of p-values were generated using the gaston R package (Version 1.6.0) (151).

Results of all 4 SMMAT tests in 7 filters were evaluated. Resulting genes were ranked based on p-values. Genes with p-values less than 10^{-3} were filtered. Each gene and their respective variants were evaluated in terms of variant quality and familial segregation.

3.5 Pathway Enrichment Analysis

Pathway enrichment analysis with candidate genes obtained from SMMAT results was performed with Enrichr (103). Pathway information was obtained from Gene Ontology Biological Process and Molecular Function (2023) (160,161) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Human (2021) (162) databases. Candidate genes were also evaluated based on Mouse Genome Informatics (MGI) Mammalian Phenotype (Level 4) (163). Results were ranked based on p-values. Pathways and phenotypes with both p-values and adjusted p-values less than 0.05 were evaluated.

Enrichr is an online pathway analysis tool that provides access to over 100 pathway, expression profile, disease, and ontology databases. Enrichr performs PEA from non-ranked gene lists and provides three different scores. The first score is the p-value, calculated using Fisher's exact test under the assumption of a binomial distribution and independence within the gene list. The second score is the z-score, an adjusted p-value that measures the deviation from the expected rank. The third score, combined score (c-score), multiplies the logarithm of the p-value by the z-score. Enrichr features a user-friendly web interface that is easy to use and allows for result visualization with bar graphs and clustergrams. It is widely utilized for PEA in RNA sequencing, GWASs, and independent gene sets from various omics studies (103).

The Gene Ontology (GO) database is a comprehensive resource that provides detailed functional information of genes across all organisms. Widely used in biological experiments and research, GO is organized into three main categories. The GO Molecular Function describes the activities performed by a gene product at the molecular level. The GO Cellular Component category details the locations within cellular structures where these molecular functions occur. Finally, the GO Biological Process ategory outlines the processes and pathways involving a gene and its products (160,161). KEGG is a manually curated, taxonomy-based database that offers comprehensive information on the functions and interactions of genes and molecules at both the cellular and organismal levels across all organisms. It includes databases

for genes and genomes, pathways and networks, and taxonomic information. KEGG Pathway contains manually created pathway maps for biological networks based on cellular and organism level functional characterizations of genes and their products (162). The Mammalian Phenotype Ontology (MPO) is an online tool that classifies and organizes phenotype information associated with mouse and other mammalian species. MPO has been integrated with Mouse Genome Informatics (MGI) Database, Rat Genome Database (RGD), and the Online Mendelian Inheritance in Animals (OMIA). This enables researchers to perform comparative analyses of specific phenotypes across mammalian species which supports the design of relevant experiments and aids in the discovery of candidate disease genes and pathways (163).

3.6 Polygenic Risk Score Analysis

3.6.1 Genotype imputation

Prior to PRS calculation, genotype imputation was performed to accurately evaluate the genetic variants that are not directly genotyped. A two-step imputation process was employed using known haplotype blocks from the phased Turkish Genome Panel and the 1000 Genomes Project reference datasets.

Initially, m3sav files were generated for both the Turkish Genome Panel and the 1000 Genomes Project datasets using Minimac3 (Version 2.0.4) (164). These m3vcf files were converted to msav data files using Minimac4 (Version 4.1.6) (164). The study data was phased using Eagle (Version 2.4.1) (165). The phased study data was first imputed with the Turkish Genome Panel dataset (140) and then with the 1000 Genomes Project reference dataset to complete the imputation process.

3.6.2 PRS calculation

For PRS calculation family members of both case and control individuals were excluded to avoid overfitting of the data due to high linkage disequilibrium. PLINK tool (Version 1.9) (144) was utilized to perform PRS calculation of the imputed study

dataset. The reference dataset used in this study is from a GWAS study conducted in UK Biobank dataset consist of 7973 European ancestry cases with severe obesity and 142553 European ancestry controls (94). Classical p-value thresholding and clumping method was used to calculate individual risk scores. Different p-value thresholds were assessed and p-value of 0.5 was chosen for analysis. The analysis was repeated with top genes in our study.

4 RESULTS

4.1 Quality Statistics

Variant level of graphs for quality, depth, missingness (Figure 3) and individual level of graphs for missingness, heterozygosity, depth (Figure 4) were obtained.

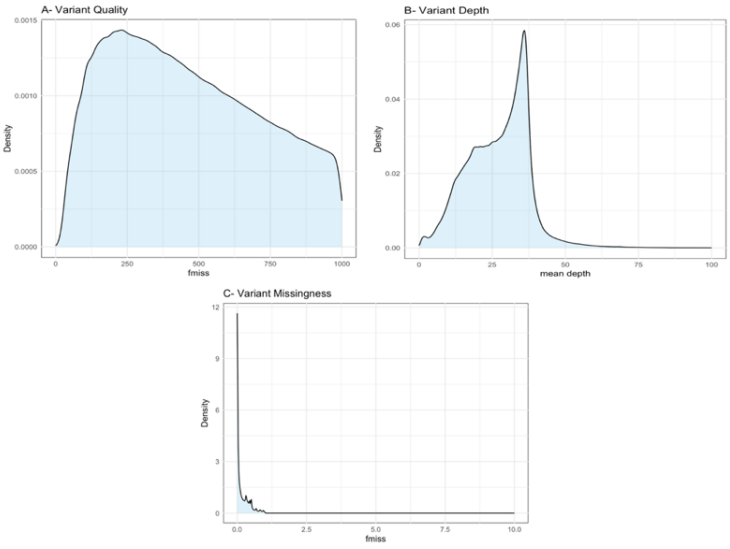

Figure 3. Variant Level Quality Graphs of Cohort Data

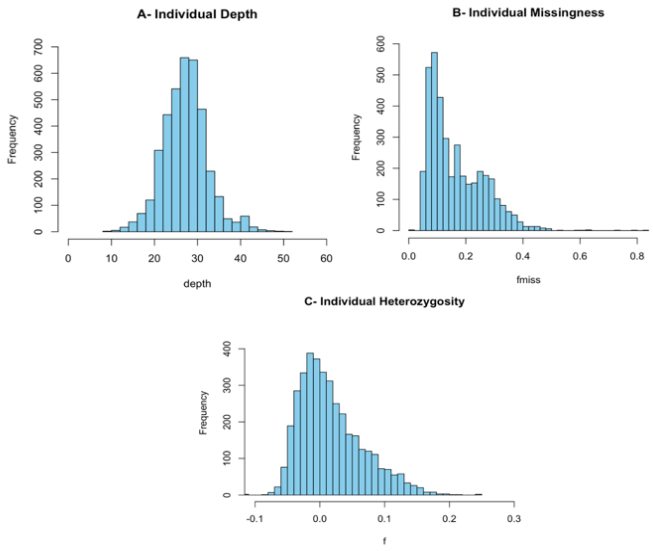

Figure 4. Individual Level Quality Graphs of Cohort Data

4.2 Principle Component Analysis

Principle Component Analysis (PCA) was conducted to evaluate population structure based on ethnicities and detect outliers (Figure 5).

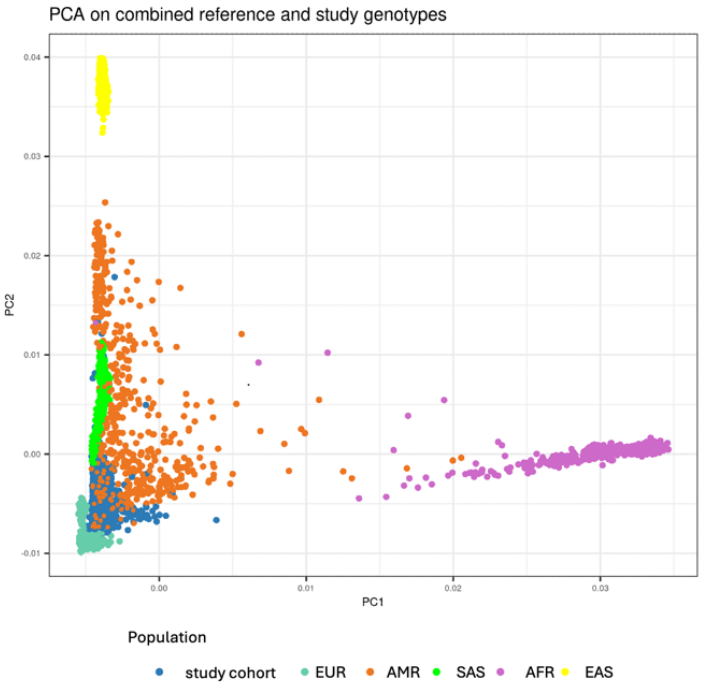

Figure 5. PCA Results of Study Cohort Merged With 1000 Genomes Data

4.3 MAF Threshold Optimization

To optimize the MAF threshold, synonymous variants filtered by six different MAF thresholds were analyzed with SMMAT-B, SMMAT-S, SMMAT-O, and SMMAT- E (Figure 6-11). MAF filtering was based on population frequencies in gnomAD exomes (Version 4.0), gnomAD genomes (Version 4.0), ExAC and ALFA databases.

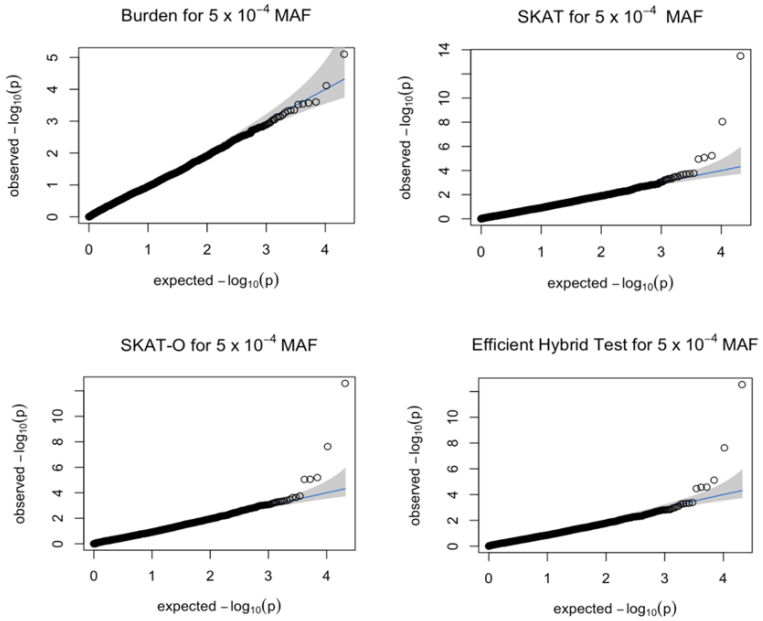

Figure 6. QQ Plots for p-values of SMMAT Models with Synonymous Variants Filtered with MAF Threshold of $5*10^{-4}$

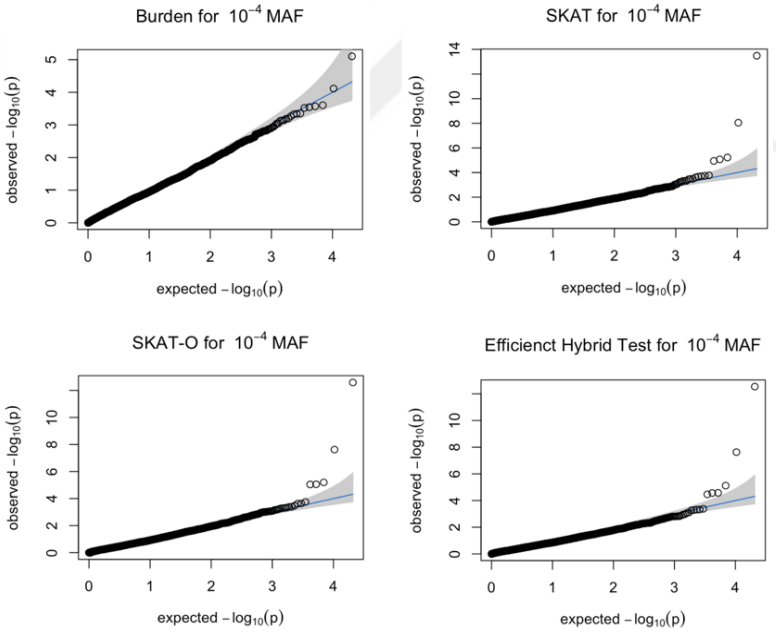

Figure 7. QQ Plots for p-values of SMMAT Models with Synonymous Variants Filtered with MAF Threshold of 10^{-4}

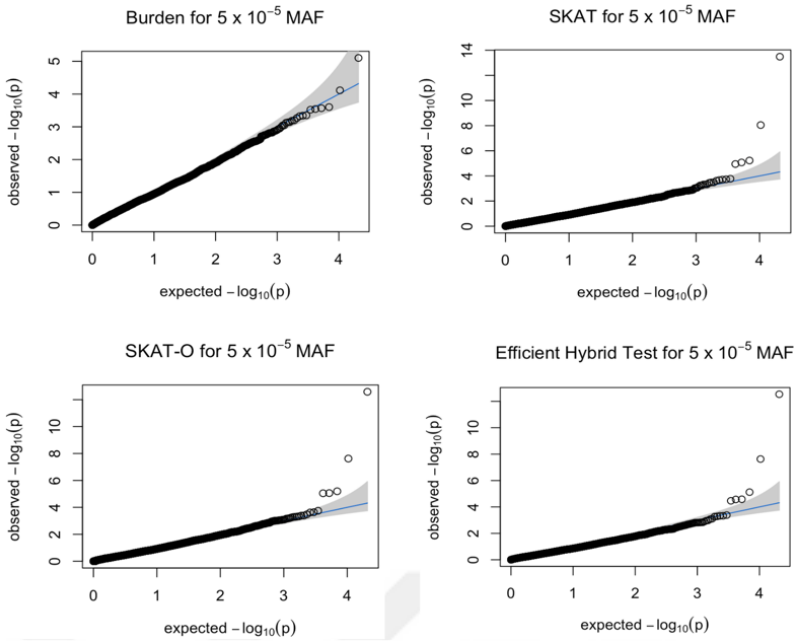

Figure 8. QQ Plots for p-values of SMMAT Models with Synonymous Variants Filtered with MAF Threshold of $5*10^{-5}$

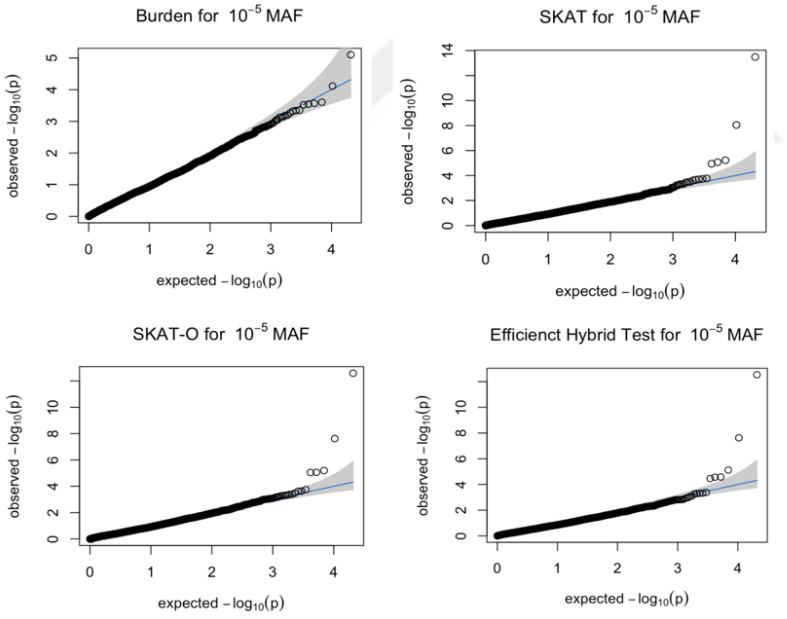

Figure 9. QQ Plots for p-values of SMMAT Models with Synonymous Variants Filtered with MAF Threshold of 10^{-5}

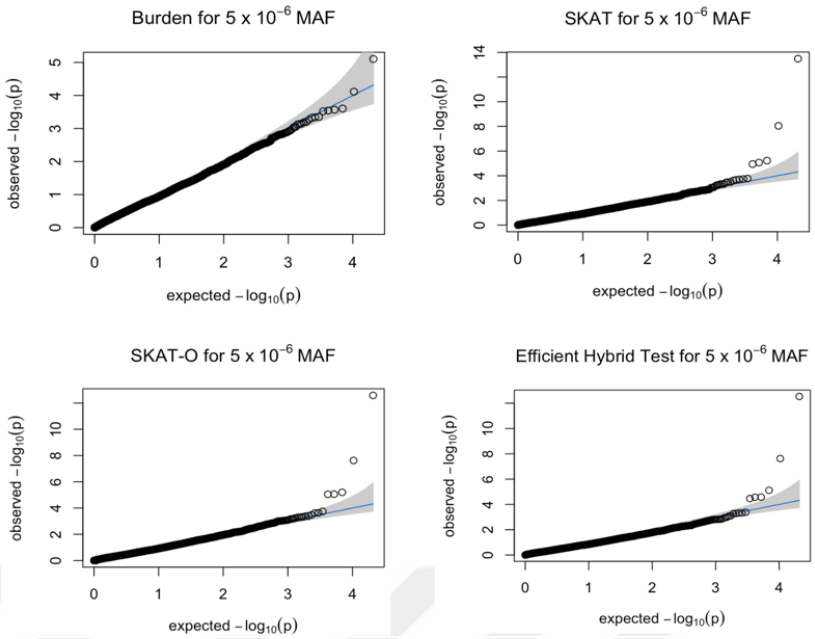

Figure 10. QQ Plots for p-values of SMMAT Models with Synonymous Variants Filtered with MAF Threshold of $5*10^{-6}$

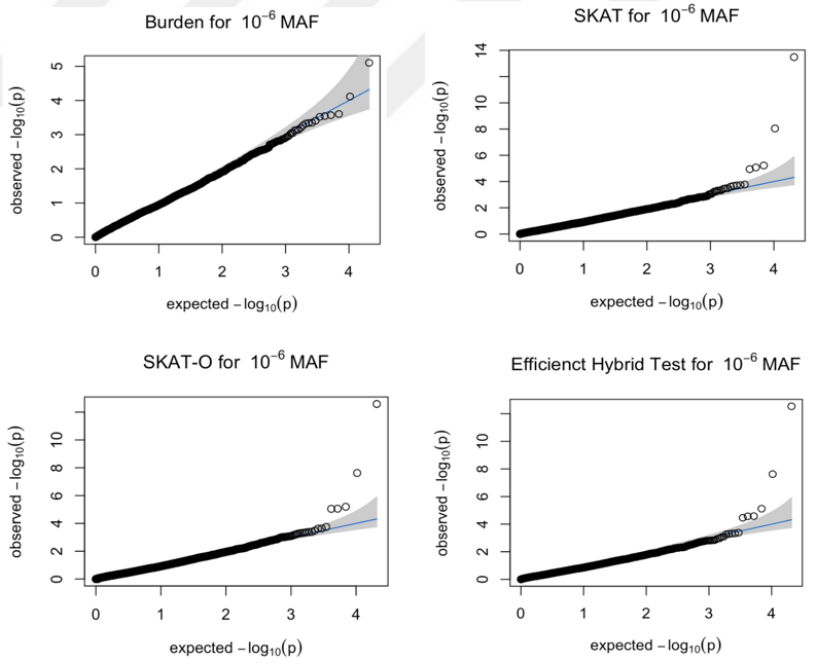

Figure 11. QQ Plots for p-values of SMMAT Models with Synonymous Variants Filtered with MAF Threshold of 10^{-6}

4.4 SMMAT Results

4.4.1 H1 filter

Variants were filtered according to H1 filter, which consist of high impact variants. All four SMMAT tests were applied. Figure 12 shows QQ plot of p-values for each test.

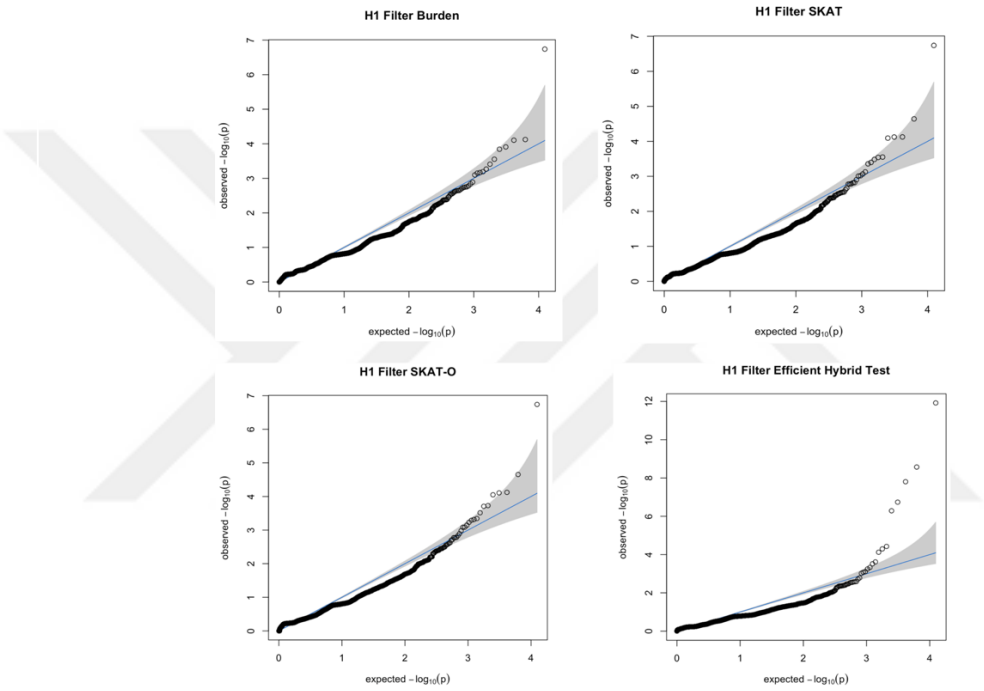

Figure 12. QQ Plots for p-values of SMMAT Models with H1 Filter

Genes were ranked based on their p-values from each statistical tests and genes with p values less than 10^{-3} were filtered. Burden test resulted in 12 genes, SKAT resulted in 14 genes, SKAT-O resulted in 15 genes and finally EHT resulted in 15 genes. In total number of 24 genes with p values less than 10^{-3} was prioritized in H1 filter. Of these genes, 17 were found significant and prioritized in multiple tests. Remaining 7 genes had significant p values only in one statistical approach. Variants in each gene was manually evaluated in terms of segregation in family pedigrees, variant annotations and quality. During this process, 11 genes were excluded and final number of 6 genes were selected for further analysis.

4.4.2 H2 filter

Variants were filtered according to H2 filter including high impact variants filtered according to LOFTEE high calls. All four SMMAT tests were applied. Figure 13 shows QQ plot of p-values for each test.

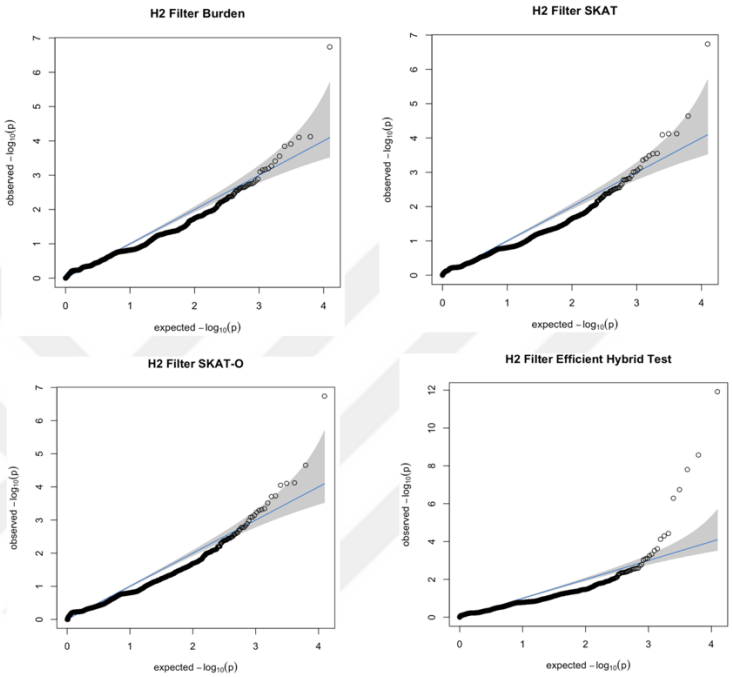

Figure 13. QQ Plots for p-values of SMMAT Models with H2 Filter

Genes were ranked based on their p-values from each statistical tests and genes with p values less than 10^{-3} were filtered. Burden test resulted in 11 genes, SKAT resulted in 15 genes, SKAT-O resulted in 15 genes and finally EHT resulted in 15 genes. In total number of 24 genes with p values less than 10^{-3} was prioritized in H2 filter. Of these genes, 18 were found significant and prioritized in multiple tests. Remaining 6 genes had significant p values only in one statistical approach. Variants in each gene was manually evaluated in terms of segregation in family pedigrees, variant annotations and quality. During this process, 11 genes were excluded and final number of 6 genes were selected for further analysis.

4.4.3 MS1 filter

In the MS1 filter, which includes high impact variants and missense variants, variants were filtered with in-silico prediction threshold scores of CADD > 25, REVEL > 0.5, BayesDel > 0.0 and Splice AI > 0.2. All four SMMAT tests were applied. Figure 14 shows QQ plot of p-values for each test.

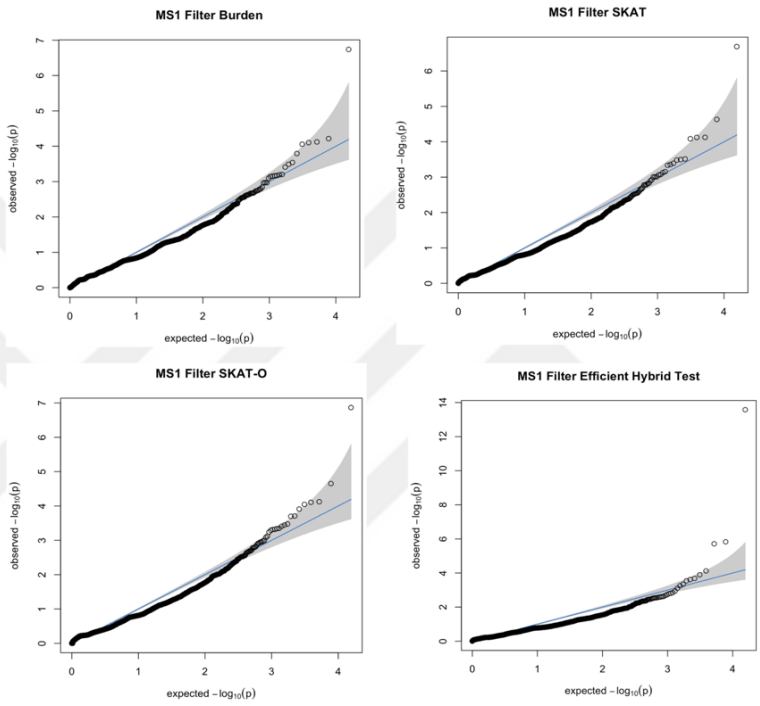

Figure 14. QQ Plots for p-values of SMMAT Models with MS1 Filter

Genes were ranked based on their p-values from each statistical tests and genes with p values less than 10^{-3} were filtered. Burden test resulted in 16 genes, SKAT resulted in 18 genes, SKAT-O resulted in 19 genes and finally EHT resulted in 11 genes. In total number of 28 genes with p values less than 10^{-3} was prioritized in MS1 filter. Of these genes, 19 were found significant and prioritized in multiple tests. Remaining 9 genes had significant p values only in one statistical approach. Variants in each gene was manually evaluated in terms of segregation in family pedigrees, variant annotations and quality. During this process, 17 genes were excluded and final number of 11 genes were selected for further analysis.

4.4.4 MS2 filter

In the MS2 filter, which includes high impact variants and missense variants, variants were filtered with in-silico prediction threshold scores of CADD > 25, PolyPhen 2 (HVAR) rank score > 0.8, SIFT converted rankscore > 0.8. All four SMMAT tests were applied. Figure 15 shows QQ plot of p-values for each test.

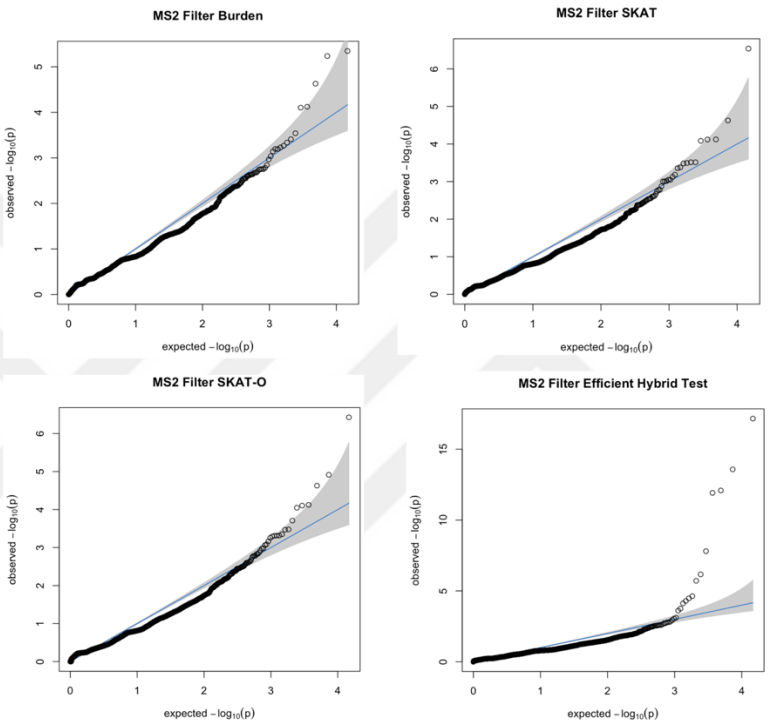

Figure 15. QQ Plots for p-values of SMMAT Models with MS2 Filter

Genes were ranked based on their p-values from each statistical tests and genes with p values less than 10^{-3} were filtered. Burden test resulted in 14 genes, SKAT resulted in 18 genes, SKAT-O resulted in 18 genes and finally EHT resulted in 15 genes. In total number of 32 genes with p values less than 10^{-3} was prioritized in MS2 filter. Of these genes, 18 were found significant and prioritized in multiple tests. Remaining 14 genes had significant p values only in one statistical approach. Variants in each gene was manually evaluated in terms of segregation in family pedigrees, variant annotations and quality. During this process, 21 genes were excluded and final number of 11 genes were selected for further analysis.

4.4.5 MM1 filter

In the MM1 filter, which includes only missense variants, variants were filtered with in-silico prediction threshold scores of CADD > 25, REVEL > 0.5, BayesDel > 0.0 and Splice AI > 0.2. All four SMMAT tests were applied. Figure 16 shows QQ plot of p-values for each test.

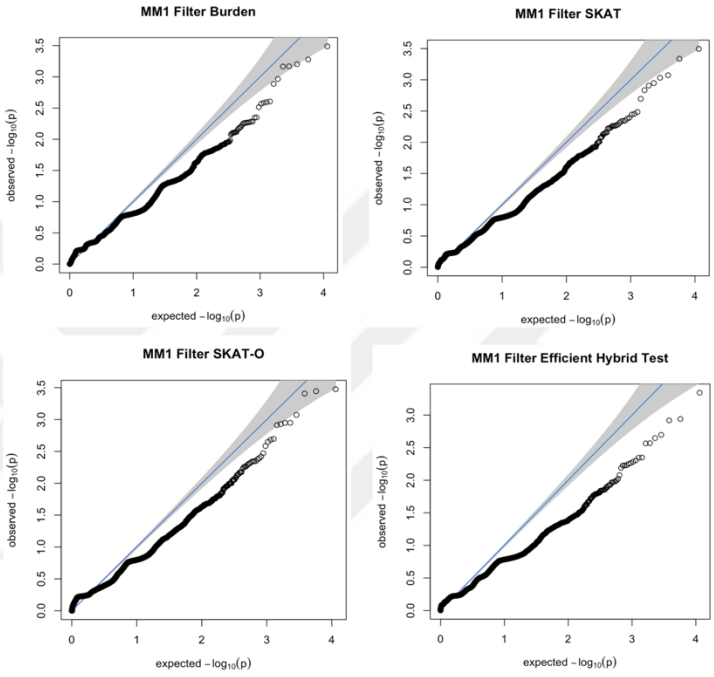

Figure 16. QQ Plots for p-values of SMMAT Models with MM1 Filter

Genes were ranked based on their p-values from each statistical tests and genes with p values less than 10^{-3} were filtered. Burden test resulted in 5 genes, SKAT resulted in 4 genes, SKAT-O resulted in 4 genes and finally EHT resulted in only one gene. In total number of 7 genes with p values less than 10^{-3} was prioritized in MM1 filter. Of these genes, 4 were found significant and prioritized in multiple tests. Remaining 3 genes had significant p values only in one statistical approach. Variants in each gene was manually evaluated in terms of segregation in family pedigrees, variant annotations and quality. During this process, one gene was excluded and final number of 6 genes were selected for further analysis.

4.4.6 MM2 filter

In the MM2 filter, which includes only missense variants, variants were filtered with in-silico prediction threshold scores of CADD > 25, PolyPhen 2 (HVAR) rank score > 0.8, SIFT converted rankscore > 0.8. All four SMMAT tests were applied. Figure 17 shows QQ plot of p-values for each test.

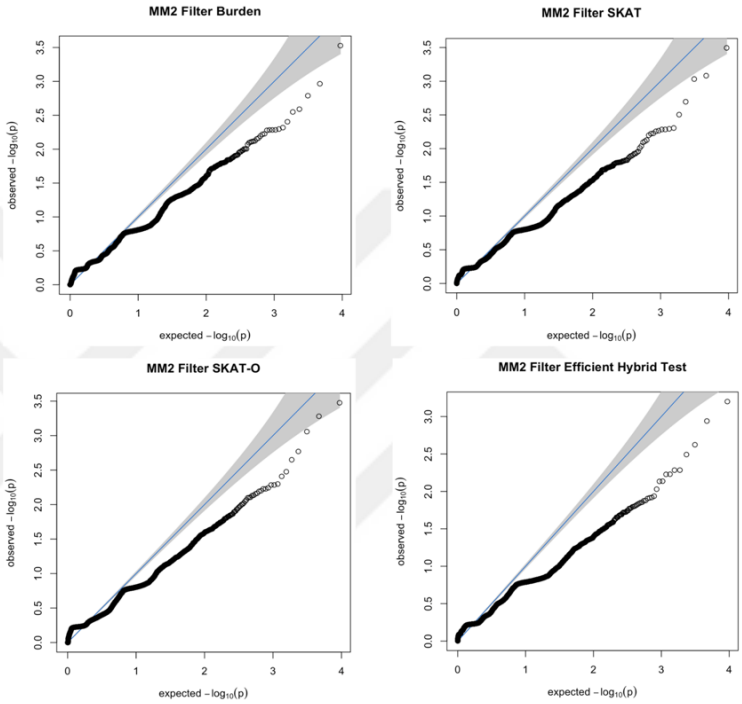

Figure 17. QQ Plots for p-values of SMMAT Models with MM2 Filter

Genes were ranked based on their p-values from each statistical tests and genes with p values less than 10^{-3} were filtered. Burden test resulted in only one gene, SKAT resulted in 3 genes, SKAT-O resulted in 3 genes and finally EHT resulted in only one gene. In total number of 4 genes with p values less than 10^{-3} was prioritized in MM2 filter. Of these genes, 3 were found significant and prioritized in multiple tests. Only one gene had significant p values in a single statistical approach. Variants in each gene was manually evaluated in terms of segregation in family pedigrees, variant annotations and quality. During this process, 3 genes were excluded and one gene was selected for further analysis.

4.4.7 MB1 filter

In the MB1 filter, which includes high and moderate impact variants, variants were filtered with in-silico prediction threshold scores of CADD > 20, REVEL > 0.2, BayesDel > -0.25 and Splice AI > 0.1. All four SMMAT tests were applied. Figure 18 shows QQ plot of p-values for each test.

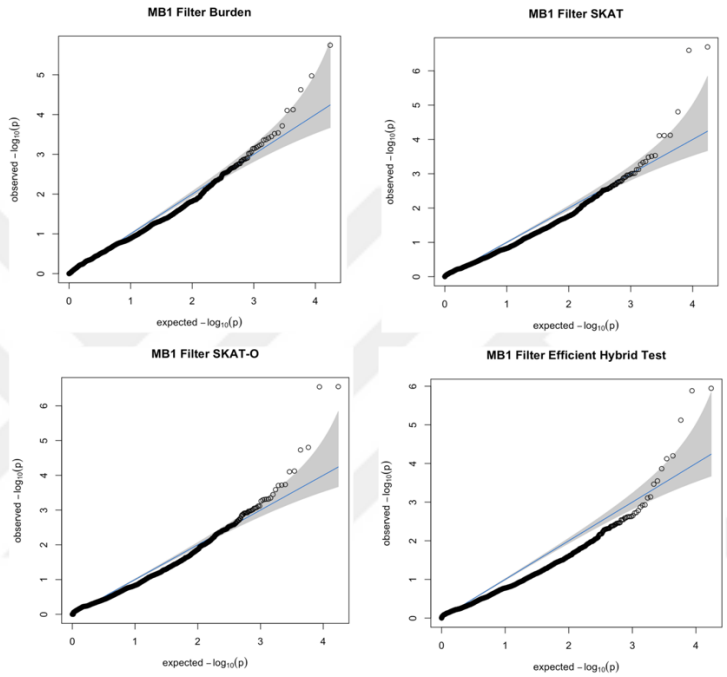

Figure 18. QQ Plots for p-values of SMMAT Models with MB1 Filter

Genes were ranked based on their p-values from each statistical tests and genes with p values less than 10^{-3} were filtered. Burden test resulted in 21 genes, SKAT resulted in 17 genes, SKAT-O resulted in 23 genes and finally EHT resulted in 10 genes. In total number of 31 genes with p values less than 10^{-3} was prioritized in MB1 filter. Of these genes, 23 were found significant and prioritized in multiple tests. Remaining 8 genes had significant p values in a single statistical approach. Variants in each gene was manually evaluated in terms of segregation in family pedigrees, variant annotations and quality. During this process, 21 genes were excluded and 10 genes were selected for further analysis.

4.4.8 Candidate gene list

Results of all 4 SMMAT tests in 7 filters were evaluated. Resulting genes were ranked based on p-values. Genes with p-values less than 10^{-3} were filtered and total 53 genes were obtained. Each gene and their respective variants were evaluated in terms of variant quality and familial segregation. Genes with inaccurate population frequency annotation due to multi-allelic sites, genes that fall to low-mappability genomic regions which includes pseudogenes, highly similar orthologs, repeat regions which accounts to 24 genes in total were excluded from the analysis after manual evaluation. Remaining 29 genes were selected for pathway enrichment analysis.

4.5 Pathway Enrichment Analysis Results

Pathway enrichment analysis of 29 genes was performed with Enrichr with pathway and phenotype information from GO Biological Process, GO Molecular Function (2023), KEGG Human (2021), and MGI Mammalian Phenotype Level 4 (2021).

4.5.1 PEA with GO database

Bar graph of top five enrichment results with significant p-values (<0.05) for GO Biological Process (2023) was shown in Figure 19. Table 2 shows p-value, z-score and c-score for the pathways. Three pathways including Protein K27-linked Ubiquitination (GO:0044314), Protein K6-linked Ubiquitination (GO:0085020) and Regulation Of Insulin Secretion (GO:0050796) also have a significant adjusted p-value, which is z-score.

Top five enrichment results with significant p-values (<0.05) for GO Molecular Function (2023) was shown in Figure 20. Table 3 shows p-value, z-score and c-score for the pathways. One pathway which is Serotonin Receptor Activity (GO:0099589), also has a significant adjusted p-value, which is z-score.

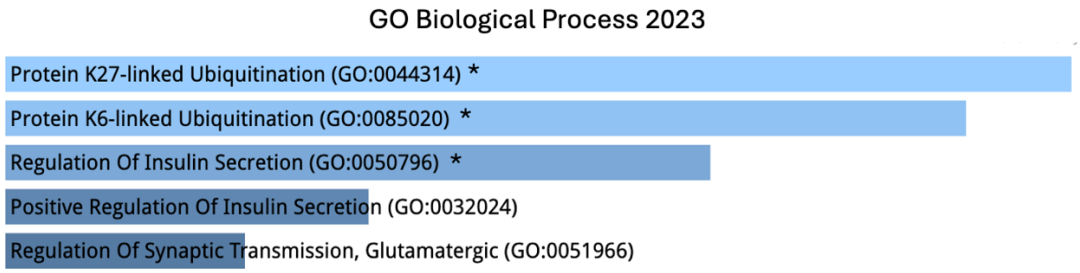

Figure 19. Bar Graph of Top Five PEA Results for GO Biological Process Terms are ranked based on p-values. An asterisk (*) next to the term indicates it also has a significant z-score (<0.05).

Table 2. Enrichr Statistics for Top 5 Terms from GO-Biological Process

Term	Overlap	P-value	Z-score	C-score
Protein K27-linked Ubiquitination (GO:0044314)	2/7	0.00004244	0.01314	2977.88
Protein K6-linked Ubiquitination (GO:0085020)	2/9	0.00007262	0.01314	2013.35
Regulation Of Insulin Secretion (GO:0050796)	3/87	0.0002677	0.03230	224.70
Positive Regulation Of Insulin Secretion (GO:0032024)	2/40	0.001530	0.08334	251.88
Regulation Of Synaptic Transmission, Glutamatergic (GO:0051966)	2/55	0.002874	0.08334	162.90

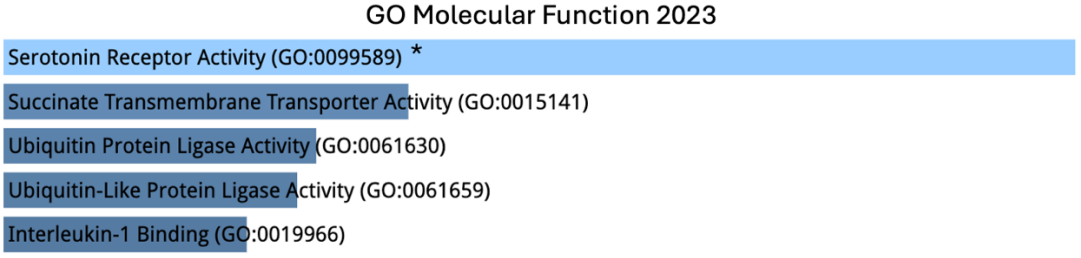

Figure 20. Bar Graph of Top Five PEA Results for GO Molecular Function Terms are ranked based on p-values. An asterisk (*) next to the term indicates it also has a significant z-score (<0.05).

Table 3. Enrichr Statistics for Top 5 Terms from GO-Molecular Function

Term	Overlap	P-value	Z-score	C-score
Serotonin Receptor Activity (GO:0099589)	2/26	0.0006457	0.04649	452.21
Succinate Transmembrane Transporter Activity (GO:0015141)	1/5	0.007230	0.1195	878.83
Ubiquitin Protein Ligase Activity (GO:0061630)	3/311	0.01009	0.1195	33.86
Ubiquitin-Like Protein Ligase Activity (GO:0061659)	3/319	0.01081	0.1195	32.49
Interleukin-1 Binding (GO:0019966)	1/9	0.01298	0.1195	387.19

4.5.2 PEA with KEGG database

Top five enrichment results with significant p-values (<0.05) for KEGG Human (2021) database was shown in Figure 21. Table 4 shows p-value, z-score and c-score for the pathways. One pathway, which is Taste transduction also has a significant adjusted p-value, which is z-score.

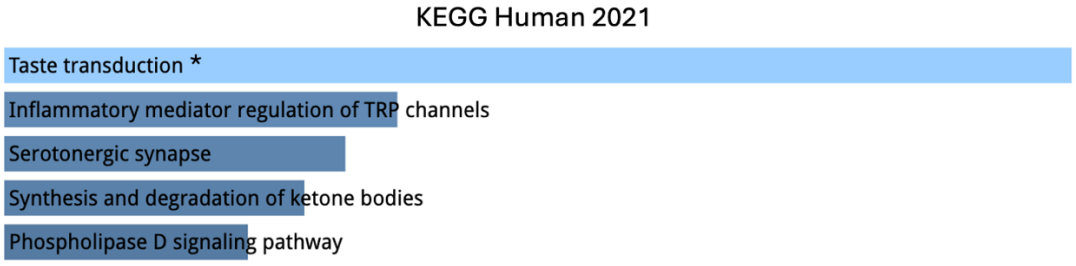

Figure 21. Bar Graph of Top Five PEA Results for KEGG Human Terms are ranked based on p-values. An asterisk (*) next to the term indicates it also has a significant z-score (<0.05).

Table 4. Enrichr Statistics for Top 5 Pathways from KEGG Human

Term	Overlap	P-value	Z-score	C-score
Taste transduction	3/86	0.0002587	0.02225	228.37
Inflammatory mediator regulation of TRP channels	2/98	0.008853	0.2600	72.49
Serotonergic synapse	2/113	0.01163	0.2600	59.03
Synthesis and degradation of ketone bodies	1/10	0.01441	0.2600	335.86
Phospholipase D signaling pathway	2/148	0.01938	0.2600	39.67

4.5.3 PEA with MGI mammalian phenotype

Top five enrichment results with significant p-values (<0.05) for MGI Mammalian Phenotype Level 4 (2021) database was shown in Figure 22. Table 5 shows p-value, z-score and c-score for the pathways. One term, which is increased body mass index (MP:0006087), also has a significant z-score.

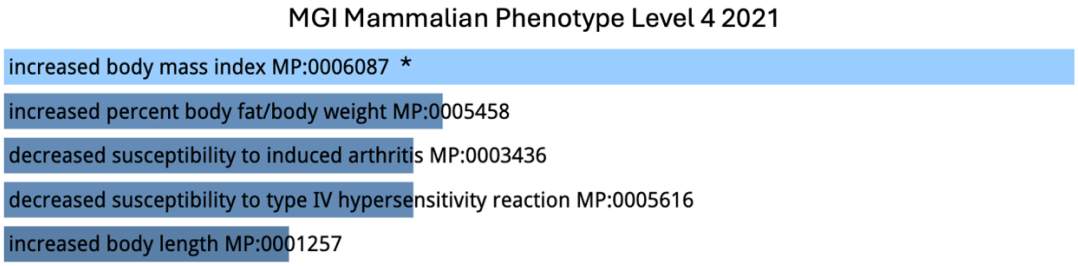

Figure 22. Bar Graph of Top Five PEA results for MGI Mammalian Phenotype Terms are ranked based on p-values. An asterisk (*) next to the term indicates it also has a significant z-score (<0.05).

Table 5. Enrichr Statistics for Top 5 Terms from MGI Mammalian Phenotype

Term	Overlap	P-value	Z-score	C-score
increased body mass index MP:0006087	2/14	0.0001827	0.04148	1060.46
increased percent body fat/body weight MP:0005458	2/49	0.002289	0.1048	190.91
decreased susceptibility to induced arthritis MP:0003436	2/52	0.002574	0.1048	175.97
decreased susceptibility to type IV hypersensitivity reaction MP:0005616	2/52	0.002574	0.1048	175.97
increased body length MP:0001257	2/67	0.004234	0.1048	123.96

4.6 Polygenic Risk Score Analysis Results

PRS analysis was performed with UK-Biobank severe obesity GWAS summary statistics and p-value threshold was chosen as 0.5. Figure 23 shows boxplot of PRS in obesity cohort and control cohort.

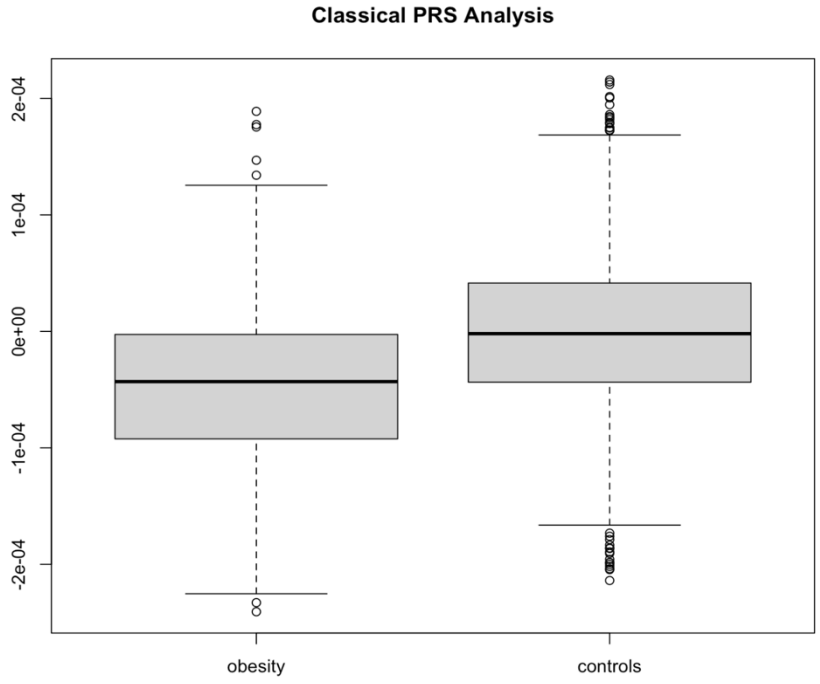

Figure 23. Comparative Boxplots for PRS of Obesity and Control Cohorts

Subsequently, PRS analysis was performed with the 29 candidate genes that are obtained from rare variant association testing SMMAT results. Variants in these genes and their p values and odd ratio scores were filtered in UK-Biobank severe obesity GWAS summary statistics dataset. The calculation of polygenic risk scores was carried out using the workflow outlined in the Materials and Methods section. However, the process did not yield significant results for either the obesity or the control group. This failure was attributed to the absence of significant values during the classical p-value thresholding and clumping steps. Consequently, the polygenic risk score calculations could not be successfully completed for these groups. Several factors could contribute to this outcome, and these are explored in detail in the Discussion section.

5 DISCUSSION

The variant and individual-level statistics of our cohort demonstrate the presence of high-quality data, making it well-suited for exome-wide studies based on common and rare variants. Our analysis reveals that the average sequencing depth per individual and per variant is high, with a significant proportion of variants exhibiting a high Phred quality score. These metrics indicate the effectiveness of our pre-processing quality filtering steps, ensuring sufficient variant coverage and minimizing the likelihood of false positives in variant detection. Furthermore, assessing missingness is crucial for NGS data as high rates can lead to false negatives and compromise downstream analyses and associations. Our dataset exhibits very low rates of missingness at both the individual and variant levels, which indicates the reliability of the analysis. Together these metrics indicate that our cohort possesses high quality data that is suitable for reliable downstream analysis.

Exome and genome wide association analyses primarily depend on in-house comparison of individuals, previous association studies, and population datasets. Given the comparative nature of these studies, population structure emerges as crucial bottleneck. Majority of the available data in large population sets such as gnomAD, 1000 Genomes, and UK Biobank predominantly represent individuals of European ancestr. A study dataset comprising individuals with diverse ethnicities could potentially lead to false associations and misinterpretation of results. PCA analysis showed that there is no significant genetic divergence in our cohort, which is an important indicator for suitability for association analyses. Furthermore, ethnicity-based comparison with 1000 Genomes data revealed that our cohort is relatively close to European ancestry, which is expected for Turkish population. Overall, our cohort dataset is found suitable for rare variant association analysis as well as PRS analysis based on UK-Biobank European ancestry individuals.

Accurate annotation of dataset is one of the most important factors that is often overlooked in large-scale studies. Our study showed that discrepancies between databases and differences in data formatting may cause inaccuracies in annotation

which would lead to missing information for variants and ineffective filtering. One significant example for this multiallelic variants, especially indels, which often have more than one forms such as a 3-base insertion, a 4-base deletion, and a 5-base insertion occurring in the same site. Presence of these variants adds a layer of complexity to genetic analyses due to complexity of annotation and interpretation. Databases may have discrepancies and differences for annotation of these variants. However, excluding all multiallelic frameshift and non-frameshift indels can lead to the loss of true positive variants. To avoid this, we did not exclude these variants and manually evaluated them after obtaining SMMAT results. However, presence of these variants may also lead to inaccuracies in test results. For instance, a variant with high MAF may still be present in dataset after filtering due to discrepancies in annotation source. In summary, developing advanced statistical methods and a fine-tuned approach is essential for further studies to effectively manage multiallelic indels.

Another important aspect of annotation is inclusion of splicing prediction tools. Ensembl vep adds a "splicing" tag for missense and synonymous variants within the last 5 bases of exons. Despite their benign missense prediction scores in common computational tools for missense variants and no scores for synonymous variants, splicing prediction tool Splice AI showed that these variants potentially affect splicing and may cause loss of function. Including Splice AI tool as a filtering criterion prevented excluding these potentially important variants based on missense predictions. Hence, variant filtering shouldn't solemnly depend on missense pathogenicity prediction tools.

Different variant filtering strategies in our rare variant association analyses yielded various results. Filtering only high impact variants enables to perform analyses with LoF variants that have higher potential of causing disease. Comparing the results of H1 and H2 filters demonstrated that LOFTEE filtering is crucial for excluding potential non-LoF variants, therefore enhancing the quality of filtered dataset by retaining only potential LoF variants. MM1 and MM2 filters included only missense variants, while MS1 and MS2 filters contained both LoF and missense variants. MB1 filter included LoF, missense, and in-frame indel variants. In MS1, MS2, MM1 and

MM2 filters, strict pathogenicity thresholds for computational tools were applied, whereas in MB1 filter lower thresholds were chosen. Applying higher thresholds results in fewer remaining variants, but those selected have a higher likelihood of being disruptive. However, this approach may also exclude potentially relevant variants that computational tools fail to accurately assess. As previously mentioned, pathogenicity prediction tools can only "predict" a variant's deleteriousness based on statistical models built on existing data. While each filter identified a few common top-prioritized genes, they also highlighted different genes. These results underscore the importance of utilizing various approaches with different tools and thresholds, as well as incorporating variants with varying impacts when building models. This comprehensive strategy enhances the identification of potentially significant variants and genes.

One of the limitations of this study was that it was conducted as a case-control study rather than incorporating BMI as a continuous trait due to missing BMI information for the control cohort. A significant portion of the obesity cohort, approximately 86%, consists of probands with severe obesity. This focus on severe obesity was intentional, aiming to reduce the influence of potential environmental factors that might contribute to the disorder, thereby establishing a stronger case-control group.

Including BMI as a continuous trait in future studies could potentially improve the results. By obtaining and incorporating BMI data for the control cohort, more nuanced analyses that consider the full spectrum of BMI values can be performed. This approach would allow for the detection of genetic variants associated with a range of BMI values, rather than just distinguishing between obese and non-obese individuals. It would also enable the examination of dose-response relationships, where the effect of genetic variants on BMI could be assessed across its entire range.

Generalized linear mixed model framework provided by SMMATs enabled us to address the relatedness within our dataset effectively. This framework allowed us to account for familial relationships, which is crucial in genetic studies to avoid

confounding effects. In our study, designed as a case-control investigation with obesity as a binary trait, the use of SMMATs proved particularly advantageous. Traditional methods often struggle to account for the complexities introduced by relatedness or population structure, which can lead to increase in error rates and spurious associations. The GLMM approach, however, incorporates random effects to model the correlation structure among related individuals, thereby providing a more robust analysis framework. Moreover, the flexibility of SMMATs framework is highly, allowing us to integrate various sources of information and to test for associations under different statistical models. Burden test assumes all variants in a gene have effects in same direction, while SKAT assumes effects may be in both directions. Combined tests leverage the advantages of both approaches. Obesity is a complex multifactorial disorder that involves many genes and biological processes, and we lack prior information on the specific effects of individual genes. Employing four different tests within the SMMAT framework allowed us to account for potential variations in in the effects of genes. This comprehensive approach not only improves the power of detecting true associations but also reduces the risk of false positives as well as false negatives. Combining the strengths of different statistical approaches enabled us to capture a wider range of genetic effects, whether they act uniformly or in opposing directions within a gene. Overall, the application of the GLMM framework through SMMATs has enhanced the robustness of our study. This methodological approach underscores the importance of using different statistical techniques to accurately address the complexities and confounding factors in genetic datasets, particularly in studies involving related individuals.

Integrating relatedness into the statistical models enhances the tests by distinguishing segregating and non-segregating variants, which leads to more robust results. Family data plays a critical role in all genome studies, whether it is diagnostic testing for a rare disorder or an association study that asses a disease or a phenotype, the first step after testing probands is evaluation of segregation of the variant in affected and unaffected family members. This segregation information is critical for confirming the biological relevance of candidate genes and variants. Complex traits such as obesity is no exception to this. Accounting for family relationships with both

affected and unaffected individuals within the same pedigree enables the development of more reliable statistical models, which are essential for identifying and validating candidate genes and variants with biological relevance.

Of 29 top ranked genes based on p-values in rare variant association analyses, 2 genes were prioritized in 5 filters, 11 genes were in top of 3 filters, 9 genes were in top of 2 filters and finally 7 genes were amongst top genes in only one filter. These results further indicate the importance of utilizing various approaches in variant filtering and employing different statistical tests. Nevertheless, it's important to recognize that p-values primarily indicate the statistical significance rather than biological relevance. Therefore, while these results are promising, further evaluation and experimental studies are needed to assess the biological significance of genes prioritized in the tests.

Pathway enrichment analysis based on top 29 genes from rare variant association analyses resulted in 48 statistically significant (p-value <0.05) pathways and terms. However, with the filtering by adjusted p-values, 4 pathways and 2 terms remained. As recommended, it is important to utilize adjusted p-values to avoid erroneous inferences due to multiple testing.

PEA with GO Biological Process prioritized three pathways, which are Protein K27-linked Ubiquitination, Protein K6-linked Ubiquitination, and Regulation Of Insulin Secretion. Protein ubiquitination pathways are common biological processes which many genes and their products contribute. There has been no previous studies that associate ubiquitination with obesity. Further studies are needed to assess the association of this pathway and weight gain.

The association between insulin hormone and obesity has been extensively demonstrated in previous research with insulin resistance being one of the key factors observed in individuals with obesity. Insulin plays a critical role not only in postprandial metabolism but also in nutritional feedback. It acts on hypothalamic neurons to enhance satiety and increases dopamine levels, consequently influencing reward signaling. Impaired insulin signaling in the brain affects the regulation of

hunger and satiety, leading to an increased caloric intake and further weight gain. This disruption in feedback mechanisms can create a cycle where obesity exacerbates insulin resistance, and insulin resistance, in turn, promotes further obesity. Additionally, the change in dopamine levels due to impaired insulin signaling potentially increases the signals for overeating, particularly of high-calorie, palatable foods. Recent studies indicate that insulin mediated nutritional-feedback may be impaired in obesity (110).

PEA with GO Molecular Function prioritized the function “Serotonin Receptor Activity”. As discussed previously, serotonin is a neurotransmitter that plays a crucial role in signaling food-intake control to the brain's homeostatic and reward regulatory systems, which are the main pathways regulating food intake behavior. Elevated serotonin signaling has been linked to reduced food intake, whereas decreased serotonin signaling is associated with excessive eating and weight gain. Consequently, impairments in genes and their products involved in serotonin pathways have the potential to be associated with obesity (119).

PEA with KEGG Human database resulted in enrichment of genes in taste transduction pathway. Taste receptors and transmission of taste signals have major contribution in controlling appetite, satiety and food intake. One of the brain's major pathways involved in obesity is the hedonistic or reward pathway, which is significantly influenced by pleasure signals triggered by food consumption. Therefore, sense and signaling of taste are suggested to play a crucial role in development of obesity. Multiple studies have associated increased BMI with changes in eating habits resulting from reduced taste sensitivity, particularly for sweet foods. This further suggest that impaired taste signaling may contribute to weight gain and ultimately obesity (166).

PEA with MGI Mammalian Phenotype data revealed resulted in enrichment of genes associated with increased body mass index. As previously mentioned, primary candidate genes and pathways for obesity were discovered initially in mice studies (35,36). Subsequent research in mice has provided valuable insights into the numerous

biological mechanisms and processes involved in body weight regulation and food intake. Therefore, these studies serve as crucial indicators of the biological significance and relevant associations of genes of interest with the trait.

Evaluation of PRS analysis revealed no significant difference in PRS between individuals with obesity and control group. This discrepancy is likely due to the fact that our target data is WES, whereas the base data used for PRS calculation consists of summary statistics from a genome-wide association study. Our target data consists of WES, which focuses on sequencing the coding regions of the genome. In contrast, the base data used for calculating PRS originates from a GWAS that includes summary statistics covering the entire genome. GWAS typically identifies variants associated with a trait, including those in non-coding regions such as introns and intergenic regions, which are not targeted by WES. Upon evaluation, it was observed that the top variants showing high statistical significance in the GWAS are predominantly located in deep intronic and intergenic regions, which are not covered in study dataset. This discrepancy in genomic coverage between the base data and our sequencing approach may explain the lack of significant differences observed in the PRS analysis.

7 CONCLUSION

Common and rare variant association analyses using various statistical models was applied to WES data of 929 obesity patients from 782 families and 2924 controls. Subsequently, pathway enrichment analysis for most significant results was performed. Our primary objective was to identify candidate genes and pathways that could provide insights into the molecular mechanisms underlying obesity.

Our study has shown that proper data pre-processing and annotation are key elements that would significantly affect the results. While building the models for rare variant association testing, a comprehensive approach that is empowered by different filtering strategies enhances the identification of potentially significant variants and genes. For complex traits such as obesity, utilizing various statistical approaches that leverage family data contributes to improving the accuracy and reliability of the results.

Pathway enrichment analysis of top candidate genes showed the enrichment in taste transduction, regulation of insulin secretion, and serotonin signaling pathways. Each of these pathways has been previously implicated in various mechanisms contributing to obesity. Additionally, PEA with Mouse Genome Informatics database resulted in enrichment of genes associated with increased BMI in mice. These findings suggest that our top candidate genes may play significant roles in weight gain and obesity. Further research is required to evaluate the specific roles of these genes and their associations with obesity.

Common variant analysis for polygenic risk score analysis indicated that using WES data may not be suitable for PRS calculation. This limitation arises because the majority of summary statistics used for PRS calculation are derived from GWAS studies, where the significant variants are often located in deep intronic and intergenic regions not covered by WES data.

8 REFERENCES

1. Lin X, Li H. Obesity: Epidemiology, Pathophysiology, and Therapeutics. Front Endocrinol. 2021 Sep 6;12.

2. Mahmoud R, Kimonis V, Butler MG. Genetics of Obesity in Humans: A Clinical Review. Int J Mol Sci. 2022 Oct 1;23(19).

3. Obesity and overweight [Internet]. [cited 2024 Apr 8]. Available from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

4. World Obesity Federation, World Obesity Atlas 2023 [Internet]. [cited 2024 Apr 8]. Available from: https://data.worldobesity.org/publications/?cat=19

5. Kloock S, Ziegler CG, Dischinger U. Obesity and its comorbidities, current treatment options and future perspectives: Challenging bariatric surgery? Pharmacol Ther. 2023 Nov 1;251:108549.

6. Collaborators G 2015 O. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N Engl J Med. 2017 Jul 6;377(1):13.

7. Abbafati C, Abbas KM, Abbasi-Kangevari M, Abd-Allah F, Abdelalim A, Abdollahi M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020 Oct 10;396(10258):1223.

8. Sarwer DB, Polonsky HM. The Psychosocial Burden of Obesity. Endocrinol Metab Clin North Am. 2016 Sep 1;45(3):677.

9. Jones-Corneille LR, Wadden TA, Sarwer DB, Faulconbridge LF, Fabricatore AN, Stack RM, et al. Axis I Psychopathology in Bariatric Surgery Candidates with and without Binge Eating Disorder: Results of Structured Clinical Interviews. Obes. Surg. 2012 Mar;22(3):389.

10. Kalarchian MA, Marcus MD, Levine MD, Courcoulas AP, Pilkonis PA, Ringham RM, et al. Psychiatric disorders among bariatric surgery candidates: relationship to obesity and functional health status. Am J Psychiatry. 2007;164(2):328–34.

11. Legenbauer T, De Zwaan M, Benecke A, Mühlhans B, Petrak F, Herpertz S. Depression and anxiety: their predictive function for weight loss in obese individuals. Obes Facts. 2009 Sep;2(4):227–34.

12. Mitchell JE, Selzer F, Kalarchian MA, Devlin MJ, Strain GW, Elder KA, et al. Psychopathology before surgery in the longitudinal assessment of bariatric surgery-3 (LABS-3) psychosocial study. Surg Obes Relat Dis. 2012 Sep;8(5):533–41.

13. Rosenberger PH, Henderson KE, Grilo CM. Correlates of body image dissatisfaction in extremely obese female bariatric surgery candidates. Obes Surg. 2006 Oct 1;16(10):1331–6.

14. Latzer Y, Stein D. A review of the psychological and familial perspectives of childhood obesity. J Eat Disord. 2013 Feb 25;1(1):1–13.

15. Tremmel M, Gerdtham UG, Nilsson PM, Saha S. Economic Burden of Obesity: A Systematic Literature Review. Int J Environ Res Public Health. 2017 Apr 19;14(4).

16. Okunogbe A, Nugent R, Spencer G, Ralston J, Wilding J. Economic impacts of overweight and obesity: current and future estimates for eight countries. BMJ Glob Health. 2021 Nov 4;6(10):6351.

17. Baqai N, Wilding JPH. Pathophysiology and aetiology of obesity. Medicine. 2015 Feb 1;43(2):73–6.

18. González-Muniesa P, Mártinez-González MA, Hu FB, Després JP, Matsuzawa Y, Loos RJF, et al. Obesity. Nature Reviews Disease Primers. 2017 Jun 15;3(1):1–18.

19. Mahmoud R, Kimonis V, Butler MG. Genetics of Obesity in Humans: A Clinical Review. Int J Mol Sci. 2022 Oct 1;23(19).

20. Wardle J, Carnell S, Haworth CMA, Plomin R. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. Am J Clin Nutr. 2008 Feb 1;87(2):398–404.

21. Stunkard AJ, Foch TT, Hrubec Z. A Twin Study of Human Obesity. JAMA. 1986 Jul 4;256(1):51–4.

22. Singh RK, Kumar P, Mahalingam K. Molecular genetics of human obesity: A comprehensive review. C R Biol. 2017 Feb 1;340(2):87–108.

23. Duis J, Butler MG. Syndromic and Nonsyndromic Obesity: Underlying Genetic Causes in Humans. Adv Biol. 2022 Oct 1;6(10):2101154.

24. Butler MG. Single Gene and Syndromic Causes of Obesity: Illustrative Examples. Prog Mol Biol Transl Sci. 2016 Jan 1;140:1.

25. Carvalho LML, Jorge AA de L, Bertola DR, Krepischi ACV, Rosenberg C. A Comprehensive Review of Syndromic Forms of Obesity: Genetic Etiology, Clinical Features and Molecular Diagnosis. Curr Obes Rep. 2024 Jan 26;1–25.

26. Carvalho LML, D'Angelo CS, Villela D, da Costa SS, de Lima Jorge AA, da Silva IT, et al. Genetic investigation of syndromic forms of obesity. International Journal of Obesity. 2022 May 21;46(9):1582–6.

27. Kaur Y, de Souza RJ, Gibson WT, Meyre D. A systematic review of genetic syndromes with obesity. Obesity Reviews. 2017 Jun 1;18(6):603–34.

28. Tirthani E, Said MS, Rehman A. Genetics and Obesity. StatPearls [Internet]. 2023 Jul 31 [cited 2024 Apr 9];

29. Forsyth R, Gunay-Aygun M. Bardet-Biedl Syndrome Overview. Genetics and Genomics of Eye Disease: Advancing to Precision Medicine. 2023 Mar 23;117–36.

30. Seo S, Guo DF, Bugge K, Morgan DA, Rahmouni K, Sheffield VC. Requirement of Bardet-Biedl syndrome proteins for leptin receptor signaling. Hum Mol Genet. 2009 Apr 4;18(7):1323.

31. Pomeroy J, Krentz AD, Richardson JG, Berg RL, VanWormer JJ, Haws RM. Bardet-Biedl syndrome: Weight patterns and genetics in a rare obesity syndrome. Pediatr Obes. 2021 Feb 1;16(2).

32. Sohn YB. The genetics of obesity: A narrative review. Precision and Future Medicine. 2022 Dec 31;6(4):226–32.

33. Ingalls AM, Dickie MM, Snell GD. Obese, a new mutation in the house mouse*. Journal of Heredity. 1950 Dec 1;41(12):317–8.

34. Hummel KP, Dickie MM, Coleman DL. Diabetes, a New Mutation in the Mouse. Science. 1966 Sep 2;153(3740):1127–8.

35. Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM. Positional cloning of the mouse obese gene and its human homologue. Nature. 1994 Dec 1;372(6505):425–32.

36. Chen H, Charlat O, Tartaglia LA, Woolf EA, Weng X, Ellis SJ, et al. Evidence that the diabetes gene encodes the leptin receptor: Identification of a mutation in the leptin receptor gene in db/db mice. Cell. 1996 Feb 9;84(3):491–5.

37. Lu D, Willard D, Patel IR, Kadwell S, Overton L, Kost T, et al. Agouti protein is an antagonist of the melanocyte stimulating hormone receptor. Nature. 1994;371(6500):799–802.

38. Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ, et al. Congenital leptin deficiency is associated with severe early-onset obesity in humans. Nature. 1997 Jun 26;387(6636):903–8.

39. Clément K, Vaisse C, Lahlou N, Cabrol S, Pelloux V, Cassuto D, et al. A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. Nature. 1998 Mar 26;392(6674):398–401.

40. Jackson RS, Creemers JWM, Ohagi S, Raffin-Sanson ML, Sanders L, Montague CT, et al. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. Nature Genetics. 1997;16(3):303–6.

41. Yeo GSH, Farooqi IS, Aminian S, Halsall DJ, Stanhope RG, O'Rahilly S. A frameshift mutation in MC4R associated with dominantly inherited human obesity. Nature Genetics. 1998;20(2):111–2.

42. Krude H, Biebermann H, Luck W, Horn R, Brabant G, Grüters A. Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. Nature Genetics. 1998;19(2):155–7.

43. Saxton RA, Caveney NA, Moya-Garzon MD, Householder KD, Rodriguez GE, Burdsall KA, et al. Structural insights into the mechanism of leptin receptor activation. Nature Communications. 2023 Mar 31;14(1):1–10.

44. Dornbush S, Aeddula NR. Physiology, Leptin. StatPearls [Internet]. 2023 Apr 10 [cited 2024 Apr 10]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK537038/

45. Heymsfield SB, Greenberg AS, Fujioka K, Dixon RM, Kushner R, Hunt T, et al. Recombinant Leptin for Weight Loss in Obese and Lean Adults: A Randomized, Controlled, Dose-Escalation Trial. JAMA. 1999 Oct 27;282(16):1568–75.

46. Enriori PJ, Evans AE, Sinnayah P, Jobst EE, Tonelli-Lemos L, Billes SK, et al. Diet-Induced Obesity Causes Severe but Reversible Leptin Resistance in Arcuate Melanocortin Neurons. Cell Metab. 2007 Mar 7;5(3):181–94.

47. Lindberg I, Fricker LD. Obesity, POMC, and POMC-processing Enzymes: Surprising Results From Animal Models. Endocrinology. 2021 Dec 1;162(12).

48. Gregoric N, Groselj U, Bratina N, Debeljak M, Zerjav Tansek M, Suput Omladic J, et al. Two Cases With an Early Presented Proopiomelanocortin Deficiency—A Long-Term Follow-Up and Systematic Literature Review. Front Endocrinol. 2021 Jun 9;12:1.

49. Fatima MT, Ahmed I, Fakhro KA, Akil ASAS. Melanocortin-4 receptor complexity in energy homeostasis,obesity and drug development strategies. Diabetes Obes Metab. 2022 Apr 1;24(4):583.

50. Stutzmann F, Tan K, Vatin V, Dina C, Jouret B, Tichet J, et al. Prevalence of Melanocortin-4 Receptor Deficiency in Europeans and Their Age-Dependent Penetrance in Multigenerational Pedigrees. Diabetes. 2008 Sep 1;57(9):2511–8.

51. Ramos-Molina B, Martin MG, Lindberg I. PCSK1 Variants and Human Obesity. Prog Mol Biol Transl Sci. 2016 Jan 1;140:47.

52. Wilschanski M, Abbasi M, Blanco E, Lindberg I, Yourshaw M, Zangen D, et al. A Novel Familial Mutation in the PCSK1 Gene That Alters the Oxyanion Hole Residue of Proprotein Convertase 1/3 and Impairs Its Enzymatic Activity. PLoS One. 2014 Oct 1;9(10).

53. Saeed S, Bonnefond A, Manzoor J, Shabir F, Ayesha H, Philippe J, et al. Genetic variants in LEP, LEPR, and MC4R explain 30% of severe obesity in children from a consanguineous population. Obesity. 2015 Aug 1;23(8):1687–95.

54. van der Klaauw AA, Croizier S, Mendes de Oliveira E, Stadler LKJ, Park S, Kong Y, et al. Human Semaphorin 3 Variants Link Melanocortin Circuit Development and Energy Balance. Cell. 2019 Feb 7;176(4):729-742.e18.

55. Loos RJF, Yeo GSH. The genetics of obesity: from discovery to biology. Nat Rev Genet. 2022 Feb 1;23(2):120.

56. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science. 2007 May 5;316(5826):889.

57. Lan N, Lu Y, Zhang Y, Pu S, Xi H, Nie X, et al. FTO – A Common Genetic Basis for Obesity and Cancer. Front Genet. 2020 Nov 16;11:559138.

58. Dina C, Meyre D, Gallina S, Durand E, Körner A, Jacobson P, et al. Variation in FTO contributes to childhood obesity and severe adult obesity. Nat Genet. 2007 Jun;39(6):724–6.

59. Haupt A, Thamer C, Machann J, Kirchhoff K, Stefan N, Tschritter O, et al. Impact of variation in the FTO gene on whole body fat distribution, ectopic fat, and weight loss. Obesity (Silver Spring). 2008 Aug;16(8):1969–72.

60. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet. 2007 Jul;3(7):1200–10.

61. Cauchi S, Stutzmann F, Cavalcanti-Proença C, Durand E, Pouta A, Hartikainen AL, et al. Combined effects of MC4R and FTO common genetic variants on obesity in European general populations. J Mol Med (Berl). 2009 May;87(5):537–46.

62. Wang D, Wu Z, Zhou J, Zhang X. Rs9939609 polymorphism of the fat mass and obesity-associated (FTO) gene and metabolic syndrome susceptibility in the Chinese population: a meta-analysis. Endocrine. 2020 Aug 1;69(2):278–85.

63. Wen W, Cho YS, Zheng W, Dorajoo R, Kato N, Qi L, et al. Meta-analysis identifies common variants associated with body mass index in east Asians. Nat Genet. 2012 Mar;44(3):307–11.

64. Reuter CP, Rosane De Moura Valim A, Gaya AR, Borges TS, Klinger EI, Possuelo LG, et al. FTO polymorphism, cardiorespiratory fitness, and obesity in Brazilian youth. Am J Hum Biol. 2016 May 1;28(3):381–6.

65. Mehrdad M, Fardaei M, Fararouei M, Eftekhari MH. The association between FTO rs9939609 gene polymorphism and anthropometric indices in adults. J Physiol Anthropol. 2020 May 12;39(1).

66. Shahid A, Rana S, Saeed S, Imran M, Afzal N, Mahmood S. Common variant of FTO Gene, rs9939609, and obesity in Pakistani females. Biomed Res Int. 2013;2013.

67. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014;507(7492):371–5.

68. Stratigopoulos G, Padilla SL, LeDuc CA, Watson E, Hattersley AT, McCarthy MI, et al. Regulation of Fto/Ftm gene expression in mice and humans. Am J Physiol Regul Integr Comp Physiol. 2008;294(4):1185–96.

69. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med. 2015 Sep 9;373(10):895.

70. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015 Feb 11;518(7538):197–206.

71. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. PLoS Genet. 2015;11(10):e1005378.

72. Hakanen M, Raitakari OT, Lehtimäki T, Peltonen N, Pahkala K, Sillanmäki L, et al. FTO genotype is associated with body mass index after the age of seven years but not with energy intake or leisure-time physical activity. J Clin Endocrinol Metab. 2009;94(4):1281–7.

73. Jonsson A, Renström F, Lyssenko V, Brito EC, Isomaa B, Berglund G, et al. Assessing the effect of interaction between an FTO variant (rs9939609) and physical activity on obesity in 15,925 Swedish and 2,511 Finnish adults. Diabetologia. 2009;52(7):1334–8.

74. Stutzmann F, Vatin V, Cauchi S, Morandi A, Jouret B, Landt O, et al. Non-synonymous polymorphisms in melanocortin-4 receptor protect against obesity: the two facets of a Janus obesity gene. Hum Mol Genet. 2007 Aug 1;16(15):1837–44.

75. Loos RJF, Lindgren CM, Li S, Wheeler E, Hua Zhao J, Prokopenko I, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nature Genetics. 2008 May 4;40(6):768–75.

76. Chami N, Preuss M, Walker RW, Moscati A, Loos RJF. The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population. PLoS Med. 2020 Jul 1;17(7).

77. Li S, Chen W, Srinivasan SR, Boerwinkle E, Berenson GS. Influence of lipoprotein lipase gene Ser447Stop and β1-adrenergic receptor gene Arg389Gly polymorphisms and their interaction on obesity from childhood to adulthood: the Bogalusa Heart Study. International Journal of Obesity. 2006 Mar 14;30(8):1183–8.

78. Linné Y, Dahlman I, Hoffstedt J. β1-Adrenoceptor gene polymorphism predicts long-term changes in body weight. International Journal of Obesity. 2005 Feb 1;29(5):458–62.

79. Macho-Azcarate T, Marti A, González A, Martinez JA, Ibañez J. Gln27Glu polymorphism in the beta2 adrenergic receptor gene and lipid metabolism during exercise in obese women. International Journal of Obesity. 2002 Nov 13;26(11):1434–41.

80. Corbalán MS, Marti A, Forga L, Martínez-González MA, Martínez JA. β2-adrenergic receptor mutation and abdominal obesity risk: Effect modification by gender and HDL-cholesterol. Eur J Nutr. 2002 Feb 1;41(3):114–8.

81. Park HS, Kim Y, Lee C. Single nucleotide variants in the β2-adrenergic and β3-adrenergic receptor genes explained 18.3% of adolescent obesity variation. Journal of Human Genetics. 2005 Jul 1;50(7):365–9.

82. Ochoa MC, Marti A, Azcona C, Chueca M, Oyarzábal M, Pelach R, et al. Gene–gene interaction between PPARγ2 and ADRβ3 increases obesity risk in children and adolescents. International Journal of Obesity. 2004 Nov 15;28(3):S37–41.

83. Carmen Ochoa M del, Martí A, Alfredo Martínez J. [Obesity studies in candidate genes]. Med Clin (Barc). 2004 Jan;122(14):542–51.

84. Oktavianthi S, Trimarsanto H, Febinia CA, Suastika K, Saraswati MR, Dwipayana P, et al. Uncoupling protein 2 gene polymorphisms are associated with obesity. Cardiovasc Diabetol. 2012 Apr 25;11(1):1–10.

85. Suviolahti E, Oksanen LJ, Öhman M, Cantor RM, Ridderstråle M, Tuomi T, et al. The SLC6A14 gene shows evidence of association with obesity. J Clin Invest. 2003 Dec 1;112(11):1762–72.

86. Durand E, Boutin P, Meyre D, Charles MA, Clement K, Dina C, et al. Polymorphisms in the amino acid transporter solute carrier family 6 (neurotransmitter transporter) member 14 gene contribute to polygenic obesity in French Caucasians. Diabetes. 2004 Sep;53(9):2483–6.

87. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. The American Journal of Human Genetics. 2023 Feb 2;110(2):179–94.

88. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1(1):1–21.

89. Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. Genome Med. 2020 May 18;12(1):1–11.

90. Choi SW, Mak TSH, O'Reilly PF. A guide to performing Polygenic Risk Score analyses. Nat Protoc. 2020 Sep 9;15(9):2759.

91. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature. 2021 Mar 11;591(7849):211.

92. Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. Annu Rev Biomed Data Sci. 2022 Aug 8;5:293.

93. Collister JA, Liu X, Clifton L. Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists. Front Genet. 2022 Feb 18;13:818574.

94. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021 Oct 18;599(7886):628–34.

95. Auer PL, Lettre G. Rare variant association studies: Considerations, challenges and opportunities. Genome Med. 2015 Feb 23;7(1):1–11.

96. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014 Jul 7;95(1):5.

97. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. Journal of Human Genetics. 2020 Sep 18;66(1):11–23.

98. 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;91(2):224.

99. Chen W, Coombes BJ, Larson NB. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. Front Genet. 2022 Oct 6;13.

100. Chicco D, Agapito G. Nine quick tips for pathway enrichment analysis. PLoS Comput Biol. 2022 Aug 1;18(8).

101. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nature Protocols. 2019 Jan 21;14(2):482–517.

102. Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, et al. Integrative pathway enrichment analysis of multivariate omics data. Nature Communications. 2020 Feb 5;11(1):1–16.

103. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles G V., et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013 Apr 15;14.

104. Schwartz MW, Seeley RJ, Zeltser LM, Drewnowski A, Ravussin E, Redman LM, et al. Obesity Pathogenesis: An Endocrine Society Scientific Statement. Endocr Rev. 2017 Aug 8;38(4):267.

105. Leibel RL, Rosenbaum M, Hirsch J. Changes in energy expenditure resulting from altered body weight. N Engl J Med. 1995 Mar 9;332(10):621–8.

106. Banks WA, Kastin AJ, Huang W, Jaspan JB, Maness LM. Leptin enters the brain by a saturable system independent of insulin. Peptides (NY). 1996 Jan 1;17(2):305–11.

107. Elmquist JK, Ahima RS, Maratos-Flier E, Flier JS, Saper CB. Leptin Activates Neurons in Ventrobasal Hypothalamus and Brainstem. Endocrinology. 1997 Feb 1;138(2):839–42.

108. Dornbush S, Aeddula NR. Physiology, Leptin. StatPearls [Internet]. 2023 Apr 10 [cited 2024 Apr 23]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK537038/

109. Yeo GSH, Chao DHM, Siegert AM, Koerperich ZM, Ericson MD, Simonds SE, et al. The melanocortin pathway and energy homeostasis: From discovery to obesity therapy. Mol Metab. 2021 Jun 1;48:101206.

110. Oussaada SM, van Galen KA, Cooiman MI, Kleinendorst L, Hazebroek EJ, van Haelst MM, et al. The pathogenesis of obesity. Metabolism. 2019 Mar 1;92:26–36.

111. Lavoie O, Michael NJ, Caron A. A critical update on the leptin-melanocortin system. J Neurochem. 2023 May 1;165(4):467–86.

112. Huo L, Gamber K, Greeley S, Silva J, Huntoon N, Leng XH, et al. Leptin-Dependent Control of Glucose Balance and Locomotor Activity by POMC Neurons. Cell Metab. 2009 May 14;9(6):537–47.

113. Berglund ED, Liu C, Sohn JW, Liu T, Kim MH, Lee CE, et al. Serotonin 2C receptors in pro-opiomelanocortin neurons regulate energy and glucose homeostasis. J Clin Invest. 2013 Dec 2;123(12):5061–70.

114. Hill JW, Elias CF, Fukuda M, Williams KW, Berglund ED, Holland WL, et al. Direct Insulin and Leptin Action on Pro-opiomelanocortin Neurons Is Required for Normal Glucose Homeostasis and Fertility. Cell Metab. 2010 Apr 7;11(4):286–97.

115. Steuernagel L, Lam BYH, Klemm P, Dowsett GKC, Bauder CA, Tadross JA, et al. HypoMap—a unified single-cell gene expression atlas of the murine hypothalamus. Nature Metabolism. 2022 Oct 20;4(10):1402–19.

116. Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, Tenen D, et al. A molecular census of arcuate hypothalamus and median eminence cell types. Nature Neuroscience. 2017 Feb 6;20(3):484–96.

117. Gropp E, Shanabrough M, Borok E, Xu AW, Janoschek R, Buch T, et al. Agouti-related peptide–expressing neurons are mandatory for feeding. Nature Neuroscience. 2005 Sep 11;8(10):1289–91.

118. Lam DD, Garfield AS, Marston OJ, Shaw J, Heisler LK. Brain serotonin system in the coordination of food intake and body weight. Pharmacol Biochem Behav. 2010 Nov 1;97(1):84–91.

119. van Galen KA, ter Horst KW, Serlie MJ. Serotonin, food intake, and obesity. Obesity Reviews. 2021 Jul 1;22(7).

120. La Fleur SE, Serlie MJ. The interaction between nutrition and the brain and its consequences for body weight gain and metabolism; studies in rodents and men. Best Pract Res Clin Endocrinol Metab. 2014 Oct 1;28(5):649–59.

121. Farhadipour M, Depoortere I. The Function of Gastrointestinal Hormones in Obesity—Implications for the Regulation of Energy Intake. Nutrients. 2021;13(6).

122. Lean MEJ, Malkova D. Altered gut and adipose tissue hormones in overweight and obese individuals: cause or consequence? International Journal of Obesity. 2015 Oct 26;40(4):622–32.

123. Liu BN, Liu XT, Liang ZH, Wang JH. Gut microbiota in obesity. World J Gastroenterol. 2021 Jul 7;27(25):3837.

124. Torres-Fuentes C, Schellekens H, Dinan TG, Cryan JF. The microbiota–gut–brain axis in obesity. Lancet Gastroenterol Hepatol. 2017 Oct 1;2(10):747–56.

125. Bäckhed F, Ding H, Wang T, Hooper L V., Gou YK, Nagy A, et al. The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A. 2004 Nov 2;101(44):15718–23.

126. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci USA. 2005 Aug 2;102(31):11070–5.

127. Indiani CMDSP, Rizzardi KF, Castelo PM, Ferraz LFC, Darrieux M, Parisotto TM. Childhood Obesity and Firmicutes/Bacteroidetes Ratio in the Gut Microbiota: A Systematic Review. Child Obes. 2018 Nov 1;14(8):501–9.

128. 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.

129. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008 Sep 12;83(3):311–21.

130. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007;63(4):1079–88.

131. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 2008 Feb 8;82(2):386–97.

132. Larson NB, Chen J, Schaid DJ. A Review of Kernel Methods for Genetic Association Studies. Genet Epidemiol. 2019 Mar 1;43(2):122.

133. 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 7;89(1):82.

134. Davies RB. The Distribution of a Linear Combination of $\chi 2$ Random Variables. J R Stat Soc Ser C Appl Stat. 1980 Nov 1;29(3):323–33.

135. Kuonen D. Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables. Biometrika. 1999 Dec 1;86(4):929–35.

136. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013 Feb;37(2):196–204.

137. Onifade M, Roy-Gagnon MH, Parent MÉ, Burkett KM. Comparison of mixed model based approaches for correcting for population substructure with application to extreme phenotype sampling. BMC Genomics. 2022 Dec 1;23(1):1–12.

138. 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. The American Journal of Human Genetics. 2019 Feb 7;104(2):260–74.

139. Zhao K, Rhee SY. Interpreting omics data with pathway enrichment analysis. Trends in Genetics. 2023 Apr 1;39(4):308–19.

140. Kars ME, Başak AN, Onat OE, Bilguvar K, Choi J, Itan Y, Çağlar C, Palvadeau R, Casanova JL, Cooper DN, Stenson PD, Yavuz A, Buluş H, Günel M, Friedman JM, Özçelik T. The genetic structure of the Turkish population reveals high levels of variation and admixture. Proc Natl Acad Sci U S A. 2021 Sep 7;118(36):e2026076118.

141. 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. Current protocols in bioinformatics. 2013;11(1110):11.10.1.

142. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 8;27(15):2156.

143. Wickham H. ggplot2. ggplot2. 2009;

144. 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;81(3):559.

145. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Jun 6;17(1).

146. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Jul 3;38(16):e164.

147. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020 Dec 1;12(1):1–8.

148. 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: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012 Apr 1;6(2):80–92.

149. Hartmann T, Schröder C, Kuthe E, Lähnemann D, Köster J. Insane in the vembrane: filtering and transforming VCF/BCF files. Bioinformatics. 2023 Jan 1;39(1).

150. 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.

151. Lorenzo Bermejo J, Marcella Devoto H, Christine Fischer R. 46th European Mathematical Genetics Meeting (EMGM) 2018, Cagliari, Italy, April 18-20, 2018: Abstracts. Hum Hered. 2018 Jun 26;83(1):1–29.

152. Walker LC, Hoya M de la, Wiggins GAR, Lindy A, Vincent LM, Parsons MT, et al. Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. Am J Hum Genet. 2023 Jul 6;110(7):1046–67.

153. Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022 Dec 12;109(12):2163.

154. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan 1;47(Database issue):D886.

155. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct 10;99(4):877.

156. Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017 Mar 1;38(3):243–51.

157. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan 24;176(3):535-548.e24.

158. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Current protocols in human genetics / editorial board, Jonathan L Haines . [et al]. 2013;0 7(SUPPL.76):Unit7.20.

159. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009 Jun 25;4(7):1073–81.

160. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25.

161. Aleksander SA, Balhoff J, Carbon S, Michael Cherry J, Drabkin HJ, Ebert D, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224(1).

162. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023 Jan 6;51(D1):D587–92.

163. Smith CL, Eppig JT. The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med. 2009 Nov;1(3):390.

164. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016 Aug 29;48(10):1284–7.

165. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics. 2016 Oct 3;48(11):1443–8.

166. Rohde K, Schamarek I, Blüher M. Consequences of Obesity on the Sense of Taste: Taste Buds as Treatment Targets? Diabetes Metab J. 2020;44(4):509.

9 CURRICULUM VITAE