Title
A Genome Wide Association Study of Gene by Smoking Interaction on Type II Diabetes

Permalink
https://escholarship.org/uc/item/4282c29c

Author
Ferdos, Sepideh

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

A Genome Wide Association Study of Gene by Smoking Interaction on Type II Diabetes

A Thesis submitted in partial satisfaction of the requirements for the Master's degree

in

Public Health

by

Sepideh Ferdos

Committee in charge:

Professor Rany Salem, Chair
Professor Richard Garfein
Professor David Strong

2020

The Thesis of Sepideh Ferdos is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

Table of Contents

List of Figures

List of Tables

ABSTRACT OF THE THESIS

A Genome Wide Association Study of Gene by Smoking Interaction on Type II Diabetes

by

Sepideh Naggash Ferdos

Master of Public Health

University of California San Diego, 2020

Professor Rany Salem, Chair

Type 2 diabetes is a multifactorial disease that is the result of glycemic disfunction in the body. Risk for type 2 diabetes is influenced by the interplay of multiple lifestyle, environmental, and genetic factors. Prior genome wide association studies have identified hundreds of genetic variants associated with type 2 diabetes. However, the majority of these genetic studies have failed to consider the role of environmental risk factors on disease risk, e.g. gene x environment interaction analyses. The role of smoking as a modulator in the association between genetic variants and type 2 diabetes is of particular importance, since smoking is a modifiable risk factor and has been established as a causal risk factor for type 2 diabetes. Conducting a gene by smoking interaction analysis on type 2 diabetes allows consideration of both genetic and environmental risk factors and has the potential to unveil novel single-nucleotide polymorphisms (SNPs) that influence disease vulnerability typically missed by traditional analytic approaches. To conduct a large-scale gene x environment study, phenotypic and genotypic data on 345,955 individuals was utilized from 14 studies within two biomedical repositories, dbGaP and UK Biobank. Phenotypic data extracted from the studies were reviewed and harmonized prior to performing a gene by smoking genome-wide association study (GWAS) for each study. Individual GWAS statistics were combined through meta-analysis, stratified by racial/ethnic groups. This study included the following sample of individuals: 324,834 of European ancestry, 9,040 of African ancestry, 6,125 of Hispanic ancestry, and 5,956 of Asian ancestry. The study sample includes 40, 994 diabetics and 34,764 current smokers, and majority females (55%, 177,369). Results from the meta-analysis produced genome-wide significant SNPs for the European ancestry main effect and the Asian ancestry interaction effect. The European ancestry main effect results revealed a genome wide significant (p-value $9.59\text{x}10^{-33}$) signal in the *TCF7L2* gene, a commonly replicated SNP in several GWAS's of type 2 diabetes. The smoking by SNP results found SNPs at the genome-wide suggestive (p-value $\leq 1\text{x}10^{-5}$) level. The Asian ancestry interaction results revealed inflated results driven by two studies with modest sample size that we considered false positives, and therefore not included in the final results. This work underscores the importance of conducting gene by smoking interaction analyses on type 2 diabetes to identify genetic variants that influence disease vulnerability.

Chapter I:

Introduction

Type 2 Diabetes Background

Type 2 diabetes (T2D) is a chronic metabolic disorder characterized by glycemic disfunction.[1] T2D is one of the leading causes of death and responsible for increased morbidity and healthcare costs in the United States.[2] Individuals with T2D require lifelong clinical treatment, management, and supervision.[2] Medical care costs for individuals living with T2D are two times higher than those without T2D.[3] A study conducted by the American Diabetes Association (ADA) in 2017 estimated that the economic impact of diabetic illnesses results in an annual cost of diagnosed conditions at 327 billion dollars, an increase of 82 billion dollars (26%) since last examined in 2012.[3] The individual and national economic burden of this condition continues to increase linearly, placing it as a top health priority.[3,4] The International Diabetes Federation (IDF) places T2D as an illness of growing public health importance, due to its increasing prevalence and significant cause morbidity, death, and economic costs.[4]

The Centers for Disease Control and Prevention's (CDC) 2020 National Diabetes Statistics Report indicates that 1 in 10 individuals have T2D and 1 in 3 individuals are prediabetic in the United States.[5] Globally, prevalence of type 2 diabetes will increase to 693 million by 2045, a 50% increase from the year 2017.[4] Investigation into the distribution of T2D disease burden by sociodemographic status reveals that 80% of individuals living with diabetes are from low to middle income countries.[6] The CDC report notes that 15% of diabetic individuals are current smokers, while 36% are former smokers.[5] Comparatively, the CDC reported in 2018 that 13.7% of the US population (≥18 years old) are current smokers.[7]

Furthermore, the 2014 Surgeon General's Report, titled 'The Health Consequences of Smoking—50 Years of Progress', concludes that smoking is causally associated to T2D, increasing risk of T2D by 30% to 40% in active smokers vs. nonsmokers.[8] It is essential to consider the possibility that many diabetic individuals stop smoking due to their condition, for that reason the risk associated with smoking on T2D may be greater than observed. The report states that smoking directly affects glucose regulation, with a positive dose-response relationship between quantity of cigarette intake and risk of T2D.[8] The report underscores the gap in knowledge on the biological explanation behind this relationship and suggested the need for further research to understand the mechanistic and epidemiological impact of this relationship.[8]

Pathophysiology of Type 2 Diabetes

T2D, also known as adult-onset diabetes, is characterized by glycemic dysfunction and elevated serum glucose levels. There are multiple pathophysiological mechanisms that can lead to glycemic dysfunction, however, there are two common pathways. The first common mechanism occurs when pancreatic β-cell function is impaired, and insulin secretion fails to sufficiently respond to elevated glucose levels leading to dysregulated serum glucose control.[1,9,10] A second common mechanism is insulin resistance and occurs when cells in the body fail to respond to insulin and do not uptake glucose from the blood, resulting in high levels of blood glucose.[1,9,10] Insulin resistance is a physiological dysfunction commonly observed amongst obese individuals. Clinical manifestations of T2D include unexplained weight loss, fatigue, polyuria, repeated infections, dry mouth, decreased vision, irritability, and sexual dysfunction.[11] Diagnosis of T2D is determined through a blood test involving a metabolic panel (serum glucose, insulin, or HbA1c levels).[12,13] Treatment options available include anti-diabetes medication, and lifestyle changes, such as diet and exercise.[14] Patients with long-term progressive disease states are at an elevated risk for a multitude of comorbidities, including, diabetic kidney disease, end stage renal disease, diabetic retinopathy, diabetic neuropathy, cardiovascular complications.[15]

Risk Factors of Type 2 Diabetes

Development of T2D often involves the interplay of environmental, lifestyle factors, and genetic traits. Risk factors of T2D can be compartmentalized into two broad categories: modifiable, and non-modifiable risk factors. Modifiable risk factors include smoking, sleep quality, stress, lack of physical activity, sedentary lifestyle, and poor diet.[11,15–17] Non-modifiable risk factors of T2D include age, race/ethnicity, family history of diabetes, gestational diabetes, and genetic variants.[11,15–17] Risk factors such as, visceral obesity or ectopic fat, high blood pressure, low high-density lipoproteins (HDL) or high triglycerides can be considered as modifiable and non-modifiable, as they are influenced by behavioral and genetic influence.[11,15–17] The racial or ethnic groups that experience a higher prevalence of T2D include African Americans, Hispanics, American Indian/Alaskan Natives, Hawaiians, Pacific Islander, and Asians.[18,19] Prevalence of T2D among these racial groups is ~10% higher than European ancestry populations overall.[18]

Genetics of Type 2 Diabetes

T2D has long been observed to run in families, suggesting a genetic component.[20] Risk of T2D increases with number of affected parents, with a 40% increase in those with one diabetic parent and 70% increases in those with two diabetic parents.[20] T2D has strong genetic heritability, which is the proportion of phenotypic variation that is attributable to genetic factors, ranging from 20% to 80%.[20] Therefore, comprehensive understanding of T2D requires consideration of the role of genetic variation on disease susceptibility.[20]

Genetics have been instrumental providing insights into T2D pathophysiology.[21] Early genetic studies, comprised of linkage and candidate gene studies, had limited success.[22] The emergence of GWAS has been pivotal in shedding insight into the genetic basis of many complex or chronic diseases, including T2D.[21–24] The expansion of GWAS over the past decade has provided essential information in identifying genetic variants and genes associated with disease susceptibility, insights into personalized medicine, and discovery of novel pharmacological treatments.[21–23] GWAS's are a highly ubiquitous study design in genetic epidemiology and involve the analysis of millions of single-nucleotide polymorphisms (SNPs) genetic variants with a phenotype of interest (e.g. disease status or quantitative trait).[21,24–26] With the use of GWAS studies, researchers have identified risk loci for hundreds of traits and diseases of public health importance, including asthma, T2D, coronary artery disease, and several types of cancers.[27,28]

GWAS of T2D have identified hundreds of loci. The National Institute of Health (NIH) Genetics Reference reports prior studies have identified approximately 150 gene variants associated with the risk of developing T2D.[29] A study conducted in 2013 elaborated on initial gene discoveries from GWAS; such include the *TCF7L2* gene, first discovered by Sladek et al., which is the most consistently replicated gene associated with T2D globally.[30] While some genes are seen in diverse populations globally, the *HHEX* gene has been found in several GWAS's of T2D significant among European and Asian ancestry populationst.[30] A review published by Olokoba et al. in 2012 reported several other genes with strong association to the development of

T2D, which include the following: *PPARG, FTO, KCNJ11, NOTCH2, WFS1, CDKAL1, IGF2BP2, SLC30A8*, and *JAZF1*.[31]

Although several studies have identified and replicated dominant gene variants related to T2D, recent literature focuses on analyzing millions of genetic variants to provide a more complete picture of genes that influence disease susceptibility. Large-scale GWAS studies have provided insights into both common and rare genetic variants that influence T2D susceptibility. A GWAS meta-analysis conducted by Xue et al. in 2018 identified 139 common variants at a p-value level of $<5x10^{-8}$ and 4 rare variants at the $<5x10^{-9}$ p-value level (higher threshold required to control genome-wide false positive rate) associated with T2D.[32] Another study published in 2018 by Mahajan et al. used similar methodology to conduct a GWAS, however, they expanded the study through combining data from 31 GWAS's to include 74,124 T2D cases and 824,006 controls.[33] The study discovered 231 loci with genome-wide significance in the BMI-unadjusted analysis and 152 loci in the BMI-adjusted analysis.[33] This study performed BMI-adjusted and unadjusted GWAS to identify T2D risk effects driven primarily by BMI or adipose tissue, as BMI is an effect modifier for T2D.[33] This study identified 135 novel T2D risk.[33] Flannick et al. investigated the role of rare variants on T2D risk through use of whole-genome and exome sequencing of genetic data.[34] This study examined the association of over 27 million SNPs, indels, and gene variants with T2D.[34] Another study by Flannick et al. conducted an exome-sequencing of 20,791 cases of T2D and 24,440 controls to identify rare gene associations with a minor allele frequency (MAF) of 0.5%.[35] This study identified 4 genes at the exome-wide significance level and 30 *SLC30A8* gene alleles that indicated protection against T2D.[35] These studies emphasizes the importance of evaluating rare variants and their contribution to risk of complex diseases such as T2D, as whole genome-sequencing allows to capture more novel and rare variants missed in GWAS studies.[34,35]

While genetic studies have allowed for the discovery of various genetic polymorphisms involved in development of T2D, the influence of environmental risk factors on genetic variations involved in disease susceptibility is largely unknown. GWAS studies analyzing only main genetic effects entail an important study limitation. The majority of GWAS's consider environmental measures as nuisance parameters to be factored out and do not account for the role of these factors on disease susceptibility.[36] To address this limitation, there is a need to consider gene by environment analyses at scale to gain insights on the shared genetic effect of genetic variants and environmental factors on disease susceptibility.

The objective of gene by environment interaction studies is to examine the joint impact of both genes and environmental influences on disease susceptibility.[37] Gene by environment interactions may be of particular value in the study of complex disease as they directly examine the interplay of environmental factors and genetic variants not captured in standard genetic analyses.[37] Gene by smoking interaction studies are emerging as an important genetic epidemiology study framework, with studies considering a broad set of disease outcomes such as chronic obstructive pulmonary disease (COPD), colorectal cancer, serum lipids, pulmonary function, coronary heart disease, coronary artery calcification, and hypertension.[38–44] A literature review on gene by environment interactions on T2D identified several studies that investigated a variety of exposures, including alcohol consumption, physical activity, and lifestyle changes.[45–47] As of date, there is only one article regarding gene by smoking interaction on T2D, however, this paper has included only a handful of studies (n=74,583) and did not incorporate a sex stratified analysis. Sex stratified analyses are of importance since there are significant difference in smoking prevalence by sex as noted in the Healthy People 2020 survey.[48,49]

Research Aims and Goals

The research question being addressed in this study considers whether smoking influences genetic susceptibility of individuals to T2D. The aim of this study is to perform a large-scale gene by smoking interaction on T2D, using studies retrieved from two biorepositories, the Database of Genotypes and Phenotypes (dbGaP), and UK Biobank.[50,51] dbGaP is an NIH sponsored biorepository with genetic and phenotype data managed by the National Center for Biotechnology Information (NCBI).[19,52] UK Biobank is a study based in the United Kingdom collecting genetic data to evaluate disease risk.[53] These biorepositories provide a resource for investigators to conduct genetic studies on phenotypes and disease endpoints of interest. This study is comprised of three main goals. The first goal of this study is to extract variables of interest from datasets within dbGaP and UK Biobank and harmonize the variables of interest according to established phenotype definitions and categorizations. The second goal is to conduct a genome-wide association study to analyze the interaction between smoking and T2D through analysis of single-nucleotide polymorphisms. The final goal of this study is to identify genetic variants associated with the impact of smoking on T2D.

Chapter II:

Materials and Methods

Study Samples

This project leverages studies that were previously retrieved from the database for Genotypes and Phenotypes (dbGaP), an NIH sponsored biorepository created and managed by The National Center for Biotechnology Information, and UK Biobank study supported by the National Health Service (NHS).[50–54] All NIH funded genetic studies (e.g. genetic, genome, genotyping arrays and sequencing data) are required to submit both genetic and phenotypic data to dbGaP. The dbGaP collection contains extensive and diverse amount of individual-level data regarding variables, datasets, and molecular assays data.[54] UK Biobank is a biobank scale prospective cohort study supported by the NHS, designed and conducted with data sharing in mind.[55] UK Biobank collected data on a variety of phenotypic information and biological samples from ~500,000 individuals in the United Kingdom between the ages of 40 to 69 from 2006 to 2010.[55] The objective of both dbGaP and UK Biobank is to make available extensive individual level phenotypic and genotypic data to the research community and allow investigators to explore a wide variety of questions in the relationship between human genetic variation on disease susceptibility.[50,55] The Salem lab has previously requested and retrieved data for >150 dbGaP studies and UK Biobank.

Fourteen studies were included in this project after comprehensive review to identify studies with genome-wide genotyping array data and relevant phenotype data availability (e.g. smoking exposure and diabetes disease status). Thirteen studies were acquired via dbGaP for this project, include: The National Institute on Aging Long Life Study (LLFS – phs000397, n=1,800), The Research Program on Genes, Environment, & Health (RPGEH – phs000788, n=74,303), Geneva Diabetes Study (GENEVA – phs000091, n=5,552), GWAS on Cataract and HDL (CATARACT – phs000170, n=2,177), Northwestern Nugene Project: T2D (NW – phs000237, n=1,288), Development and Use of Network Infrastructure for GWAS (DUNI – phs000234, n=1,438), GWAS of Peripheral Artery Disease (PAD – phs000203, n=3,048), Catheterization Genetics (CATHGEN – phs000703, n=1,152), National Institute of Diabetes & Digestive Kidney Disease (NIDDK – phs000524, n=3,367), Atherosclerosis Risk in Communities (ARIC – phs000280, n=10,137), Cardiovascular Health Studies (CARDIA – phs000285, n=2,347), Cardiovascular Health Studies (CHS – phs000287, n=3,754), Multi-Ethnic Study of Atherosclerosis (MESA – phs000209, n=6,164). The fourteenth study, UK Biobank (UKBB), included 229,428 participants of European ancestry, after excluded individuals to break related pairs (below 1st cousin pairs).

Phenotype Definitions and Harmonization

The fourteen studies consisted of a mix of study types, including cross-sectional studies, case-control studies, case set studies, and longitudinal studies. There is significant heterogeneity in availability of relevant phenotypes and variability in modes of measure due to study type and original intent of variable collection. Phenotype harmonization and standardization is the primary challenge in secondary data analysis across multiple studies. Phenotype definitions and categorizations (detailed below) were established prior to data extraction and harmonization. A few challenges of the phenotype harmonization in this project include: 1) assortment of naming schema unique to each study, 2) multiple variables for the same phenotype, 3) ambiguous phenotype labels and 4) varying units. To tackle these challenges, a detailed catalog of each study was performed to identify variables of interest, and use of the statistical software, R Studios, to rename variables, carefully review variable categorizations, measurement units, missingness, and generate a clean harmonized variable set of each of the fourteen studies.

T2D was defined by a through a broad set of measures, including lab tests, metabolic panel, or, self-report of medical history and/or medication. The multiple criteria facilitate defining T2D across a large set of studies with heterogeneity in available phenotypic data. The aim of using lab tests/metabolic panel and report data allows maximization the of study inclusion, while establishing clear cutoffs. Use of self-report data increases susceptibility to misclassification, though disease status was checked and cross-referenced using different diabetes variables where available in each study. Table 3 presents the criteria used to define T2D in this project and was adapted from reference to the American Diabetes Association guidelines.[12] T2D status was defined in a hierarchal order: (1) an Oral Glucose Tolerance Test (OGTT) of ≥200 mg/dl, (2) a Fasting Plasma Glucose (FPG) of ≥126 mg/dl, (3) physician diagnosis, (4) self-report, and (5) T2D medication report. The first two types of tests are administered through a lab test or metabolic panel. Oral glucose tolerance tests are typically considered the gold standard of glucose measurement as it provides the most accurate assessment. The FPG test is the most commonly utilized, since it is faster and easier as subjects are only required to complete a post fasting blood draw, versus timed collection post standard glucose intake in OGTT. The OGTT and FPG test cut-offs were based on the 2020 version of the American Diabetes Association T2D classifications.[12] Determination of T2D status through physician diagnosis, self-report, or T2D medication status, are all considered a form of self-report. Participants involved in the study fill out questionnaires or assessments of medical history to determine patient history. Medications that were included in the datasets and considered T2D prescriptions included metformin, sulfonylureas, thiazolidinediones, meglitinides, biguanides, alpha-glucosidase inhibitors, DPP-4 inhibitors, and insulin.

As Table 4 illustrates, smoking status is divided into following three categories, never smokers, current smokers, and former smokers.[56] The category of never smokers are defined as individuals who report being non-smokers and/or have smoked ≤100 cigarettes (<0.02 pack years) in their lifetime.[56] Categories include an "and/or" to indicate the use of different variables used to form the categorizations of "current", "former", and "never". Current smokers are defined as individuals who report smoking now and/or have been smoking at least within the past one month or more, and/or have smoked >100 cigarettes (>0.02 pack years) in their lifetime. Former smokers have smoked ≥100 cigarettes (≥0.02 pack years) in their lifetime and are defined as individuals who self-report as former smokers, and/or not current smoker. For the purposes of this study, extraction of smoking status data focused on current smokers and never smokers for the analysis. Smoking related measures collected in each study differed considerably, a significant source of heterogeneity in quantifying and harmonizing this behavioral factor. Information on smoking encompassed a range of measurements such as report of smoking status, pack years, having ever smoked 100 cigarettes in a lifetime, number of cigarettes per day, number of cigarettes per week, number of years an individual smoked, age of individual when they started smoking, average number of daily cigarette use, smoking history, and cigarettes per week. The majority of studies had limited data on smoking status and comprised of categorical variable definitions: never smoker, current smoker, and former smoker. Variable manipulation was conducted to format smoking categories into the gross categories of current smokers and never smokers. The former smoker category was established for the purposes of distinguishing smoking status and avoiding misclassification.

Genotype Quality Control and Imputation

dbGaP studies have been genotyped on a broad set of genotyping arrays and platforms, creating heterogeneity in terms of genetic data across studies in terms of both total variants and genetic coverage. Moreover, dbGaP does not require or employ a systematic methodology for standardizing and quality controlling (QC) genetic data. A small fraction of dbGaP studies provide quality-controlled genetic data, but unfortunately, each study performs their own unique QC, resulting in heterogeneity between studies. To address this issue, the Salem lab developed and applied a stringent genotype quality control protocol for raw genetic data from each dbGaP study. In brief, the quality control protocol included variant call and subject call rate filters, test of heterozygosity, Hardy-Weinberg Equilibrium, allele frequency checks, strand checks and assignment of standardized variant ID. The QC protocol utilized custom UNIX and R code, in addition to Plink, EIGENSTRAT, and KING program. The program, Plink, was used for the purpose of data manipulation of genotype files, while the program EIGENSTRAT was used for Principal Component Analysis, and KING was used to identify the unrelated subset of participants.

To expand genomic coverage and enable meta-analysis of results from individual studies, genotype imputation was performed for each dbGaP study. Genotype imputation is a statistical technique that leverages directly genotyped variants and a reference panel to infer ungenotyped variants. Imputation was performed by the Salem lab for each dbGaP study, stratified by genotyping array and ethnic/racial group using the NIH-funded Michigan Imputation server.[57] Prior to genotype imputation each dbGaP study had between 400k-1.5M variants, however, post imputation included ~43 million genetic variants. In addition to facilitating meta-analysis across studies, genotype imputation also results in increased power for GWAS.[58]

Genetic data from UK Biobank was genotyped on two highly similar arrays.[55] The majority of participants (*n*=438,427) were genotyped using the Applied Biosystems UK Biobank Axiom Array, and a subset of 49,950 participants were genotyped using Applied Biosystems UK BiLEVE Axiom Array.[55] To facilitate use of the UKBB resource by the research community, genotype QC and imputation were performed centrally by primary UKBB investigators. Prior to imputation, genetic data from two arrays was combined and a stringent QC procedure was performed.

Table 1. Type II Diabetes Categorization

Type 2 Diabetes	
Lab Tests/Metabolic Panel	1. Oral Glucose Tolerance Test (OGTT): ≥200 mg/dl
	2. Fasting Plasma Glucose (FPG): ≥ 126 mg/dl
Reported	3. Physician Diagnosis
	4. Self-Report
	5. Medication Report

Table 2. Smoking Status Definition

Smoking Status Definition		
Never Smokers	Self-reports never smokers	
	AND/OR	Not current smoker & pack-year <0.02
	AND/OR	Smoked <100 cigarettes in lifetime
Current Smokers	Self-reports current smoker	
	AND/OR	Smoked ≥ Past 1 month
	AND/OR	Smoked >100 cigarettes in lifetime
Former Smokers	Self-reports former-smokers	
	AND/OR	Not current smoker & pack-year ≥0.02
	AND/OR	Smoked ≥100 cigarettes in lifetime

Statistical Analyses

The primary aim of this study is to assess the relationship between gene x smoking interactions on T2D. We implemented a statistical model with a joint framework that entails a single regression model with both genetic main effects and gene by smoking interaction among the smokers vs. non-smokers. The statistical analysis structure requires a multiple logistic regression with interaction as shown in the following equation:

$$logit = \alpha + \beta_1 x_1 + B_2 x_2 + \beta_1 \times \beta_2$$

The exposure variables are smoking and SNP, and the outcome of interest is T2D. Potential confounding factors included in the model are gender, age, race, case-control status, and the first 10 principal component analysis (PCA) of GWAS data for each racial group. All analyses were performed by racial/ethnic group. Additionally, analyses were further stratified on gender, case-control status (if applicable), and genotyping array. To account for biased sample recruitment in case-controls studies, analyses were stratified by case-control stratum. The variables of age and principal component were adjusted for in the analysis process. The statistical model utilized for the purpose of this study is as follows:

$$T2D = \alpha + \beta_1 * SNP + \beta_2 * SMOKE + \beta_3 * SNP * SMOKE + COVARIATES$$

Genome-Wide Association Study (GWAS)

GWAS is a commonly used methodology in genetic epidemiology to analyze a genome-wide sets of variants for diseases or phenotypes of interest.[59] GWAS studies allow researchers to efficiently analyze millions of variants, the majority of which are single-nucleotide polymorphisms (SNP), across the genome against a complex trait or disease traits of interest.[59] Results of a GWAS identifies genetic variants, genes, and genomic regions that are associated with a disease of interest, and in turn, to provide insights into gene biology, disease pathophysiology, and guide pharmaceutical development.[28] In this study, GWAS was performed using SNPTEST, a program created by researchers at the University of Oxford.[60] SNPTEST is a computationally efficient program that implements a broad set of statistical methods commonly used in genetic epidemiology and GWAS analyses, including interaction analyses.[60] All GWAS analyses were all conducted on the NRNB cluster.

Meta-analysis

A meta-analysis study is a type of quantitative epidemiological study design that systematically leverages data from a variety of previous research studies to form analyses and derive conclusions from them.[61] Meta-analysis studies are a powerful methodology, which increase study power and precision, and provide a better association estimate within populations.[61,62] A limitation of meta-analysis is the challenge of dealing with heterogeneity among the different studies being analyzed, and the presence of potentially small sample sizes within some of the included studies.[62]

In this project, meta-analysis will be used to combine GWAS summary statistics from individual studies and subsets to produce one set of GWAS summary statistic for each of the 4 racial groups (European ancestry, African Ancestry, Hispanic Ancestry, and Asian Ancestry). Stratification by race is commonly utilized in genetic epidemiology studies to account for differences in allele frequencies between populations, a significant source of bias in genetic studies., Moreover, race/ethnic group stratification minimizes differences in disease prevalence, environmental and lifestyle factors, such as diet and lifestyle factors within the group. METAL is a widely used meta-analysis program in genetic epidemiology and has been optimized to be computationally efficient meta-analyses of dozens across millions of genetic variants common in genetic studies.[63] This program uses a fixed effects model to perform an inverse variance-weighted average meta-analysis.[63,64] Prior to conducting the meta-analysis, SNPTEST GWAS outputs for individual chromosomes are merged together and a QC protocol is applied to summary statistics. The summary statistics QC protocol involves checks for null values, removing implausible beta statistics and p-values, and application of a minor allele count (MAC) 10 filter. The MAC 10 filter is used to specify that the cases and controls each have a minimum of 10 copies of each allele, which has been shown to reduce false positives and p-value inflation. Finally, the QC protocol applies a genomic control, to correct for inflation in GWAS summary statistics due to potential sample admixture. Genomic control estimates inflation in GWAS summary statistics by calculating the lambda (λ), and the median chi-square value for all SNPs in the GWAS. When the study sample and number of variants is large, λ follows a chi-square distribution ($\chi^2{}_1$i with 1 degree of freedom). λ values greater than 1 indicate inflation and a correction is applied to all variant test statistics (deflation factor).[65] Meta-analysis is performed on QC'ed GWAS summary statistics by running through METAL stratified by race, and generate combined p-values, odds-ratios, allele frequency, and test of heterogeneity. Subsequently, a filter is applied to the meta-analysis output which requires SNP results to be derived from at least two studies. Finally, to identify independent regions and variants associated with the outcome of interest, the meta-analysis results are clumped. 'Clumping' is a statistical procedure in which GWAS results are filtered to identify top independent signals in a region, while taking into account the linkage disequilibrium (correlation) between variants. This involves identification of the most significant SNP for each haplotype block based on a specified p-value threshold. Clumping was performed considering two p-value thresholds, genome-wide significant ($5x10^{-8}$) and genome-wide suggestive ($1x10^{-5}$).

Chapter III:

Results

Study Datasets

The sample includes 345,955 individuals across 14 studies and contains 324,834 individuals of European ancestry, 9,040 of African ancestry, 6,125 of Hispanic ancestry, and 5,956 of Asian ancestry. The study includes 40,994 individuals with T2D and 301,940 without T2D, as well as 34,764 current smokers and 184,060 never smokers. The sex distribution of the sample involves 156,212 males and 189,743 females. Approximately 65% of the sample population is derived from UK Biobank, while the rest is from dbGaP. The different study types from the sample consisted of cross-sectional, case-control, case set, and longitudinal studies. Table 3 is a breakdown of the demographics for each study by racial ancestry. Table 4 provides the study population and encompasses a list of the studies, the study label, and total number of participants in each study.

Table 3. Study Characteristics

Ancestry	N	Gender		Type 2 Diabetes (*n*)		Smoking (*n*)	
		Male (*n*)	Female (*n*)	Non-Diabetic	Diabetic	Never	Current
European	324834	147465 (45%)	177369 (54%)	286442 (88%)	35382 (11%)	171443 (52%)	32230 (10%)
African	9040	3746 (41%)	5294 (58%)	5987 (66%)	3044 (33%)	4724 (52%)	1902 (21%)
Hispanic	6125	2516 (41%)	3609 (59%)	4603 (75%)	1521 (35%)	3595 (58%)	430 (7%)
Asian	5956	2485 (41%)	3471 (58%)	4908 (82%)	1047 (17%)	4298 (72%)	202 (3%)
Total	345955	156212 (45%)	189743 (55%)	301940 (88%)	40994 (12%)	184060 (53%)	34764 (10%)

Table 4. Study Population

Study Type	Study	Race	N	Male(*n*)	Female(*n*)	Type 2 Diabetes(*n*)		Smoking (*n*)	
						Non-Diabetic	Diabetic	Never	Current
Cross-Sectional	The National Institute on Aging Long Life Study (LLFS)								
		European	1800	842	958	1557	240	1189	137
Cross-Sectional	Research Program on Genes, Environment and Health (RPGEH)								
		European	62,827	25,826	37,001	54,987	7,840	34,039	2,668
		Black	1931	711	1,220	1470	461	1061	143
		Asian	5193	2,111	3,082	4,348	845	3,727	159
		Hispanic	4352	1,577	2,775	3,558	794	2,648	219
Case Control	Geneva Diabetes Study (GENEVA)								
		European	5552	2,366	3,186	2,929	2,499	2,488	574
Case Control	Genome-Wie Association Study on Cataract and HDL (CATARACT)								
		European	2177	891	1,286	1,473	704	1,092	188
Case Control	Northwestern Nugene project: Type 2 Diabetes (NW)								
		European	1288	640	648	623	604	566	102
Case Control	Development and Use of Network Infrastructure for GWAS (DUNI)								
		European	1438	594	844	1,121	317	607	58
Case Control	GWAS of Peripheral Artery Disease (PAD)								
		European	3048	1,900	1,148	2,537	511	713	1,194
Case set	Catheterization Genetics (CATHGEN)								
		European	882	456	426	663	219	430	452
		Black	270	129	141	162	108	148	122
Case set	National Institute of Diabetes and Digestive Kidney Diseases (NIDDK)								
		European	1517	914	603	906	611	666	141
		Black	1451	704	747	701	750	619	283
		Hispanic	399	242	157	124	275	211	26
Longitudinal	Atherosclerosis Risk in Communities (ARIC)								
		European	7885	3728	4157	6087	1798	3118	1969
		Black	2252	870	1382	1348	904	1032	672
Longitudinal	Coronary Artery Risk Development in Young Adults Study (CARDIA)								
		European	1343	635	708	1,236	105	766	332
		Black	1004	393	611	892	106	906	301
Longitudinal	Cardiovascular Health Study (CHS)								
		European	3192	1268	1924	2,224	968	1511	339
		Black	562	209	353	350	211	256	88
Longitudinal	Multi-Ethnic Study of Atherosclerosis (MESA)								
		European	2457	1,186	1,271	2,091	362	1,086	278
		Black	1570	730	840	1,064	504	702	293
		Hispanic	1374	697	677	921	452	736	185
		Asian	763	374	389	560	202	571	43
Longitudinal	UK Biobank (UKBB)								
		European	229428	106,219	123,209	208,008	18,604	123,172	23,798
Total			345955	156212	189743	301940	40994	184060	34764

GWAS Analyses and Results

GWAS analyses was first conducted individually for each study, using the joint framework to analyze both genetic main and interaction effects. Studies were stratified by race, gender, case control status, and genotype array, while also adjusted for PCA's and age, to obtain an unbiased estimate. Studies with familial relatedness were broken up to create an unrelated subset based on a specified pihat value dependent on each study. Once studies were individually analyzed, a meta-analysis was performed to combine study statistics and produce one set of results for each racial ancestry (European, African, Hispanic, Asian). The final meta-analysis results generated a total sample size of 345,955 individuals. Top SNPs from the meta-analysis results are provided stratified by racial ancestry as well as the full list of SNP results. The top SNPs are chosen based on a p-value cutoff (≤ 0.05), SNPs associated T2D established through prior literature, presence of the SNP in two or more racial ancestries at a p-value of ≤ 0.05, consistency in direction of effect, and degree of heterogeneity.

Results are presented separately for interaction and main effect analyses, including quantile-quantile plots (QQ) and Manhattan plots. A Manhattan plot illustrates the p-value of SNPs on a negative log scale for the y-axis and SNP genome position (by chromosome and position) on the x-axis. The Manhattan plot is used to visually identify genomic position with significance and problematic variant signals or regions. QQ plots are a useful tool to graphically compare the observed versus the expected p-value distributions. The QQ plots use a negative logarithmic scale (for both x- and y-axes), annotated with λ (inflation factor), and a 95% confidence interval to help identify problematic results (e.g. severe inflation) and studies that require further investigation. Clumping was performed using two p-value thresholds: genome-wide significant threshold (p-value $<5\text{x}10^{-8}$) and genome wide suggestive variants (p-value

<$1\text{x}10^{-5}$). Moreover, variants were organized by minor allele frequency (MAF) into two categories, common and uncommon SNPs. SNPs with an MAF <1% were categorized as uncommon variants. To reduce potential false positives, particularly for smaller non-European results, genome-wide suggestive variants were prioritized for consideration if they were nominally significant (p-value <0.05) in another racial group result. Finally, variants in the uncommon genome-wide suggestive category are not presented in the results or discussion sections below due to high probability of being false positives.

European Ancestry: Interaction Effect

The Manhattan plots for the European interaction effect (Figure 1) shows several SNPs that have reached genome-wide suggestive level, along with three positions on chromosomes 2, 3, and 8, that are close to genome-wide significance level, which may have been possible to attain with a larger sample size. The European interaction QQ plot (Figure 2) displays a line that indicates the observed values have shown to follow within the expected range within the 95% confidence interval.

Clumping of the European ancestry interaction effect results revealed zero variants at the genome-wide significant threshold (p-value <$5\text{x}10^{-8}$) and 55 genome-wide suggestive SNPs ($1\text{x}10^{-5}$). Of the suggestive significance loci, 43 were common variants and 3 had significance (p-value <0.05) in another racial ancestry group. The full list of common SNPs are provided in Table 7 and top SNPs are provided in Table 5. Two of the SNPs are on an intergenic region, the first of which is the rs6826172 on chromosome 4, and the second is the rs10915300 on chromosome 1. The third SNP is rs261227 on chromosome 5 and is on the *LOC101929710* gene with a p-value of 2.57×10^{-6}. The *LOC101929710* gene is uncharacterized and the function of this gene is currently unknown.[66]

European Ancestry: Main Effect

The Manhattan Plot for European Ancestry Main Effect (Figure 3) shows a highly significant p-value on chromosome 10, which is the *TCF7L2* gene. The Manhattan plots shows several other genomic markers have reached genome-wide significance, however, after filtering out uncommon SNPs, these markers were not considered in the final results. Figure 4 displays the QQ plot for the European Ancestry main effect. The QQ plot is graphed on a negative logarithmic scale and shows the observed outcome (the plotted black line) as higher than the expected range or the 95% confidence interval (the two red lines). The λ value (1.21) is greater than 1, which indicates that the observed p-values are more significant than expected and slight genomic inflation.

The meta-analyzed GWAS on individuals of European ancestry identified 34 genome-wide significant SNPs from the main effect (p-value $<5 \times 10^{-8}$). Results from main effect also displayed varying levels of frequencies, similar to the interaction effect, and as a result the uncommon SNPs were filtered out. Once filtered and categorized, the only SNP that remained as a common significant variant is the rs7903146 on chromosome 10, which has a p-value of 9.59×10^{-33}. This SNP is located on the *TCF7L2* gene, an established gene known for being involved in the pathophysiology of causing T2D.[22,67] The transcription factor-7 like two gene (*TCF7L2*) is widely proven in literature as a gene with high impact on disease susceptibility since it effects sensitivity of β-cell to incretins, which are metabolic hormones that induce a decrease in blood glucose.[67] Results on the top SNPs for the main effect are provided in Table 8.

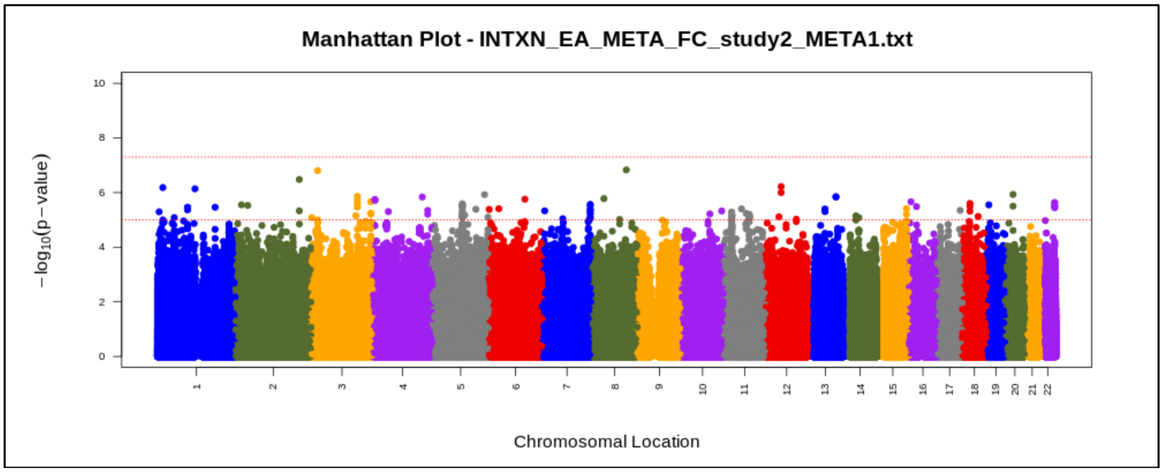

Figure 1. Manhattan Plot: European Ancestry Interaction Effect. Figure 1 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the interaction effect from the meta-analysis GWAS of Europeans. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold. The plot indicates several positions that have reached genome-wide suggestive level. The color sequence is shown to visually differentiate chromosome locations.

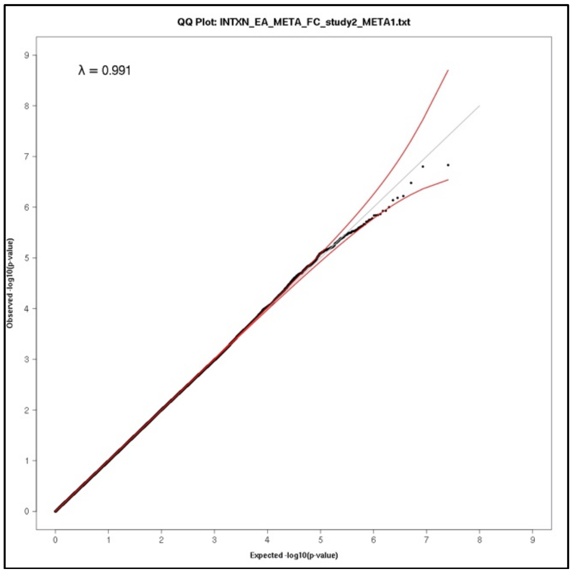

Figure 2. QQ Plot: European Ancestry Interaction Effect. Figure 2 is a quantile-quantile (QQ) plot for interaction effect of individuals from European ancestry. This QQ plot provides a graphical method to assess the distribution of GWAS p-values. The figure plots observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value.

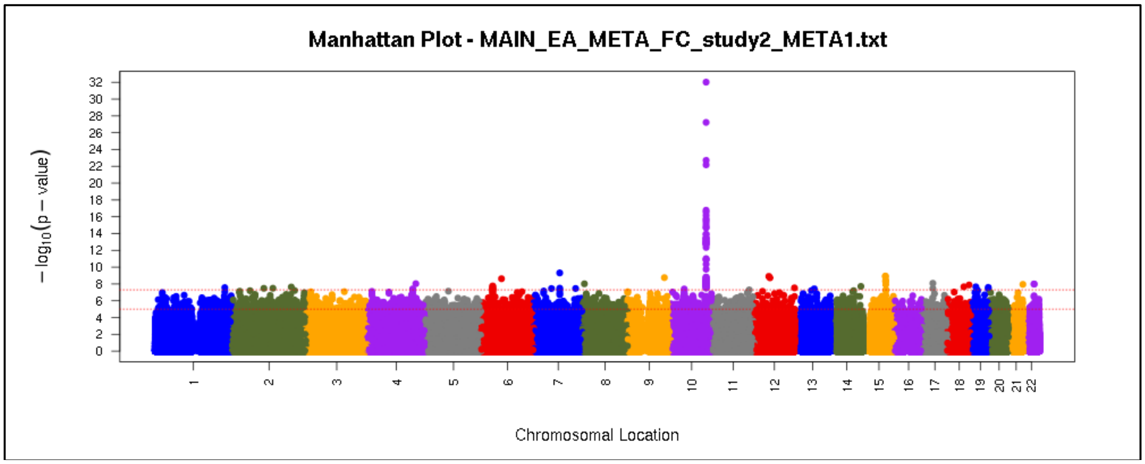

Figure 3. Manhattan Plot: European Ancestry Main Effect. Figure 3 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the main effect from the meta-analysis GWAS of Europeans. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold. The plot displays 36 positions with genome-wide significance ($5x10^{-8}$) with the highest p-value pertaining to rs7903146 on chromosome 10.

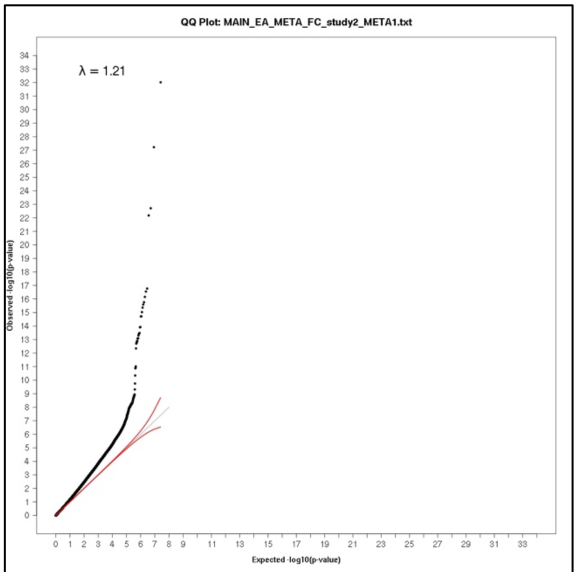

Figure 4. QQ Plot: European Ancestry Main Effect. Figure 4 is the quantile-quantile (QQ) plot for the European ancestry main effect. The figure plots observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value. This figure displays the observed value (plotted black line) as higher than the expected value.

African Ancestry: Interaction Effect

Figure 5 is the Manhattan plot for the African ancestry interaction effect illustrating several markers at the genome-wide suggestive level. The QQ plot is presented in Figure 6, which shows the observed outcome well below the expected range with a λ less than one (λ=0.946), indicating the sample is underpowered due to modest overall sample size in African ancestry studies.

Interaction results from the African ancestry analysis generated a list of zero genome-wide significant (p-value of 5x10^{-8}) SNPs, and 18 SNPs at the genome-wide suggestive level (p-value of 1x10^{-5}). The 18 SNPs still remained after being filtered for an MAF ≥1% for common SNPs (Table 9). Since the SNP results are genome-wide suggestive, the 18 SNPs were further examined for significance across at least one other racial ancestry. Only 1 SNP showed significance at a p-value threshold of <0.05 among the African and Asian ancestries. The SNP is rs13100451on chromosome 3, with a p-value of 6.45x10^{-6} and lies on the *ST6GAL1* gene. The *ST6GAL1* gene, known as ST6 beta-galactoside alpha-2 6-sialyltransferase 1, is a type II membrane protein that catalyzes the transfer of sialic acid to galactose-containing substrates.[68] Rudman et al. described the *ST6GAL1* gene as a novel candidate risk gene for T2D from prior GWAS studies published on the risk associated with individuals of European and South Asian ancestry.[69]

African Ancestry: Main Effect

Figure 7 is the Manhattan plot for the African Ancestry main effect illustrating several markers that have reached the genome-wide suggestive level, along with three positions on chromosomes 7 and 14 that are quite close to the genome-wide significant p-value of 5x10^{-8}.

Similar to the interaction effect, the QQ plot (Figure 8) reveals a lower distribution of p-values than expected, with a λ less than 1 (λ=0.977). This is likely reflecting lack of power due to modest overall sample size in African ancestry studies.

Results from the main effect analysis of African ancestry identified zero genome-wide significant (p-value of $5x10^{-8}$) SNPs and 37 SNPs at a genome-wide suggestive level (p-value of $1x10^{-5}$). All 37 SNPs were common and reported in Table 10, similar to the African ancestry interaction effect results. Since the SNP results were genome-wide suggestive, the 37 SNPs were filtered based on significance (p-value of <0.05) in at least one other racial ancestry from the GWAS analysis, leaving 3 SNPs. The first SNP, rs73337298 is located on chromosome 7, in an intergenic region. The second SNP is rs9866900 on chromosome 3, with a p-value of $1.97x10^{-6}$ and is on the *NAALADL2* gene. The *NAALADL2* gene was significant in the African ancestry, Hispanic ancestry, and Asian ancestry results. The *NAALADL2* gene, otherwise known as N-acetylated alpha-linked acidic dipeptidase like 2, is part of the glutamate carboxypeptidase II family.[70] A study by Berndt et al. found this gene has been associated with prostate cancer aggressiveness and is related to the prostate-specific membrane antigen, a diagnostic and drug target for prostate cancer.[70] The *NAALADL2* gene has also been reported in a study by Zhang et al. that conducted a quantitative trait locus association analyses on influencing pleiotropy of Metabolic Syndrome.[71] *NAALADL2* gene mutation is associated to visceral fat and insulin responsiveness in the metabolic syndrome association study.[71] The third SNP is rs16954017 on chromosome 15, with a p-value of $8.21x10^{-6}$ and is on the *TMED3* gene. The *TMED3* variant was significant in the African ancestry and Asian ancestry results. The *TMED3* gene, also known as transmembrane p24 trafficking protein 3, has primarily been discovered to be a potential drug target in metastatic suppressor in colon cancer.[72,73] However, a study conducted by Hall et al. on the effects of high glucose exposure on global gene expression and DNA methylation in human pancreatic islets, found that the *TMED3* gene showed differences in mRNA expression between high glucose treated pancreatic islets and control treated islets in the pancreas.[74]

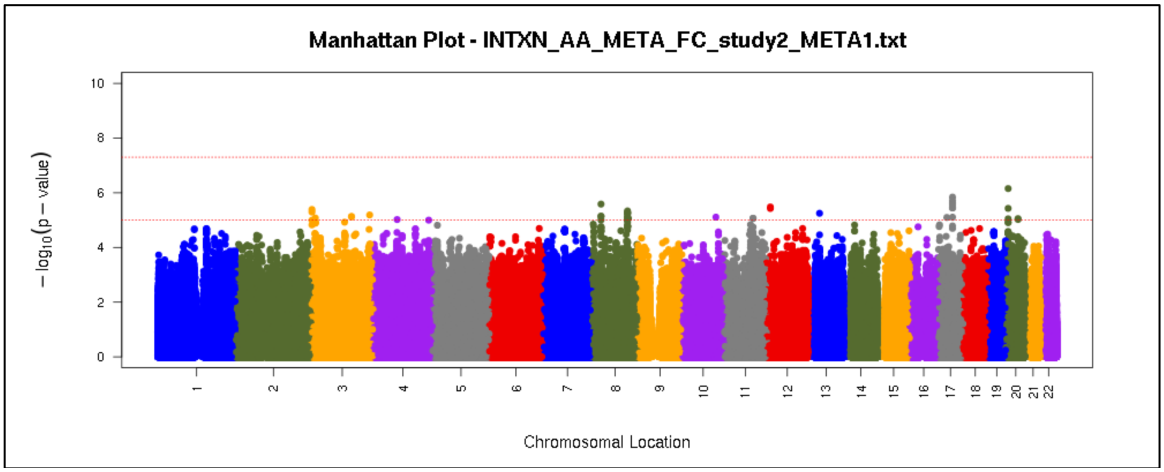

Figure 5. Manhattan Plot: African Ancestry Interaction Effect. Figure 5 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the interaction effect from the meta-analysis GWAS of Africans. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold. The graph shows 18 SNPs that have reached a genome-wide suggestive level (1x10^{-5}).

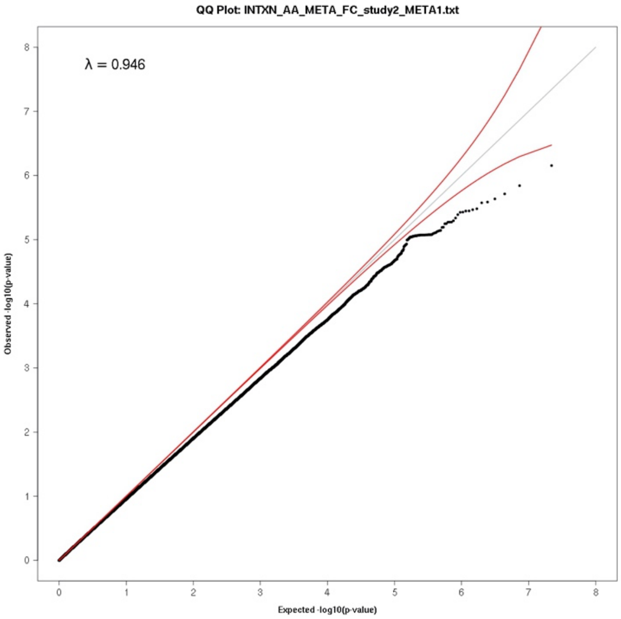

Figure 6. QQ Plot: African Ancestry Interaction Effect. Figure 6 is a quantile-quantile (QQ) plot of the interaction effect from individuals of African ancestry. The plot provides a graphical representation of observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value.

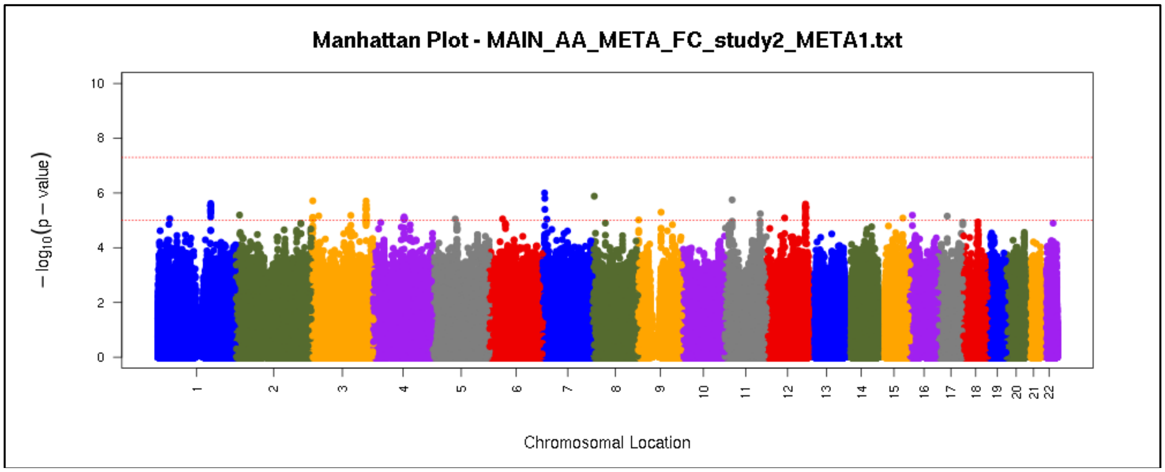

Figure 7. Manhattan Plot: African Ancestry Main Effect. Figure 7 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the main effect from the meta-analysis GWAS of Africans. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold. The graph shows 37 SNPs that have reached a genome-wide suggestive level.

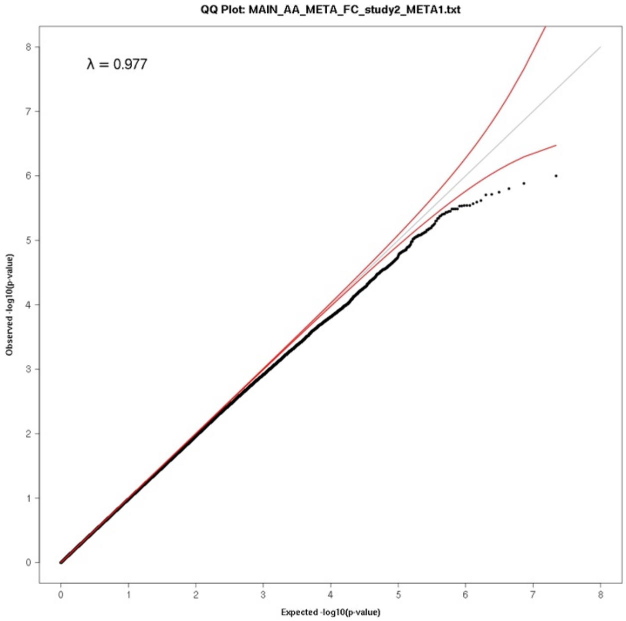

Figure 8. QQ Plot: African Ancestry Main Effect. Figure 8 is a quantile-quantile (QQ) plot of the main effect from individuals of African ancestry. The plot provides a graphical representation of observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value.

Hispanic Ancestry: Interaction Effect

The Manhattan plot for the Hispanic Ancestry interaction effect (Figure 9) displays several SNPs at the genome-wide suggestive level, along with three particular positions on chromosomes 7 and 14 that are quite close to the genome-wide significance level, which may have been attained given a larger sample size. Figure 10 provides the QQ plot for the interaction effect, which shows the observed line is slightly lower than the expected line, with a λ below 1 (λ=0.937). The figure displays variability towards the top end of the graph from the rest of the line.

Results from the meta-analyzed GWAS of Hispanic ancestry for the interaction effect found zero genome-wide significant (p-value of $5x10^{-8}$) SNPs and 5 genome-wide suggestive SNPs (p-value of $1x10^{-5}$). All 5 SNPs are common, with an MAF ≥1% (see Table 11). Since results are genome-wide suggestive, the SNPs were filtered based on significance (p-value <0.05) in at least one other racial group from the GWAS analysis, resulting in 1 SNP remaining. The SNP is rs73209286 on chromosome 8, with a p-value of $6.14x10^{-6}$ and is on the *BLK* gene, which was significant in Hispanic and European ancestry studies. *BLK*, or B lymphocyte kinase, is a gene that encodes a nonreceptor tyrosine-kinase among the family of proto-oncogenes involved in cell proliferation and differentiation.[75] The function of this protein in B-cell development and stimulates insulin synthesis and secretion in response to glucose, while it also enhances the expression of pancreatic beta-cell transcription factors.[75] A study conducted by Borowiec et al. in 2009 discovered mutations on the *BLK* locus is associated with mature onset diabetes of the young and described *BLK* as a modulator of beta-cell function.[76]

Hispanic Ancestry: Main Effect

The Manhattan Plot for Hispanic Ancestry main effect (Figure 10) shows several SNPs that have reached past the genome-wide suggestive p-value cut-off value on the negative logarithmic scale. The QQ plot for the main effect (Figure 12) shows the observed values are mostly within the expected range, with a λ of a slightly over 1 (λ=1.015).

The meta-analyzed GWAS on individuals of Hispanic ancestry identified zero genome-wide significant (p-value of $5x10^{-8}$) SNPs and 26 genome-wide suggestive SNPs (p-value of $1x10^{-5}$) for the main effect, all of which were common and provided in Table 12. Since results are genome-wide suggestive, SNPs were filtered based on significance (p-value <0.05) in at least one other racial ancestry from GWAS analysis, and 3 SNPs remained. Two of the SNPs are on an intergenic region: the rs113240724 SNP on chromosome 2, and the rs2416722 SNP on chromosome 9. The third SNP is rs11264442 on chromosome 1, with a p-value of $6.12x10^{-6}$ and falls on the *LMNA* gene. The *LMNA* indicated to be significant among Hispanic and European ancestral studies. The *LMNA* gene, known as lamin A/C, provide instructions for the proteins lamin A and lamin C, which are intermediate filaments that provide cells with stability and strength.[77] Two studies have investigated the role of *LMNA* gene mutations in increasing susceptibility to T2D and found an association with disease risk.[78,79]

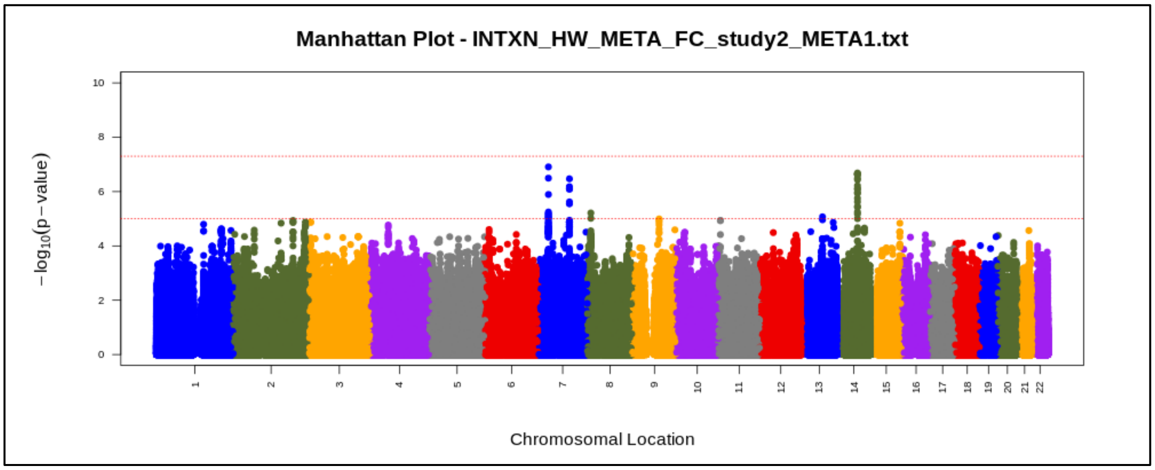

Figure 9. Manhattan Plot: Hispanic Ancestry Interaction Effect. Figure 9 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the interaction effect from the meta-analysis GWAS of individuals with Hispanic ancestry. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold.

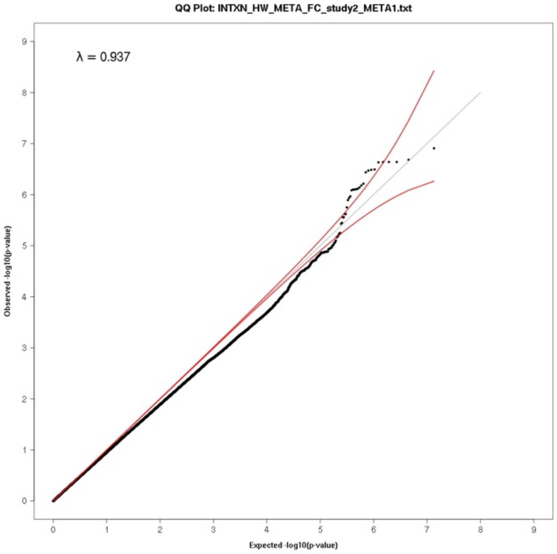

Figure 10. QQ Plot: Hispanic Ancestry Interaction Effect. Figure 10 is a quantile-quantile (QQ) plot of the interaction effect from individuals of Hispanic ancestry. The plot provides a graphical representation of observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value.

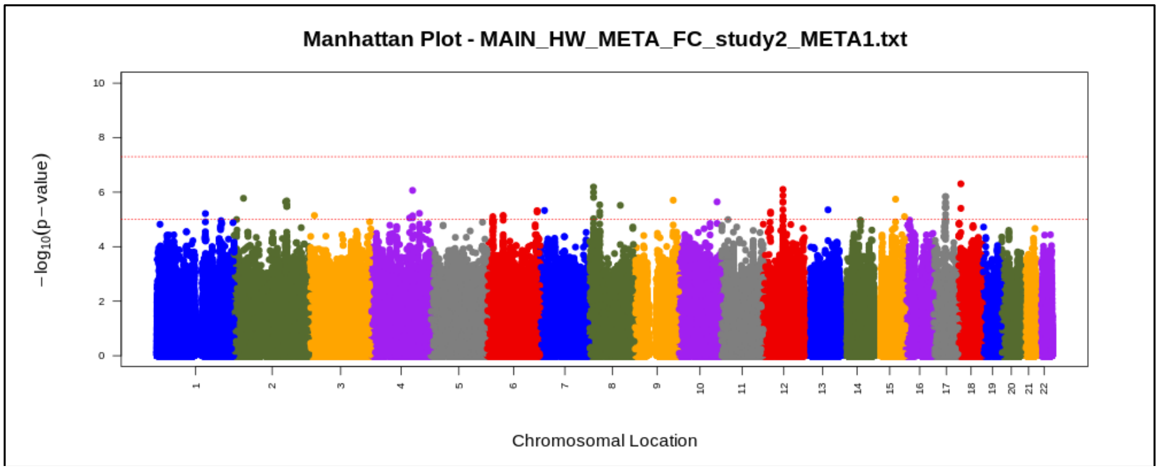

Figure 11. Manhattan Plot: Hispanic Ancestry Main Effect. Figure 11 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the main effect from the meta-analysis GWAS of Hispanic individuals. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold. The figure shows 26 significant SNPs that are at the genome-wide suggestive level.

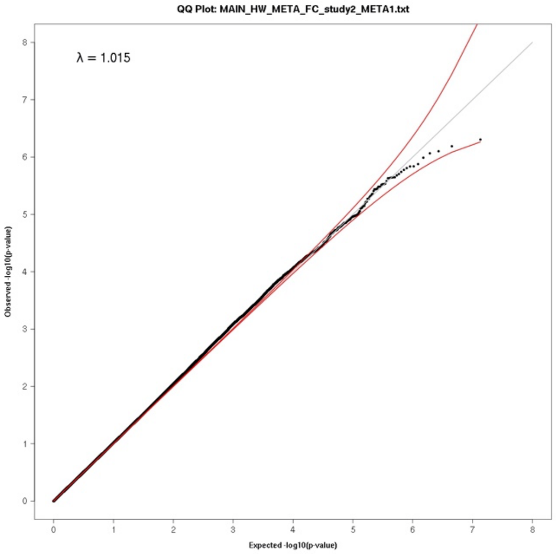

Figure 12. QQ Plot: Hispanic Ancestry Main Effect. Figure 12 is a quantile-quantile (QQ) plot of the interaction effect from individuals of Hispanic ancestry. The plot provides a graphical representation of observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value.

Asian Ancestry: Interaction Effect

The Manhattan Plot for the Asian ancestry interaction effect (Figure 13) shows several genetic markers with spectacularly low p-values on chromosomes 2, 3, 8, and 13, which do not reflect typical Manhattan plots that normally display a skyscraper shape on highly significant genomic positions, instead of single dots on the figure. The QQ plot (Figure 14) for the interaction effect of Asian Ancestry appears to follow a normal distribution except for the set of unexpected variants with extremely low p-values.

The GWAS interaction results on individuals of Asian ancestry found 6 genome-wide significant SNPs (p-value $\leq 5x10^{-8}$). The observed genome-wide significant p-values ($9.51x10^{-188}$, $2.67x10^{-38}$, $1.03x10^{-36}$, $7.27x10^{-30}$, $1.01x10^{-12}$, $1.08x10^{-11}$) are highly unexpected given the small sample size of Asian ancestry samples (n=5,956, smokers=1,047, T2D=202) in comparison to what was observed in other racial groups. We investigated the unexpectedly low variants by reviewing the study specific results and finding the results are mainly driven by two studies (MESA & RPGEH) with low sample sizes. We conclude the observed genome-wide significant loci are very likely false positives and drop them from further consideration or discussion. The results are reported in Table 5 as part of the top SNPs and in Table 13 for all common SNPs, to be fully comprehensive and transparent, however should be viewed with high skepticism.

Asian Ancestry: Main Effect

The Manhattan Plot for the Asian ancestry main effect (Figure 15), shows several genomic markers at genome-wide suggestive level. Markers on chromosomes 4, 5, and 7 are particularly high and are close to the genome-wide significant cut-off. The QQ plot (Figure 16) displays a normal line with the observed values slightly higher than the expected near the top of the graph, while having a λ a little below one (λ=0.997).

The meta-analyzed GWAS on individuals of Asian ancestry found 23 genome-wide suggestive SNPs (p-value of $1x10^{-5}$) for the main effect, all of which were common (see Table 14). Since results are genome-wide suggestive, the 23 SNPs were filtered based on a significant p-value threshold (≤0.05) across at least two racial ancestries. As a result, 2 SNPs remained. The first SNP is rs78355386 on chromosome 9, which is on an intergenic region. The second SNP is rs1671407 on chromosome 8, with a p-value of $8.59x10^{-6}$ and is on the *DLC1* gene. The *DLC1* variant was significant in Hispanic and Asian ancestry results. The *DLC1* gene, known as *DLC1* (Deleted in Human Liver Cancer 1) Rho GTPase activating protein, is a member of the rhoGAP family, which are involved in regulation of GTP-binding proteins.[80,81] The function of this gene acts as a tumor suppressor in prevalent cancers such as prostate cancer, lung cancer, colorectal cancer, and breast cancer.[80] Interestingly, a GWAS by Matoba et al. of smoking behavior in 165,436 Japanese individuals found that the *DLC1* gene was associated with smoking initiation.[82]

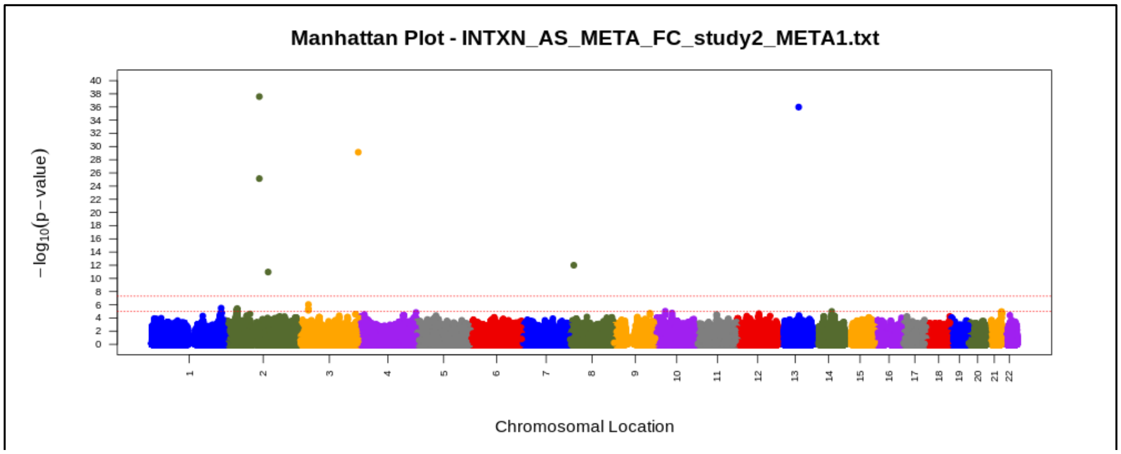

Figure 13. Manhattan Plot: Asian Ancestry Interaction Effect. Figure 13 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the interaction effect from the meta-analysis GWAS of Asian ancestry. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold.

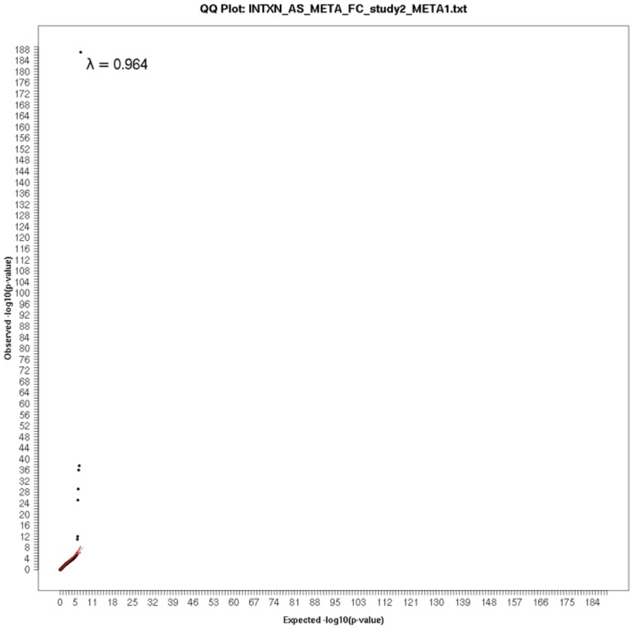

Figure 14. QQ Plot: Asian Ancestry Interaction Effect. Figure 14 is a quantile-quantile (QQ) plot of the interaction effect from individuals of Asian ancestry. The plot provides a graphical representation of observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value.

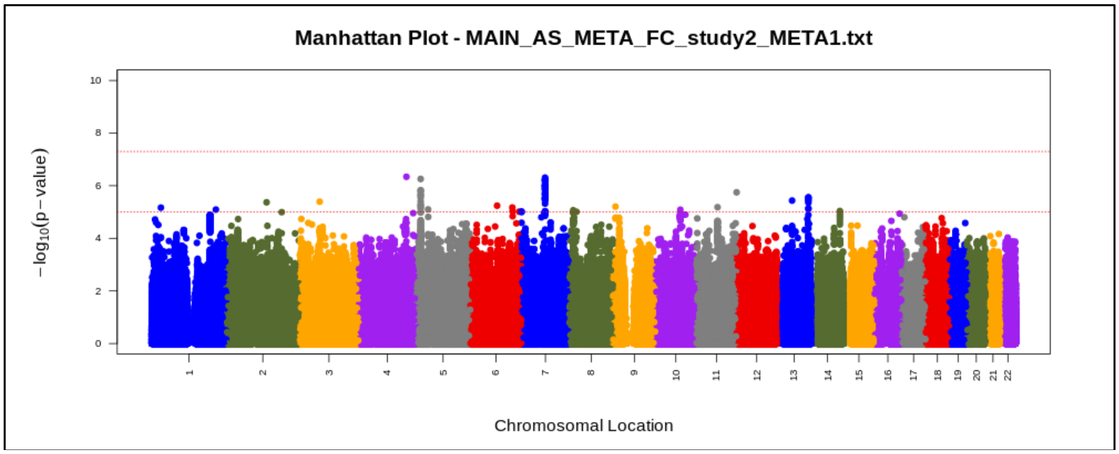

Figure 15. Manhattan Plot: Asian Ancestry Main Effect. Figure 15 illustrates a Manhattan plot that allows visual inspection of the distribution of p-values of the main effect from the meta-analysis GWAS of Asian ancestry. The x-axis represents the genomic position, and the y-axis represents a negative logarithmic scale of p-values. The top red dotted line is the genome-wide significant threshold while the bottom red dotted line is the genome-wide suggestive threshold.

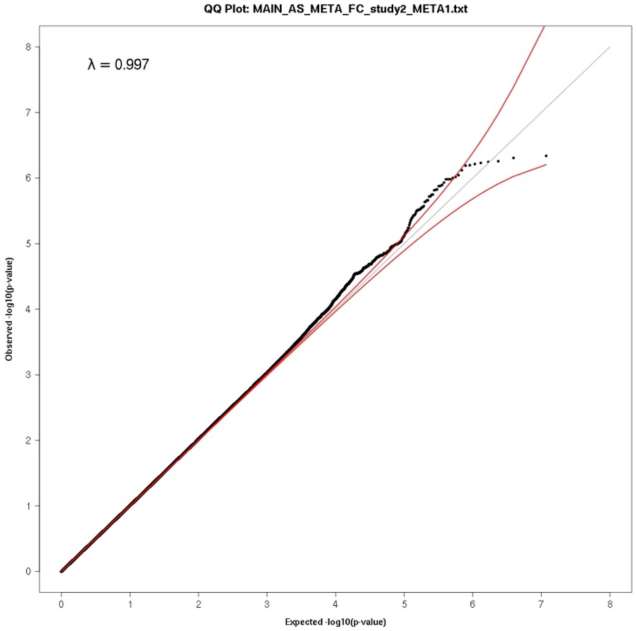

Figure 16. QQ Plot: Asian Ancestry Main Effect. Figure 16 is a quantile-quantile (QQ) plot of the interaction effect from individuals of Asian ancestry. The plot provides a graphical representation of observed p-values versus expected p-values on a negative log scale. The red lines represent the 95% confidence interval and the plotted black lines is the p-value.

Cross Ancestry Comparison

Meta-analysis SNP results from each racial ancestry were investigated for significance at a p-value threshold of <0.05 across the different racial group analyses to prioritize loci. Results from the genome-wide significant common SNPs outputs (European ancestry main effect and Asian ancestry interaction effect) were examined in European, African, Hispanic, and Asian ancestral analyses by genomic position for significance at a p-value threshold of 0.05 in at least two or more studies. The Asian ancestry interaction SNP results were not significant in other racial studies. The *TCF7L2* gene on chromosome 10 (rs7903146) from the European ancestry results is also nominally significant in individuals of African (p-value of 3.16×10^{-3}) and Hispanic (p-value of 1.25×10^{-2}) ancestries. The p-values of significant SNPs across different racial groups are provided in Table 15 for the interaction effect and Table 16 for the main effect. A breakdown of that SNP for each racial group is provided in Table 17 for the interaction effect and Table 18 for the main effect. The *TCF7L2* gene is the only genome-wide significant result shown to also be significant in other races.

Genome-wide suggestive common SNPs were evaluated for significance at a p-value of <0.05 cut-off in at least two racial studies, a total of 13 genes were identified in this cross-ancestry look-up. Six of the SNPs are on an intergenic region. The *ST6GAL1* gene, previously mentioned in the African ancestry interaction effect, is also significant in the Asian (p-value of 2.26×10^{-2}) GWAS results. The *NAALADL2* gene from the African ancestry results was significant in Hispanic ancestry results (p-value of 2.25×10^{-2}) and Asian ancestry results (p-value of 1.54×10^{-2}) as well. The *TMED3* gene from the African results showed significance in the Asian ancestry results (p-value of 3.73×10^{-3}). The *BLK* gene, discussed in Hispanic ancestry results, was significant in European ancestry results (p-value of 1.40×10^{-2}). The *LMNA* gene discussed in the Hispanic ancestry results was significant in individuals of European ancestry (p-value of 4.06x10^{-2}) ancestry. From the Asian ancestry results, the *DLC1* gene showed nominal significance in Hispanic ancestry results (p-value of 4.55x10^{-2}).

Table 5. Top SNPs from Interaction Effect

Top SNPs: Interaction Effect									
Ancestry	CHR	Marker	SNP	Gene	EA	NEA	Freq	P-value	OR
European	4	4:158480125	rs6826172	Intergenic	A	G	0.8768	1.46E-06	0.81
European	5	5:95902989	rs261227	LOC101929710	A	C	0.607	2.57E-06	1.15
European	1	1:5026921	rs10915300	Intergenic	A	G	0.2369	5.06E-06	1.17
African	3	3:186653218	rs13100451	ST6GAL1	A	T	0.4341	6.45E-06	1.61
Hispanic	8	8:11374228	rs73209286	BLK	A	C	0.0445	6.14E-06	4.51
Asian	2	2:108747865	rs9917181	Intergenic	T	C	0.6197	2.67E-38	26.47
Asian	2	2:138093273	rs1463279	THSD7B	T	C	0.7018	1.08E-11	0.18
Asian	3	3:194772806	rs4677798	Intergenic	A	G	0.7266	7.27E-30	21.04
Asian	8	8:21562155	rs7826525	GFRA2	A	G	0.5542	9.51E-188	0.01
Asian	8	8:12515242	rs62488764	LOC729732	T	C	0.5243	1.01E-12	0.07
Asian	13	13:69233302	rs12871979	Intergenic	A	G	0.7464	1.03E-36	0.06

*Abbreviations: CHR: chromosome, SNP: single-nucleotide polymorphism, EA: effect allele, NEA: non-effect allele, Freq: frequency, OR: odds ratio.

Table 6. Top SNPs from Main Effect.

Top SNPs: Main Effect									
Ancestry	CHR	Marker	SNP	Gene	EA	NEA	Freq	P-value	OR
European	10	10:114758349	rs7903146	TCF7L2	T	C	0.2934	9.59E-33	1.29
African	3	3:174808476	rs9866900	NAALADL2	A	T	0.1713	1.97E-06	1.44
African	15	15:79630253	rs16954017	TMED3	A	T	0.6133	8.21E-06	1.30
African	7	7:6328946	rs73337298	Intergenic	A	G	0.1071	1.58E-06	1.58
Hispanic	1	1:156104375	rs11264442	LMNA	A	G	0.0357	6.12E-06	2.54
Hispanic	9	9:122590178	rs2416722	Intergenic	A	T	0.0991	1.97E-06	1.75
Hispanic	2	2:164851254	rs113240724	Intergenic	T	C	0.9679	2.25E-06	0.34
Asian	8	8:13222219	rs1671407	DLC1	A	T	0.2374	8.59E-06	1.43
Asian	9	9:7686600	rs78355386	Intergenic	A	T	0.9615	6.23E-06	0.35

*Abbreviations: CHR: chromosome, POS: genomic position, SNP: single-nucleotide polymorphism, EA: effect allele, NEA: non-effect allele, Freq: frequency, OR: odds ratio.

Table 7. European Ancestry Interaction Effect Common Variants

European Ancestry Interaction Effect: Genome-Wide Suggestive & Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
1	1:16806683	rs7520744	CROCCP3	A	G	0.1191	6.55E-07	1.26
1	1:184508413	rs144420111	C1orf21	A	G	0.9895	3.44E-06	0.48
1	1:96126649	rs11165462	LOC101928219	T	C	0.4119	3.47E-06	1.15
1	1:5026921	rs10915300	Intergenic	A	G	0.2369	5.06E-06	1.17
1	1:53263826	rs4351588	ZYG11B	A	T	0.6273	8.29E-06	0.88
2	2:205366177	rs78262818	Intergenic	A	C	0.9738	3.32E-07	0.63
2	2:19952783	rs11692297	Intergenic	T	C	0.1214	2.81E-06	1.23
2	2:39911151	rs13427136	TMEM178A	T	C	0.9329	2.95E-06	0.63
3	3:20754619	rs62234997	Intergenic	A	G	0.7406	1.58E-07	0.84
3	3:148353411	rs1602698	Intergenic	C	G	0.8761	1.37E-06	0.80
3	3:191405080	rs13323801	Intergenic	A	T	0.0987	2.17E-06	1.32
3	3:2211932	rs6789159	CNTN4	A	G	0.3081	8.27E-06	1.19
4	4:158480125	rs6826172	Intergenic	A	G	0.8768	1.46E-06	0.81
4	4:6751596	rs75775756	Intergenic	C	G	0.9138	1.78E-06	0.78
4	4:49642993	rs10452406	Intergenic	T	G	0.3331	4.98E-06	1.25
5	5:167079454	rs112850392	TENM2	A	G	0.9873	1.19E-06	0.52
5	5:95902989	rs261227	LOC101929710	A	C	0.607	2.57E-06	1.15
5	5:139066854	rs34897167	Intergenic	A	C	0.1103	4.09E-06	1.26
5	5:103679062	rs75368393	Intergenic	A	G	0.9598	6.75E-06	0.72
5	5:178088566	rs7730583	Intergenic	A	G	0.0721	7.91E-06	1.35
6	6:115574669	rs368832345	Intergenic	T	C	0.9667	1.76E-06	0.63
6	6:32087258	rs62402721	ATF6B	T	C	0.0253	3.89E-06	2.25
7	7:154206637	rs78773933	DPP6	T	G	0.8859	2.70E-06	0.80
8	8:38845890	rs149162978	HTRA4/TM2D2	T	C	0.0774	1.66E-06	1.30
8	8:89239596	rs72675183	MMP16	T	C	0.0417	9.56E-06	1.40
10	10:129100209	rs985872861	DOCK1	CA	C	0.0278	4.70E-06	1.73
11	11:57211797	rs150617079	Intergenic	A	C	0.0147	3.95E-06	1.76
11	11:25398182	rs2404078	Intergenic	A	T	0.2484	5.20E-06	0.86
11	11:72407884	rs2306613	ARAP1	T	C	0.9036	5.82E-06	0.80

Table 7. European Ancestry Interaction Effect Common Variants (*continued)*

European Ancestry Interaction Effect: Genome-Wide Suggestive & Common Variants *(continued)*								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
11	11:81509968	rs1903240	Intergenic	A	C	0.1445	6.33E-06	1.26
11	11:83524658	rs12362201	DLG2	T	C	0.0582	8.30E-06	1.33
12	12:49312681	rs117646559	CCDC65	T	G	0.0118	6.03E-07	4.22
12	12:42298802	rs1328324106	Intergenic	T	ACTT	0.3639	7.72E-06	0.83
13	13:55868470	rs77608207	Intergenic	T	G	0.9722	4.01E-06	1.64
14	14:40046061	rs75360298	LOC105370461	A	C	0.9766	7.11E-06	0.63
14	14:49390161	rs72690222	LOC105378178	A	T	0.6956	8.26E-06	0.87
15	15:94945704	rs7180682	MCTP2	A	G	0.5982	3.88E-06	0.87
16	16:24685446	rs80237910	TNRC6A	A	C	0.0245	3.25E-06	1.58
17	17:74311627	*	*	A	ACTT	0.0112	4.49E-06	2.03
18	18:24592493	rs62082096	CHST9	A	C	0.0167	2.51E-06	1.68
18	18:50887302	rs11663173	DCC	T	C	0.0108	7.50E-06	5.02
19	19:7030780	rs1690412	MBD3L5	A	C	0.7336	2.82E-06	0.81
22	22:47944748	rs10460765	Intergenic	T	C	0.7144	2.33E-06	0.86

*Position not affiliated with any SNP or gene.

Table 8. European Ancestry Main Effect Common Variants

European Ancestry Main Effect: Genome-Wide Significant and Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
10	10:114758349	rs7903146	TCF7L2	T	C	0.2934	9.59E-33	1.29

Table 9. African Ancestry Interaction Common Variants

African Ancestry Interaction Effect: Genome-Wide Suggestive & Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
3	3:1669255	rs280127	Intergenic	A	G	0.802	4.08E-06	0.55
3	3:186653218	rs13100451	ST6GAL1	A	T	0.4341	6.45E-06	1.61
3	3:128540990	rs111882733	Intergenic	A	G	0.7501	7.37E-06	1.73
3	3:12532057	rs299640	TSEN2	T	C	0.3707	8.35E-06	1.59
4	4:76814405	rs113631320	PPEF2	A	T	0.061	9.49E-06	2.86
8	8:28002281	rs12542344	ELP3	A	C	0.1411	2.60E-06	0.46
8	8:112742197	rs1904365	Intergenic	A	T	0.7848	4.60E-06	1.78
10	10:108560037	rs822000	SORCS1	T	C	0.2959	7.77E-06	1.65
11	11:92163754	rs495762	FAT3	A	G	0.1231	8.55E-06	2.05
12	12:12575171	rs11054898	BORCS5	T	C	0.2173	3.30E-06	0.55
13	13:36401746	rs9574698	DCLK1	T	C	0.7143	5.62E-06	0.60
17	17:47227275	rs62079771	B4GALNT2	A	G	0.1364	1.44E-06	0.45
17	17:29009905	rs7222253	LOC105371723	C	G	0.4849	7.97E-06	0.63
20	20:7090913	rs77095026	Intergenic	A	C	0.0401	7.01E-07	4.67
20	20:7123709	rs115561333	Intergenic	T	C	0.972	3.72E-06	0.10
20	20:38633460	rs6071895	LINC01370	T	C	0.1745	8.81E-06	1.78
20	20:7092065	rs78069445	Intergenic	C	G	0.9589	8.97E-06	0.26

Table 10. African Ancestry Main Effect Common Variants

African Ancestry Main Effect: Genome-Wide Suggestive & Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
1	1:169502942	rs147838710	F5	A	G	0.0227	7.48E-06	4.55
1	1:169497628	rs76904241	F5	T	G	0.023	4.96E-06	4.50
1	1:169492236	rs114542453	F5	A	G	0.0232	2.88E-06	4.53
1	1:169491048	rs370366327	F5	A	G	0.0232	2.88E-06	4.53
1	1:169483175	rs76510731	F5	C	G	0.0232	3.90E-06	4.37
1	1:169494232	rs115191744	F5	A	G	0.0234	2.42E-06	4.61
1	1:169472898	rs115195224	Intergenic	A	C	0.0241	5.52E-06	3.95
1	1:38609717	rs74390957	LOC105378654	A	G	0.9375	8.73E-06	0.57
1	1:169475850	rs116773669	Intergenic	A	C	0.9762	2.93E-06	0.24
1	1:169482063	rs75764442	F5	C	G	0.9767	4.24E-06	0.23
1	1:169490119	rs77136555	F5	C	G	0.9768	2.95E-06	0.22
1	1:169495651	rs114407237	F5	T	C	0.9768	4.53E-06	0.22
2	2:12612540	rs10929814	MIR3681HG	A	G	0.5961	6.39E-06	0.74
3	3:125847878	rs78299669	ALDH1L1	A	T	0.0317	6.59E-06	2.95
3	3:3977058	rs116727036	Intergenic	A	T	0.0484	7.89E-06	1.91
3	3:3967931	rs75364484	Intergenic	T	G	0.0505	7.77E-06	1.89
3	3:174808476	rs9866900	NAALADL2	A	T	0.1713	1.97E-06	1.44
3	3:23146682	rs6783710	Intergenic	T	C	0.4196	6.88E-06	1.30
3	3:3995440	rs113610982	Intergenic	A	C	0.949	9.37E-06	0.54
3	3:3964094	rs79094344	Intergenic	A	G	0.9491	1.94E-06	0.51
4	4:98587845	rs17026853	STPG2	A	G	0.0932	8.73E-06	1.54
4	4:98654023	rs74807654	STPG2	A	G	0.8939	7.45E-06	0.66
5	5:71421903	rs115247276	MAP1B	T	G	0.9634	8.95E-06	0.37
6	6:42211424	rs9394890	TRERF1	T	C	0.9567	8.86E-06	0.37
7	7:5491335	rs56017129	Intergenic	A	G	0.0674	1.00E-06	2.00
7	7:6328946	rs73337298	Intergenic	A	G	0.1071	1.58E-06	1.58
7	7:12764983	rs4719329	Intergenic	T	C	0.5982	9.10E-06	1.31
8	8:5501607	rs2189887	Intergenic	T	G	0.0777	1.31E-06	1.75
9	9:1488655	rs10961710	Intergenic	A	C	0.0753	9.63E-06	0.57
9	9:73183023	rs12351121	KLF9-DT/TRPM3	T	C	0.8272	5.05E-06	1.43
11	11:114323972	rs548236	Intergenic	A	T	0.5487	5.73E-06	0.77
11	11:24100122	rs59465145	Intergenic	A	T	0.9304	1.79E-06	0.55
12	12:124202498	rs112325438	ATP6V0A2	A	G	0.0392	2.56E-06	2.28
12	12:57354514	rs11524050	RDH16	T	C	0.0478	8.17E-06	1.90
15	15:79630253	rs16954017	TMED3	A	T	0.6133	8.21E-06	1.30
16	16:7744600	rs2191133	RBFOX1	A	G	0.7628	6.48E-06	0.74
17	17:29010103	rs11653547	LOC105371723	T	C	0.6169	6.99E-06	0.76

Table 11. Hispanic Ancestry Interaction Effect Common Variants

Hispanic Ancestry Interaction Effect: Genome-Wide Suggestive & Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
7	7:33430008	rs3779241	BBS9	T	C	0.0889	1.24E-07	19.75
7	7:101358045	rs56195308	Intergenic	T	G	0.9266	3.38E-07	3.02
8	8:11374228	rs73209286	BLK	A	C	0.0445	6.14E-06	4.51
13	13:67823434	rs115242679	Intergenic	A	G	0.0524	8.49E-06	7.38
14	14:65687536	rs10144720	LINC02324	A	G	0.2268	2.08E-07	0.15

Table 12. Hispanic Ancestry Main Effect Common Variants

Hispanic Ancestry Main Effect: Genome-Wide Suggestive & Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
1	1:156104375	rs11264442	LMNA	A	G	0.0357	6.12E-06	2.54
2	2:29497921	rs149039367	ALK	T	C	0.0228	1.67E-06	3.15
2	2:164851254	rs113240724	Intergenic	T	C	0.9679	2.25E-06	0.34
2	2:168774587	rs112814375	B3GALT1-AS1	A	G	0.9745	2.09E-06	0.24
3	3:14411415	rs13074307	Intergenic	A	G	0.0659	7.17E-06	1.77
4	4:122658940	rs17051397	Intergenic	C	G	0.9644	8.83E-06	0.42
4	4:154394813	rs144751590	TMEM131L	T	G	0.9741	5.98E-06	0.36
4	4:132408019	rs11935040	Intergenic	T	C	0.9816	8.61E-07	0.29
6	6:52108777	rs2294835	IL17F	T	C	0.4957	7.13E-06	1.35
6	6:18921885	rs10807636	Intergenic	T	C	0.653	7.89E-06	0.73
6	6:161181133	rs116480834	Intergenic	A	C	0.9587	4.80E-06	0.44
7	7:14003327	rs149210567	ETV1	A	G	0.0215	4.70E-06	3.28
8	8:12610570	rs146382316	LONRF1	A	C	0.0753	6.48E-07	1.83
8	8:32515469	rs2439326	NRG1	T	G	0.0885	2.92E-06	1.69
8	8:12645580	rs3935195	LOC340357	A	G	0.4057	9.39E-06	1.37
8	8:12623103	rs151087974	LOC340357	T	C	0.9416	1.03E-06	0.50
8	8:98870322	rs114784956	Intergenic	T	G	0.9511	3.04E-06	0.52
9	9:122590178	rs2416722	Intergenic	A	T	0.0991	1.97E-06	1.75
10	10:122470170	rs147766911	LOC105378516	T	C	0.0269	2.27E-06	3.01
12	12:24294258	rs10734733	SOX5	A	G	0.3389	5.35E-06	1.38
12	12:64000369	rs11533673	DPY19L2	A	C	0.3802	7.91E-07	0.69
13	13:75086836	rs75501548	Intergenic	T	C	0.0619	4.43E-06	1.85
15	15:70098557	rs433460	Intergenic	C	G	0.1867	1.81E-06	0.63
15	15:98768093	rs1442802	Intergenic	A	G	0.3848	7.86E-06	1.36
17	17:38101719	rs117381273	LRRC3C	A	C	0.0833	1.45E-06	1.74
18	18:6130594	rs34650640	L3MBTL4	T	C	0.9748	4.95E-07	0.29

Table 13. Asian Ancestry Interaction Effect Common Variants

Asian Ancestry Interaction Effect: Genome-Wide Significant & Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
2	2:108747865	rs9917181	Intergenic	T	C	0.6197	2.67E-38	26.47
2	2:138093273	rs1463279	THSD7B	T	C	0.7018	1.08E-11	0.18
3	3:194772806	rs4677798	Intergenic	A	G	0.7266	7.27E-30	21.04
8	8:12515242	rs62488764	LOC729732	T	C	0.5243	1.01E-12	0.07
8	8:21562155	rs7826525	GFRA2	A	G	0.5542	9.51E-188	0.01
13	13:69233302	rs12871979	Intergenic	A	G	0.7464	1.03E-36	0.06

Table 14. Asian Ancestry Main Effect Common Variants

Asian Ancestry Main Effect: Genome-Wide Suggestive & Common Variants								
CHR	Position	SNP	Gene	EA	NEA	Freq	P-value	OR
1	1:214316775	rs145984379	Intergenic	T	C	0.0148	8.10E-06	3.16
1	1:31011987	rs422927	Intergenic	T	C	0.6513	6.83E-06	1.40
2	2:134090939	rs80236647	NCKAP5	C	G	0.0213	4.28E-06	3.41
3	3:67861633	rs62254903	SUCLG2-AS1	T	C	0.9155	4.05E-06	0.56
4	4:159063054	rs571859818	GASK1B	C	G	0.9858	4.59E-07	0.27
5	5:15612315	rs2964270	FBXL7	T	C	0.5781	5.54E-07	1.43
5	5:40384153	rs147171514	Intergenic	T	G	0.9847	7.96E-06	0.32
6	6:140386532	rs78948484	LOC100507477	C	G	0.0157	6.75E-06	3.30
6	6:162736476	rs7760647	PRKN	A	G	0.2755	9.71E-06	1.41
6	6:89593312	rs141960395	RNGTT	A	G	0.9727	5.76E-06	0.37
6	6:141310398	rs147077736	Intergenic	A	G	0.9834	9.36E-06	0.31
7	7:78440709	rs73369413	MAGI2	A	G	0.0239	4.93E-07	2.84
7	7:78442847	rs246461	MAGI2	A	G	0.0677	6.12E-07	2.32
7	7:152640	rs12718117	LOC100507642	T	C	0.4449	9.53E-06	0.62
8	8:23875440	rs73559223	LOC107986931	A	G	0.1489	9.80E-06	1.53
8	8:13222219	rs1671407	DLC1	A	T	0.2374	8.59E-06	1.43
9	9:7686600	rs78355386	Intergenic	A	T	0.9615	6.23E-06	0.35
10	10:82693296	rs56772124	Intergenic	T	C	0.8832	8.22E-06	0.64
11	11:71241268	rs80183776	Intergenic	A	G	0.0573	6.56E-06	1.89
11	11:134723650	rs74398592	Intergenic	A	G	0.9711	1.79E-06	0.40
13	13:50795587	rs118080265	DLEU1	A	G	0.0532	3.68E-06	2.04
13	13:105127442	rs117293597	Intergenic	A	G	0.1275	2.78E-06	1.56
14	14:95029852	rs5508	SERPINA4	T	G	0.1546	9.16E-06	1.48

Table 15. Cross Ancestry Significant SNPs: Interaction Effect

Gene & Position		Interaction Effect P-value			
GENE	POS	EA	AA	HW	AS
Intergenic	1:5026921	**5.06E-06**	9.92E-01	2.44E-01	**1.81E-02**
Intergenic	4:158480125	**1.46E-06**	6.51E-02	3.92E-01	**4.54E-02**
LOC101929710	5:95902989	**2.57E-06**	**2.29E-02**	5.16E-02	3.99E-01
ST6GAL1	3:186653218	3.39E-01	**6.45E-06**	5.65E-01	**2.26E-02**
BLK	8:11374228	**1.40E-02**	7.49E-02	**6.14E-06**	9.20E-01

Abbreviations: POS: genomic position, AA: African American ancestry, AS: Asian ancestry, EA: European ancestry, HW: Hispanic ancestry.

Table 16. Cross Ancestry Significant SNPs: Main Effect

Gene & Position		Main Effect P-value			
GENE	POS	EA	AA	HW	AS
*TCF7L2	*10:114758349	**9.59E-33**	**3.16E-03**	**1.25E-02**	9.89E-01
NAALADL2	3:174808476	2.30E-01	**1.97E-06**	**2.25E-02**	**1.54E-02**
Intergenic	7:6328946	2.85E-01	**1.58E-06**	**4.93E-02**	9.89E-01
TMED3	15:79630253	5.42E-01	**8.21E-06**	1.04E-01	**3.73E-03**
LMNA	1:156104375	**4.06E-02**	4.52E-01	**6.12E-06**	
Intergenic	2:164851254	**1.62E-02**	6.02E-02	**2.25E-06**	6.47E-01
Intergenic	9:122590178	5.81E-01	**3.44E-02**	**1.97E-06**	1.82E-01
DLC1	8:13222219	5.25E-01	1.41E-01	**4.55E-02**	**8.59E-06**
Intergenic	9:7686600	9.27E-02	7.49E-01	**4.29E-02**	**6.23E-06**

*Genomic position is genome-wide significant.

Abbreviations: POS: genomic position, AA: African American ancestry, AS: Asian ancestry, EA: European ancestry, HW: Hispanic ancestry.

Table 17. Breakdown of Cross-Ancestry Significant SNPs: Interaction Effect

Cross-Ancestry SNPs: Interaction Effect							
Position	Gene	Race	Freq	Beta	StdErr	P-value	HetPVal
5:95902989	LOC101929710	EA	0.607	0.140	0.030	2.57E-06	1.14E-01
		AA	0.759	-0.261	0.115	2.29E-02	7.68E-01
		HW	0.736	0.417	0.214	5.16E-02	6.08E-01
		AS	0.852	0.321	0.381	3.99E-01	6.95E-01
3:186653218	ST6GAL1	EA	0.192	-0.036	0.037	3.39E-01	3.12E-01
		AA	0.434	0.473	0.105	6.45E-06	8.24E-03
		HW	0.233	-0.124	0.216	5.65E-01	7.90E-01
		AS	0.388	0.699	0.307	2.26E-02	2.51E-01
8:11374228	BLK	EA	0.128	0.107	0.044	1.40E-02	7.52E-01
		AA	0.081	0.355	0.199	7.49E-02	2.29E-01
		HW	0.089	1.507	0.333	6.14E-06	5.89E-01
		AS	0.264	0.028	0.277	9.20E-01	9.97E-01

*Abbreviations: Freq: Frequency, StdErr: Standard Error, HetPVal: heterogeneity p-value.

Table 18. Breakdown of Cross-Ancestry Significant SNPs: Main Effect

Cross-Ancestry SNPs: Main Effect							
Position	Gene	Race	Freq	Beta	StdErr	P-value	HetPVal
10:114758349	TCF7L2	EA	0.293	0.252	0.021	9.59E-33	2.24E-02
		AA	0.301	0.189	0.064	3.16E-03	7.72E-02
		HW	0.265	0.185	0.074	1.25E-02	1.61E-01
		AS	0.073	-0.003	0.184	9.89E-01	2.93E-01
3:174808476	NAALADL2	EA	0.170	0.020	0.017	2.30E-01	6.03E-01
		AA	0.171	0.365	0.077	1.97E-06	1.32E-01
		HW	0.200	0.185	0.081	2.25E-02	5.63E-01
		AS	0.330	0.176	0.073	1.54E-02	6.39E-01
15:79630253	TMED3	EA	0.731	-0.009	0.015	5.42E-01	6.89E-03
		AA	0.613	0.262	0.059	8.21E-06	7.98E-01
		HW	0.748	-0.122	0.075	1.04E-01	7.00E-02
		AS	0.841	-0.262	0.090	3.73E-03	9.59E-02
1:156104375	LMNA	EA	0.023	-0.097	0.048	4.06E-02	8.15E-01
		AA	0.176	-0.057	0.075	4.52E-01	1.69E-01
		HW	0.036	0.931	0.206	6.12E-06	6.23E-01
8:13222219	DLC1	EA	0.615	0.008	0.013	5.25E-01	6.25E-01
		AA	0.337	-0.089	0.061	1.41E-01	2.87E-01
		HW	0.510	0.134	0.067	4.55E-02	7.49E-01
		AS	0.237	0.359	0.081	8.59E-06	7.70E-01

*Abbreviations: Freq: Frequency, StdErr: Standard Error, HetPVal: heterogeneity p-value.

Chapter IV:

Discussion

We completed a large-scale genome-wide gene by environment study of in four racial/ethnic groups. The gene by smoking interaction on T2D analyses identified the involvement of 20 SNPs. The most significant main effect SNP was identified in European ancestry analyses, the *TCF7L2* gene, is encouraging of accuracy of phenotype definitions and meta-analysis, as it is one of the most commonly replicated gene with T2D.

Study Rationale

The motivation to conduct this study stems from the gap in literature on the joint influences of environment (smoking) and genetics on T2D susceptibility. While it has been established that smoking is a causal risk factor for T2D, the biological and genetic mechanisms of this relationship is largely unknown. Smoking is hypothesized to modulate the effect genetic variation on T2D. Thus, conducting gene by environment interaction studies to analyze the influence of genes and environment on disease development provides insight into the biological mechanisms at play.

European Ancestry Findings

Results from the European ancestry interaction effect meta-analysis yielded genome-wide suggestive SNPs and results from the main effect meta-analysis found genome-wide significant SNPs. After SNPs were filtered by minor allele frequency for common variants, the interaction effect consisted 43 SNPs and the main effect consisted of only one SNP. Since the interaction results were genome-wide suggestive, another filter was applied for SNPs to have nominal significance (p-value <0.05) in another racial ancestry, which resulted in three SNPs. Of the three SNPs from the interaction results, two were on an intergenic region, and one was the *LOC101929710* gene. The significant output of the TCF7L2 gene from the main effect results provided confirmation of accuracy in the analysis as it is an established gene strongly associated in the pathophysiology of T2D.

African Ancestry Findings

Meta-analysis results for individuals of African ancestry produced genome-wide suggestive results for both the interaction effect and the main effect. The suggestive level genes for both the interaction and main effect were filtered based on allele frequency and significance in at least two racial cohorts. There were three top SNPs from the main effect, which included one intergenic gene, the *NAALADL2* gene, and the *TMED3* gene. Mutations in the *NAALADL2* gene has been previously reported to be associated to visceral fat and insulin responsiveness in the metabolic syndrome association study.[71] While the *TMED3* gene was found to show differences in mRNA expression between high glucose treated pancreatic islets and control islets.[74] In the interaction effect, there was one top SNP after filtering, the *ST6GAL1* gene. As noted previously, the *ST6GAL1* gene is a novel candidate risk gene for T2D from GWAS of European and South Asian ancestry.[69] In this study it was found to be a genome-wide suggestive SNP in the relationship between the *ST6GAL1* gene and smoking on the outcome of T2D.

Hispanic Ancestry Findings

The meta-analyzed GWAS on individuals of Hispanic ancestry yielded genome-wide suggestive results for both the interaction and main effect. Once the results were filtered based on allele frequency and referenced across the racial ancestries, two SNPs remained for the interaction effect and two for the main effect. The top SNPs from the interaction effect include a SNP on an intergenic region, and the *BLK* gene. Mutations in the *BLK* gene have been associated with mature onset diabetes of the young.[76] The top SNPs for the main effect included two genes, one on an intergenic region, and the *LMNA* gene. *LMNA* gene mutations have previously been shown to increase risk of T2D.[78]

Asian Ancestry Findings

Meta-analyzed GWAS results for Asian ancestry produced genome-wide significant results for the interaction effect, and genome-wide suggestive results for the main effect. Both the main effect and interaction effect were filtered based on allele frequency. The main effect SNPs were also filtered by having nominal significance in at least two studies. The 6 genome-wide significant SNPs from the interaction effect meta-analysis GWAS were primarily driven by two of the sub-studies, including in the meta-analysis, and are most likely false-positive results as the sample size for these cohorts is small (5,956 individuals with only 1,047 diabetic individuals, and only 200 current smokers). The QC protocol of the GWAS summary statistics needs to be modified to implement a thorough check of the QQ plots for each individual sub-study in the meta-analysis, implement a sample size filter (n>500) and consider alternative approaches (e.g. interaction analysis approximation via analysis stratify by smoking status).

Results Comparison

Results across the different racial ancestries were compared to locate important SNPs present in each of these analyses. Performing a cross-ancestry search of SNPs is essential to discover genes involved in influencing disease susceptibility across various racial or ethnic backgrounds. This process also helped provide additional filter for the genome-wide suggestive SNPs, since they hold a less significant p-value than genome-wide significant SNPs. Regardless, both genome-wide significant and genome-wide suggestive results were investigated across the individual GWAS racial groups. As a result, the significant cross-ancestry results included 1 SNP from the genome-wide significant results (*TCF7L2* gene) and 12 SNPs from the genome-wide suggestive results. A SNP from the genome-wide suggestive results, the *NAALADL2* gene, was significant in three ancestry results, and the *LMNA* and *BLK* genes were significant in the European ancestry results.

Conclusion

This work makes important contributions to understanding the interplay of smoking and genetic variants on risk of T2D. An important limitation of this project is the relatively small sample size for certain racial-ethnic groups, such as the Asian and Hispanic ancestries. Furthermore, the meta-analysis GWAS conducted for the Asian ancestry resulted in implausibly significant p-values that were driven by two of the individual studies contributing to the meta-analysis. Finally, the analysis conducted in SNPTEST did not allow the joint meta-analysis of main and interaction effect simultaneously. We identified the involvement of 20 SNPs in a GWAS study that places emphasis on a gene by environment analysis and interaction on the development of T2D. The results provided genomic markers (SNPs) involved in the relationship of gene by smoking interaction as the exposure, on Type II Diabetes development. The most significant SNP was identified in European ancestry analyses, the *TCF7L2* gene, is encouraging of accuracy of phenotype definitions and meta-analysis, as it is one of the most commonly replicated gene with T2D. Variations in allele frequency levels of the European results and extremely significant markers affirmed the need for a larger sample size. The Hispanic Ancestry results displayed several SNPs that were quite close to the genome-wide significant threshold. Furthermore, Asian ancestry analysis produced error resulting in unreliable results. Results from all ancestral analyses indicate that maximization of sample size would be beneficial to increase significance threshold in Hispanic and African ancestry results, while also providing more stable results for the Asian ancestry analysis.

Future Directions

This study findings provide insights into the relationship between the genetic variation and smoking on susceptibility to T2D, as it is one of the first GWAS's to examine this relationship. The results of this study suggest the need for several future directions. It is essential increased sample size and thereby the power of GWAS studies, particularly for non-European ancestry samples in the final meta-analysis. This entails maximization of sample size, improvement of power and accuracy of results, helping to identify truly associated variants amongst the list of genome-wide suggestive variants. Moreover, increasing sample size will improve the reliability of the findings, particularly in the non-European groups. Additionally, studies with detailed longitudinal follow-up data could be leveraged to consider temporal timing (e.g. latency period) of smoking by gene interactions on T2D. Finally, sensitivity analyses should be conducted to assess different smoking statuses (e.g. former smoker and ever smokers) and T2D susceptibility; this would shed light on the temporal relationship between smoking and genetic susceptibility on T2D development as we only included current smokers in this study. It would also maximize the study sample as former smokers account for a large portion of smokers in several of the studies that data was derived from.

References

1. Scheen AJ. PATHOPHYSIOLOGY OF TYPE 2 DIABETES. *Acta Clinica Belgica*. 2003;58(6):335-341. doi:10.1179/acb.2003.58.6.001

2. Heron M. National Vital Statistics Reports - Deaths: Leading Causes for 2017. *CDC*. 68:77.

3. American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care*. 2018;41(5):917-928. doi:10.2337/dci18-0007

4. Cho NH, Shaw JE, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*. 2018;138:271-281. doi:10.1016/j.diabres.2018.02.023

5. National Diabetes Statistics Report 2020. Estimates of diabetes and its burden in the United States. Published online 2020:32.

6. Diabetes Data and statistics. World Health Organization. Published May 26, 2020. Accessed May 25, 2020. http://www.euro.who.int/en/health-topics/noncommunicable-diseases/diabetes/data-and-statistics

7. CDC. Current Cigarette Smoking Among Adults in the United States. Centers for Disease Control and Prevention. Published November 18, 2019. Accessed August 22, 2020. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

8. United States Surgeon General. The Health Consequences of Smoking -- 50 Years of progress: A Report of the Surgeon General: (510072014-001). Published online 2014. doi:10.1037/e510072014-001

9. Aj S. Pathophysiology of Type 2 Diabetes. Acta clinica Belgica. doi:10.1179/acb.2003.58.6.001

10. Brunton S. Pathophysiology of Type 2 Diabetes: The Evolution of Our Understanding. *J Fam Pract*. 2016;65(4 Suppl).

11. Clark NG, Fox KM, Grandy S. Symptoms of Diabetes and Their Association With the Risk and Presence of Diabetes: Findings from the Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes (SHIELD). *Diabetes Care*. 2007;30(11):2868-2873. doi:10.2337/dc07-0816

12. Association AD. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020. *Diabetes Care*. 2020;43(Supplement 1):S14-S31. doi:10.2337/dc20-S002

13. Diagnosis | ADA. Accessed July 28, 2020. https://www.diabetes.org/a1c/diagnosis

14. Marín-Peñalver JJ, Martín-Timón I, Sevillano-Collantes C, del Cañizo-Gómez FJ. Update on the treatment of type 2 diabetes mellitus. *World J Diabetes*. 2016;7(17):354-395. doi:10.4239/wjd.v7.i17.354

15. Risk Factors for Type 2 Diabetes | NIDDK. National Institute of Diabetes and Digestive and Kidney Diseases. Accessed May 25, 2020. https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes

16. Understand Your Risk for Diabetes. American Heart Association. Accessed July 28, 2020. https://www.heart.org/en/health-topics/diabetes/understand-your-risk-for-diabetes

17. Bellou V, Belbasis L, Tzoulaki I, Evangelou E. Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses. *PLoS One*. 2018;13(3). doi:10.1371/journal.pone.0194127

18. Cheng YJ, Kanaya AM, Araneta MRG, et al. Prevalence of Diabetes by Race and Ethnicity in the United States, 2011-2016. *JAMA*. 2019;322(24):2389-2398. doi:10.1001/jama.2019.19365

19. Race/Ethnic Difference in Diabetes and Diabetic Complications | SpringerLink. Accessed May 25, 2020. https://link.springer.com/article/10.1007/s11892-013-0421-9

20. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia*. 1999;42(2):139-145. doi:10.1007/s001250051131

21. Billings LK, Florez JC. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci*. 2010;1212:59-77. doi:10.1111/j.1749-6632.2010.05838.x

22. Ali O. Genetics of type 2 diabetes. *World J Diabetes*. 2013;4(4):114-123. doi:10.4239/wjd.v4.i4.114

23. Sun X, Yu W, Hu C. Genetics of Type 2 Diabetes: Insights into the Pathogenesis and Its Clinical Application. BioMed Research International. doi:https://doi.org/10.1155/2014/926713

24. Ahlqvist E, Ahluwalia TS, Groop L. Genetics of Type 2 Diabetes. *Clin Chem*. 2011;57(2):241-254. doi:10.1373/clinchem.2010.157016

25. Genome Wide Association Studies (GWAS). Genetics Generation. Accessed May 25, 2020. https://knowgenetics.org/genome-wide-association-studies-gwas/

26. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*. 2012;8(12). doi:10.1371/journal.pcbi.1002822

27. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet*. 2019;20(8):467-484. doi:10.1038/s41576-019-0127-1

28. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*. 2008;9(5):356-369. doi:10.1038/nrg2344

29. Genetics Home Reference. Type 2 diabetes. NIH Genetics Home Reference. Accessed July 27, 2020. https://ghr.nlm.nih.gov/condition/type-2-diabetes

30. Ali O. Genetics of type 2 diabetes. *WJD*. 2013;4(4):114. doi:10.4239/wjd.v4.i4.114

31. Olokoba AB, Obateru OA, Olokoba LB. Type 2 Diabetes Mellitus: A Review of Current Trends. *Oman Med J*. 2012;27(4):269-273. doi:10.5001/omj.2012.68

32. Xue A, Wu Y, Zhu Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun*. 2018;9(1):2941. doi:10.1038/s41467-018-04951-w

33. Mahajan A, Taliun D, Thurner M. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps | Nature Genetics. *Nature Genetics*. Published online October 8, 2018. doi:https://doi.org/10.1038/s41588-018-0241-6

34. Flannick J, Fuchsberger C, Mahajan A, et al. Sequence data and association statistics from 12,940 type 2 diabetes cases and controls. *Sci Data*. 2017;4. doi:10.1038/sdata.2017.179

35. Flannick J, Mercader JM, Fuchsberger C, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*. 2019;570(7759):71-76. doi:10.1038/s41586-019-1231-2

36. Lyssenko V, Laakso M. Genetic Screening for the Risk of Type 2 Diabetes: Worthless or valuable? *Diabetes Care*. 2013;36(Supplement 2):S120-S126. doi:10.2337/dcS13-2009

37. Hunter DJ. Gene–environment interactions in human diseases. *Nat Rev Genet*. 2005;6(4):287-298. doi:10.1038/nrg1578

38. Park B, Koo S-M, An J, et al. Genome-wide assessment of gene-by-smoking interactions in COPD. *Sci Rep*. 2018;8(1):1-11. doi:10.1038/s41598-018-27463-5

39. Song N, Shin A, Jung HS, Oh JH, Kim J. Effects of interactions between common genetic variants and smoking on colorectal cancer. *BMC Cancer*. 2017;17. doi:10.1186/s12885-017-3886-0

40. Bentley AR, Sung YJ, Brown MR, et al. Multi-ancestry genome-wide gene–smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nature Genetics*. 2019;51(4):636-648. doi:10.1038/s41588-019-0378-y

41. Aschard H, Tobin MD, Hancock DB, et al. Evidence for large-scale gene-by-smoking interaction effects on pulmonary function. *Int J Epidemiol*. 2017;46(3):894-904. doi:10.1093/ije/dyw318

42. Polfus LM, Smith JA, Shimmin LC, et al. Genome-Wide Association Study of Gene by Smoking Interactions in Coronary Artery Calcification. *PLOS ONE*. 2013;8(10):e74642. doi:10.1371/journal.pone.0074642

43. Montasser M, Shimmin L, Hanis C, Boerwinkle E, Hixson J. Gene by Smoking Interaction in Hypertension: Identification of a Major QTL on Chromosome 15q for Systolic Blood Pressure in Mexican Americans. *Journal of hypertension*. 2009;27:491-501. doi:10.1097/HJH.0b013e32831ef54f

44. Talmud PJ. Gene-environment interaction and its impact on coronary heart disease risk. *Nutr Metab Cardiovasc Dis*. 2007;17(2):148-152. doi:10.1016/j.numecd.2006.01.008

45. Yu H, Wang T, Zhang R, et al. Alcohol consumption and its interaction with genetic variants are strongly associated with the risk of type 2 diabetes: a prospective cohort study. *Nutr Metab (Lond)*. 2019;16:64. doi:10.1186/s12986-019-0396-x

46. Qi L, Hu FB, Hu G. Genes, environment, and interactions in prevention of type 2 diabetes: a focus on physical activity and lifestyle changes. *Curr Mol Med*. 2008;8(6):519-532. doi:10.2174/156652408785747915

47. Dietrich S, Jacobs S, Zheng J-S, Meidtner K, Schwingshackl L, Schulze MB. Gene-lifestyle interaction on risk of type 2 diabetes: A systematic review. *Obesity Reviews*. 2019;20(11):1557-1571. doi:10.1111/obr.12921

48. Wu P, Rybin D, Bielak LF, et al. Smoking-by-genotype interaction in type 2 diabetes risk and fasting glucose. *PLoS One*. 2020;15(5). doi:10.1371/journal.pone.0230815

49. Tobacco | Healthy People 2020. Accessed August 23, 2020. https://www.healthypeople.gov/2020/leading-health-indicators/2020-lhi-topics/Tobacco/data#1

50. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181-1186. doi:10.1038/ng1007-1181

51. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*. 2015;12(3). doi:10.1371/journal.pmed.1001779

52. dbGaP Overview. dbGaP. Accessed May 26, 2020. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.cgi

53. About UK Biobank | UK Biobank. Accessed May 26, 2020. https://www.ukbiobank.ac.uk/about-biobank-uk/

54. Tryka KA, Hao L, Sturcke A, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res*. 2014;42(Database issue):D975-D979. doi:10.1093/nar/gkt1211

55. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

56. NHIS - Adult Tobacco Use - Glossary. Published May 10, 2019. Accessed August 23, 2020. https://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm

57. Michigan Imputation Server. Accessed July 30, 2020. https://imputationserver.sph.umich.edu/index.html#!

58. Li Y, Willer C, Sanna S, Abecasis G. Genotype Imputation. *Annu Rev Genomics Hum Genet*. 2009;10:387-406. doi:10.1146/annurev.genom.9.081307.164242

59. Genome-Wide Association Studies Fact Sheet. Genome.gov. Accessed May 26, 2020. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

60. Matchini J, Band G. SNPTEST. SNPTEST. Accessed May 26, 2020. https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

61. Haidich AB. Meta-analysis in medical research. *Hippokratia*. 2010;14(Suppl 1):29-37.

62. Stone DL, Rosopa PJ. The Advantages and Limitations of Using Meta-analysis in Human Resource Management Research. *Human Resource Management Review*. 2017;27(1):1-7. doi:10.1016/j.hrmr.2016.09.001

63. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191. doi:10.1093/bioinformatics/btq340

64. Lee CH, Cook S, Lee JS, Han B. Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. *Genomics Inform*. 2016;14(4):173-180. doi:10.5808/GI.2016.14.4.173

65. Bacanu S-A, Devlin B, Roeder K. The Power of Genomic Control. *The American Journal of Human Genetics*. 2000;66(6):1933-1944. doi:10.1086/302929

66. LOC101929710 [Homo sapiens (human)] - Gene - NCBI. NCBI. Accessed July 31, 2020. https://www.ncbi.nlm.nih.gov/gene/101929710

67. Gloyn AL, Braun M, Rorsman P. Type 2 Diabetes Susceptibility Gene TCF7L2 and Its Role in β-Cell Function. *Diabetes*. 2009;58(4):800-802. doi:10.2337/db09-0099

68. PubChem. ST6GAL1 - ST6 beta-galactoside alpha-2,6-sialyltransferase 1 (human). NIH National Library of Medicine. Accessed July 4, 2020. https://pubchem.ncbi.nlm.nih.gov/gene/ST6GAL1/human

69. Rudman N, Gornik O, Lauc G. Altered N-glycosylation profiles as potential biomarkers and drug targets in diabetes. *FEBS Letters*. 2019;593(13):1598-1615. doi:10.1002/1873-3468.13495

70. Berndt SI, Wang Z, Yeager M, et al. Two susceptibility loci identified for prostate cancer aggressiveness. *Nature Communications*. 2015;6(1):6889. doi:10.1038/ncomms7889

71. Zhang Y, Kent JW, Olivier M, et al. QTL-based association analyses reveal novel genes influencing pleiotropy of Metabolic Syndrome (MetS). *Obesity (Silver Spring)*. 2013;21(10):2099-2111. doi:10.1002/oby.20324

72. PubChem. TMED3 - transmembrane p24 trafficking protein 3 (human). NIH National Library of Medicine. Accessed July 31, 2020. https://pubchem.ncbi.nlm.nih.gov/gene/TMED3/human

73. Zheng H, Yang Y, Han J, et al. TMED3 promotes hepatocellular carcinoma progression via IL-11/STAT3 signaling. *Scientific Reports*. 2016;6(1):37070. doi:10.1038/srep37070

74. Hall E, Dekker Nitert M, Volkov P, et al. The effects of high glucose exposure on global gene expression and DNA methylation in human pancreatic islets. *Molecular and Cellular Endocrinology*. 2018;472:57-67. doi:10.1016/j.mce.2017.11.019

75. Genetics Home Reference. BLK gene. NIH U.S. National Library of Medicine. Accessed July 5, 2020. https://ghr.nlm.nih.gov/gene/BLK

76. Mutations at the BLK Locus Linked to Maturity Onset Diabetes of the Young and Beta-Cell Dysfunction - PubMed. Accessed July 5, 2020. https://pubmed.ncbi.nlm.nih.gov/19667185/

77. Genetics Home Reference. LMNA gene. NIH U.S. National Library of Medicine. Accessed July 5, 2020. https://ghr.nlm.nih.gov/gene/LMNA

78. Owen KR, Groves CJ, Hanson RL, et al. Common variation in the LMNA gene (encoding lamin A/C) and type 2 diabetes: association analyses in 9,518 subjects. *Diabetes*. 2007;56(3):879-883. doi:10.2337/db06-0930

79. Mesa JL, Loos RJF, Franks PW, et al. Lamin A/C Polymorphisms, Type 2 Diabetes, and the Metabolic Syndrome. *Diabetes*. 2007;56(3):884-889. doi:10.2337/db06-1055

80. Genetics Home Reference. DLC1 gene. NIH National Library of Medicine. Accessed July 31, 2020. https://ghr.nlm.nih.gov/gene/DLC1

81. Yuan B-Z, Miller MJ, Keck CL, Zimonjic DB, Thorgeirsson SS, Popescu NC. Cloning, Characterization, and Chromosomal Localization of a Gene Frequently Deleted in Human Liver Cancer (DLC-1) Homologous to Rat RhoGAP. *Cancer Res*. 1998;58(10):2196-2199.

82. Matoba N, Akiyama M, Ishigaki K, et al. GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. *Nature Human Behaviour*. 2019;3(5):471-477. doi:10.1038/s41562-019-0557-y