國立臺灣大學共同教育中心國際學院
智慧醫療與健康資訊碩士學位學程

碩士論文

Master's Program in Smart Medicine and Health Informatics
Center of General Education, International College
National Taiwan University

Master's Thesis

台灣族群膽結石疾病遺傳易感性研究：基因-表型組學方法與多基因風險評分模型

Genetic Susceptibility of Gallstone Disease in Taiwanese Population: A Genome-Phenome Approach with Polygenic Risk Score

蔡岳君

Tsai, Yueh-Chun

指導教授：廖世偉 博士 與 李建璋 博士

Advisor: Shih-Wei, Liao, Ph.D.& Chien-Chang, Lee, Ph.D.

中華民國 114 年 6 月

June, 2025

doi:10.6342/NTU202500806

國立臺灣大學碩士學位論文
口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

台灣族群膽結石疾病遺傳易感性研究：
基因-表型組學方法與多基因風險評分模型
Genetic Susceptibility of
Gallstone Disease in Taiwanese Population:
A Genome-Phenome Approach with Polygenic Risk Score

The undersigned, appointed by the Department / Graduate Institute of Master's Program in Smart Medicine and Health Informatics on 29 (date) 04 (month) 2025 (year) have examined a Master's Thesis entitled above presented by Tsai, Yueh-Chun 蔡岳君 (name) R12H45002 (student ID) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

庾世偉	雷建琦	
(指導教授 Advisor)	(共同指導教授 Co-Advisor)	
郭育良	黃冬富	

系（所、學位學程）主管 Director: 國際學院看醫健康資訊學程主任 林澤

誌謝

感謝我的共同指導教授李建璋老師，以及隸屬之實驗室所有同仁耐心指導與研究資源，讓我順利完成碩士論文的撰寫與國際期刊投稿，也感謝我的指導教授廖世偉老師與保安扶輪社，研究過程中，提供我教育生活協助與支柱。感謝 StanCode 與人工智慧醫療課程中結識的兩群朋友們，讓我對程式設計有熱情與自信。感謝郭育良教授與林澤系主任，提供我專案研究的機會。感謝智慧醫療學系的行政團隊，提拔我成為斐陶斐學會新榮譽會員候選人。感謝我的摯友、高小飛、華岳君與父母，兩年間無條件的支持。最後要謝謝陳建佑和我自己，努力將自己的選擇變成最好的決定，也願我們在未來都能夠更從容不迫的優雅面對任何挑戰。

doi:10.6342/NTU202500806

Acknowledgement

I would like to express my heartfelt gratitude to my co-advisor, Professor Chien-Chang Lee, and all the members of his laboratory for their patient guidance and research resources, which enabled me to successfully complete my master's thesis and submit it to an international journal. Meanwhile, I am also deeply thankful to my advisor, Professor Shih-Wei Liao, and the Baoan Rotary Club for their support and assistance throughout my academic journey.

I sincerely thank the two groups of friends I met through StanCode and the AI in Healthcare course, who inspired my passion and confidence in programming. My appreciation also goes to Professor Leon Kuo and Professor Che Lin for offering me opportunities to participate in research projects. I am grateful to the administrative team of the Department of SMARTMHI for nominating me as a candidate for the Phi Tau Phi Scholastic Honor Society.

I would also like to thank my Instagram close friends, Fly and Hua, as well as my parents, for their unconditional support over the past two years. Lastly, I want to thank Zander the Ted and myself—for the dedication to make the choice right. May we continue to face all future challenges with grace and confidence.

doi:10.6342/NTU202500806

摘要

膽結石是種常見消化系統疾病，復發率與衍生相關疾病造成醫療系統極大負擔。膽結石形成與膽汁酸代謝、膽固醇代謝失衡、基因背景與種族差異密切相關，但大多數膽結石疾病基因研究主採用西方人群數據庫，限制對東方人群治病性生理機轉的理解。因此，本研究利用台灣人體生物資料庫，經品質控制後共 108,403 名具漢族受試者進行全基因體關聯分析（GWAS），探討與膽結石相關的遺傳變異，亦進行 LDSC 分析驗證通膨程度（λ GC = 1.0741；intercept = 1.0067），結果顯示通膨現象主要源自多基因性而非其他混雜因子。建構多基因風險預測模型（Polygenic Risk Score, PRS），PRS-CS 展現極高預測力（AUC = 0.98）。本研究鑑定出 57 個顯著基因變異位點，其中有 38 個新發現，計算每個顯著基因變異位點的治病機率，以 rs80217587（*TM4SF4*）具有最高的後驗包含機率（PIP = 0.202），顯示其為最具潛力的致病變異位點。本研究發現的 7 個 Lead SNP, 分別對應 1 個控制區（RP11-626H12.1）與 6 個鄰近基因（*TM4SF4*、*LRBA*、*UBXN2B*、*CYP7A1*、*HNF4A* 和 *ANO1*），其中 *TM4SF4*、*LRBA*、*UBXN2B* 和 *ANO1* 為首次在東亞族群提出與膽結石易染性的關聯基因。我們的研究結果指出 *TM4SF4* 可作為膽結石疾病早期偵測與預防性醫療管理的潛在生物標記與治療標的。。透過路徑富集分析，我們也提出三組血中生物標記群（GGT、ALT、CRP、膽酸/膽紅素前驅物等），可望作為早期非侵入性風險預測工具。搭配多基因風險分數與個人飲食改善計畫建議，有助於建立針對膽結石高風險族群之精準預防策略。

關鍵詞：膽結石；全基因組關聯研究；表型全關聯研究；台灣人體生物資料庫；多基因風險評分；蛋白質-蛋白質交互作用；血液檢測；預防醫學。

doi:10.6342/NTU202500806

Abstract

Gallstone disease (GSD) is a prevalent gastrointestinal disorder with high recurrence and substantial healthcare burden. Its etiology involves bile acid and cholesterol metabolism, genetic predisposition, and ethnic variation. However, most genetic studies focused on Western populations, limiting insights into East Asian-specific mechanisms. This study leveraged data from 108,403 Han Chinese individuals in the Taiwan Biobank to perform a genome-wide association study (GWAS) on GSD. After quality control and adjustment for population stratification, linkage disequilibrium score regression (LDSC) confirmed minimal inflation (λGC = 1.0741; intercept = 1.0067), suggesting that observed signals are due to polygenicity. Polygenic risk prediction using PRS-CS demonstrated strong performance (AUC = 0.98). We identified 57 significant SNPs, including 38 novel variants. Bayesian fine-mapping revealed rs80217587 (*TM4SF4*) as the top causal candidate (PIP = 0.202), along with 7 lead SNPs mapped to *TM4SF4, LRBA, UBXN2B, CYP7A1, HNF4A, ANO1*, and a non-coding region (RP11-626H12.1). *TM4SF4, LRBA, UBXN2B* and *ANO1* are first mentioned in East Asian population. We highlight TM4SF4 as a potential biomarker and therapeutic target for the early detection and preventive management of GSD. Moreover, pathway enrichment analysis identified candidate biomarkers (GGT, ALT, CRP, bile acid precursors) for potential early, non-invasive risk laboratory test screening. By integrating PRS with personalized dietary intervention plans, we propose a precision prevention strategy tailored for individuals at high risk of GSD.

Keywords: Gallstone; GWAS; phenome-wide association study; Taiwan Biobank; polygenic risk score; Protein-protein interaction network; laboratory test; prevention medicine.

doi:10.6342/NTU202500806

Contents

doi:10.6342/NTU202500806

doi:10.6342/NTU202500806

doi:10.6342/NTU202500806

List of Figures

doi:10.6342/NTU202500806

List of Tables

doi:10.6342/NTU202500806

Abbreviation

ABCG8	ATP-Binding Cassette Subfamily G Member 8
ALT	Alanine Transaminase
AST	Aspartate aminotransferase
ALP	Alkaline phosphatase
ANO1	Anoctamin 1
AUC	Area Under the Receiver Operating Characteristic Curve
BP	Base-Pair Position
CHR	Chromosome
CRP	C-Reactive Protein
CYP7A1	Cytochrome P450 Family 7 Subfamily A Member 1
EA	Effect Allele
EFO	Experimental Factor Ontology
FDR	False Discovery Rate
GCP	Good Clinical Practice
GGT	Gamma-Glutamyl Transferase (γ-Glutamyl Transferase)
GO	Gene Ontology
GSD	Gallstone Disease
GWAS	Genome-Wide Association Studies
HP	Human Phenotype Ontology
HDL	High-Density Lipoprotein
Hg37	Human Genome Build 37 (also known as GRCh37)

doi:10.6342/NTU202500806

HNF4A	Hepatocyte Nuclear Factor 4 Alpha
HWE	Hardy–Weinberg Equilibrium
ICD-10	International Classification of Diseases, 10th Revision
LD	Linkage Disequilibrium
LDSC	Linkage Disequilibrium Score Regression
λGC	Genomic Inflation Factor
LDL	Low-Density Lipoprotein
LRBA	Lipopolysaccharide Responsive Beige-Like Anchor Protein
MAF	Minor Allele Frequency
MAP	Frequency of the Effect Allele
maj	Major Allele / Effect Allele
min	Minor Allele / Non-Effect Allele
MCMC	Markov Chain Monte Carlo
NAFLD	Non-Alcoholic Fatty Liver Disease
NEA	Non-Effect Allele
NHI	Taiwan's National Health Insurance
OR	Odds Ratio
PCA	Principal Component Analysis
PheWAS	Phenome-Wide Association Studies
PIP	Posterior Inclusion Probability
PPI	Protein-Protein Interaction (Network)
PRS	Polygenic Risk Score
PRS-CS	Polygenic Risk Score with Continuous Shrinkage Priors

doi:10.6342/NTU202500806

PRS-CT	Polygenic Risk Score via Clumping and Thresholding
QC	Quality Control
Q-Q plot	Quantile–Quantile Plot
RAME	Rare Alleles of Major Effect
SE	Standard Error
SLCO1A2	Solute Carrier Organic Anion Transporter Family Member 1A2
SLCO1B1	Solute Carrier Organic Anion Transporter Family Member 1B1
SNP	Single Nucleotide Polymorphism
TM4SF4	Transmembrane 4 L Six Family Member 4
TWB	Taiwan Biobank
UBXN2B	UBX Domain Protein 2B

doi:10.6342/NTU202500806

Chapter 1

Introduction

1.1 Research Background

Gallstone disease (GSD), or cholelithiasis, significantly impacts public health and healthcare systems worldwide, 15% of the population with an annual incidence of 0.6% [1]. It is a disease with a complication rate of 5–10% [2,3]. Moreover, GSD incidence in individuals aged 60+ has risen by 15% over the past decade, showing its tendency to become an aging chronic disease [4]. In Taiwan, GSD ranks among the top 15 diseases, with a prevalence of 10%, National Health Insurance system spends over $1.6 billion on related GSD disorders in 2022 [5]. About 1–2% of patients experience acute symptoms and visit the emergency room due to severe pain [1].

1.2 Problem Statement

Although many cases are asymptomatic or mild, about 1–2% of patients experience acute symptoms and visit the emergency room due to severe pain. Medications may not ensure complete clearance, and surgical cholecystectomy requires skilled surgeons for optimal results [4]. This increases the risk of recurrence or leads to life-threatening complications, such as acute pancreatitis, sepsis, or even gastrointestinal cancer [1, 2]. The GSD-related complications and comorbidities are also a burden in healthcare system. Moreover, GSD in Asian populations presents a higher prevalence of pigment gallstones compared to the cholesterol gallstones commonly observed in Western populations [5]. This raises the question of pathophysiological differences in GSD across ethnic groups. Despite substantial progress in understanding the genetic underpinnings of GSD, research remains heavily biased toward European ancestry, with limited studies on Asian populations. Moreover, GSD also displays context-environmental trait in clinical practice observation.

doi:10.6342/NTU202500806

1.3 Research Objectives

This study explores genetic differences in GSD between Eastern and Western populations, addressing the predominant research focus on European ancestry. Utilizing the Taiwan Biobank (TWB), a genetically homogeneous resource, we conducted genome-wide association studies (GWAS) to identify significant single nucleotide polymorphism (SNP) associated with GSD in the Taiwanese population. We applied polygenic risk score (PRS) framework to develop a robust GSD risk prediction model. Additionally, to aid clinical interpretation, we employed various genetic analysis tools from SNP-level statistical analysis to gene-level interpretation. Our goal is to enhance early detection through a non-invasive screening strategy develop genome-driven prevention and drug strategies by examining the relationship between phenotypic traits and genetic risk factors, thus tailored to the Taiwanese population for better health outcomes.

1.4 Thesis Organization

This thesis is structured into seven chapters. Chapter 2 provides a genome-focused literature review to contextualize the study. Chapter 3 introduces the dataset and describes the analytical tools and methods used. Chapter 4 presents the main results alongside clinical interpretations. Chapter 5 discusses the implications of the findings and concludes the study. Chapter 6 includes an ethnic background declaration. Bibliography provides the full list of references cited throughout the thesis. Supplemental materials are in the appendix.

doi:10.6342/NTU202500806

Chapter 2

Literature Review

2.1 Diagnostic Criteria of Gallstone Disease

Gallstone Disease (GSD), also known as cholelithiasis, the ICD-10 code is K80. Its diagnosis is based on a combination of clinical symptoms from inquiry or personal history and physical examinations such as imaging findings, and laboratory tests [2, 6]. The most common symptom is biliary colic, characterized by episodic right upper quadrant or epigastric pain, often triggered by fatty meals and lasting from 30 minutes to several hours. Other symptoms include nausea, vomiting, bloating, and dyspepsia. In asymptomatic cases, gallstones are often detected incidentally during routine imaging. Ultrasonography is the gold standard for GSD diagnosis, offering high sensitivity (84–89%) and specificity (99%). It detects gallstones as echogenic foci with posterior acoustic shadowing. Image examinations, computed tomography (CT) and magnetic resonance cholangiopancreatography (MRCP), are useful for detecting complicated cases, such as choledocholithiasis. Meanwhile, endoscopic ultrasound (EUS) provides higher resolution for small stones or sludge. Laboratory blood tests, including liver function tests (LFTs), amylase, and lipase, help assess biliary obstruction or pancreatitis. A pattern of elevated bilirubin, ALP, and GGT may indicate involvement of the common bile duct, particularly in the setting of obstructive biliary pathology. In clinical practice, the Tokyo Guidelines (TG18) aid in diagnosing and grading acute cholecystitis severity. Patients are stratified into five clinical levels based on symptom severity and treatment needs: Level 0 (asymptomatic or mild symptoms), Level 1 (pain management with NSAIDs or opioids), Level 2 (conservative therapy with dissolution treatment), Level 3 (surgical intervention via laparoscopic cholecystectomy), and those with severe complications. Recurrence or complications came after inappropriate treatments.

doi:10.6342/NTU202500806

2.2 Genetic and Biological Mechanisms of Gallstone Formation

Gallstone formation involves three major pathophysiological processes: cholesterol supersaturation, shortened nucleation time, and gallbladder hypomotility [2]. When hepatic cholesterol secretion exceeds bile acid solubilization, cholesterol precipitates into crystals, initiating lithogenesis. Previous Taiwan Biobank (TWB) studies mentioned that HNF4A regulates bile acid homeostasis, with rs1800961 linked to increased gallstone risk, particularly in obese males [7-8]. ANO1, which affects gallbladder motility, also harbors risk variants such as rs56363382 [8].

2.3 Genetic Association with Metabolic and Environmental Contexts

Beyond cholesterol metabolism, FUT2 contributes to GSD heritability through its influence on lipid levels and gut microbiota [6,9–10]. Notably, FUT2 variant rs601338 is also associated with gastrointestinal infections and microbial shifts, which linked to cholelithiasis [6,9]. Environmental and dietary contexts also modulate gallstone susceptibility. A Mediterranean-style diet—rich in vegetable oils, fruits, fish, legumes, nuts, whole grains, and lean proteins—was found to reduce symptomatic GSD risk by 33% compared to high-fat diets [17, 18]. Moreover, GSD also displays ethnic variability: pigment stones predominate in Asian populations, likely due to high-carbohydrate diets, whereas cholesterol stones are more common in Western populations with obesity-related dyslipidemia [11,12]. These findings underscore the multifactorial nature of GSD, involving genetics, diet and lifestyle. This indicates that both genetic and lifestyle factors play a role in disease formation, which makes gallstone disease a complex target for prevention and management in different population. These findings highlight the interaction between host genetics, diet, and the gut microbiome in shaping gallstone disease risk.

doi:10.6342/NTU202500806

Chapter 3

Methods

3.1 Data Sources and Quality Control

This study utilized data from the TWB, including 148,567 Taiwanese Han Chinese subjects while excluding individuals of Indigenous or non-East Asian ancestry. TWB is a prospective cohort study integrating genomic data to explore genome science, lifestyle factors, drug development, and disease progression [12]. As of August 2021, TWB covers 9,814,944 SNPs, with participants aged 30–70 years recruited from 29 community-based centers across Taiwan. Individuals with kinship or a history of psychiatric disorders, substance abuse, or alcoholism were excluded. All participants completed 1,048 questionnaire items and 178 clinical assessments. TWB is managed under Taiwan's National Health Insurance (NHI) system [13], and the study was approved by the Research Ethics Committee (IRB No. 201904080RINB) in compliance with the Declaration of Helsinki and Good Clinical Practice (GCP) guidelines. Data details are available on the TWB website. We performed quality control (QC) procedures on the genotype data using PLINK (v1.9) [15]. This included removing samples with a missing call rate > 5%. Our criteria of quality control on SNP data: (i) minor allele frequency (MAF) < 0.1% or imputation accuracy score < 0.3, (ii) violation of Hardy–Weinberg equilibrium (HWE) < 1×10^{-12}, and (iii) individuals in related pairs with a kinship coefficient ≥ 8.84%. (iv) SNPs on the X and Y chromosomes and mitochondrial SNPs were excluded from analyses.

3.2 Genome-Wide Association Studies (GWAS)

GWAS is a linear regression statistical model to analyze the association between each SNP and a trait or phenotype [15]. The effect size beta (β) indicates how strongly a particular SNP contributes to the disease. We identified GSD-associated variants at genome-wide

doi:10.6342/NTU202500806

significance ($P < 5\times10^{-8}$) and defined the phenotype as self-report questionnaire gallstone = 1 [12]. General analysis was conducted using PLINK (v1.9) and R 4.3.2. Lead SNPs were defined as the most statistically significant SNPs nearest to candidate genes [21]. All genomic positions in this study are reported based on the GRCh37 (hg19) reference genome. This build was selected for consistency with TWB genotyping array used during the initial data collection period. The majority studies and databases are also based on GRCh37, facilitating downstream annotation and cross-study comparison.

3.3 Quality Control Analysis

To correct for population stratification and ensure proper statistical calibration, we performed principal component analysis (PCA) on the genome-wide genotype matrix. The top five principal components, capturing the major axes of genetic variation, were included alongside age and sex as covariates in a logistic regression model under an additive genetic framework. Association summary statistics were generated for each SNP, and significant signals were visualized using a Manhattan plot with the qqman R package [16]. To assess potential confounding and estimate the genetic contribution to GSD, we performed linkage disequilibrium score regression (LDSC) using GWAS summary statistics and LD (Linkage Disequilibrium) scores from the 1000 Genomes East Asian reference panel [25]. LDSC estimates the genomic inflation factor (λGC) to evaluate test statistic inflation, where values >1 may reflect population stratification or other biases. The LDSC intercept helps distinguish true polygenicity from confounding, with values near 1 indicating minimal bias. SNP-based heritability (h^2) was also calculated to quantify the proportion of phenotypic variance explained by common variants. A QQ plot [16] and a scatter plot of absolute effect sizes versus minor allele frequency were generated to confirm calibration and consistency with LDSC results. We included SNPs with $p < 5 \times 10^{-5}$ to improve resolution in visualization.

doi:10.6342/NTU202500806

3.4 SNP-Level Statistical Analysis

In Section 3.4, we introduce two complementary locus-centric strategies for SNP-level prioritization. They differ in purpose: the regional association plot highlights clusters of associated SNPs and secondary peaks, but it sometimes cannot distinguish a causal SNP from a densely linked tag. By contrast, Bayesian fine-mapping with SuSiE computes a posterior inclusion probability (PIP) for each SNP and assembles minimal 95 % credible sets. Variants with high PIPs are far more likely to be driving the association, allowing us to downweigh mere proxies and highlight true functional candidates. To complement these locus-specific approaches, we perform a genome-wide enrichment analysis to categorize our SNPs into exclusive functional classes to reveal our top signals are generally over-represented in which functional mechanism.

3.4.1 Regional Association Plots

Regional association plots are used to visualize the association signals of SNPs within a genomic region, typically centered around a lead SNP identified through GWAS [22]. These plots highlight both the statistical significance of each SNP (represented by $-\log_{10}$(p-value) on the y-axis) and the local LD pattern, which reflects how SNPs are inherited together. The x-axis represents the genomic position in base pairs, and each point is color-coded based on its r^2 value, indicating the strength of LD in each SNP with the lead SNP. To generate these plots, we used PLINK v1.9 to compute the LD matrix for selected SNPs. Using GWAS summary statistics, the computed LD matrix, and the 1000 Genomes East Asian reference panel, we constructed regional association plots with the SusieR package to investigate possible signal clustering for credible set analysis. It highlights clusters of associated SNPs and their LD patterns, offering a broader view of signal enrichment.

doi:10.6342/NTU202500806

3.4.2 Functional Fine-Mapping

To identify lead variants and distinguish causal SNPs from tagging SNPs, we performed functional fine-mapping using the Bayesian regression-based framework implemented in SusieR, which estimates the posterior inclusion probability (PIP) for each SNP. The PIP reflects the probability that a variant is causally responsible for the observed association signal. In genomic loci with dense LD, GWAS often identifies clusters of SNPs with similarly low p-values, complicating causal inference. Fine-mapping resolves this ambiguity by leveraging the LD structure to separate overlapping signals. For LD estimation, we used GWAS summary statistics (z-scores) and LD matrices derived from the 1000 Genomes Project East Asian reference panel. Sample size information was extracted from GWAS summary report. We constructed 95% credible sets, defined as the smallest set of SNPs whose cumulative PIP reaches 95%. Each credible set was characterized by the number of SNPs included (C) and the average LD (R) among these SNPs, with R-values closer to 1 indicating stronger internal correlation within the credible set. SusieR plots were generated to visualize PIP values across chromosomal positions, facilitating biologically informed SNP prioritization beyond simple p-value rankings [22].

3.4.3 Functional Annotation Enrichment

We first derived a non-redundant set of gallstone-associated loci by LD-pruning all genome-wide significant SNPs ($p < 5 \times 10^{-8}$) to $r^2 < 0.1$ using PLINK v1.9. Each remaining lead SNP was annotated with Ensembl VEP (v104, GRCh37) [32] and assigned—without overlap—to one of eleven functional classes (intergenic; intronic; upstream ≤2 kb of the TSS; downstream ≤2 kb of the TES; exonic; 5′ UTR; 3′ UTR; canonical splice ± 2 bp; ncRNA_exonic; ncRNA_intronic; ncRNA_splicing). To quantify enrichment, we compared the observed fraction of lead SNPs in each category against the fraction expected under a genome-wide background of ~4.23 million QC-passed variants. For each functional class, we

doi:10.6342/NTU202500806

built a 2 × 2 contingency table (lead vs. background SNPs × in-category vs. out-of-category) and performed Fisher's exact test to assess over- or under-representation. Fold enrichment was defined as $\log_2$(observed / expected) and plotted with bar shading proportional to its magnitude. Statistical significance was denoted by * for nominal $p < 0.05$ and ** after Bonferroni correction for 11 tests ($p < 0.05 / 11$). This framework [33] furnishes a clear portrait of which genomic elements are preferentially targeted by GSD-associated variation.

3.5 Polygenic Risk Score (PRS)

We applied two established methods to construct polygenic risk scores (PRS) [16-18]: PRS-CT and PRS-CS. PRS-CT (Clumping and Thresholding), implemented using PLINK v1.9, selects SNPs based on linkage disequilibrium (LD) structure and p-value thresholds. The procedure involves: (i) merging SNPs within a 250 kb window to avoid redundant associations, (ii) applying p-value thresholds ($P < 0.5$ to $P < 0.001$) for SNP inclusion, and (iii) filtering based on LD ($r^2 > 0.1$). Clumped SNPs across all chromosomes were aggregated, and individual-level PRS was calculated as the sum of weighted genotype dosages. PRS-CS (Bayesian Continuous Shrinkage), executed via the Python package [28], estimates SNP effect sizes by leveraging LD information and GWAS summary statistics without requiring predefined p-value thresholds, making it more suitable for highly polygenic traits. The process included: (i) use of the 1000 Genomes East Asian reference panel for LD structure, (ii) input of GWAS summary statistics, (iii) Markov chain Monte Carlo (MCMC) sampling parameters (1,000 iterations, 500 burn-in, thinning every 5 steps), (iv) specification of gamma priors for effect size sparsity ($a = 1$, $b = 0.5$), and (v) LD block construction. For model refinement, we included age and sex as covariates in the PRS modeling framework. Predictive performance was evaluated using Area Under the ROC (Receiver Operating Characteristic) Curve (AUC) metrics in R. We conducted a sensitivity analysis to explore the relationship between sex, age, and PRS-CS measurements. For categorical variables, age was

doi:10.6342/NTU202500806

divided into physiological male and physiological female. For continuous variables, we categorized age into 30-40, 41-59, and 60-70, and generated a forest plot. The sensitivity analysis in this study aimed to assess the potential impact of sex and age on the PRS-CS measurement results, ensuring the model's stability and reliability across different populations. By subdividing age groups, we were able to further understand the differences in performance among various age groups in polygenic risk assessment, thereby improving the accuracy of the prediction model and its clinical application value. Afterall, we also investigate on subgroup PRS to validate the roboutness of the predicting power.

3.6 SNP-to-Phenotype Association Analysis

In Section 3.6, we selected Phenome-wide association study (PheWAS) and GWAS Catalog to understand the relationship between SNP and pheotype. PheWAS takes a SNP-centric approach to explore associations with a range of diseases, allowing for the discovery of novel pleiotropic effects. Meanwhile, GWAS Catalog provides previously reported SNP–phenotype associations, helping to validate whether newly identified SNPs align with existing evidence and enhancing their biological credibility. By integrating both resources, we assess SNPs from the perspectives of known evidence and novel associations, offering a more comprehensive interpretation of their phenotypic relevance.

3.6.1 Phenome-Wide Association Study (PheWAS)

We implemented PheWAS using the official PheWAS R package and Plink (v1.9). This approach explores the association of a single SNP across a broad spectrum of clinical phenotypes [29]. In our study, we established 1,378 phenotype codes (phecodes) derived from aggregated ICD codes in TWB, aiming to identify novel gene-phenotype associations and pleiotropic effects. Covariates included age, sex, and the first five PCA to control for potential confounding factors. To account for multiple testing, false discovery rate (FDR) correction was applied.

doi:10.6342/NTU202500806

3.6.2 GWAS Catalog

We queried the NHGRI-EBI GWAS Catalog (www.ebi.ac.uk/gwas) where integrates GWAS trait annotations, Polygenic Score (PGS) Catalog and harmonized data in standardized formats to enable cross-study comparability. We cross-referenced SNPs identified in our study with existing literature and GWAS reports via the GWAS Catalog to identify validated phenotypic traits, shared risk loci, or overlapping biological pathways relevant to gallstone disease and related metabolic phenotypes.

3.7 Gene-Level Interpretation

3.7.1 Protein-protein Interaction Network

Protein-protein interaction (PPI) networks provide a systems-level view of how gene products interact within biological processes [27]. Disruptions in PPI networks are known to contribute to complex disease mechanisms and can aid in identifying therapeutic targets. We constructed a PPI network using the STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins; https://string-db.org), which integrates known and predicted protein–protein associations derived from experimental data, computational prediction, and text mining. A pre-selected gene set (UGT1A1, SLCO1A2, ABCG8, FUT2, CYP7A1, and SLCO1B1) was used as the input based on previous literature studies [11-14] for initial network construction. Additional genes identified from GWAS findings were subsequently added to enrich the network.

3.7.2 Pathway Enrichment Analysis

We conducted pathway enrichment analysis using the functional enrichment tools integrated within the STRING platform [27], including annotations from different pathway enrichment databases. Both default gene sets and GWAS-identified gene sets were analyzed within STRING to detect pathways significantly enriched with a false discovery rate (FDR)-adjusted p-value < 0.05. The enrichment results were summarized by several key metrics:

doi:10.6342/NTU202500806

Term ID provides a unique identifier representing each biological term or pathway within its specific database; Term description explains the biological processes, clinical biomarkers, or phenotypic characteristics associated with each term; Strength denotes the statistical enrichment level of each pathway, with higher values indicating stronger relevance; Signal indicates the intensity or significance of the evidence supporting each pathway.

3.8 Data Availability

The datasets used and analyzed in this study are not publicly accessible due to ethical confidentiality restrictions. However, they can be obtained upon reasonable request from the corresponding author, subject to approval from the ethics committee.

doi:10.6342/NTU202500806

Chapter 4

Results

4.1 Study Pipeline of The Genome-to-Phenome Approach

Our study (Figure 1) developed an end-to-end, multi-scale pipeline that starts with a rigorously QC'd GWAS in 108,403 Han participants from the Taiwan Biobank and drills all the way down to system-level biology. We performed a genome-wide scan with stratified heritability and LDSC summary checks. By weaving together locus-centric fine-mapping, genome-wide enrichment, SNP-phenome cross study comparison and protein-to-protein network analyses, this framework translates individual GWAS hits into coherent molecular mechanisms and clinical insights.

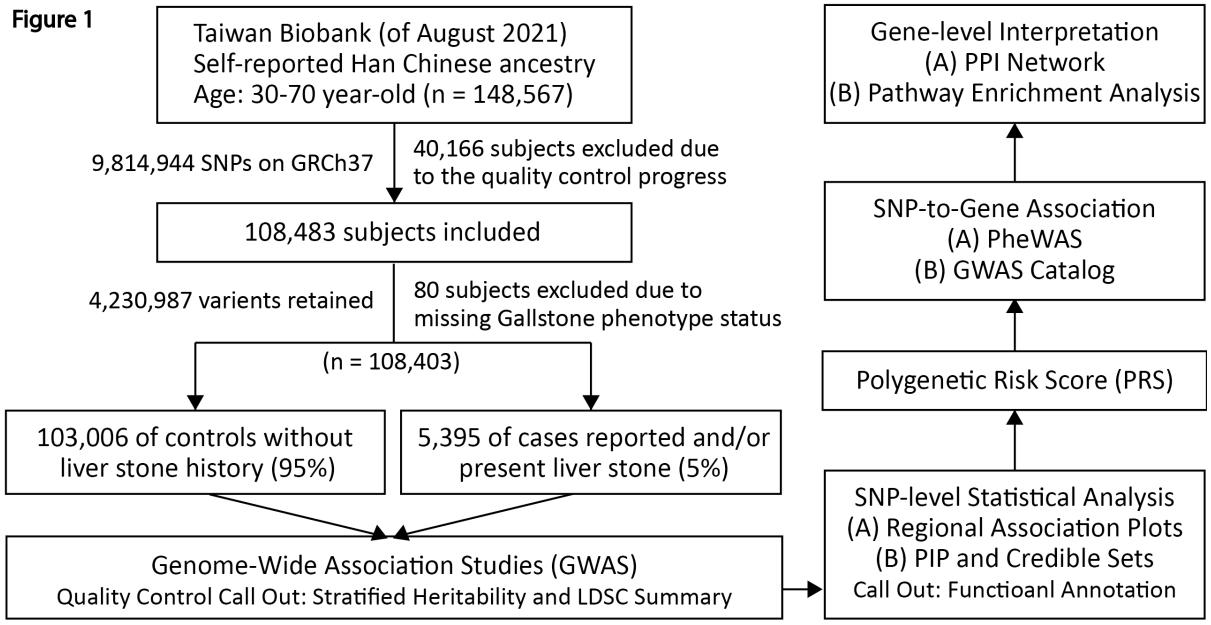


Figure 1. A flowchart of a Genome-to-Phenome approach with PRS Model in our study.

From the 148,576 participants, we excluded 80 due to contamination or close relatives and removed 40,166 based on quality control, leaving 108,483 for analysis. Among them, 5,395 had liver stone disease (4.97% prevalence), while 103,006 had no history of the condition. After filtering out low-quality SNPs, a total of 4,230,987 SNPs were retained. Cohort characteristics of participants are provided in supplemental Table 1.

doi:10.6342/NTU202500806

4.2 Genome-Wide Association Studies (GWAS)

Significant association signals were observed on chromosomes 3, 4, 8, 11, and 20, as visualized in the Manhattan plot (Figure 2). Table 1 summarizes 57 SNPs associated with GSD, including 38 novel variants. Each row in Table 1 includes the SNP ID, chromosomal location, alleles (minor/major), and minor allele frequency (MAF), with the minor allele used as the effect allele. Odds ratio (OR) and p-value represent the strength and significance of association, while beta and standard error (SE) indicate the effect size and its precision. Additional columns, including PIP and gene annotations.

Figure 2

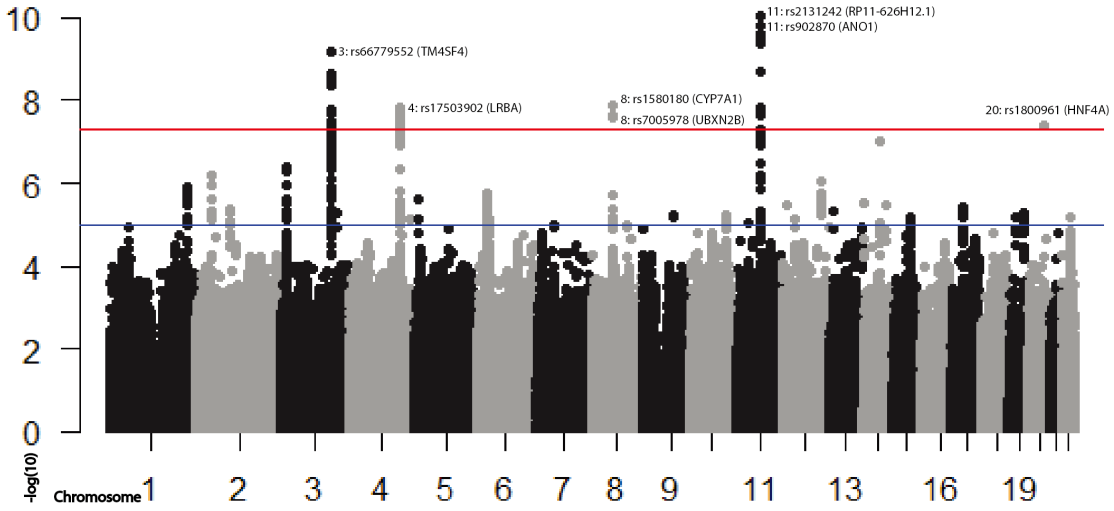


Figure 2. Manhattan plot for gallstone disease association results (N_{cases} = 5,395 and $N_{controls}$ = 103,008). Variants are plotted based on their chromosomal position (x-axis) and -log10 P-values (y-axis). The blue line represents the genome-wide significance threshold of 5×10^{-5}, while the red line is that of 5×10^{-8}. In summary, 7 lead SNPs were highlighted and marked with nearest genes or non-coding region.

Figure 3A

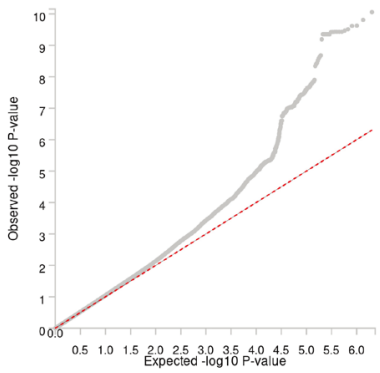


Figure 3B

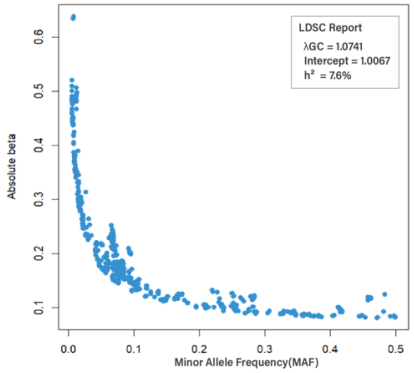


Figure 3. Evaluation on statistical calibration and polygenic contribution. (A) The Q-Q plot: the x-axis represents the -log10 of the expected P-values of the association from the chi-square distribution, and the y-axis represents the -log10 of the observed P-values from the observed chi-square distribution. (B) Scatter plot for the minor allele frequency (MAF) and absolute effect size(β) with LDSC summary report.

doi:10.6342/NTU202500806

Table 1. GWAS summary statistics of 57 Significant SNPs and fine mapping of posterior inclusion probabilities (PIPs) with 95% credible sets. The lead SNPs are marked with an asterisk, **were sourced from the GWAS Catalog and *** were novel SNPs in our findings.

Chr	Hg37 Position	SNP	min/maj	MAF	OR	p Value	Nearest Gene	beta	std	PIP	PMID (other reference)
3	149222751	rs80217587***	T/C	0.067	0.777	1.74E-08	TM4SF4	-0.253	0.045	0.202568	Novel
3	149222569	rs114173880***	A/G	0.067	0.778	2.09E-08	TM4SF4	-0.251	0.045	0.19998	Novel
3	149216778	rs80092430***	T/G	0.067	0.783	3.05E-08	TM4SF4	-0.244	0.044	0.196101	Novel
3	149208266	rs78927161***	T/A	0.069	0.787	3.78E-08	TM4SF4	-0.240	0.044	0.194437	Novel
3	149218466	rs4441598***	A/G	0.068	0.784	3.43E-08	TM4SF4	-0.243	0.044	0.194413	Novel
3	149221328	rs66779552*/***	A/G	0.488	1.133	6.43E-10	TM4SF4	0.125	0.020	0.131761	Novel
3	149209610	rs9870457***	A/G	0.484	1.127	2.32E-09	TM4SF4	0.120	0.020	0.114705	Novel
3	149212076	rs4681515**	A/G	0.485	1.126	2.63E-09	TM4SF4	0.119	0.020	0.114641	37705021
3	149211897	rs6774253**	G/C	0.485	1.125	3E-09	TM4SF4	0.118	0.020	0.113551	38116116
3	149212125	rs4681516	G/C	0.485	1.124	4E-09	TM4SF4	0.117	0.020	0.111633	1015-9584
3	149210443	rs9857970**	T/C	0.484	1.125	3.63E-09	TM4SF4	0.118	0.020	0.111215	29403010
3	149211512	rs12633863**	G/A	0.485	1.124	4.54E-09	TM4SF4	0.117	0.020	0.110437	30504769
3	149207934	rs9289788***	T/C	0.484	1.121	1.6E-08	TM4SF4	0.114	0.020	0.101057	Novel
3	149207968	rs9289789***	T/C	0.484	1.121	1.64E-08	TM4SF4	0.114	0.020	0.100883	Novel
4	151214005	rs17503902*/***	G/T	0.238	1.131	1.46E-08	LRBA	0.123	0.022	0.071083	Novel
4	151191830	rs4696660***	T/G	0.238	1.129	2.09E-08	LRBA	0.121	0.022	0.069053	Novel
4	151193201	rs1824406***	T/C	0.238	1.129	2.23E-08	LRBA	0.121	0.022	0.0687	Novel
4	150270468	rs4696659***	C/A	0.283	1.128	2.48E-08	LRBA	0.120	0.022	0.068093	Novel
4	151210344	rs7656639***	G/T	0.238	1.127	3.06E-08	LRBA	0.120	0.022	0.066945	Novel
4	151206319	rs6817154	C/T	0.239	1.127	3.36E-08	LRBA	0.120	0.022	0.066438	TreeWAS
4	151205643	rs2085722	C/T	0.238	1.127	3.51E-08	LRBA	0.120	0.022	0.066184	GRASP
4	151189878	rs1080195***	G/A	0.238	1.127	3.65E-08	LRBA	0.120	0.022	0.065962	Novel
4	151182830	rs4696646	A/T	0.24	1.128	3.65E-08	LRBA	0.120	0.022	0.065962	In Press
4	151189320	rs7686693***	A/C	0.238	1.127	3.72E-08	LRBA	0.120	0.022	0.065868	Novel
4	151183334	rs6852279***	A/G	0.236	1.128	3.78E-08	LRBA	0.120	0.022	0.065773	Novel
4	151207895	rs10520053***	A/G	0.238	1.126	4.03E-08	LRBA	0.119	0.022	0.065425	Novel
4	151193980	rs62344508***	C/A	0.238	1.127	4.07E-08	LRBA	0.120	0.022	0.065362	Novel
4	151188522	rs55728314***	C/T	0.238	1.126	4.46E-08	LRBA	0.119	0.022	0.064858	Novel
4	151199080	rs2290846**	A/G	0.237	1.126	4.94E-08	LRBA	0.119	0.022	0.064293	34651315

doi:10.6342/NTU202500806

(cont.) Table 1. GWAS summary statistics of 57 Significant SNPs and fine mapping of posterior inclusion probabilities (PIPs) with 95% credible sets.

Chr	Hg37 Position	SNP	min/maj	MAF	OR	p Value	Nearest Gene	beta	std	PIP	PMID (other reference)
8	59387337	rs1580180*	T/C	0.243	1.142	1.27E-08	CYP7A1	0.133	0.023	0.200762	10.21037/hbsn.2019.02.04
8	59382715	rs7005978*/**	A/G	0.246	1.139	2.24E-08	UBXN2B	0.130	0.023	0.199974	37563310
8	59392737	rs10107182**	C/T	0.245	1.137	2.58E-08	CYP7A1	0.128	0.023	0.199766	32042192
8	59385919	rs2326077**	C/T	0.244	1.139	2.59E-08	CYP7A1	0.130	0.023	0.199757	35213538
8	59384181	rs983812**	C/T	0.244	1.139	2.62E-08	CYP7A1	0.130	0.023	0.199741	34651315
11	69839305	rs2131242*/***	A/G	0.104	1.233	8.78E-11	RP11-626H12.1	0.209	0.032	0.055452	Novel
11	69839359	rs2131241***	A/T	0.24	1.135	2.27E-08	RP11-626H12.1	-	-	-	Novel
11	69838386	rs902870*/**	A/G	0.104	1.228	1.54E-10	RP11-626H12.1	0.205	0.032	0.052377	34651315
11	69834473	rs56363382**	T/C	0.104	1.225	2.36E-10	RP11-626H12.1	0.203	0.032	0.050124	34651315
11	69989443	rs72933806***	A/G	0.093	1.225	2.41E-10	RP11-626H12.2	0.203	0.032	0.050021	Novel
11	69838295	rs4603323***	A/G	0.107	1.222	3.08E-10	RP11-626H12.1	0.200	0.032	0.048733	Novel
11	69834127	rs72931793***	A/C	0.1	1.223	3.4E-10	RP11-626H12.1	0.201	0.032	0.048197	Novel
11	69837563	rs872838***	T/G	0.107	1.221	3.71E-10	RP11-626H12.1	0.200	0.032	0.047764	Novel
11	69837423	rs10501402	G/T	0.107	1.221	3.71E-10	RP11-626H12.1	0.200	0.032	0.047764	EWAS
11	69837095	rs74616534***	C/T	0.107	1.221	3.71E-10	RP11-626H12.1	0.200	0.032	0.047764	Novel
11	69837564	rs872839***	T/C	0.107	1.221	3.71E-10	RP11-626H12.1	0.200	0.032	0.047764	Novel
11	69835877	rs55691279***	A/G	0.107	1.221	3.75E-10	RP11-626H12.1	0.200	0.032	0.047698	Novel
11	69835887	rs56402676***	T/C	0.107	1.221	3.75E-10	RP11-626H12.1	0.200	0.032	0.047698	Novel
11	69834983	rs10501401***	G/A	0.107	1.22	4.4E-10	RP11-626H12.1	0.199	0.032	0.046873	Novel
11	69835461	rs17160502***	A/G	0.107	1.22	4.4E-10	RP11-626H12.1	0.199	0.032	0.046873	Novel
11	69835577	rs58680271***	G/T	0.107	1.22	4.4E-10	RP11-626H12.1	0.199	0.032	0.046873	Novel
11	69833853	rs72931788***	A/G	0.107	1.22	4.4E-10	RP11-626H12.1	0.199	0.032	0.046873	Novel
11	69834917	rs72931796***	T/C	0.107	1.22	4.4E-10	RP11-626H12.1	0.199	0.032	0.046873	Novel
11	69835266	rs72933804***	A/T	0.107	1.22	4.4E-10	RP11-626H12.1	0.199	0.032	0.046873	Novel
11	69840586	rs12290642	A/G	0.081	1.24	2.08E-09	RP11-626H12.1	0.215	0.036	0.039175	In Press
11	69841267	rs12282444***	A/G	0.083	1.229	1.42E-08	RP11-626H12.1	0.206	0.036	0.030528	Novel
11	69842821	rs12284684***	A/G	0.077	1.227	1.98E-08	RP11-626H12.1	0.205	0.036	0.029137	Novel
20	43042364	rs1800961*/**	T/C	0.016	1.477	3.89E-08	HNF4A	-	-	-	34887591

doi:10.6342/NTU202500806

4.3 Quality Control Analysis

To evaluate the presence of genomic inflation and ensure the validity of the association signals, we conducted a downstream sensitivity analysis using LDSC. As shown Figure 3A, the observed p-values deviate from the null expectation (red line), particularly in the tail region, indicating the presence of true associations rather than random noise. This initial inflation may raise concerns about confounding, such as population stratification. However, the LDSC intercept (1.0067, SE (Standard Error) = 0.0092) remains close to 1, and the λGC (1.0741) reflects only minimal inflation. These values suggest that the deviation is primarily driven by polygenicity rather than confounding bias. Furthermore, the estimated h^2 was 7.6%, supporting a polygenic architecture for gallstone disease. The mean chi-squared statistic was 1.0879, further confirming adequate statistical calibration. Population structure was also quantified, with population stratification explaining only 7.57% of the total variance (SE = 0.1052), indicating minimal impact. Figure 3B shows a trend where SNPs with lower MAF tend to exhibit larger effect sizes. This inverse relationship supports the biological relevance of the identified variants and strengthens the evidence for their contribution to gallstone susceptibility. Together, these results validate the reliability of our GWAS findings and confirm that the observed inflation is due to the underlying genetic complexity of GSD.

4.4 SNP-Level Statistical Analysis

We utilized regional association plots and functional fine-mapping methods to identify and prioritize SNP-level genetic signals. Instead of GWAS statistical lead SNP, from the observation of Figure 4, we prioritized 7 lead genetic-centric SNPs (rs66779552, rs17503902, rs1580180, rs7005978, rs2131242, rs902870, rs1800961). To further refine causal inference, we compute posterior inclusion probabilities (PIP) for each significant SNP and summarized in Table 1. SNP rs1800961 was excluded from PIP analysis due to the absence of additional significant SNPs on chromosome 20.

doi:10.6342/NTU202500806

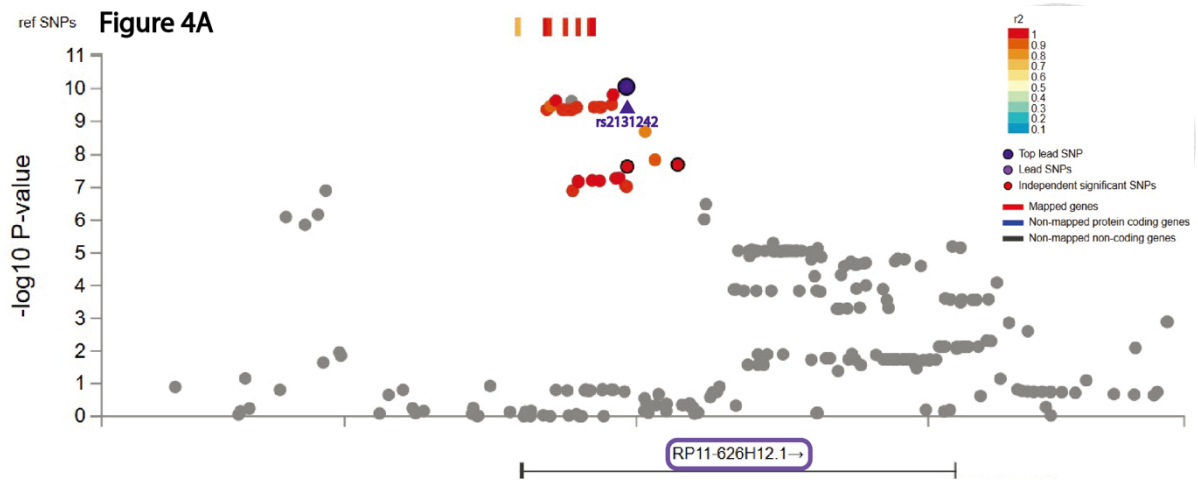


Figure 4A. Regional association plot for rs2131242 on chromosome 11. It shows 3 independent significant SNPs. The nearest SNP cluster is RP11-626H12.1 gene region.

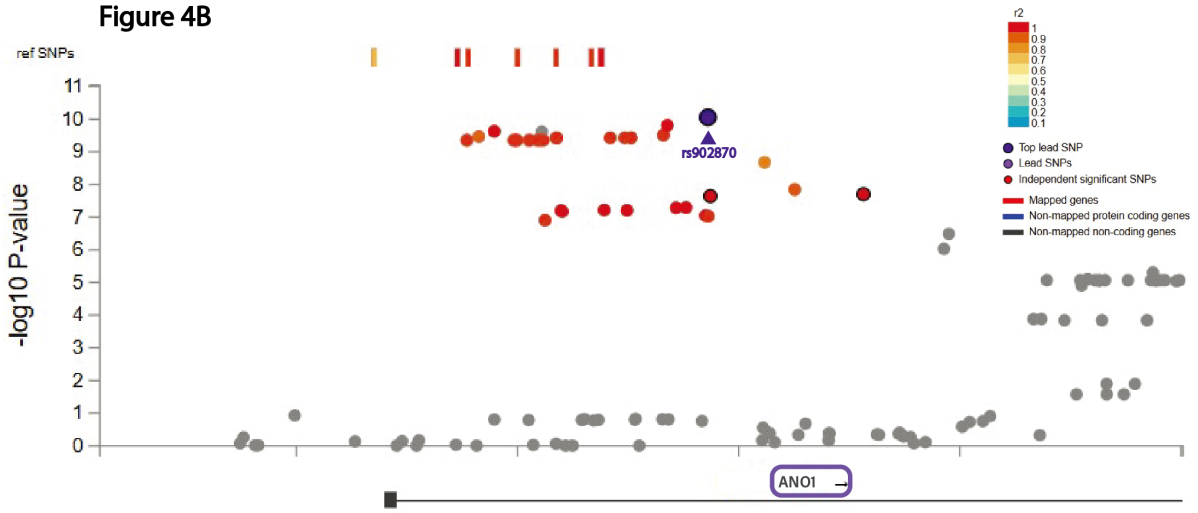


Figure 4B. Regional association plot for rs902870 on chromosome 11. It shows 3 independent significant SNP. The nearest gene annotation is ANO1.

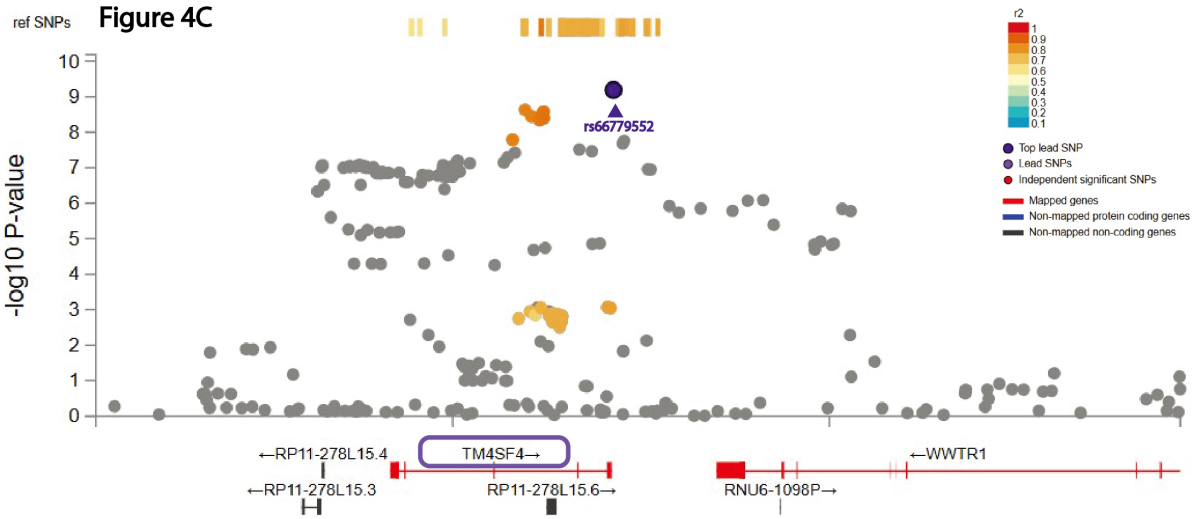


Figure 4C. Regional association plot for rs66779552 on chromosome 3. It shows one independent significant SNP. The nearest gene annotation is TM4SF4.

doi:10.6342/NTU202500806

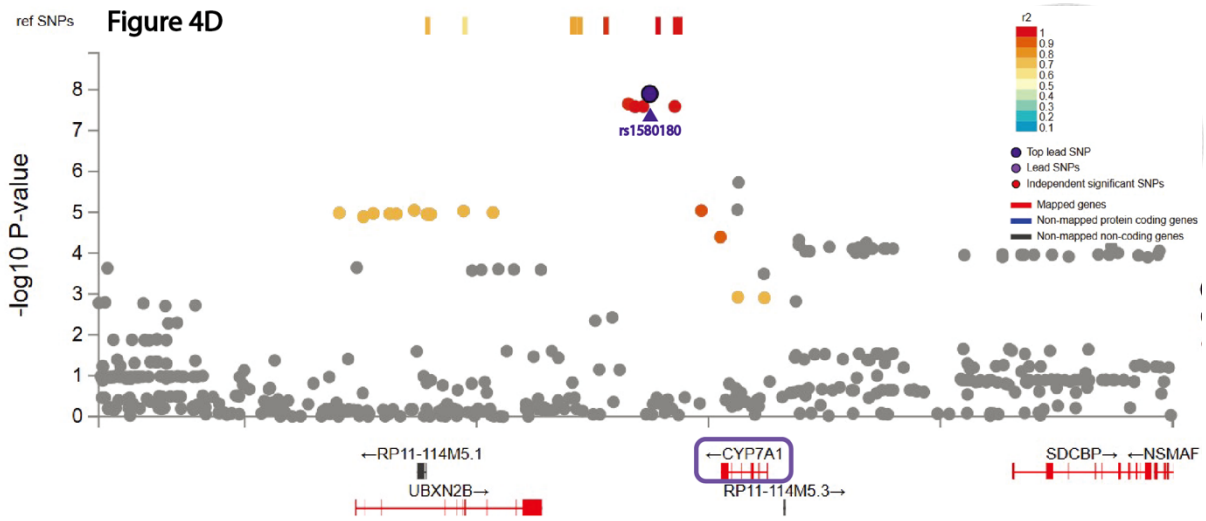


Figure 4D. Regional association plot for rs1580180 on chromosome 8. It shows one independent significant SNP. The nearest gene annotation is CYP7A1.

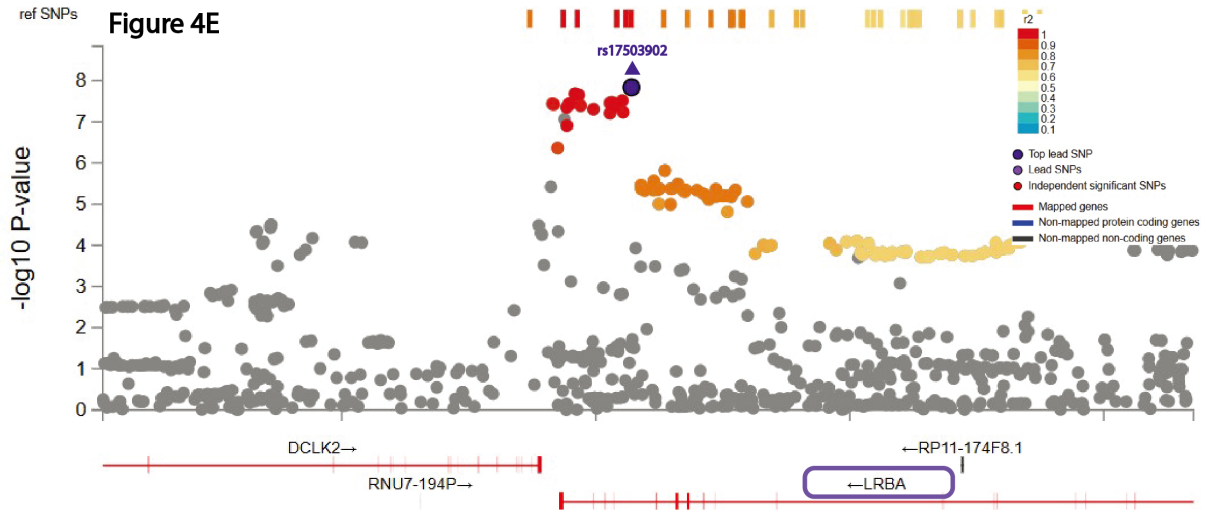


Figure 4E. Regional association plot for rs17503902 on chromosome 4. It shows one independent significant SNP. The nearest gene annotation is LRBA.

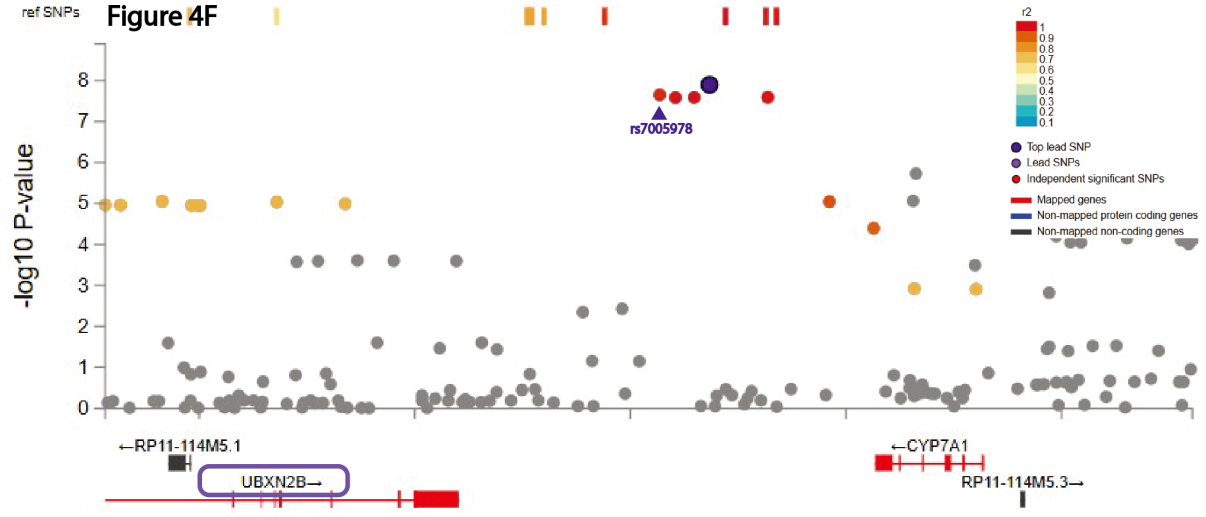


Figure 4F. Regional association plot for rs7005978 on chromosome 8. It shows one independent significant SNP. The nearest gene annotation is UBXN2B.

doi:10.6342/NTU202500806

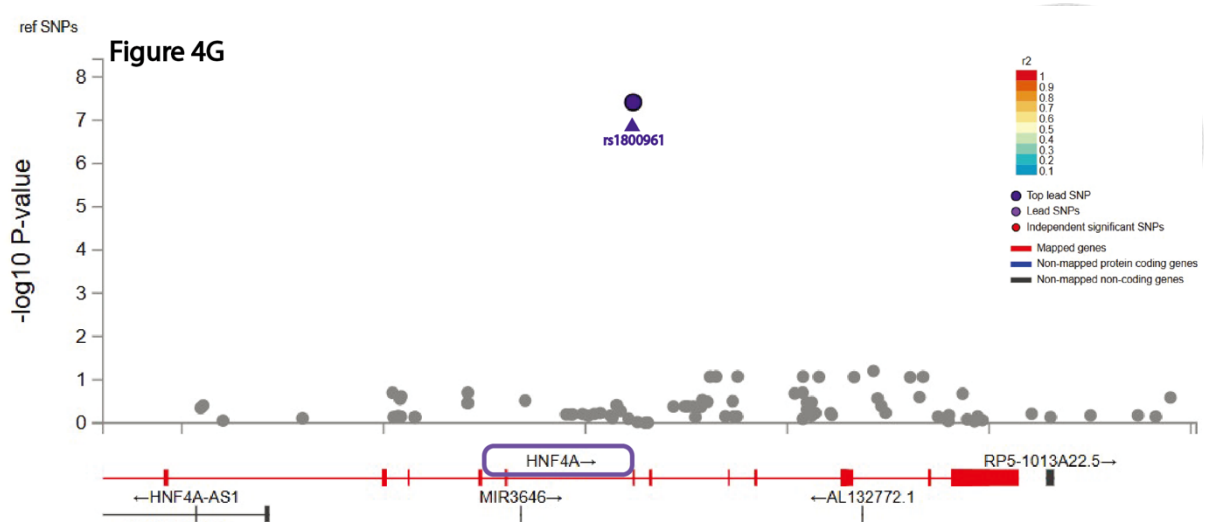


Figure 4G. Regional association plot for rs1800961 on chromosome 20. It shows one independent significant SNP. The nearest gene annotation is HNF4A.

To further refine causal inference, we compute posterior inclusion probabilities (PIP) for each significant SNP and summarized in Table 1. SNP rs1800961 was excluded from PIP analysis due to the absence of additional significant SNPs on chromosome 20. Figure 5A comprised 21 SNPs on chromosome 11 around ~69.9 Mbp, with moderate PIPs (~0.05) and relatively lower average LD (R = 0.698), potentially indicative of multiple weak-effect variants or allelic heterogeneity within this locus. The highest prioritized SNP on chromosome 4 was rs17503902, with a PIP of 0.071 (Figure 5B). SNP rs80217587 on chromosome 3 exhibited the highest overall PIP (0.2026), implicating it as a likely causal variant for GSD, which the finding differed from earlier analyses that prioritized rs66779552 as the lead SNP based solely on GWAS p-values and effect sizes (Figure 5C). Significant SNPs on chromosome 8 (Figure 5D) presented a defined credible set (R = 0.943) characterized by consistently high PIP values (~0.20). As shown in Figure 5E, the identified SNPs are most strongly enriched in intronic regions (~63%, $p < 0.05$), with additional significant enrichment in intergenic (~17%, $p < 0.05$) and exonic non-coding RNA regions (~5%, $p < 0.05$), consistent with a role in gene regulatory mechanisms.

doi:10.6342/NTU202500806

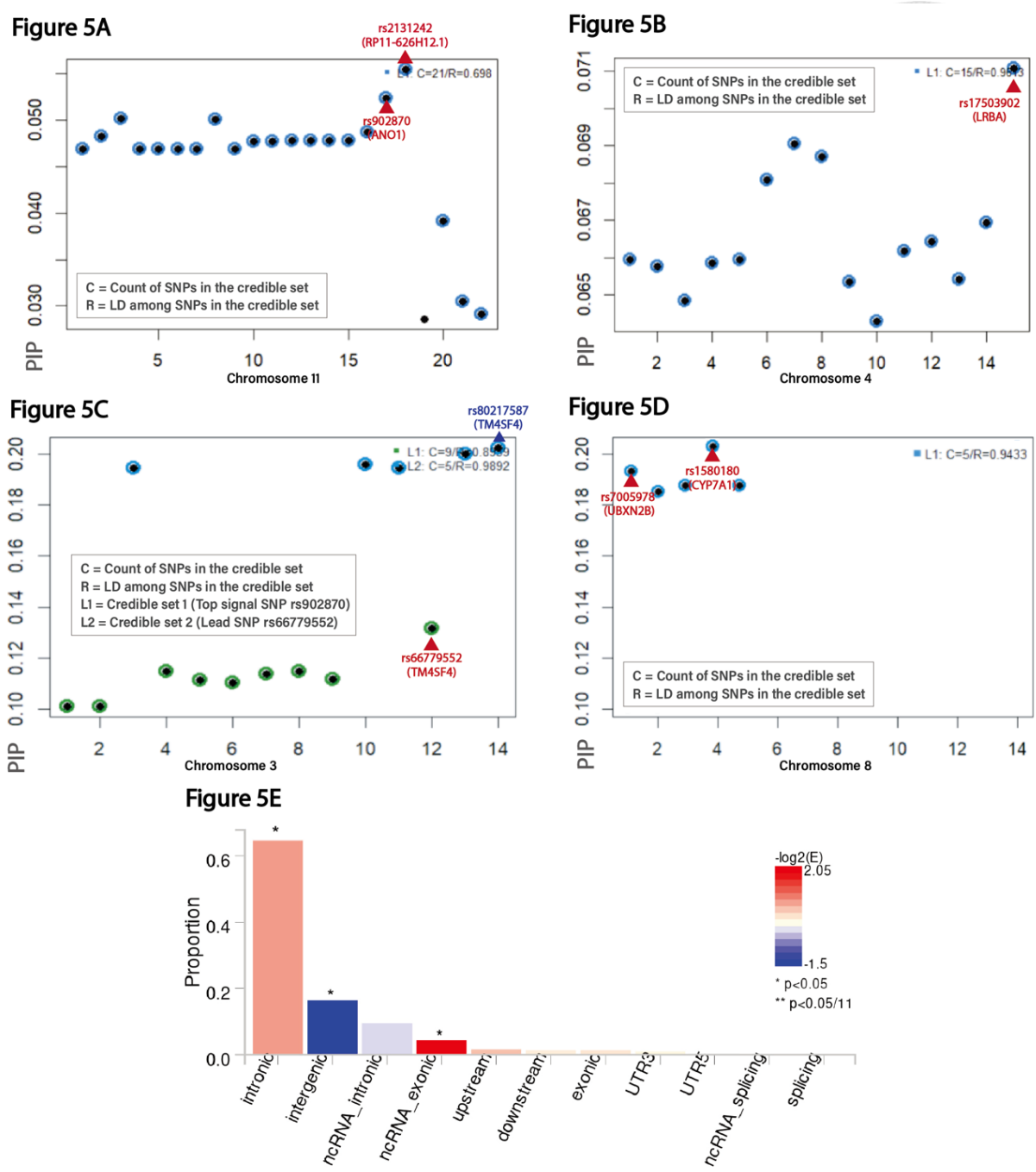


Figure 5. Fine-mapping results of posterior inclusion probability (PIP) for SNP identifiers in credible sets across chromosomes. The x-axis indicates each SNP identifier, and the y-axis represents its respective PIP. The lead SNP for each credible set is highlighted in red, with SNP rs80217587 (marked in blue) shown in blue due to its highest PIP. Each credible set is summarized by the total number of SNPs (C) included and the average linkage disequilibrium (R, average LD) among SNPs within the credible set, with values closer to 1 indicating stronger internal correlation. (A) Credible set on chromosome 11 (Lead SNP: rs2131242; C=21, R=0.698). (B) Credible set on chromosome 4 (Lead SNP: rs17503902; C=15, R=0.9543). (C) Two credible sets on chromosome 3: set 1 (Lead SNP: rs902870; C=9, R=0.8992) and set 2 (Lead SNP: rs66779552; C=5, R=0.9892). SNP rs80217587 (blue), annotated to TM4SF4, exhibited the highest PIP among all SNPs identified. (D) Credible set on chromosome 8 (Lead SNP: rs1580180; C=5, R=0.9433). (E) Distribution of GWAS variants across functional categories with bars shaded according to their $\log_2$ enrichment values. Asterisks indicate categories significantly enriched or depleted relative to all tested variants via Fisher's exact test.

doi:10.6342/NTU202500806

4.4 Polygenic Risk Score (PRS)

Table 2 compares the predictive performance of five polygenic risk score (PRS) models, including four PRS-CT models with varying p-value thresholds and one PRS-CS model. Among PRS-CT models, looser thresholds ($P < 0.1$, $P < 0.05$) yielded moderate discriminative ability, with the best adjusted model ($P < 0.05$) achieving an AUC of 0.71. In contrast, more stringent thresholds ($P < 0.001$) reduced predictive accuracy, likely due to the exclusion of informative SNPs. This suggests that PRS-CT performance is sensitive to parameter tuning and benefits from appropriate covariate adjustment. Figure 6A illustrates the improvement gained through covariate adjustment, where the crude AUC of 0.66 increases to 0.71 after adjustment, highlighting the influence of non-genetic factors. However, Figure 6B shows that PRS-CS, a Bayesian-based model that adaptively shrinks SNP effect sizes based on LD structure, achieved superior and stable predictive power with an AUC of 0.98 in both crude and adjusted models. This indicates that PRS-CS is more robust to confounding and parameter sensitivity. Collectively, these results demonstrate that GSD risk can be effectively predicted using polygenic scores, particularly through the PRS-CS model. The improved performance in confounder-adjusted models further supports that both genetic and contextual factors contribute to GSD susceptibility in Han Chinese.

PRS Methods	Tuning Parameters	AUC (95% CI)	
		Crude	Cofounder
PRS-CT	$P < 0.5$	0.59 (0.5809, 0.5971)	0.66 (0.6529, 0.6676)
	$P < 0.1$	0.66 (0.6478, 0.6630)	0.70 (0.6970, 0.7109)
	$P < 0.05$*	0.66 (0.6518, 0.6668)	0.71 (0.7010, 0.7148)
	$P < 0.001$	0.57 (0.5649, 0.5807)	0.65 (0.6456, 0.6599)
PRS-CS**	adaptive shrinkage	0.98(0.9789-0.9819)	0.98(0.9819-0.9846)

Table 2. Comparison of the predictive performance of different Polygenic Risk Score (PRS) models. This table compares the performance of various PRS methods, including Clumping and Thresholding (C+T) with different p-value thresholds and the Visualization of this table will be *Figure 6A and **Figure 6B.

doi:10.6342/NTU202500806

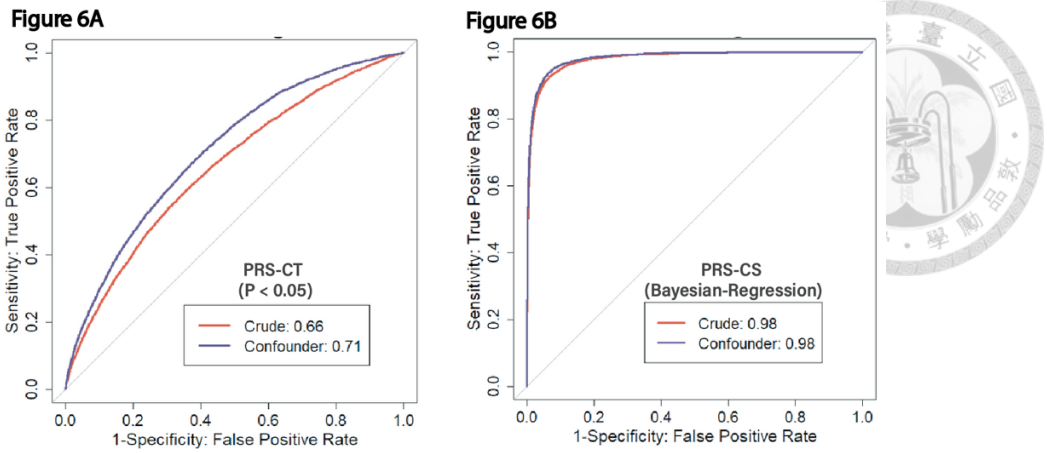
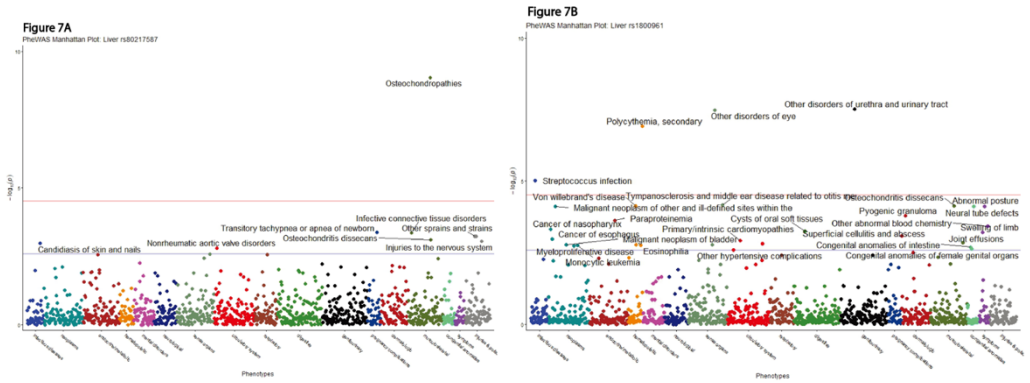


Figure 6. Receiver Operating Characteristic (ROC) Curves Comparing Predicting GSD PRS Models with and without Confounders. The ROC curve compares the crude model (red line), which uses PRS alone, and the adjusted model (blue line), which incorporates age and sex as confounders. (A) The adjusted model shows an improvement in PRS-CT performance, with the Area Under the Curve (AUC) increasing from 0.66 to 0.71. (B) The ROC curve demonstrates the performance of the Continuous Shrinkage (PRS-CS) method, achieving near-perfect AUC values of 0.98 for both crude and adjusted models.

4.6 SNP-to-Phenotype Association

We identified TM4SF4, LRBA, UBXN2B, CYP7A1, HNF4A, and ANO1 as the nearest genes mapped to the 7 lead SNPs, as well as the SNP rs80217587 with the highest PIP(Figure 5C). We then applied PheWAS and leveraged the GWAS Catalog to investigate phenotypic associations for these SNPs. In the PheWAS analysis, SNP rs80217587 (mapped to TM4SF4, Figure 7A) showed associations osteochondropathies. For the lead SNP rs1800961 (mapped to HNF4A, Figure 7B), we observed a genome-wide significant association with streptococcus infection, polycythemia, disorders of urethra and urinary tract, and eye disorders. It showed GSD relates to heterogeneous multisystemic disorders. PheWAS results for other lead SNPs are provided in Supplemental Figure 2.



doi:10.6342/NTU202500806

Figure 7. Phenome-Wide Association Study (PheWAS) Disease Model Interpretation. The blue line represents the genome-wide significance threshold of 5×10^{-5}, while the red line indicates the threshold of 5×10^{-8}. (A) rs80217587. (B) rs1800961.

To further investigate the pleiotropic effects of these SNPs, we cross-referenced our findings with the GWAS Catalog. Among the 57 genome-wide significant SNPs identified, 12 had prior associations documented in the Catalog (Table 3 and Supplementary Tables 2A–2E). Notably, rs1800961 in HNF4A showed a well-established link with high-density lipoprotein (HDL) cholesterol ($P = 1.0 \times 10^{-268}$), consistent with its role in lipid metabolism and liver function. Another lead SNP, rs7005978, was linked to fibroblast growth factor 19 (FGF19) levels, a hormone involved in bile acid regulation, further supporting its biological relevance to gallstone pathogenesis. By integrating this catalog data with Figure 6C and Table 1, we observed two distinct credible sets at the TM4SF4 locus with opposite effect directions and the novel signal related to GSD susceptibility and musculoskeletal phenotypes (Figure 7A). L1, led by rs80217587 with a negative β, includes only novel SNPs not previously reported in the GWAS Catalog, suggesting a potentially East Asian–specific protective signal. In contrast, L2, led by rs66779552 with a positive β, includes multiple SNPs that have been previously associated with GSD and liver enzyme levels, indicating a known risk signal supported by existing GWAS evidence. Overall, recurrent associations of TM4SF4, CYP7A1, and HNF4A across our study and prior GWAS highlight their potential as key regulators of bile acid metabolism, lipid homeostasis, and gallstone disease pathogenesis in both Western and East Asian populations.

chr	Risk Allele	p-Value	Mapped Genes	Trait Name
3	**rs4681515**	1.0E-56	TM4SF4	GSD or coronary artery disease
3	**rs12633863**	4.0E-30	TM4SF4	GSD
3	**rs9857970**	2.0E-20	TM4SF4	Gamma glutamyl transferase levels
3	**rs6774253**	5.0E-17	TM4SF4	Gamma glutamyl transferase levels
4	**rs2290846-A**	5.0E-46	LRBA	GSD

doi:10.6342/NTU202500806

8	**rs983812-T**	1.0E-42	UBXN2B, CYP7A1	GSD
8	**rs10107182-T**	3.0E-22	CYP7A1, UBXN2B	Sex hormone-binding globulin levels
8	**rs7005978-A***	1.0E-17	CYP7A1, UBXN2B	Fibroblast growth factor 19 levels
8	**rs2326077-C**	4.0E-12	UBXN2B, CYP7A1	Concentration of VLDL particles
11	**rs56363382-T**	7.0E-16	ANO1	GSD or coronary artery disease
11	**rs902870***	5.0E-09	ANO1	cholelithiasis
20	**rs1800961***	1.0E-268	HNF4A	High density lipoprotein cholesterol

Table 3. 12 Reported SNPs in GWAS Catalog and Its Most Significant Trait. The lead SNPs are marked with an asterisk (*).

4.7 Gene Level Interpretation

The repeated identification of TM4SF4, CYP7A1, and HNF4A in our previous analysis highlights their potential as key regulators in GSD. We constructed a PPI network combining well-known GSD gene set and target gene set. TM4SF4 was selected due to its SNP rs80217587, which showed the highest posterior inclusion probability (PIP = 0.202568) across all variants. Two credible sets were mapped to this locus with opposite effect directions—rs80217587 indicating a protective effect, and rs66779552 associated with increased risk—suggesting allelic heterogeneity and cis-regulatory divergence, and underscoring TM4SF4 as a functionally important locus in gallstone disease (GSD). CYP7A1 was prioritized based on the top credible set on chromosome 8 (average PIP = 0.1998), including the lead SNP rs1580180 which established role in bile acid metabolism makes it a strong biological candidate for GSD. HNF4A was included for its SNP rs1800961, which showed pleiotropic associations in the PheWAS and has been reported in prior TWB-based GSD studies. As a hepatic transcriptional regulator, HNF4A adds both statistical and functional support to GSD pathogenesis. The resulting PPI network contained 13 edges, an average node degree of 3.25, and a clustering coefficient of 0.633, indicating a modular architecture.

doi:10.6342/NTU202500806

Figure 8

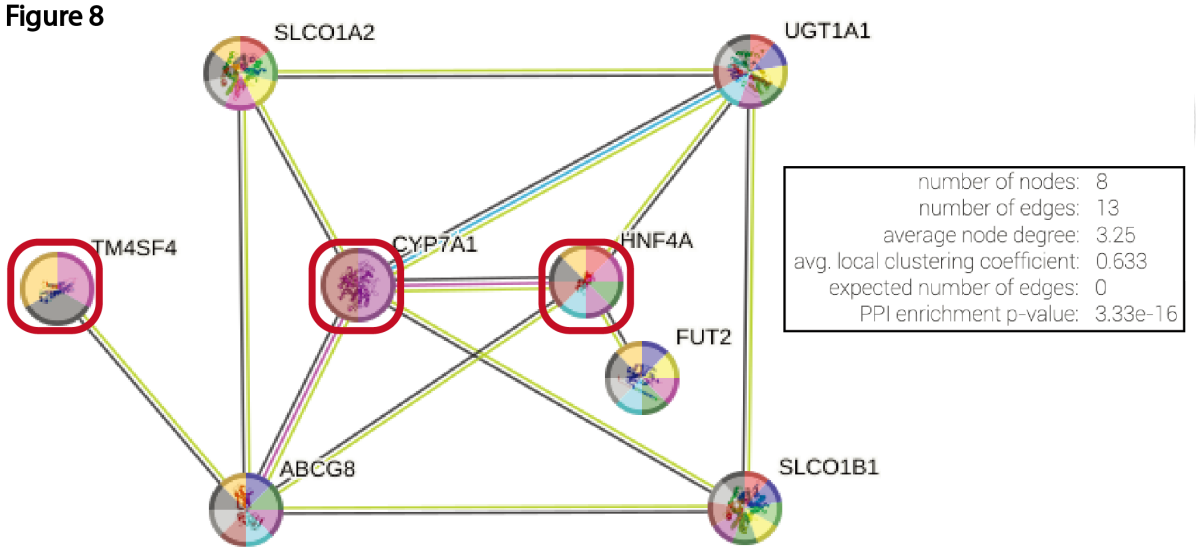


Figure 8. The protein-to-protein network diagram. Target genes are in the red circles and the rest of the genes are default gene set mentioned in section method - The protein-to-protein (PPI) network.

The highly significant PPI enrichment p-value (3.33×10^{-16}) further supported the presence of meaningful interactions. TM4SF4 interacted directly with key bile acid metabolism genes (CYP7A1 and ABCG8), suggesting its potential involvement in cholesterol metabolism and bile acid regulation within liver physiology. CYP7A1, a well-established rate-limiting enzyme in bile acid synthesis, served as a central hub directly connected with HNF4A, UGT1A1, and ABCG8, emphasizing its critical role in cholesterol metabolism and gallstone pathogenesis. Gene HNF4A, as a key transcription factor involved in liver-specific gene regulation and lipid homeostasis [7-8], reinforced its role as a regulatory node within the gallstone-associated gene network.

Term ID	Term Description	Strength	Signal	FDR	Target analyzing genes involved [f]		
					CYP7A1	HNF4A	TM4SF4
GO:0042632	Cholesterol homeostasis	1.92	0.9	0.028	V	V	
EFO:0005278	Cardiovascular disease biomarker measurement	0.89	0.64	0.007		V	V
EFO:0004532	Serum gamma-glutamyl transferase measurement	1.49	1.03	0.008	V	V	
HP:0001392	Abnormality of the liver	1.14	0.76	0.013	V	V	

doi:10.6342/NTU202500806

HP:0032180	Abnormal circulating metabolite concentration	1.09	0.67	0.022	V	V	
EFO:0004582	Liver enzyme measurement	1.04	0.63	0.026		V	V
BTO:0000759	Liver	0.91	0.71	0.003	V	V	V

Table 4. Pathway enrichment analysis summary. The column "Target analyzing genes involved" indicates whether specific genes are directly associated with the identified pathways, marked by a "V" to denote relevance. TM4SF4 is the target gene we focused on.

doi:10.6342/NTU202500806

Chapter 5

Discussion

5.1 Novel Findings

We identified 57 GWAS significant SNPs, among which 38 represent novel associations not previously reported in both western and eastern literatures. In Figure 4, we found 4 novel genes in Asian population (TM4SF4, LRBA, UBXN2B and ANO1), also SNP rs1800961 reported in previous TWB clinical study [7]. In Figure 5C, we identified rs80217587 the most promising disease SNP because of its highest PIP value of all. Plus, we observed 2 credible sets in this locus, in Table 1 which exhibit opposite effect directions: with rs80217587 (blue set) showing a negative effect, and rs66779552 (green set) vice versa. This finding suggests that the TM4SF4 locus may influence GSD susceptibility through multiple regulatory mechanisms, potentially by: (i) acting on distinct regulatory elements (e.g., promoters or enhancers), or (ii) affecting alternative transcripts or isoforms of TM4SF4.

5.2 Clinical Significance and Intervention Strategy

PRS-CS model (AUC = 0.98) provides robust evidence supporting its application in genetic screening for GSD, potentially enabling the identification of high-risk individuals and facilitating timely intervention and personalized management strategies. The results from Table 4 revealed biomarkers associated with our candidate genes: GGT(EFO:0004532), reflecting liver function and oxidative stress; cholesterol markers(GO:0042632), including TC, LDL, and HDL, indicating lipid metabolism status; bile acids, bilirubin precursors, and oxysterols (HP:0032180), directly marking bile acid synthesis and metabolic balance; and classic liver enzymes (ALT, AST and ALP, EFO:0004582, HP:0001392), indicative of hepatobiliary inflammation and stress. These suggest the feasibility of developing laboratory test or early detection strategies by integrating genetic markers [30, 31]. The biomarkers

doi:10.6342/NTU202500806

emphasize the intertwined roles of metabolic dysregulation and hepatic function in the pathogenesis of GSD. These findings suggest that modifying metabolic and hepatic pathways may help prevent disease onset. Diet directly affects lipid metabolism, diets high in saturated fats and refined sugars promote insulin resistance and hepatic cholesterol production, increasing bile cholesterol saturation. In contrast, fiber-rich and plant-based diets enhance lipid clearance and reduce inflammation. The liver, as the central organ in bile acid synthesis and cholesterol regulation, responds dynamically to dietary inputs. For instance, CYP7A1, is transcriptionally regulated by dietary cholesterol and bile acid levels [6, 8, 29]. liver enzymes such as ALT and GGT, commonly used as blood biomarkers, reflect liver stress induced by poor dietary habits. Thus, leveraging gene–diet interventions present a clinically relevant pathway to GSD precisive medicine prevention [10, 11, 31]. In supplemental figure 3, we noticed he occurrence of gallstones is closely related to age, with risk generally increasing as age advances due to the accumulation of exposure factors such as diet, obesity, and metabolic issues. However, the relative influence of genetic risk is usually more pronounced in younger populations because they are less affected by environmental and lifestyle factors, making genetics a larger contributing factor [32]. Gallstone incidence is generally higher in females than in males, partly due to the influence of female hormones on cholesterol metabolism. Nevertheless, male patients often carry a stronger genetic predisposition, while female GSD formation is more influenced by environmental and hormonal factors such as pregnancy [33].

5.3 Study Limitation

One limitation of this study lies in the uncertainty of causal inference due to potential mismatches between the study population and the reference LD panel, which may obscure population-specific genetic signals as Figure 5C and the GWAS summary result show this paradox. Second, the scope of the PheWAS was limited to 1,378 phenotype codes, despite newer versions including over 1,800 diseases. This constraint may have hindered the

doi:10.6342/NTU202500806

identification of additional pleiotropic associations. Future studies incorporating a broader 'phecode' range could uncover additional genotype–phenotype links. Lastly, we observed a notable difference between PRS models. Incorporating covariates such as sex and age improved AUC in the PRS-CT model, but had no effect in the PRS-CS model. This discrepancy likely reflects differences in model architecture highlighting the importance of selecting appropriate PRS modeling strategies based on study goals and population characteristics.

5.4 Future Study Suggestion

Functional validation of TM4SF4 is essential. Tools such as CRISPR-Cas9 knockout or overexpression systems could clarify its role in bile metabolism, epithelial barrier integrity, and oxidative stress, providing valuable insights for the development of evidence-based diagnostic or screening tools. Additionally, clinical studies should assess the utility of blood-based biomarkers, including GGT, CRP, ALT, and bile acid profiles, for the early detection of GSD. These biomarkers may prove particularly useful for individuals carrying high-risk variants in TM4SF4, CYP7A1, or HNF4A. Furthermore, further investigation into novel genes identified in this study, such as UBXN2B, LRBA, and ANO1, is warranted, especially as they were first reported in East Asian populations and may represent population-specific susceptibility loci. Notably, based on our gene-centric definition of lead SNPs, rs7005978 (nearest to UBXN2B) and rs1580180 (nearest to CYP7A1) were each identified as the most statistically significant variants proximal to their respective genes. However, both SNPs fall within the same credible set and exhibit a high linkage disequilibrium (R = 0.9433), suggesting a likely shared association signal. Functional studies are needed to determine whether CYP7A1, UBXN2B, or both contribute to GSD pathogenesis. It is plausible that one gene (CYP7A1) may act as the functional effector, while the other (UBXN2B) serves as a proximal tag in East Asian populations. This underscores the need for downstream validation

doi:10.6342/NTU202500806

to identify the true causal gene driving the observed risk association. Finally, to validate the accuracy of the PRS-CS risk score across populations, we could utilize Asian cohorts from international databases, such as AllofMe, which includes 3% Asian individuals.

5.5 Conclusion

Collectively, these results support TM4SF4 as a regulatory hub, harboring both risk- and protective-associated variants within the same genomic region. Such bidirectional genetic effects underscore the presence of allelic heterogeneity and possible cis-regulatory divergence, highlighting the complex regulatory landscape of this locus and reinforcing TM4SF4 as a functionally relevant candidate gene for gallstone disease.

doi:10.6342/NTU202500806

Chapter 6

Ethics Statement

6.1 Competing interests and Ethic Approval and Patient Consent

The authors declare that they have no competing interests. This research was approved by the ethics committee of National Taiwan University Hospital Institutional Review Board. The study was conducted with the principles of the Declaration of Helsinki and the Good Clinical Practice Guidelines, and all the participants were informed consent.

6.2 Contribution

Chien-Chang Lee supervised and designed the study, obtained funding, drafted the analytical plan, guided the statistical analysis, interpreted the data. Chin-Hua Su and Yueh-Chun Tsai performed the statistical analysis. Yueh-Chun Tsai drafted the manuscript, interpreted the data, and critically revised the manuscript. Chien-Chang Lee and Shih-Wei Liao contributed to the revision of the manuscript and provided insights into the contents.

doi:10.6342/NTU202500806

Chapter 7

Bibliography

[1] Peery, A. F., Crockett, S. D., Murphy, C. C., Jensen, E. T., Kim, H. P., Egberg, M. D., Lund, J. L., Moon, A. M., Pate, V., Barnes, E. L., Schlusser, C. L., Baron, T. H., Shaheen, N. J., & Sandler, R. S. (2022). Burden and Cost of Gastrointestinal, Liver, and Pancreatic Diseases in the United States: Update 2021. Gastroenterology, 162(2), 621–644. https://doi.org/10.1053/j.gastro.2021.10.017

[2] Shabanzadeh, D. M. (2018). Incidence of gallstone disease and complications. Current Opinion in Gastroenterology, 34(2), 81–89. https://doi.org/10.1097/MOG.0000000000000418

[3] Di Ciaula, A., Wang, D. Q.-H., & Portincasa, P. (2019). Cholesterol cholelithiasis: part of a systemic metabolic disease, prone to primary prevention. Expert Review of Gastroenterology & Hepatology, 13(2), 157–171. https://doi.org/10.1080/17474124.2019.1549988

[4] Loor, M. M., Morancy, J. D., Glover, J. K., Beilman, G. J., & Statz, C. L. (2017). Single-setting endoscopic retrograde cholangiopancreatography (ERCP) and cholecystectomy improve the rate of surgical site infection. Surgical Endoscopy, 31(12), 5135–5142. https://doi.org/10.1007/s00464-017-5579-9

[5] Welfare, N. H. I. A. M. of H. and. (n.d.). Handbook of Taiwan's National Health Insurance. National Health Insurance Administration Ministry of Health and Welfare. Retrieved January 27, 2025, from https://www.nhi.gov.tw/en/lp-60-2.html

[6] Ferkingstad, E., Oddsson, A., Gretarsdottir, S., Benonisdottir, S., Thorleifsson, G., Deaton, A. M., Jonsson, S., Stefansson, O. A., Norddahl, G. L., Zink, F., Arnadottir, G. A., Gunnarsson, B., Halldorsson, G. H., Helgadottir, A., Jensson, B. O., Kristjansson, R.

doi:10.6342/NTU202500806

P., Sveinbjornsson, G., Sverrisson, D. A., Masson, G., … Stefansson, K. (2018). Genome-wide association meta-analysis yields 20 loci associated with gallstone disease. Nature Communications, 9(1), 5101. https://doi.org/10.1038/s41467-018-07460-y

[7] Lin, Y.-C., Chen, I.-C., Chen, Y.-J., Lin, C.-T., Chang, J.-C., Wang, T.-J., Chen, Y.-M., & Lin, C.-H. (2024). Association between HNF4A rs1800961 polymorphisms and gallstones in a Taiwanese population. Journal of Gastroenterology and Hepatology, 39(2), 305–311. https://doi.org/10.1111/jgh.16426

[8] Fairfield, C. J., Drake, T. M., Pius, R., Bretherick, A. D., Campbell, A., Clark, D. W., Fallowfield, J. A., Hayward, C., Henderson, N. C., Iakovliev, A., Joshi, P. K., Mills, N. L., Porteous, D. J., Ramachandran, P., Semple, R. K., Shaw, C. A., Sudlow, C. L. W., Timmers, P. R. H. J., Wilson, J. F., … Harrison, E. M. (2022). Genome-wide analysis identifies gallstone-susceptibility loci including genes regulating gastrointestinal motility. Hepatology (Baltimore, Md.), 75(5), 1081–1094. https://doi.org/10.1002/hep.32199

[9] Zhu, S. J., & Ding, Z. (2024). Association between gut microbiota and seven gastrointestinal diseases: A Mendelian randomized study. The Journal of Gene Medicine, 26(1), e3623. https://doi.org/10.1002/jgm.3623

[10] Borges, A. S. G. (2024). Exploring the impact of the human FUT2 gene on gut lactobacilli and bifidobacteria to improve gut health through personalised probiotics (Doctoral dissertation). Christian-Albrechts-Universität zu Kiel, Department of Microbiology and Biotechnology, Kiel, Germany.

[11] Di Ciaula, A., Garruti, G., Frühbeck, G., De Angelis, M., de Bari, O., Wang, D. Q.-H., Lammert, F., & Portincasa, P. (2019). The Role of Diet in the Pathogenesis of Cholesterol Gallstones. Current Medicinal Chemistry, 26(19), 3620–3638. https://doi.org/10.2174/0929867324666170530080636

[12] Wang, X., Yu, W., Jiang, G., Li, H., Li, S., Xie, L., Bai, X., Cui, P., Chen, Q., Lou, Y., Zou, L., Li, S., Zhou, Z., Zhang, C., Sun, P., & Mao, M. (2024). Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis. Clinical Gastroenterology and Hepatology, 22(8), 1586–1595. https://doi.org/10.1016/j.cgh.2024.01.051

[13] Wei, C.-Y.; Yang, J.-H.; Yeh, E.-C.; Tsai, M.-F.; Kao, H.-J.; Lo, C.-Z.; Chang, L.-P.; Lin, W.-J.; Hsieh, F.-J.; Belsare, S. Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. NPJ Genom. Med. 2021, 6, 1–10.

[14] Lin, L.-Y., Warren-Gash, C., Smeeth, L., & Chen, P.-C. (2018). Data resource profile: the National Health Insurance Research Database (NHIRD). Epidemiology and Health, 40, e2018062. https://doi.org/10.4178/epih.e2018062

[15] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (3), 559–575. doi:10.1086/519795

[16] Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J. Open Source Softw. 3, 731. doi:10.21105/joss.00731

[17] Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015; 47: 291–295.

[18] L Kachuri, N Chatterjee, J Hirbo, DJ Schaid, I Martin, IJ Kullo, EE Kenny, B Pasaniuc, JS Witte, T Ge. Principles and methods for transferring polygenic risk scores across global populations. Nature Reviews Genetics, 25:8-25, 2024.

doi:10.6342/NTU202500806

[19] Y Ruan, YF Lin, YCA Feng, CY Chen, M Lam, Z Guo, Stanley Global Asia Initiatives, L He, A Sawa, AR Martin, S Qin, H Huang, T Ge. Improving polygenic prediction in ancestrally diverse populations. Nature Genetics, 54:573-580, 2022.

[20] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature Communications, 10(1), 1776. https://doi.org/10.1038/s41467-019-09718-5

[21] Finucane HK, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 2015;47:1228–1235. doi: 10.1038/ng.3404.

[22] Watanabe K, Taskesen E, Bochoven AV, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. 2017;8:1826. doi: 10.1038/s41467-017-01261-5.

[23] Hebbring, S. J. (2014). The challenges, advantages and future of phenome-wide association studies. Immunology, 141(2), 157–165. https://doi.org/10.1111/imm.12195

[24] Dudbridge, F., & Gusnanto, A. (2008). "Estimation of significance thresholds for genomewide association scans." Genetic Epidemiology, 32(3), 227–234.

[25] NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource | Nucleic Acids Research | Oxford Academic. (n.d.). Retrieved February 5, 2025, from https://academic.oup.com/nar/article/51/D1/D977/6814460

[26] de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11:e1004219.

[27] Safari-Alighiarloo, Nahid et al. “Protein-protein interaction networks (PPI) and complex diseases.” Gastroenterology and hepatology from bed to bench vol. 7,1 (2014): 17-31.

[28] Li, Y., Wang, L., Qiu, J., Da, L., Tiollais, P., Li, Z., & Zhao, M. (2012). Human tetraspanin transmembrane 4 superfamily member 4 or intestinal and liver tetraspan membrane protein is overexpressed in hepatocellular carcinoma and accelerates tumor

doi:10.6342/NTU202500806

cell growth. Acta biochimica et biophysica Sinica, 44(3), 224–232. https://doi.org/10.1093/abbs/gmr124

[29] Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C. A., Vangjeli, C., Weale, M. E., Mattsson, H., Kilpeläinen, E., Sleiman, P. M. A., Reilly, D. F., McElwee, J., Maranville, J. C., Chatterjee, A. K., Bhandari, A., Nguyen, K.-D. H., Estrada, K., Reeve, M.-P., … Runz, H. (2018). Phenome-wide association studies across large population cohorts support drug target validation. Nature Communications, 9(1), 4285. https://doi.org/10.1038/s41467-018-06540-3

[30] A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis - PubMed. (n.d.). Retrieved February 8, 2025, from https://pubmed.ncbi.nlm.nih.gov/36778001/

[31] Tamber, S. S., Bansal, P., Sharma, S., Singh, R. B., & Sharma, R. (2023). Biomarkers of liver diseases. Molecular Biology Reports, 50(9), 7815–7823. https://doi.org/10.1007/s11033-023-08666-0

[32] Sun, H., Warren, J., Yip, J., Ji, Y., Hao, S., Han, W., & Ding, Y. (2022). Factors Influencing Gallstone Formation: A Review of the Literature. Biomolecules, 12(4), 550. https://doi.org/10.3390/biom12040550

[33] Jiang, L., Du, J., Wang, J. et al. Sex-specific differences in the associations of metabolic syndrome or components with gallstone disease in Chinese euthyroid population. Sci Rep 13, 1081 (2023). https://doi.org/10.1038/s41598-023-28088-z

doi:10.6342/NTU202500806

Supplemental Table 1. Cohort characteristics of Taiwan Biobank participants. Participants aged from 30 to 7reported the presence or history of liver stone.

Responders N = 108,855	Liver Stone History N= 5,395	No Liver Stone History N=103, 008	P-value
Male (%)	2,086 (38.7%)	37,468 (36.4%)	0.001
Female (%)	3309 (61.3%)	65540 (63.6%)	0.001
Age (mean +/- SD); years	54.66 +/- 9.63	49.61 +/- 10.96	<0.001
Height (mean +/- SD);cm	161.43 +/- 8.16	161.94 +/- 8.72	<0.001
Weight (mean +/- SD);kg	65.23 +/- 12.48	63.73 +/- 12.79	<0.001
Heart Rate (mean +/- SD); times/min	70.52 +/- 4.78	70.90 +/- 4.79	0.005
Systolic Blood Pressure (mean +/- SD); mmHg	123.78 +/- 18.81	120.24 +/- 18.74	<0.001
Diastolic Blood Pressure (mean +/- SD); mmHg	75.24 +/- 11.22	73.73 +/- 11.45	<0.001
Incident Allergy Events (%)	719 (13.3%)	9895 (9.6%)	<0.001
Gout (%)	327 (6.1%)	4308 (4.2%)	<0.001
Congenital Heart Disease (%)	19 (0.4%)	202 (0.2%)	0.02
Diabetes (%)	587 (10.9%)	5703 (5.5%)	<0.001

doi:10.6342/NTU202500806

Supplementary Table 2A. GWAS Catalog for SNPs annotated by HNF4A

Risk Allele	p-Value	beta	ci	Trait Name	pubmedId
rs1800961-T	8.00E-06	0.268 unit decrease	[0.15-0.39]	C-reactive protein	22939635
rs1800961-T	6.00E-10	0.06280401 unit increase	[0.043-0.083]	Hematocrit	27863252
rs1800961-T	4.00E-28	-	-	C-reactive protein levels or HDL-cholesterol levels (pleiotropy)	27286809
rs1800961-T	4.00E-14	-	-	C-reactive protein levels or LDL-cholesterol levels (pleiotropy)	27286809
rs1800961-T	1.00E-20	-	-	C-reactive protein levels or total cholesterol levels (pleiotropy)	27286809
rs1800961-T	1.00E-11	0.06967677 unit decrease	[0.049-0.09]	Sum basophil neutrophil counts	27863252
rs1800961-T	9.00E-12	0.07026628 unit decrease	[0.05-0.09]	Sum neutrophil eosinophil counts	27863252
rs1800961-T	2.00E-11	0.06908321 unit decrease	[0.049-0.089]	Granulocyte count	27863252
rs1800961-T	7.00E-10	0.06386389 unit decrease	[0.044-0.084]	Myeloid white cell count	27863252
rs1800961-T	6.00E-12	0.07093605 unit decrease	[0.051-0.091]	Neutrophil count	27863252
rs1800961-?	7.00E-09	0.0183 unit increase	-	Fibrinogen levels	28107422
rs1800961-?	2.00E-09	0.0178 unit increase	-	Fibrinogen levels	28107422
rs1800961-T	1.00E-10	0.06643613 unit decrease	[0.046-0.087]	Neutrophil percentage of white cells	27863252

doi:10.6342/NTU202500806

rs1800961-T	1.00E-09	-	-	LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df)	30698716
rs1800961-T	1.00E-06	-	-	HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df)	30698716
rs1800961-T	9.00E-17	-	-	HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df)	30698716
rs1800961-T	5.00E-64	-	-	HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df)	30698716
rs1800961-T	4.00E-47	-	-	HDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df)	30698716
rs1800961-T	2.00E-12	-	-	LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df)	30698716
rs1800961-T	1.00E-14	0.138848 unit increase	[0.1-0.17]	Metabolic syndrome	31589552
rs1800961-T	1.00E-08	-	-	LDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df)	30698716
rs1800961-T	3.00E-12	-	-	LDL cholesterol levels x alcohol consumption (regular vs non-regular drinkers) interaction (2df)	30698716
rs1800961-?	3.00E-36	0.0415029 unit decrease	[0.035-0.048]	Metabolic syndrome	39349817
rs1800961-T	3.00E-27	0.069305256 unit increase	[0.057-0.082]	Hematocrit	32888494
rs1800961-T	3.00E-44	0.074343 SD unit increase	[0.064-0.085]	Hemoglobin concentration	32888493
rs1800961-T	2.00E-46	-	-	Hematocrit	32888493

doi:10.6342/NTU202500806

rs1800961-C	1.00E-11	0.0408629 unit increase	[0.029-0.053]	Apolipoprotein B levels	32203549
rs1800961-T	5.00E-44	-	-	Hemoglobin concentration	32888493
rs1800961-T	9.00E-31	0.07351248 unit increase	[0.061-0.086]	Hemoglobin	32888494
rs1800961-T	1.00E-15	1.88 mg/dL decrease	[1.41-2.35]	HDL cholesterol	20686565
rs1800961-T	6.00E-13	4.73 mg/dL decrease	[3.44-6.02]	Cholesterol, total	20686565
rs1800961-T	8.00E-10	0.19 s.d. decrease	[0.09-0.29]	HDL cholesterol	19060906
rs1800961-C	2.00E-09	0.088 unit increase	[0.06-0.12]	C-reactive protein levels	21300955
rs1800961-T	1.00E-24	0.106 unit decrease	[NR]	Cholesterol, total	24097068
rs1800961-T	2.00E-34	0.127 unit decrease	[NR]	HDL cholesterol	24097068
rs1800961-T	2.00E-10	0.0409991 unit increase	[0.028-0.054]	Lymphocyte percentage of white cells	32888494
rs1800961-T	2.00E-19	-	-	Neutrophil count	32888493
rs1800961-T	9.00E-21	0.051942 SD unit decrease	[0.041-0.063]	Neutrophil count	32888493
rs1800961-?	6.00E-22	-	-	Red blood cell count	30595370
rs1800961-?	5.00E-10	-	-	White blood cell count	30595370
rs1800961-T	2.00E-32	0.063432 SD unit increase	[0.053-0.074]	Red blood cell count	32888493
rs1800961-T	1.00E-31	-	-	Red blood cell count	32888493

doi:10.6342/NTU202500806

rs1800961-?	2.00E-11	-	-	Type 2 diabetes	30595370
rs1800961-T	1.00E-12	0.037829 SD unit decrease	[0.027-0.048]	White blood cell count	32888493
rs1800961-T	9.00E-12	-	-	White blood cell count	32888493
rs1800961-C	1.00E-07	0.071 unit increase	-	Total cholesterol levels	29507422
rs1800961-C	5.00E-10	0.076 unit increase	-	Total cholesterol levels	29507422
rs1800961-C	8.00E-75	0.107274 unit increase	[0.096-0.119]	C-reactive protein levels	31900758
rs1800961-?	5.00E-28	0.0807 unit decrease	[0.068-0.094]	Low density lipoprotein cholesterol levels	32154731
rs1800961-C	2.00E-152	0.148995 unit increase	[0.14-0.16]	Apolipoprotein A1 levels	32203549
rs1800961-T	2.00E-46	0.076235 SD unit increase	[0.066-0.087]	Hematocrit	32888493
rs1800961-C	1.00E-140	0.139041 unit increase	[0.13-0.15]	HDL cholesterol levels	32203549
rs1800961-T	2.00E-22	-	[1.15-1.23]	Type 2 diabetes	30297969
rs1800961-T	2.00E-17	0.0339 unit decrease	[0.026-0.042]	HDL cholesterol levels	30926973
rs1800961-T	6.00E-24	0.059 unit increase	[NR]	Hemoglobin levels	32327693
rs1800961-C	4.00E-23	0.0596747 unit increase	[0.048-0.071]	LDL cholesterol levels	32203549
rs1800961-T	6.00E-36	-	-	HDL cholesterol x physical activity interaction (2df test)	30670697

doi:10.6342/NTU202500806

rs1800961-C	5.00E-20	0.149 mmol/l increase	[0.12-0.18]	HDL cholesterol	25961943
rs1800961-T	1.00E-10	0.017 unit decrease	NR	Fibrinogen levels	26561523
rs1800961-?	9.00E-12	2.227338 unit decrease	[1.59-2.87]	High density lipoprotein cholesterol levels	31217584
rs1800961-T	1.00E-12	0.045541454 unit decrease	[0.033-0.058]	Neutrophil percentage of white cells	32888494
rs1800961-C	7.00E-07	0.032 unit increase	[0.02-0.044]	Factor VII activity	30642921
rs1800961-T	5.00E-26	0.0550445 unit decrease	[0.046-0.064]	Non-HDL cholesterol levels	34887591
rs1800961-?	2.00E-30	-	-	Non-HDL cholesterol levels	34887591
rs1800961-T	2.00E-16	0.0934 unit decrease	[0.071-0.116]	Serum total cholesterol levels	38448586
rs1800961-T	1.00E-10	0.0733 unit decrease	[0.051-0.095]	Free cholesterol in small LDL	38448586
rs1800961-T	2.00E-32	0.1335 unit decrease	[0.11-0.16]	Apolipoprotein A-I levels	38448586
rs1800961-T	4.00E-11	0.0753 unit decrease	[0.053-0.098]	Polyunsaturated fatty acids	38448586
rs1800961-T	5.00E-13	0.0816 unit decrease	[0.059-0.104]	Phosphatidylcholine and other choline levels	38448586
rs1800961-T	1.00E-09	0.0666 unit increase	[0.045-0.088]	Triglycerides to total lipids ratio in small HDL	38448586
rs1800961-T	7.00E-10	0.0677 unit increase	[0.046-0.089]	Triglycerides to total lipids ratio in small VLDL	38448586
rs1800961-T	4.00E-12	0.079 unit decrease	[0.057-0.101]	Phospholipids in small LDL	38448586
rs1800961-T	4.00E-09	0.0371 unit decrease	[0.025-0.049]	Serum alkaline phosphatase levels	34594039

doi:10.6342/NTU202500806

rs1800961-T	5.00E-12	0.0783 unit decrease	[0.056-0.1]	Tyrosine levels	38448586
rs1800961-T	5.00E-14	0.0857 unit decrease	[0.063-0.108]	Total cholines levels	38448586
rs1800961-?	4.00E-10	-	-	Coronary artery disease or fibrinogen levels (pleiotropy)	35285134
rs1800961-T	2.00E-11	0.0767 unit decrease	[0.054-0.099]	Omega-6 fatty acids levels	38448586
rs1800961-T	9.00E-14	0.0843 unit decrease	[0.062-0.106]	Free cholesterol levels	38448586
rs1800961-T	1.00E-15	0.0439 unit increase	[0.033-0.055]	Red blood cell count	34594039
rs1800961-T	1.00E-50	0.0933 unit decrease	[0.081-0.105]	Total cholesterol levels	34594039
rs1800961-T	2.00E-06	0.17 unit decrease	[0.092-0.248]	HDL cholesterol levels	35945198
rs1800961-T	3.00E-23	0.0618 unit increase	[0.05-0.074]	Total bilirubin levels	34594039
rs1800961-T	9.00E-14	0.2064 unit increase	[0.15-0.26]	Medication use (drugs used in diabetes)	34594039
rs1800961-T	8.00E-06	0.8567 unit increase	[0.49-1.23]	Vaginal microbiome relative abundance (s_Aerococcus christensenii)	34282934
rs1800961-T	8.00E-06	1.243 unit increase	[0.71-1.78]	Vaginal microbiome MetaCyc pathway (PWY-6901|superpathway of glucose and xylose degradation)	34282934
rs1800961-T	6.00E-35	0.0549493 unit decrease	[0.047-0.062]	Low density lipoprotein cholesterol levels	34887591
rs1800961-T	1.00E-21	-	[1.12-1.19]	Type 2 diabetes	35551307

doi:10.6342/NTU202500806

rs1800961-T	4.00E-08	0.123342 unit decrease	[0.081-0.166]	High density lipoprotein cholesterol levels	34887591
rs1800961-T	4.00E-20	0.0583865 unit increase	[0.046-0.071]	Red blood cell count	32888494
rs1800961-T	2.00E-21	0.0513 unit decrease	[0.041-0.062]	Urate levels	33356394
rs1800961-A	2.00E-07	0.1586 unit decrease	[0.099-0.218]	C-reactive protein levels	29878111
rs1800961-T	1.00E-28	0.0744 unit decrease	[0.061-0.088]	Serum urate levels	39406924
rs1800961-T	2.00E-10	0.075792 mg/dL decrease	[0.053-0.099]	Urate levels	31578528
rs1800961-T	3.00E-06	-	[1.04-1.13]	Type 2 diabetes	29632382
rs1800961-T	5.00E-08	-	[1.06-1.14]	Type 2 diabetes (adjusted for BMI)	29632382
rs1800961-T	7.00E-07	-	[1.05-1.13]	Type 2 diabetes (adjusted for BMI)	29632382
rs1800961-T	8.00E-10	0.03915662 unit decrease	[0.027-0.052]	White blood cell count	32888494
rs1800961-T	5.00E-108	0.1594 unit decrease	[0.15-0.17]	Apolipoprotein A1 levels	33462484
rs1800961-T	4.00E-29	0.0773 unit decrease	[0.064-0.091]	Low density lipoprotein cholesterol levels	33462484
rs1800961-T	3.00E-22	0.067 unit increase	[0.053-0.081]	Total bilirubin levels	33462484
rs1800961-T	1.00E-20	0.0688 unit increase	[0.054-0.083]	Direct bilirubin levels	33462484
rs1800961-C	2.00E-23	0.014741 unit decrease	[0.012-0.018]	Liver enzyme levels (gamma-glutamyl transferase)	33972514
rs1800961-C	2.00E-10	0.0726624 unit increase	[0.05-0.095]	Cholesteryl esters to total lipids ratio in large HDL	35213538

doi:10.6342/NTU202500806

rs1800961-C	2.00E-10	0.0747294 unit decrease	[0.052-0.098]	Triglycerides to total lipids ratio in large LDL	35213538
rs1800961-C	4.00E-31	0.131332 unit increase	[0.11-0.15]	Concentration of HDL particles	35213538
rs1800961-C	1.00E-28	0.122528 unit increase	[0.1-0.14]	Apolipoprotein A1 levels	35213538
rs1800961-C	3.00E-09	0.0687286 unit decrease	[0.046-0.091]	Triglycerides to total lipids ratio in IDL	35213538
rs1800961-C	1.00E-16	0.0873022 unit increase	[0.067-0.108]	Free cholesterol levels in large HDL	35213538
rs1800961-C	2.00E-20	0.0977985 unit increase	[0.077-0.118]	Concentration of large HDL particles	35213538
rs1800961-C	2.00E-08	0.0667597 unit increase	[0.043-0.09]	Cholesterol levels in large LDL	35213538
rs1800961-C	3.00E-11	0.0794412 unit decrease	[0.056-0.103]	Free cholesterol to total lipids ratio in large VLDL	35213538
rs1800961-C	7.00E-19	0.0942711 unit increase	[0.073-0.115]	Cholesterol levels in large HDL	35213538
rs1800961-C	2.00E-19	0.09593 unit increase	[0.075-0.117]	Cholesteryl ester levels in large HDL	35213538
rs1800961-C	8.00E-09	0.0682823 unit increase	[0.045-0.092]	Free cholesterol levels in large LDL	35213538
rs1800961-C	6.00E-10	0.0701112 unit increase	[0.048-0.092]	Cholesterol to total lipids ratio in large HDL	35213538
rs1800961-C	9.00E-19	0.0935851 unit increase	[0.073-0.114]	Total lipid levels in large HDL	35213538
rs1800961-C	5.00E-19	0.104213 unit increase	[0.081-0.127]	Cholesterol to total lipids ratio in large LDL	35213538
rs1800961-C	4.00E-08	0.065569 unit increase	[0.042-0.089]	Cholesteryl ester levels in large LDL	35213538
rs1800961-C	5.00E-13	0.087175 unit increase	[0.063-0.111]	Cholesteryl esters to total lipids ratio in large LDL	35213538

doi:10.6342/NTU202500806

rs1800961-C	4.00E-08	0.0621872 unit decrease	[0.04-0.084]	Triglycerides to total lipids ratio in large HDL	35213538
rs1800961-C	3.00E-18	0.0922268 unit increase	[0.071-0.113]	Phospholipid levels in large HDL	35213538
rs1800961-C	2.00E-16	0.0947155 unit increase	[0.072-0.117]	Cholesterol to total lipids ratio in medium HDL	35213538
rs1800961-C	5.00E-09	0.0705739 unit increase	[0.047-0.094]	Cholesterol to total lipids ratio in medium LDL	35213538
rs1800961-C	2.00E-17	0.102232 unit increase	[0.079-0.126]	Cholesterol levels in small HDL	35213538
rs1800961-C	8.00E-15	0.0908159 unit decrease	[0.068-0.114]	Phospholipids to total lipids ratio in large VLDL	35213538
rs1800961-C	3.00E-11	0.0777018 unit increase	[0.055-0.101]	Cholesteryl esters to total lipids ratio in medium HDL	35213538
rs1800961-C	2.00E-27	0.118948 unit increase	[0.097-0.14]	Free cholesterol levels in medium HDL	35213538
rs1800961-C	5.00E-10	0.0747575 unit increase	[0.051-0.098]	Triglycerides to total lipids ratio in large VLDL	35213538
rs1800961-C	3.00E-26	0.11321 unit increase	[0.092-0.134]	Free cholesterol to total lipids ratio in medium HDL	35213538
rs1800961-C	2.00E-24	0.116829 unit decrease	[0.094-0.139]	Phospholipids to total lipids ratio in medium HDL	35213538
rs1800961-T	3.00E-13	0.066 unit increase	[0.052-0.08]	Triglyceride to HDL cholesterol ratio	38200128
rs1800961-C	2.00E-10	0.0748822 unit increase	[0.052-0.098]	Omega-6 fatty acid levels	35213538
rs1800961-C	3.00E-20	0.104 unit increase	[0.082-0.126]	Phospholipid levels in medium HDL	35213538
rs1800961-C	4.00E-28	0.122445 unit increase	[0.1-0.14]	Cholesteryl ester levels in medium HDL	35213538
rs1800961-C	7.00E-24	0.112737 unit increase	[0.091-0.135]	Total lipid levels in medium HDL	35213538

doi:10.6342/NTU202500806

rs1800961-C	3.00E-10	0.072838 unit decrease	[0.05-0.095]	Triglycerides to total lipids ratio in medium HDL	35213538
rs1800961-C	6.00E-16	0.0920164 unit increase	[0.07-0.114]	Phosphoglycerides levels	35213538
rs1800961-C	3.00E-16	0.0921772 unit increase	[0.07-0.114]	Phosphatidylcholine levels	35213538
rs1800961-C	2.00E-26	0.114978 unit increase	[0.094-0.136]	Total lipid levels in HDL	35213538
rs1800961-C	4.00E-17	0.0954274 unit increase	[0.073-0.118]	Total cholines levels	35213538
rs1800961-C	2.00E-10	0.0750529 unit increase	[0.052-0.098]	Cholesterol to total lipids ratio in IDL	35213538
rs1800961-C	5.00E-30	0.12341 unit increase	[0.1-0.14]	Cholesteryl ester levels in HDL	35213538
rs1800961-C	2.00E-09	0.0710049 unit increase	[0.048-0.094]	Cholesteryl esters to total lipids ratio in IDL	35213538
rs1800961-C	2.00E-11	0.0717205 unit increase	[0.051-0.093]	Average diameter for HDL particles	35213538
rs1800961-C	2.00E-09	0.0725772 unit decrease	[0.049-0.096]	Phospholipids to total lipids ratio in IDL	35213538
rs1800961-C	6.00E-26	0.112565 unit increase	[0.092-0.133]	Free cholesterol levels in HDL	35213538
rs1800961-C	3.00E-23	0.108046 unit increase	[0.087-0.129]	Phospholipid levels in HDL	35213538
rs1800961-C	2.00E-18	0.103686 unit increase	[0.081-0.127]	Free cholesterol levels in small HDL	35213538
rs1800961-C	5.00E-09	0.0691488 unit increase	[0.046-0.092]	Total lipid levels in lipoprotein particles	35213538
rs1800961-C	4.00E-12	0.083131 unit increase	[0.06-0.107]	Cholesteryl esters to total lipids ratio in very large HDL	35213538

doi:10.6342/NTU202500806

rs1800961-C	6.00E-11	0.0707163 unit increase	[0.05-0.092]	Cholesteryl ester levels in very large HDL	35213538
rs1800961-C	2.00E-13	0.0787598 unit decrease	[0.058-0.1]	Free cholesterol to total lipids ratio in very large HDL	35213538
rs1800961-C	2.00E-12	0.0842253 unit increase	[0.061-0.108]	Total lipid levels in small HDL	35213538
rs1800961-C	7.00E-19	0.100042 unit increase	[0.078-0.122]	Sphingomyelin levels	35213538
rs1800961-C	1.00E-26	0.118685 unit increase	[0.097-0.14]	Concentration of medium HDL particles	35213538
rs1800961-C	1.00E-14	0.0887192 unit increase	[0.066-0.111]	Total phospholipid levels in lipoprotein particles	35213538
rs1800961-C	4.00E-09	0.0641943 unit increase	[0.043-0.086]	Cholesterol levels in very large HDL	35213538
rs1800961-C	3.00E-10	0.0683219 unit increase	[0.047-0.09]	Concentration of very large HDL particles	35213538
rs1800961-C	3.00E-28	0.122365 unit increase	[0.1-0.14]	Cholesterol levels in medium HDL	35213538
rs1800961-C	1.00E-29	0.122053 unit increase	[0.1-0.14]	HDL cholesterol levels	35213538
rs1800961-C	6.00E-10	0.0725012 unit increase	[0.05-0.095]	Polyunsaturated fatty acid levels	35213538
rs1800961-C	4.00E-08	0.0645363 unit increase	[0.041-0.088]	Total free cholesterol levels	35213538
rs1800961-C	5.00E-12	0.0805991 unit increase	[0.058-0.103]	Total cholesterol levels	35213538
rs1800961-C	1.00E-13	0.0864156 unit increase	[0.064-0.109]	Total esterified cholesterol levels	35213538
rs1800961-C	3.00E-17	0.099434 unit decrease	[0.076-0.122]	Phospholipids to total lipids ratio in very small VLDL	35213538
rs1800961-C	3.00E-15	0.095431 unit increase	[0.072-0.119]	Concentration of small HDL particles	35213538

doi:10.6342/NTU202500806

rs1800961-C	1.00E-10	0.0733759 unit decrease	[0.051-0.096]	Triglycerides to total lipids ratio in small HDL	35213538
rs1800961-C	1.00E-15	0.0965922 unit increase	[0.073-0.12]	Cholesteryl ester levels in small HDL	35213538
rs1800961-C	9.00E-12	0.0814208 unit increase	[0.058-0.105]	Phospholipid levels in small HDL	35213538
rs1800961-T	4.00E-10	0.0679 unit increase	[0.048-0.087]	Triglyceride to HDL cholesterol ratio	38200128
rs1800961-T	1.00E-08	0.0601 unit increase	[0.042-0.078]	Triglyceride to HDL cholesterol ratio	38200128
rs1800961-C	7.00E-30	0.128916 unit increase	[0.11-0.15]	Total concentration of lipoprotein particles	35213538
rs1800961-C	5.00E-10	0.0741336 unit increase	[0.051-0.098]	Tyrosine levels	35213538
rs1800961-T	1.00E-09	0.2466 unit increase	[0.17-0.33]	Cholecystitis	34594039
rs1800961-T	2.00E-38	0.3186 unit increase	[0.27-0.37]	Cholelithiasis	34594039
rs1800961-C	9.00E-29	0.121 unit increase	-	High density lipoprotein cholesterol levels	29507422
rs1800961-C	5.00E-06	0.143 unit increase	-	High density lipoprotein cholesterol levels	29507422
rs1800961-C	6.00E-36	0.126 unit increase	-	High density lipoprotein cholesterol levels	29507422
rs1800961-T	7.00E-11	0.03209 unit decrease	NR	HDL cholesterol levels	30698716
rs1800961-T	2.00E-06	0.04294 unit decrease	NR	HDL cholesterol levels	30698716
rs1800961-T	3.00E-20	0.03052 unit decrease	NR	HDL cholesterol levels	30698716
rs1800961-T	4.00E-06	2.153 unit decrease	NR	LDL cholesterol levels in current drinkers	30698716

doi:10.6342/NTU202500806

rs1800961-T	2.00E-12	0.052 unit decrease	NR	HDL cholesterol levels in current drinkers	30698716
rs1800961-T	5.00E-23	0.033 unit decrease	NR	HDL cholesterol levels in current drinkers	30698716
rs1800961-T	4.00E-34	0.033 unit decrease	NR	HDL cholesterol levels in current drinkers	30698716
rs1800961-T	4.00E-07	-	-	HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df)	30698716
rs1800961-T	4.00E-17	-	-	HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df)	30698716
rs1800961-T	2.00E-50	-	-	HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df)	30698716
rs1800961-T	1.00E-71	-	-	HDL cholesterol levels x alcohol consumption (drinkers vs non-drinkers) interaction (2df)	30698716
rs1800961-T	6.00E-19	0.0613 unit decrease	[0.048-0.075]	Urate levels	33462484
rs1800961-T	7.00E-54	0.1066 unit decrease	[0.093-0.12]	C-reactive protein levels	33462484
rs1800961-T	2.00E-96	0.1502 unit decrease	[0.14-0.16]	High density lipoprotein cholesterol levels	33462484
rs1800961-T	8.00E-63	0.1153 unit decrease	[0.1-0.13]	Total cholesterol levels	33462484
rs1800961-T	5.00E-17	0.0578 unit increase	[0.044-0.071]	Gamma glutamyl transferase levels	33462484
rs1800961-T	6.00E-14	0.0518 unit decrease	[0.038-0.065]	Apolipoprotein B levels	33462484
rs1800961-T	2.00E-16	0.181 unit decrease	[0.14-0.22]	HDL cholesterol levels	39414775

doi:10.6342/NTU202500806

rs1800961-T	7.00E-74	0.107 unit decrease	[0.095-0.119]	C-reactive protein levels	35459240
rs1800961-T	1.00E-19	0.0542 unit increase	[0.042-0.066]	Gamma glutamyl transpeptidase	34594039
rs1800961-T	2.00E-99	0.133 unit decrease	[0.12-0.15]	HDL cholesterol	34594039
rs1800961-T	1.00E-09	0.0691 unit decrease	[0.047-0.091]	Phospholipids in medium LDL	38448586
rs1800961-T	5.00E-22	0.0493 unit increase	[0.039-0.059]	Hemoglobin	34594039
rs1800961-C	1.00E-18	0.194 unit decrease	[0.15-0.24]	Type 2 diabetes	39379762
rs1800961-T	1.00E-23	0.1135 unit decrease	[0.091-0.136]	Total cholesterol in HDL2	38448586
rs1800961-T	9.00E-10	0.0699 unit decrease	[0.048-0.092]	Phospholipids in large LDL	38448586
rs1800961-T	6.00E-15	0.0887 unit decrease	[0.066-0.111]	Total cholesterol in HDL3	38448586
rs1800961-T	4.00E-52	0.1005 unit decrease	[0.088-0.113]	C-reactive protein	34594039
rs1800961-T	2.00E-13	0.081 unit increase	[0.059-0.103]	Triglycerides to total lipids ratio in IDL	38448586
rs1800961-T	1.00E-11	0.0746 unit increase	[0.053-0.096]	Triglycerides to total lipids ratio in large LDL	38448586
rs1800961-T	6.00E-25	0.1168 unit decrease	[0.095-0.139]	Total cholesterol levels in HDL	38448586
rs1800961-T	9.00E-10	0.0689 unit decrease	[0.047-0.091]	Total cholesterol to total lipids ratio in medium HDL	38448586
rs1800961-T	4.00E-10	0.0711 unit decrease	[0.049-0.093]	Free cholesterol in medium LDL	38448586

doi:10.6342/NTU202500806

rs1800961-?	1.00E-10	-	-	Fibrinogen levels or plasminogen activator inhibitor 1 levels (pleiotropy)	35285134
rs1800961-?	2.00E-13	-	-	Fibrinogen levels or factor VII levels or factor XI levels or tissue plasminogen activator levels (pleiotropy)	35285134
rs1800961-?	1.00E-10	-	-	Fibrinogen levels or tissue plasminogen activator levels (pleiotropy)	35285134
rs1800961-?	1.00E-10	-	-	Ischemic stroke or fibrinogen levels (pleiotropy)	35285134
rs1800961-?	6.00E-15	-	-	Fibrinogen levels or factor VII levels (pleiotropy)	35285134
rs1800961-T	4.00E-12	0.0791 unit decrease	[0.057-0.101]	Total phosphoglycerides levels	38448586
rs1800961-T	2.00E-22	0.0634 unit decrease	[0.051-0.076]	LDL cholesterol	34594039
rs1800961-T	3.00E-21	0.0497 unit increase	[0.04-0.06]	Hematocrit	34594039
rs1800961-T	4.00E-17	0.0975 unit decrease	[0.075-0.12]	Sphingomyelins levels	38448586
rs1800961-T	3.00E-17	0.096 unit decrease	[0.074-0.118]	Esterified cholesterol levels	38448586
rs1800961-?	2.00E-11	-	-	Venous thromboembolism or fibrinogen levels (pleiotropy)	35285134
rs1800961-T	7.00E-13	0.0468 unit decrease	[0.034-0.06]	Neutrophil count	34594039
rs1800961-T	3.00E-08	-	[1.05-1.13]	Type 2 diabetes	29632382
rs1800961-T	4.00E-10	0.2428 unit decrease	-	Membrane-bound aminopeptidase P levels	33067605

doi:10.6342/NTU202500806

rs1800961-T	2.00E-65	0.1408 mg dl-1 decrease	[0.12-0.16]	HDL cholesterol	30275531
rs1800961-T	3.00E-17	0.05392499 unit decrease	[0.041-0.066]	Neutrophil count	32888494
rs1800961-?	6.00E-26	-	[1.23-1.35]	Gallstone disease	30504769
rs1800961-?	1.00E-07	0.1271606 unit decrease	[0.08-0.174]	C-reactive protein levels	31217584
rs1800961-T	6.00E-19	0.104 unit decrease	[0.08-0.128]	C-reactive protein levels	30388399
rs1800961-T	1.00E-12	0.23739344 unit increase	[0.17-0.3]	Medication use (drugs used in diabetes)	31015401
rs1800961-?	1.00E-13	0.0901 unit increase	[0.066-0.114]	Total cholesterol levels	33339817
rs1800961-?	3.00E-17	0.1125 unit increase	[0.086-0.139]	High density lipoprotein cholesterol levels	33339817
rs1800961-C	5.00E-23	0.112 unit increase	[0.09-0.134]	C-reactive protein levels	30388399
rs1800961-T	2.00E-08	0.21016 unit decrease	[0.14-0.28]	Sterol ester (27:1/18:1) levels	37907536
rs1800961-T	3.00E-09	0.221525 unit decrease	[0.15-0.29]	Sterol ester (27:1/16:0) levels	37907536
rs1800961-T	1.00E-10	0.239553 unit decrease	[0.17-0.31]	Sterol ester (27:1/18:3) levels	37907536
rs1800961-T	2.00E-10	0.23874 unit decrease	[0.17-0.31]	Sterol ester (27:1/20:3) levels	37907536
rs1800961-T	7.00E-09	0.215559 unit decrease	[0.14-0.29]	Sterol ester (27:1/18:2) levels	37907536
rs1800961-T	4.00E-08	0.204194 unit decrease	[0.13-0.28]	Sterol ester (27:1/16:1) levels	37907536
rs1800961-T	3.00E-21	0.166 unit increase	[0.13-0.2]	Type 2 diabetes	35551307

doi:10.6342/NTU202500806

rs1800961-?	1.00E-268	-	-	High density lipoprotein cholesterol levels	34887591
rs1800961-T	6.00E-09	0.130857 unit decrease	[0.089-0.173]	High density lipoprotein cholesterol levels	34887591
rs1800961-T	3.00E-157	0.133595 unit decrease	[0.13-0.14]	High density lipoprotein cholesterol levels	34887591
rs1800961-T	1.00E-14	0.151209 unit decrease	[0.11-0.19]	High density lipoprotein cholesterol levels	34887591
rs1800961-T	1.00E-84	0.0868314 unit decrease	[0.08-0.094]	Total cholesterol levels	34887591
rs1800961-C	3.00E-10	0.074 unit increase	[0.051-0.097]	Total omega-6 fatty acid levels	35692035
rs1800961-?	2.00E-111	-	-	Total cholesterol levels	34887591
rs1800961-T	2.00E-55	0.09182 unit decrease	[0.08-0.103]	C-reactive protein levels (MTAG)	36376304
rs1800961-?	1.00E-08	1.43587 unit decrease	-	High density lipoprotein cholesterol levels	36329257
rs1800961-T	4.00E-98	0.08369 unit decrease	[0.076-0.091]	Low-density lipoprotein levels (MTAG)	36376304
rs1800961-T	2.00E-09	0.009677 unit increase	[0.0065-0.0128]	Diabetes (standard GWA)	37106081
rs1800961-T	3.00E-10	-	[1.84-2.29]	Intrahepatic cholestasis of pregnancy	35977952
rs1800961-T	4.00E-37	0.09509 unit decrease	[0.08-0.11]	High-density lipoprotein levels (MTAG)	36376304
rs1800961-?	1.00E-18	2.5605786 mg/dL decrease	[1.99-3.13]	High density lipoprotein cholesterol levels	36220816
rs1800961-T	7.00E-14	0.050505 unit decrease	[0.037-0.064]	LDL (standard GWA)	37106081

doi:10.6342/NTU202500806

rs1800961-?	3.00E-15	-	-	Multi-trait sum score	37277458
rs1800961-?	2.00E-20	-	-	Multi-trait sex score	37277458
rs1800961-?	3.00E-18	-	-	Multi-trait sex score	37277458
rs1800961-?	3.00E-41	-	-	Low density lipoprotein cholesterol levels	34887591
rs1800961-?	6.00E-11	-	[1.19-1.37]	Non-cancer illness code, self-reported: cholecystitis (UKB data field 20002_1163)	37965154
rs1800961-?	3.00E-21	-	[1.22-1.36]	Cholelithiasis (UKB data field 20002_1162)	37965154
rs1800961-?	3.00E-13	-	-	High density lipoprotein cholesterol levels	38116116
rs1800961-T	7.00E-10	0.037204 unit increase	-	Estimated glomerular filtration rate (cystatin c)	39256582
rs1800961-T	1.00E-10	0.038596 unit increase	[0.027-0.05]	Estimated glomerular filtration rate (creatinine, cystatin c)	39256582
rs1800961-T	4.00E-06	0.0276848 unit increase	-	Estimated glomerular filtration rate (creatinine)	39256582
rs1800961-?	6.00E-09	-	-	Total cholesterol levels	38116116
rs1800961-?	3.00E-13	-	-	Multi-trait sum score	37277458
rs1800961-T	4.00E-50	-	[1.31-1.36]	Gallstone disease	34651315
rs1800961-T	5.00E-21	0.120437 unit decrease	[0.095-0.146]	Total lipids in medium HDL (UKB data field 23566)	36764567

doi:10.6342/NTU202500806

rs1800961-T	2.00E-18	0.113095 unit decrease	[0.088-0.138]	Phospholipids in medium HDL (UKB data field 23567)	36764567
rs1800961-T	1.00E-23	0.127506 unit decrease	[0.1-0.15]	Cholesterol in medium HDL (UKB data field 23568)	36764567
rs1800961-T	2.00E-12	0.0945203 unit increase	[0.068-0.121]	Phospholipids to total lipids in very small VLDL percentage (UKB data field 23604)	36764567
rs1800961-T	2.00E-16	0.108573 unit increase	[0.083-0.134]	Phospholipids to total lipids in medium HDL percentage (UKB data field 23639)	36764567
rs1800961-T	6.00E-23	0.125338 unit decrease	[0.1-0.15]	Concentration of medium HDL particles (UKB data field 23565)	36764567
rs1800961-T	6.00E-14	0.0975883 unit decrease	[0.072-0.123]	Phosphoglycerides levels (UKB data field 23434)	36764567
rs1800961-T	3.00E-23	0.127015 unit decrease	[0.1-0.15]	Cholesteryl esters in medium HDL (UKB data field 23569)	36764567
rs1800961-T	1.00E-23	0.126074 unit decrease	[0.1-0.15]	Free cholesterol in medium HDL (UKB data field 23570)	36764567
rs1800961-T	6.00E-26	0.0155 unit increase	[0.013-0.018]	Gamma glutamyl transferase levels	38632349
rs1800961-T	4.00E-21	0.0696 unit decrease	[0.055-0.084]	Serum urate levels	38658550
rs1800961-T	3.00E-18	0.07 unit decrease	[0.054-0.086]	Serum urate levels	38658550
rs1800961-T	2.00E-12	0.092397 unit decrease	[0.067-0.118]	Total phospholipids in lipoprotein particles (UKB data field 23411)	36764567
rs1800961-T	3.00E-10	0.0835302 unit decrease	[0.057-0.11]	Total esterified cholesterol levels (UKB data field 23415)	36764567

doi:10.6342/NTU202500806

rs1800961-T	3.00E-24	0.128631 unit decrease	[0.1-0.15]	Apolipoprotein A1 levels (UKB data field 23440)	36764567
rs1800961-T	7.00E-25	0.133217 unit decrease	[0.11-0.16]	Concentration of HDL particles (UKB data field 23430)	36764567
rs1800961-T	1.00E-23	0.129756 unit decrease	[0.1-0.16]	Total concentration of lipoprotein particles (UKB data field 23427)	36764567
rs1800961-T	3.00E-14	0.0981675 unit decrease	[0.073-0.123]	Phosphatidylcholines levels (UKB data field 23437)	36764567
rs1800961-T	2.00E-14	0.0990861 unit decrease	[0.074-0.124]	Total cholines levels (UKB data field 23436)	36764567
rs1800961-C	5.00E-08	0.1 unit increase	[0.061-0.139]	CFHR5 plasma levels	37286573

doi:10.6342/NTU202500806

Supplementary Table 2B. GWAS Catalog for SNPs annotated by UBXN2B/CYP7A

Risk Allele	pValue	beta	ci	Trait Name	pubmedId
rs983812-T	1.00E-42	-	[0.89-0.9]	Gallstone disease	34651315
rs10107182-T	1.00E-27	0.0130067 unit increase	[0.011-0.015]	Sex hormone-binding globulin levels adjusted for BMI	32042192
rs10107182-?	1.00E-24	0.013286 unit decrease	[0.011-0.016]	Medication use for hyperlipidemia (number of purchases)	36653479
rs983812-?	1.00E-23	-	[1.09-1.14]	Cholelithiasis (UKB data field 20002_1162)	37965154
rs10107182-T	3.00E-22	0.0124022 unit increase	[0.0098-0.015]	Sex hormone-binding globulin levels	32042192
rs10107182-C	2.00E-19	-	[1.41-1.58]	Intrahepatic cholestasis of pregnancy	35977952
rs10107182-?	2.00E-18	0.0374 unit increase	[0.029-0.046]	Total cholesterol levels	33339817
rs7005978-A	1.00E-17	0.105 unit decrease	[0.081-0.129]	Fibroblast growth factor 19 levels	37563310
rs7005978-A	1.00E-17	0.105 unit decrease	[0.081-0.129]	Fibroblast growth factor 19 levels	37563310
rs10107182-T	9.00E-17	0.69 nmol/L increase	[0.53-0.85]	Sex hormone-binding globulin levels	34321204
rs10107182-?	4.00E-16	0.04 unit increase	[0.03-0.05]	Low density lipoprotein cholesterol levels	33339817
rs10107182-?	3.00E-12	-	[1.08-1.14]	Non-cancer illness code, self-reported: cholecystitis (UKB data field 20002_1163)	37965154
rs2326077-C	4.00E-12	0.0300909 unit increase	[0.022-0.039]	Concentration of VLDL particles	35213538
rs2326077-C	9.00E-12	0.0289922 unit increase	[0.021-0.037]	Polyunsaturated fatty acid levels	35213538

doi:10.6342/NTU202500806

rs2326077-C	1.00E-11	0.0293367 unit increase	[0.021-0.038]	Total fatty acid levels	35213538
rs2326077-C	4.00E-11	0.0283474 unit increase	[0.02-0.037]	Total lipid levels in lipoprotein particles	35213538
rs2326077-C	4.00E-11	0.0284114 unit increase	[0.02-0.037]	Total lipid levels in small VLDL	35213538
rs2326077-C	6.00E-11	0.0285984 unit increase	[0.02-0.037]	Free cholesterol levels in medium VLDL	35213538
rs2326077-C	1.00E-10	0.0281991 unit increase	[0.02-0.037]	Phospholipid levels in small VLDL	35213538
rs2326077-C	1.00E-10	0.0282929 unit increase	[0.02-0.037]	Cholesteryl ester levels in VLDL	35213538
rs2326077-C	5.00E-10	0.027081 unit increase	[0.019-0.036]	Remnant cholesterol (non-HDL, non-LDL -cholesterol)	35213538
rs2326077-C	2.00E-09	0.0264191 unit increase	[0.018-0.035]	Free cholesterol levels in small VLDL	35213538
rs2326077-C	2.00E-09	0.0257075 unit increase	[0.017-0.034]	Total free cholesterol levels	35213538
rs2326077-C	2.00E-09	0.0261295 unit increase	[0.018-0.035]	Cholesterol levels in small VLDL	35213538
rs2326077-C	5.00E-09	0.0255643 unit increase	[0.017-0.034]	Cholesteryl ester levels in small VLDL	35213538
rs2326077-C	3.00E-08	0.0241925 unit increase	[0.016-0.033]	Cholesterol levels in medium VLDL	35213538
rs2326077-T	3.00E-07	0.05460 unit decrease	[0.034-0.076]	X-23749 levels	36357675

doi:10.6342/NTU202500806

Supplementary Table 2C. GWAS Catalog for SNPs annotated by TM4SF4

Risk Allele	pValue	beta	ci	traitName	pubmedId
rs12633863-?	4.00E-30	1.11	[1.09-1.13]	Gallstone disease	30504769
rs12633863-A	1.00E-10	0.55	-	Serum alkaline phosphatase levels	33547301
rs12633863-G	6.00E-21	NR	[0.0038-0.0058]	Gamma glutamyl transferase levels	38632349
rs12633863-G	3.00E-19	0.448302	[0.0036-0.0056]	Liver enzyme levels (gamma-glutamyl transferase)	33972514
rs4681515-?	1.00E-56	-	-	Gall stone disease or coronary artery disease	37705021
rs4681515-?	3.00E-30	1.13	[1.1-1.15]	Cholelithiasis (UKB data field 20002_1162)	37965154
rs4681515-?	3.00E-18	NR	[0.027-0.042]	Gamma glutamyl transferase levels	33339817
rs4681515-?	4.00E-13	1.11	[1.08-1.14]	Non-cancer illness code, self-reported: cholecystitis	37965154
rs4681515-G	9.00E-53	NR	[0.11-0.14]	Cholelithiasis	34594039
rs4681515-G	9.00E-52	0.89	[0.88-0.9]	Gallstone disease	34651315

doi:10.6342/NTU202500806

rs4681516-C	3.00E-26	0.472532	[0.03-0.044]	Gamma glutamyl transpeptidase	34594039
rs6774253-?	5.00E-17	NR	-	Gamma glutamyl transferase levels	38116116
rs9857970-?	2.00E-20	NR	[0.03-0.046]	Gamma glutamyl transferase levels	29403010
rs9857970-?	8.00E-11	NR	-	Liver enzyme levels (gamma-glutamyl transferase)	36329257

doi:10.6342/NTU202500806

Supplementary Table 2D: GWAS Catalog for SNPs annotated by LRBA

Risk Allele	P Value	beta	ci	Trait Name	pubmedId
rs2290846-A	5.00E-46	-	[1.12-1.14]	Gallstone disease	34651315
rs2290846-A	1.00E-36	0.116 unit increase	[0.098-0.134]	Cholelithiasis	34594039
rs2290846-?	3.00E-31	-	[1.11-1.16]	Cholelithiasis (UKB data field 20002_1162)	37965154
rs2290846-?	5.00E-27	-	[1.09-1.14]	Gallstone disease	30504769
rs2290846-G	2.00E-26	0.00264842 unit increase	[0.0021-0.0032]	Liver enzyme levels (alkaline phosphatase)	33972514
rs2290846-A	1.00E-18	0.018804 SD unit increase	[0.015-0.023]	Neutrophil count	32888493
rs2290846-A	2.00E-18	-	-	Neutrophil count	32888493
rs2290846-G	3.00E-18	8.7 z-score increase	-	Serum alkaline phosphatase levels	33547301
rs2290846-?	3.00E-17	-	[1.11-1.18]	Non-cancer illness code, self-reported: cholecystitis (UKB data field 20002_1163)	37965154
rs2290846-A	1.00E-13	0.018120443 unit increase	[0.013-0.023]	White blood cell count	32888494
rs2290846-A	1.00E-13	0.018143129 unit increase	[0.013-0.023]	Neutrophil count	32888494
rs2290846-A	2.00E-12	0.1059 unit increase	[0.076-0.135]	Cholecystitis	34594039
rs2290846-A	3.00E-12	0.0162 unit increase	[0.012-0.021]	Neutrophil count	34594039
rs2290846-A	3.00E-11	6.66 z score increase	-	Vertex-wise sulcal depth	34910505

doi:10.6342/NTU202500806

rs2290846-?	3.00E-10	-	-	Lung function (FVC)	30595370
rs2290846-A	4.00E-10	0.0134 unit increase	[0.0093-0.0175]	White blood cell count	34594039
rs2290846-A	5.00E-10	0.013 unit increase	[0.0089-0.0171]	Blood urea nitrogen levels	34594039
rs2290846-A	6.00E-10	6.182 z score increase	-	Lung function (forced vital capacity)	36914875
rs2290846-?	4.00E-09	-	-	Total cholesterol levels	34887591
rs2290846-?	9.00E-09	-	-	High density lipoprotein cholesterol levels	34887591
rs2290846-A	2.00E-08	0.09571408 unit increase	[0.062-0.129]	Cholelithiasis	34594039
rs2290846-A	3.00E-08	0.0408 unit decrease	[0.026-0.055]	Soluble transferrin receptor concentration	39643614

doi:10.6342/NTU202500806

Supplementary Table 2E. GWAS Catalog for SNPs annotated by ANO1

Risk Allele	P Value	beta	ci	Trait Name	pubmedId
rs56363382-T	7.00E-16	-	[1.1-1.13]	Gall stone disease or coronary artery disease	**34651315**
rs56363382-T	2.00E-15	-	[0.086-0.142]	cholelithiasis	**34651315**
rs902870	5.00E-09	-	[1.07-1.15]	cholelithiasis	**34651315**

doi:10.6342/NTU202500806

Supplementary Table 3. Summary statistic of protein-to-protein enrichment.

term ID	term description	observed gene count	background gene count	strength	signal	false discovery rate	matching proteins in your network
GO:0015721	Bile acid and bile salt transport	3	32	2.36	1.4	0.0045	SLCO1B1,CYP7A1,SLCO1A2
GO:0015850	Organic hydroxy compound transport	4	163	1.78	1.26	0.0045	SLCO1B1,ABCG8,CYP7A1,SLCO1A2
GO:0006787	Porphyrin-containing compound catabolic process	2	13	2.58	0.95	0.0278	SLCO1B1,UGT1A1
GO:0006869	Lipid transport	4	329	1.48	0.8	0.0278	SLCO1B1,ABCG8,CYP7A1,SLCO1A2
GO:0038183	Bile acid signaling pathway	2	11	2.65	0.95	0.0278	ABCG8,CYP7A1
GO:0042632	Cholesterol homeostasis	3	89	1.92	0.9	0.0278	ABCG8,CYP7A1,HNF4A
GO:0043252	Sodium-independent organic anion transport	2	15	2.52	0.95	0.0279	SLCO1B1,SLCO1A2
hsa04976	Bile secretion	5	88	2.15	3.81	3.91e-08	SLCO1B1,ABCG8,CYP7A1,UGT1A1,SLCO1A2
hsa04979	Cholesterol metabolism	2	48	2.01	0.9	0.0294	ABCG8,CYP7A1

doi:10.6342/NTU202500806

hsa00140	Steroid hormone biosynthesis	2	60	1.91	0.88	0.0302	CYP7A1,UGT1A1
HSA-194068	Bile acid and bile salt metabolism	3	45	2.22	1.61	0.0017	SLCO1B1,CYP7A1,SLCO1A2
HSA-879518	Transport of organic anions	2	11	2.65	1.22	0.01	SLCO1B1,SLCO1A2
HSA-9748784	Drug ADME	3	105	1.85	1.12	0.01	SLCO1B1,UGT1A1,SLCO1A2
HSA-189483	Heme degradation	2	15	2.52	1.18	0.0112	SLCO1B1,UGT1A1
HSA-159418	Recycling of bile acids and salts	2	18	2.44	1.16	0.0121	SLCO1B1,SLCO1A2
WP2882	Nuclear receptors meta-pathway	4	312	1.5	1.17	0.0034	SLCO1B1,ABCG8,CYP7A1,UGT1A1
WP1604	Codeine and morphine metabolism	2	14	2.55	1.47	0.0037	SLCO1B1,UGT1A1
WP2289	Drug induction of bile acid pathway	2	16	2.49	1.46	0.0037	SLCO1B1,CYP7A1
WP229	Irinotecan pathway	2	12	2.61	1.47	0.0037	SLCO1B1,UGT1A1
WP2874	Liver X receptor pathway	2	10	2.69	1.48	0.0037	ABCG8,CYP7A1

doi:10.6342/NTU202500806

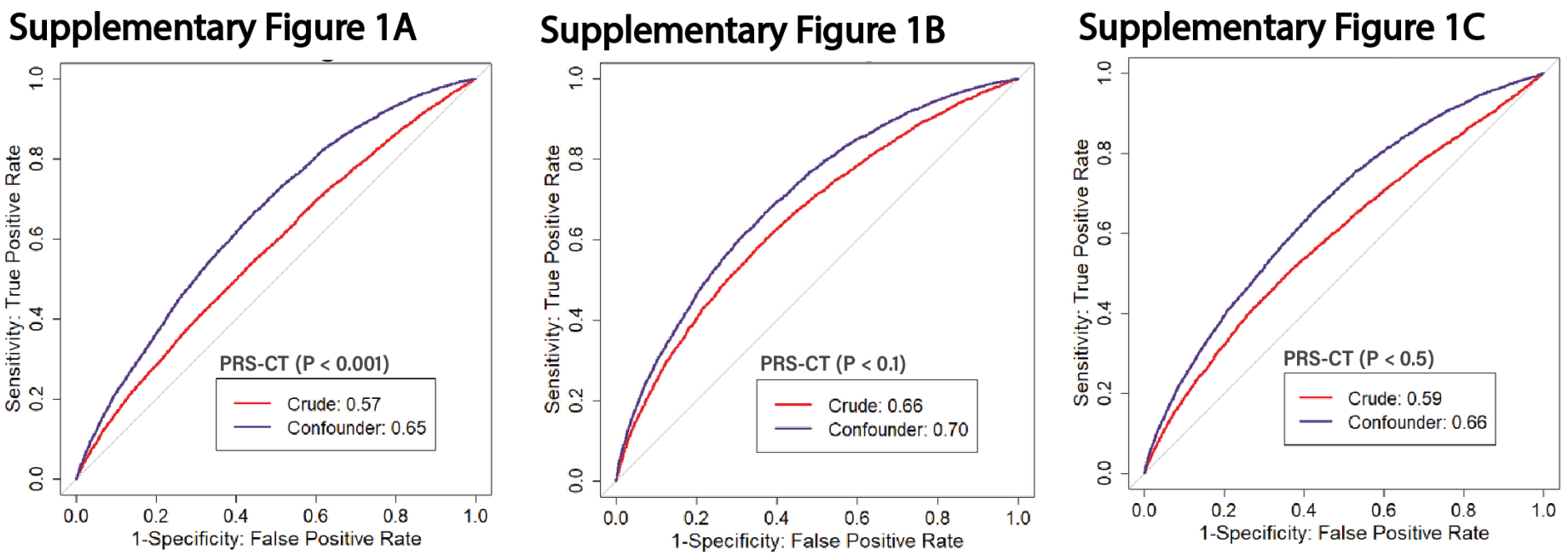
WP4389	Bile acids synthesis and enterohepatic circulation	2	12	2.61	1.47	0.0037	ABCG8,CYP7A1
WP5176	Disorders of bile acid synthesis and biliary transport	2	20	2.39	1.45	0.0037	SLCO1B1,CYP7A1
WP5238	Cholestasis	2	19	2.41	1.45	0.0037	ABCG8,SLCO1A2
WP430	Statin inhibition of cholesterol production	2	29	2.23	1.32	0.0058	ABCG8,CYP7A1
WP2876	Pregnane X receptor pathway	2	32	2.19	1.29	0.0063	SLCO1B1,UGT1A1
WP5304	Cholesterol metabolism	2	72	1.84	0.89	0.0273	ABCG8,CYP7A1
EFO:0004570	Bilirubin measurement	4	139	1.85	1.44	0.0022	SLCO1B1,UGT1A1,SLCO1A2,HNF4A
EFO:0005664	Blood metabolite measurement	4	181	1.74	1.32	0.0031	SLCO1B1,ABCG8,UGT1A1,FUT2
EFO:0004611	Low density lipoprotein cholesterol measurement	5	624	1.3	0.94	0.0069	SLCO1B1,ABCG8,UGT1A1,HNF4A,FUT2
EFO:0005278	Cardiovascular disease biomarker measurement	7	2228	0.89	0.64	0.0069	SLCO1B1,ABCG8,UGT1A1,TM4SF4,SLCO1A2,HNF4A,FUT2
EFO:0005653	Serum metabolite measurement	4	287	1.54	1.08	0.0069	SLCO1B1,UGT1A1,SLCO1A2,FUT2

doi:10.6342/NTU202500806

EFO:0004532	Serum gamma-glutamyl transferase measurement	4	317	1.49	1.03	0.008	TM4SF4,SLCO1A2,HNF4A,FUT2
EFO:0010370	Lysophosphatidylethanolamine 20:4 measurement	2	8	2.79	1.23	0.0099	SLCO1B1,SLCO1A2
HP:0001392	Abnormality of the liver	5	895	1.14	0.76	0.0133	SLCO1B1,ABCG8,CYP7A1,UGT1A1,HNF4A
HP:0032180	Abnormal circulating metabolite concentration	5	1010	1.09	0.67	0.0215	SLCO1B1,ABCG8,CYP7A1,UGT1A1,HNF4A
EFO:0004725	Metabolite measurement	5	1033	1.08	0.67	0.0218	SLCO1B1,ABCG8,UGT1A1,SLCO1A2,FUT2
EFO:0004574	Total cholesterol measurement	4	497	1.3	0.75	0.025	ABCG8,UGT1A1,HNF4A,FUT2
EFO:0004582	Liver enzyme measurement	5	1124	1.04	0.63	0.0258	ABCG8,TM4SF4,SLCO1A2,HNF4A,FUT2
DOID:10211	Cholelithiasis	2	8	2.79	0.94	0.0298	ABCG8,UGT1A1
BTO:0000759	Liver	7	2125	0.91	0.71	0.003	SLCO1B1,ABCG8,CYP7A1,UGT1A1,TM4SF4,SLCO1A2,HNF4A

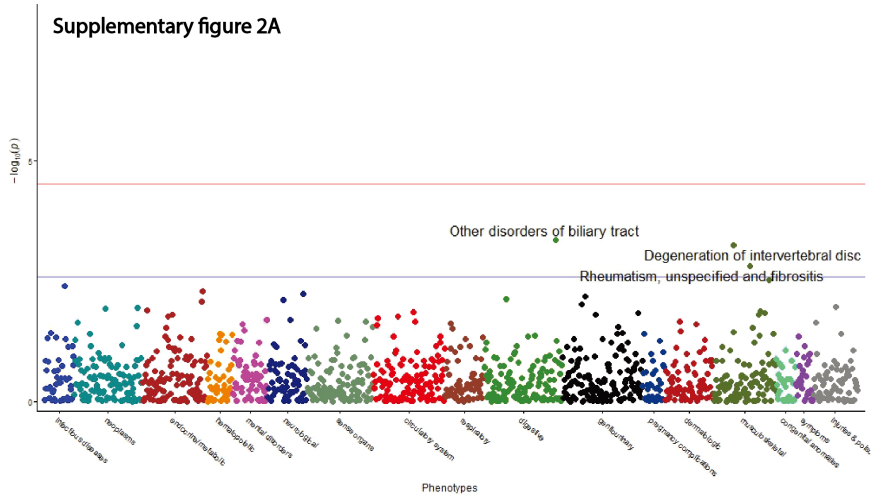
doi:10.6342/NTU202500806

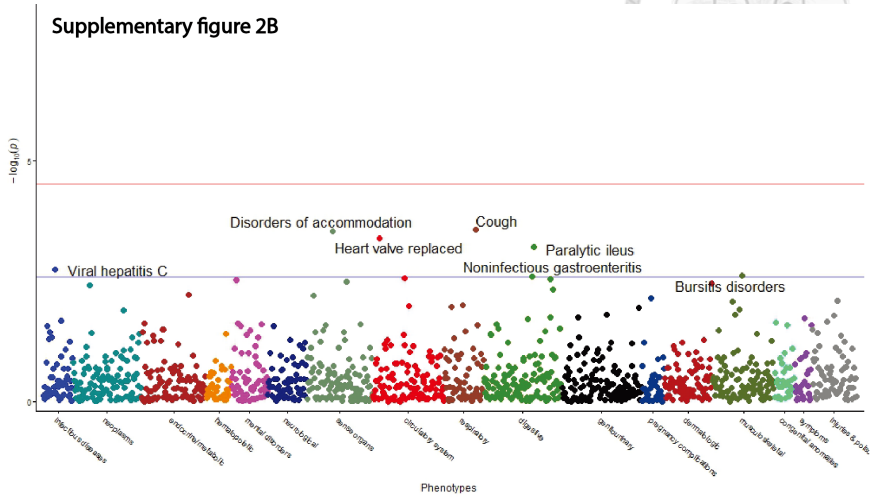
Supplementary Figure 1. Receiver Operating Characteristic (ROC) Curves Comparing Predicting GSD PRS Models with and without Confounders. The ROC curve compares the crude model (red line), which uses PRS alone, and the adjusted model (blue line), which incorporates age and sex as confounders. (A) P<0.001, (B) P<0.1, (C) P<0.5.

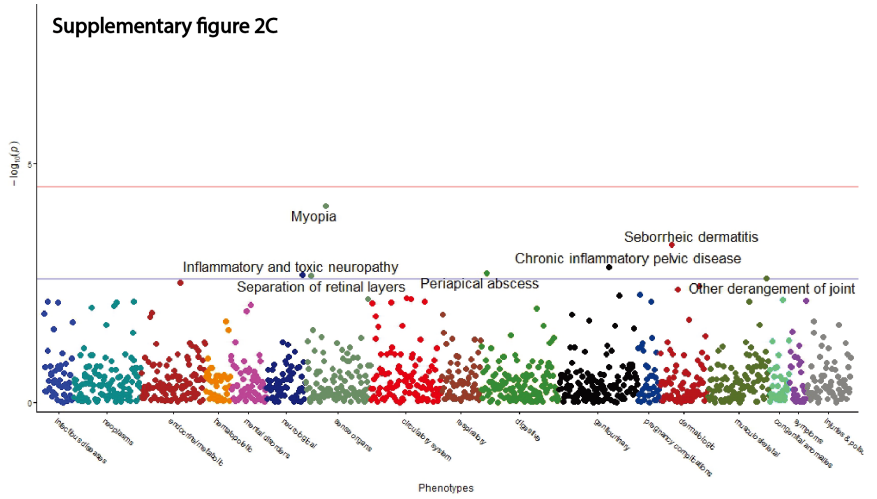


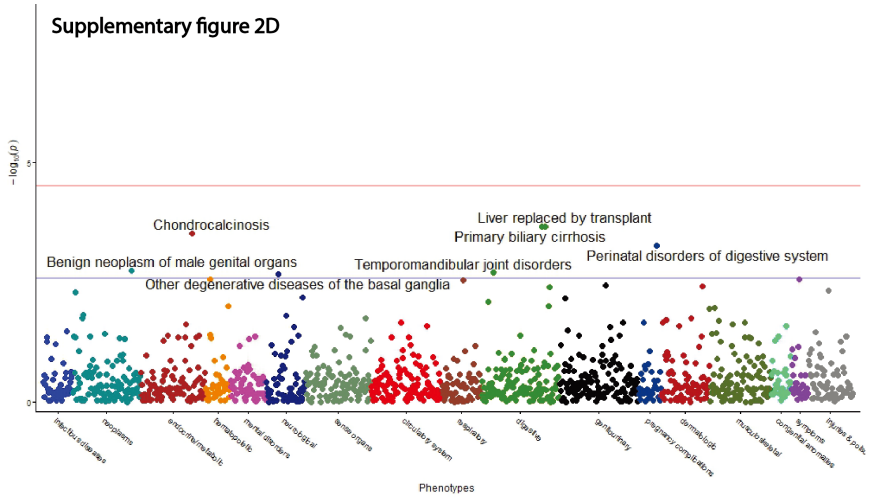
doi:10.6342/NTU202500806

Supplementary Figure 2. PheWAS Disease Model Interpretation. The blue line represents the genome-wide significance threshold of 5×10^{-5}, while the red line indicates the threshold of 5×10^{-8}. (A) rs66779552. (B) rs17503902 (C) rs1580180 (D) rs21311242.



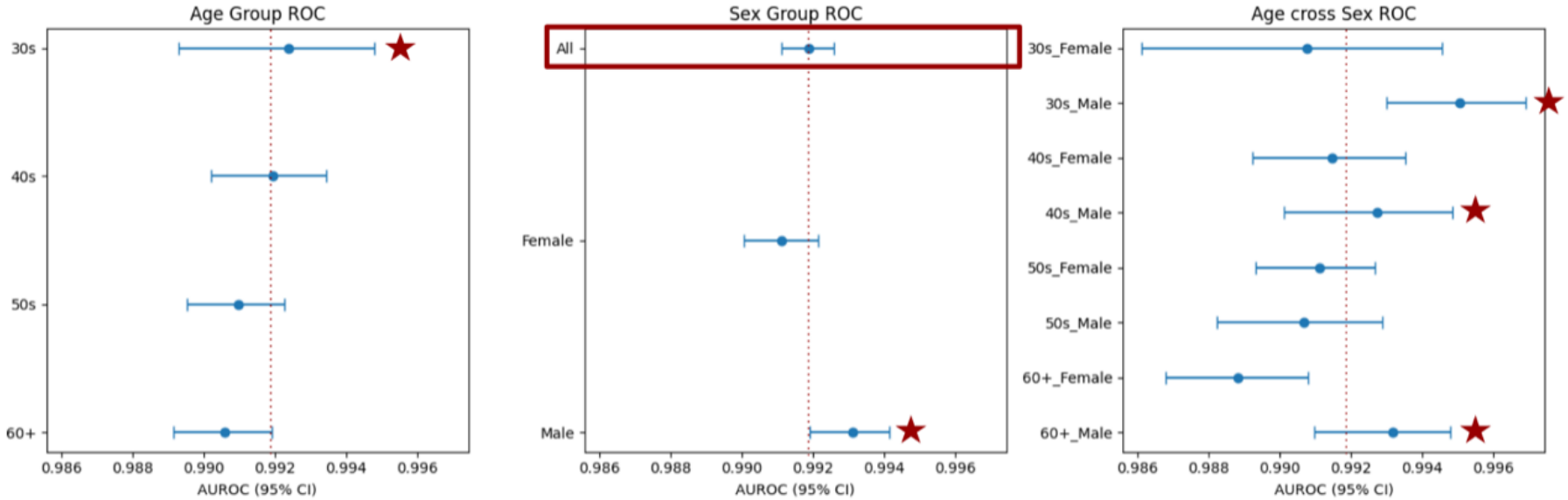






doi:10.6342/NTU202500806

Supplementary figure 3. Receiver Operating Characteristic (ROC) Curves for different subgroups. The model's predictive power is higher in males across various age groups compared to overall population. Additionally, the 30s age group demonstrates better predictive performance than the overall group.



doi:10.6342/NTU202500806