PUBLIC HEALTH

Diversity and longitudinal records: Genetic architecture of disease associations and polygenic risk in the Taiwanese Han population

Ting-Yuan Liu[1,2], Hsing-Fang Lu[1,2], Yu-Chia Chen[1,2], Chi-Chou Liao[3], Ying-Ju Lin[3,4,5], Jai-Sing Yang[3,4], Wen-Ling Liao[4,7,8], Wei-De Lin[3,4,5], Shih-Yin Chen[3,4,6], Yu-Chuen Huang[3,4,6], Wei-Yong Lin[3,7], Yu-Huei Liu[4,7,9], Kai-Cheng Hsu[10,11,12], Shih-Sheng Chang[10], Hong-Da Chen[13], Yu-Pao Chou[14], Jan-Gowth Chang[14,15]†, Chung-Hsing Wang[16], Chwen-Tzuei Chang[17], Chung-Ming Huang[18], Kai-Jieh Yeo[18], Tzu-Yuan Wang[19], Chin-Chung Yeh[20], Jiunn-Horng Chen[18], Chi-Ping Huang[20], Hsueh-Chou Lai[21], Rong-Hsing Chen[19], Hui-Ju Lin[22], Po-Yuan Wu[23], Jiu-Yao Wang[24,25], Chin-Chi Kuo[3,26,27], Der-Yang Cho[28,29], Chang-Hai Tsai[30], Fuu-Jen Tsai[3,6,31,32]*

We addressed the underrepresentation of non-European populations in genome-wide association studies (GWASs) by building HiGenome, a large-scale genetic resource for the Taiwanese Han population. Using a custom genotyping array, we integrated deidentified electronic medical records (2003 to 2021) with genomic data to enable GWASs, phenome-wide association studies, and polygenic risk score (PRS) analysis. Among 413,000 participants, 323,397 passed ancestry and quality control filtering. GWASs covered 1085 traits, focusing on diseases prevalent in Taiwan such as type 2 diabetes, chronic kidney disease, gout, and alcoholic liver damage. PRSs were calculated for 238 traits, with the strongest associations observed in musculoskeletal disorders. Incorporating PRS into clinical practice supports early risk prediction and personalized prevention. To further expand translational value, we also conducted pharmacogenomic analysis and human leukocyte antigen typing. HiGenome offers a large-scale genetic and clinical dataset from the Taiwanese Han population, supporting population-specific analyses and precision medicine development in East Asia. The hospital-based design enables continuous follow-up and longitudinal data expansion.

INTRODUCTION

Genome-wide association studies (GWASs) help researchers explore the associations between genes and specific diseases or traits (*1*). A key limitation of GWASs is the complex nature of most diseases, which result from a combination of genetic and environmental factors (*2*). In terms of genetic contributions, disease development is rarely driven by a single gene, but rather by the interplay of multiple genes and environmental influences. Polygenic risk scores (PRSs) serve as a powerful approach to summarize the cumulative effects of multiple genetic variants and can also incorporate environmental factors into the model, aiding in the assessment of disease susceptibility (*3*, *4*). Another limitation of GWASs is the underrepresentation of non-European populations, which hinders the identification of rare variants; these variants manifest as high minor allele frequencies (MAFs) in other populations. Typically, individuals' unique genetic risk factors for diseases are predominantly influenced by their ancestry (*1*, *5*). The underrepresentation of non-European populations in GWASs limits research advancements and exacerbates health disparities, particularly when the clinical applications of relevant genetic findings are primarily tailored for European populations. Heavy dependence on genetic data from a particular ancestry for the evaluation of health and disease outcomes is associated with major risks (*5*, *6*).

[1]Million-person precision medicine initiative, Department of Medical Research, China Medical University Hospital, Taichung 40402, Taiwan. [2]Master Program for Digital Health Innovation, China Medical University, Taichung 406040, Taiwan. [3]Department of Medical Research, China Medical University Hospital, Taichung 40402, Taiwan. [4]Genetic Center, Department of Medical Research, China Medical University Hospital, Taichung 40402, Taiwan. [5]School of Post Baccalaureate Chinese Medicine, China Medical University, Taichung 404333, Taiwan. [6]School of Chinese Medicine, China Medical University, Taichung 406040, Taiwan. [7]Graduate Institute of Integrated Medicine, College of Chinese Medicine, China Medical University, Taichung 406040, Taiwan. [8]Center for Personalized Medicine, Department of Medical Research, China Medical University Hospital, Taichung 406040, Taiwan. [9]Drug Development Center, China Medical University, Taichung 406040, Taiwan. [10]Artificial Intelligence Center, China Medical University Hospital, Taichung 40402, Taiwan. [11]Department of Neurology, China Medical University Hospital, Taichung 40402, Taiwan. [12]Department of Medicine, China Medical University, Taichung 406040, Taiwan. [13]Department of Laboratory Medicine, China Medical University Hospital, Taichung 40402, Taiwan. [14]Department of Bioinformatics and Medical Engineering, Asia University, Taichung 41300, Taiwan. [15]Department of Medical Laboratory Science and Biotechnology, China Medical University, Taichung, Taiwan. [16]Division of Genetics and Metabolism, Children's Hospital of China Medical University, Taichung 404447, Taiwan. [17]Intelligent Diabetes Metabolism and Exercise Center, Department of Internal Medicine, China Medical University Hospital, Taichung 40402, Taiwan. [18]Division of Immunology and Rheumatology, Department of Internal Medicine, China Medical University Hospital, Taichung 40402, Taiwan. [19]Division of Endocrine and Metabolism, China Medical University Hospital, Taichung 40402, Taiwan. [20]Department of Urology, China Medical University Hospital, Taichung 40402, Taiwan. [21]Center for Digestive Medicine, Department of Internal Medicine, China Medical University Hospital, Taichung 40402, Taiwan. [22]Department of Ophthalmology, Eye Center, China Medical University Hospital, Taichung 40402, Taiwan. [23]Department of Dermatology, China Medical University Hospital, Taichung 40402, Taiwan. [24]Center for Allergy, Immunology and Microbiome (AIM), China Medical University Hospital, Taichung 40447, Taiwan. [25]Department of Allergy, Immunology and Rheumatology (AIR), China Medical University Children's Hospital, Taichung 40447, Taiwan. [26]Big Data Center, China Medical University Hospital, Taichung 40402, Taiwan. [27]Division of Nephrology, Department of Internal Medicine, China Medical University Hospital, Taichung 40402, Taiwan. [28]Department of Neurosurgery, China Medical University Hospital, Taichung 40402, Taiwan. [29]Graduate Institute of Biomedical Sciences, China Medical University, Taichung 40402, Taiwan. [30]Division of Pediatrics Neurology, China Medical University Children's Hospital, Taichung 40447, Taiwan. [31]Division of Medical Genetics, Children's Hospital of China Medical University, Taichung 40402, Taiwan. [32]Department of Medical Laboratory Science and Biotechnology, Asia University, Taichung 41300, Taiwan.
*Corresponding author. Email: 000704@tool.caaumed.org.tw
†Present address: Show Chwan Healthcare System, 2F, No. 6-1, Lugong Rd., Lukang Township, Changhua County 505, Taiwan.

Genetic research in Asia began later than in Europe and America, leading to relatively fewer large-scale studies. Although several biobanks have been established for Han Chinese, Japanese, Korean, and Southeast Asian populations, most were developed only in recent years and still lack extensive longitudinal clinical data (*7*). Notable examples of existing biobanks in East Asia include the China Kadoorie Biobank, the Korea Biobank Array Project (*8*, *9*), the Taiwan Biobank (TWB) (*10*–*12*), and BioBank Japan (*13*, *14*). In contrast, large-scale biobanks such as the UK Biobank (UKBB) (*15*), FinnGen (*16*), and the Million Veteran Program (MVP) (*17*) integrate both patient electronic medical records (EMRs) and questionnaire-based health data, providing a more comprehensive dataset. In this study, we specifically focused on the Taiwanese Han population. Data were collected from a single institution—China Medical University Hospital (CMUH) and its affiliated branches—through a unified system encompassing nearly 19 years of EMRs. This structured and extensive dataset was designed to support both hypothesis-driven and exploratory research on various diseases, allowing for in-depth investigations across a broad spectrum of medical conditions.

In the following, we describe the design and performance of a custom Affymetrix Axiom array called the Taiwan Precision Medicine Version 1 (TPMv1) single-nucleotide polymorphism (SNP) array, also referred to as TWB SNP array version 2 (*12*). This array, tailored for the Taiwanese Han population, not only ensures genome-wide coverage, facilitating the high-quality imputation of both prevalent and infrequent variations, but also directly genotypes ~680,000 presumed variants. In this study, we conducted an imputation analysis by using 1463 whole-genome sequencing datasets from the TWB. We also used 95 distinct whole-genome sequencing datasets to validate the outcomes of our imputation. Our findings were consistent across datasets with high MAFs (*18*). After 14 pharmacogenes of clinical significance were curated from the Taiwanese Han population, data on medication prescriptions were collected. Approximately 99.9% of the population had at least one actionable pharmacogenetic phenotype, with 29% prescribed medications to which they may have had a nonstandard response (*19*).

Using the PheCode classification method developed by the Vanderbilt University Medical Center (*20*), we analyzed ~19 years' worth of EMR data for individuals who had participated in a genetic program at CMUH. By integrating data corresponding to ~14 million SNPs imputed from the TPMv1 chip, we successfully conducted GWASs for 1085 diseases and PRS analyses for 238 diseases. Concurrently, we performed robust phenome-wide association studies (PheWASs).

In this study, a comprehensive platform was established for GWASs, PheWASs, and PRS analyses among Han individuals in Taiwan. This initiative effectively addressed the global shortage of extensive genetic data from Asian populations by providing valuable insights into genetic associations with diseases in Southeast Asia. Our study also highlighted genetic differences between our cohort and the UKBB.

RESULTS

HiGenome features

HiGenome is a pan-Taiwanese genomic database established by CMUH. This database was launched in 2018 with the primary objective of determining the associations between genes and diseases prevalent among the Taiwanese Han population and calculating the PRSs for disease incidence prediction in asymptomatic individuals, with the ultimate goal of achieving precision in health care. In this study, to complete the genomic database, we used the TPMv1 SNP array, which is specifically designed for the Taiwanese Han population. Whole-genome sequencing data pertaining to Taiwanese Han individuals, obtained from the TWB, were used to develop a platform for imputation analysis (*18*). Various platforms were established for pharmacogenomic analysis (*19*, *21*, *22*), human leukocyte antigen (HLA) typing (*23*), kinship verification (*24*), ancestral analysis, and PRS modeling (*25*) (Fig. 1).

CMUH is an extensive academic medical center located in Taichung, Taiwan. It collaborates with a network of hospitals across the northern, central, and southern regions of the country. In this study, patients were enrolled from highly populated towns and districts across Taiwan from 2018 to 2021 (Fig. 2A). Because both genotypic and phenotypic (clinical) data are essential for identifying disease-gene associations in GWASs, genotypic data were obtained using the TPMv1 SNP array, supplemented by whole-genome sequencing data obtained from the Taiwanese population, and enhanced using imputation algorithms. This approach expanded our dataset from ~680,000 SNPs to nearly 14 million reference points, consistent with the reference sequence of the Taiwanese Han population. Clinical data were collected from patient EMRs, which were matched with relevant PheCodes. After the participants were categorized into case and control groups on the basis of 1085 phenotypes, preliminary stratification was performed by age and sex, with each disease assigned to either a base or target dataset. After genetic and phenotypic data were combined, the base dataset was used to conduct GWASs and PheWASs for all phenotypes, and the target dataset was used to calculate the PRSs of 238 diseases (fig. S1).

Clinical characteristics of the HiGenome cohort

The ages of our participants ranged from 0 to 111 years, with a male-to-female ratio of 45.3:54.7. In table S1, the mean age of male (47.89 ± 21.72) participants was slightly higher than that of female (46.37 ± 21.07) participants. Retrospective analysis of patient EMRs revealed that ~85.9% of the participants were followed up for more than 1 year, 65.3% were followed up for more than 5 years, 46.3% were followed up for more than 10 years, and 27.9% were followed up for more than 15 years (Fig. 2B, left). In 2003, the total number of diagnostic instances reached 800,000; this number increased to ~7 million by 2021 (annual average: 3 million; Fig. 2B, right). Notably, 43% of the participants received treatment in the hospital's internal medicine departments, including the pediatric (14%) and surgical (10%) departments (Fig. 2C). Diagnostic analysis revealed that the patients primarily sought treatment for neoplasms and for diseases affecting their circulatory, endocrine, metabolic, genitourinary, or digestive system (Fig. 2D). Figure 2E depicts the age distribution of all 1085 traits. In this figure, the median age of the disease group is plotted on the *x* axis, and the median age of the control group is plotted on the *y* axis. As observed, few traits were distributed along the reference line, indicating a consistent age distribution for these traits. Notably, most of the traits were associated with a higher median age in the disease group than in the control group, confirming that the incidence of most diseases increased with age and time. Figure 2F depicts the gender ratio of the traits. At the macro level, traits exclusive to males (coordinates at 1-1) or females (coordinates at 0-0) were observed (Fig. 2F, left). For most traits, the gender ratio of male in the control group consistently ranged between 0.49 and 0.42, reflecting our cohort's overall gender distribution. In the case

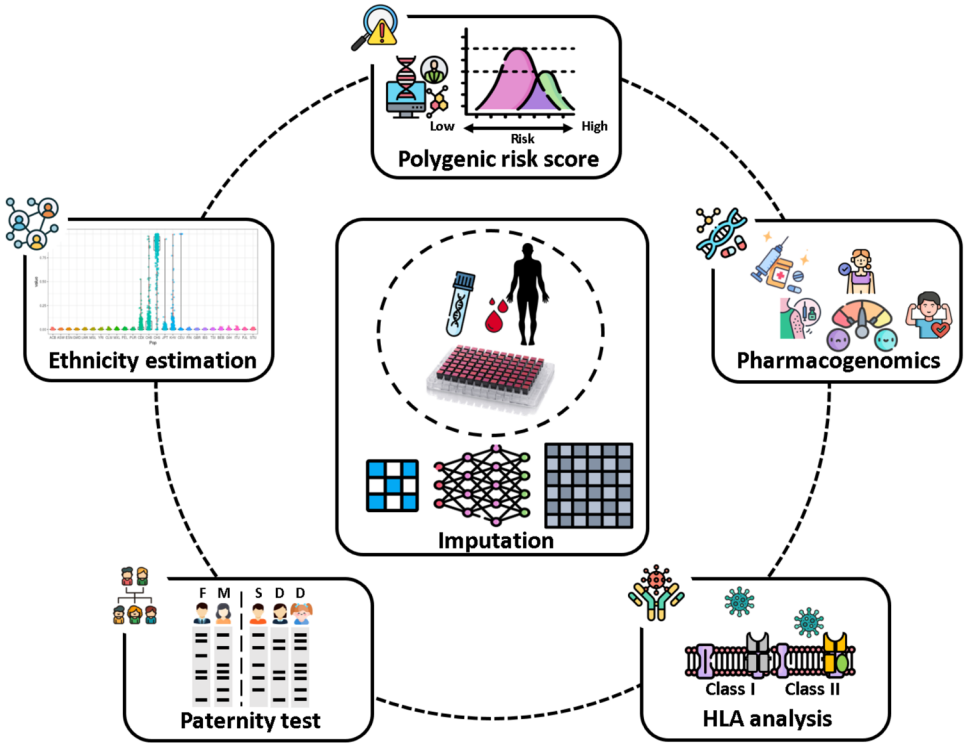


Fig. 1. Analysis platforms for our genotyping chip. The center of the schematic depicts our foundational data derived using the TPMv1 chip, revealing variants identified using blood DNA samples. We conducted an imputation analysis to enhance the data's richness, preparing the dataset for future integrative analyses for other databases. Around the center are our extended analytical platforms: pharmacogenomics, human leukocyte antigen typing, parentage testing, ancestry analysis, and PRS modeling.

group, traits in the endocrine or metabolic categories were less prevalent among male participants, and certain neoplasm traits were more prevalent among male participants (Fig. 2F, right). For more details, please refer to table S1. Overall, these observations prompted us to integrate age and gender adjustment into our subsequent GWAS analyses. Notably, our data included long-term follow-up information. Patients with chronic hepatitis B virus (HBV) infection had a higher incidence of liver cancer compared with those without HBV infection (fig. S2A). In addition, long-term diabetes was associated with an increased risk of diabetic retinopathy. Moreover, female participants were more susceptible than were male participants to diabetic retinopathy, with a highly significant *P* value (6×10^{-68}; fig. S2B).

Ancestral distribution in the HiGenome cohort

To analyze ancestry, we conducted a principal components analysis (PCA), which resulted in the extraction of principal components 1 and 2. Data from the 1000 Genomes Project were used as a reference. After aligning the HiGenome cohort with the East Asian (EAS) cohort from the 1000 Genomes Project (Fig. 3A), we found that a subset of our participants exhibited values exceeding the three quartiles of principal components 1 and 2. These participants were characterized as a potentially non-EAS subset. To ensure homogeneity, these individuals were excluded from further analyses. As shown in Fig. 3B, the quality-controlled subset of the HiGenome cohort was consistent with the reference cohort in terms of principal component 1 and 2 distribution. Further exploration of the distribution of EAS individuals revealed a subset of participants exhibiting substantial deviations from the typical EAS distribution (Fig. 3C). However, after excluding this cohort on the basis of our established criteria, we found that the distribution of most participants was consistent with that of EAS individuals, thereby confirming the robustness of our quality control process (Fig. 3D). Predominant ancestral lineages were mapped to Southern Han Chinese individuals, followed by (in descending order of frequency) Han Chinese individuals from Beijing, Chinese Dai individuals from Xishuangbanna, Kinh individuals from Ho Chi Minh, Vietnam, Japanese individuals from Tokyo, Japan, and a small subset of individuals resembling the residents of Utah with northern or western European ancestry (Fig. 3E). Although most of the participants exhibited >50% single-source ancestry (Fig. 3E, top), some did not. These participants were of mixed EAS descent, and they were retained in subsequent analyses (Fig. 3E, bottom). These findings prompted us to include PCA adjustment in our subsequent GWAS analyses.

Analysis of HLA distribution and pharmacogenomic associations

To further explore the distribution of HLA and its subsequent integration into pharmacogenomic analysis, we established a predictive model for HLA. We used the results of HLA typing and chip genotyping from the TWB as input for model training. This model is capable of identifying various HLA types, including *HLA-A* with 35 types (1824 SNPs), *HLA-B* with 75 types (2007 SNPs), *HLA-C* with 34 types (2151 SNPs), *HLA-DPA1* with 6 types (816 SNPs), *HLA-DPB1* with 33 types (1713 SNPs), *HLA-DQA1* with 17 types (2190 SNPs), *HLA-DQB1* with 19 types (1778 SNPs), and *HLA-DRB1* with 44 types (2277 SNPs). Across these eight subtypes, an average

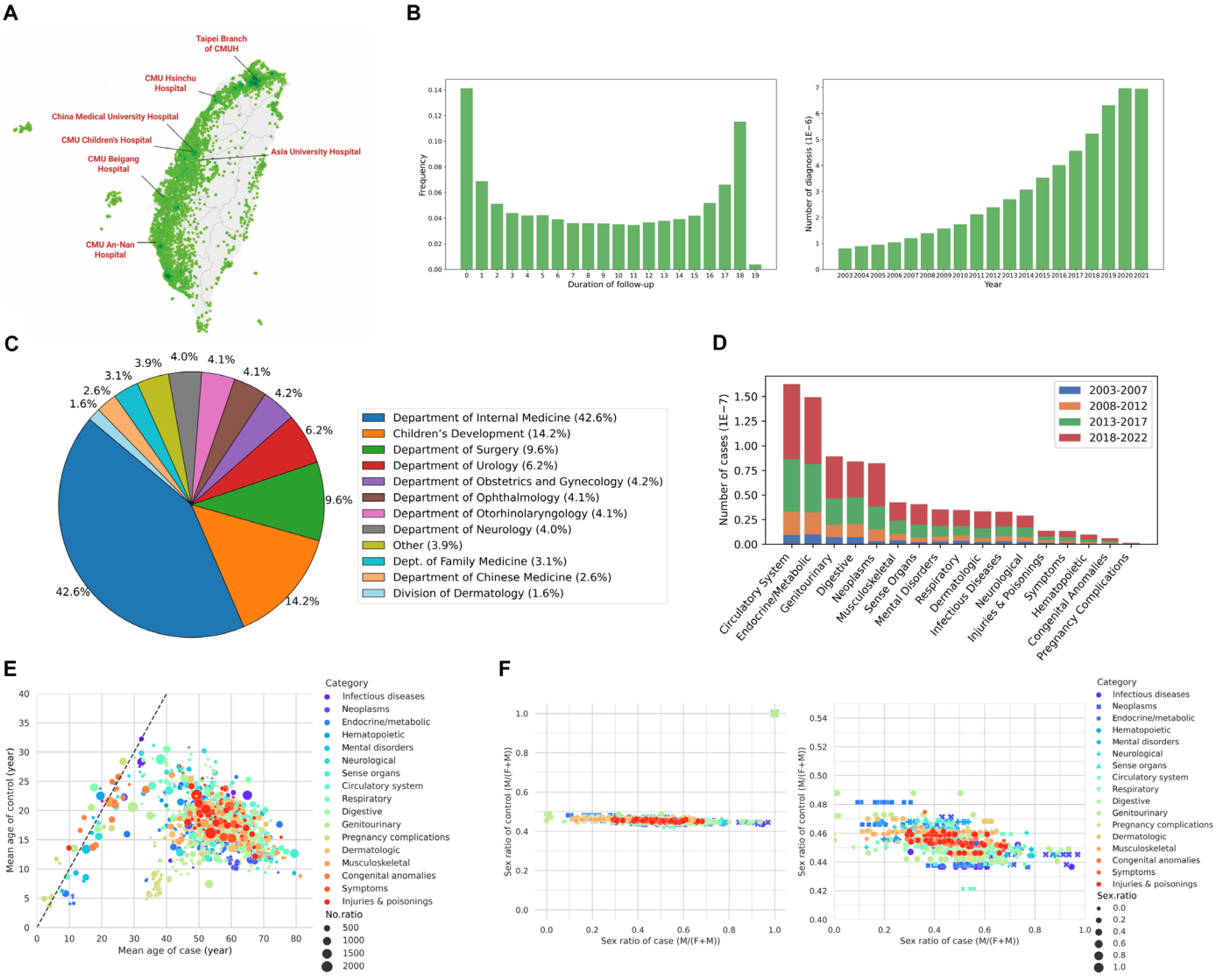


Fig. 2. HiGenome cohort clinicodemographic data. (**A**) HiGenome contains data from individuals residing in densely populated residential areas in Taiwan. These data were primarily collected by CMUH and its affiliated institutions. (**B**) The left part presents the duration of follow-up, indicating a predominance of patients who were followed up for less than 1 year up to 18 years. The right part presents the annual distribution of diagnoses identified from the patients' EMRs, indicating a gradual increase in the number of diagnoses. (**C**) In terms of patient recruitment, most patients were enrolled from the hospital's internal medicine department. (**D**) Diagnoses were classified using PheCodes. Most diagnoses were related to the circulatory system. (**E**) Age distribution for each trait, with the *x* axis representing the median age of the case group and the *y* axis representing the median age of the control group. Each color represents a unique category, with the size of the legend reflecting the number of participants. The reference line indicates equal age proportions between groups; the right half demonstrates the gender distribution for each trait. (**F**) The right half is an enlarged view focusing on the control group with a male proportion ranging between 0.4 and 0.54. The *x* axis indicates the male proportion in the case group, whereas the *y* axis indicates the male proportion in the control group. The left half indicates traits with exclusively female (lower left) or male (upper right) participants. This focused view on the right half clearly demonstrates gender proportion disparities. The male proportion in the control group ranges between 0.5 and 0.42, with notable variances in the case group due to disease characteristics.

out-of-bag accuracy of 96.86% was detected, ranging from a minimum of 92.14% (*HLA-B*) to a maximum of 99.69% (*HLA-DPA1*), with a high accuracy observed in HMC class II (fig. S3A and table S2). Haplotype analysis of HLA revealed that *HLA-A*11:01* (33.16%), *HLA-B*40:01* (26.50%), and *HLA-C*07:02* (20.87%) were the most frequent combinations, and diplotype analysis revealed that *HLA-A*11:01/24:02* (12.94%), *HLA-B*40:01/46:01* (8.47%), and *HLA-C*01:02/07:02* (8.68%) were the most frequent combinations (fig. S3, B and C, and table S3). Our model also revealed varying association distances between each HLA subtype. For instance, *HLA-A*01:01* and *HLA-A*68:01* exhibited the most distant relationship, with HLA-B*15:13 and HLA-B*27:05 being the farthest apart. Additional associations are detailed in fig. S4. Given our previous ancestry analysis results (Fig. 3), we observed significant differences and a higher proportion of certain HLA subtypes in Southern Han Chinese individuals than in Han Chinese individuals from Beijing, including *HLA-A*11:01*($P < 7 \times 10-8$) and *HLA-B*40:01*($P < 7 \times 10-70$) (fig. S5 and table S4).

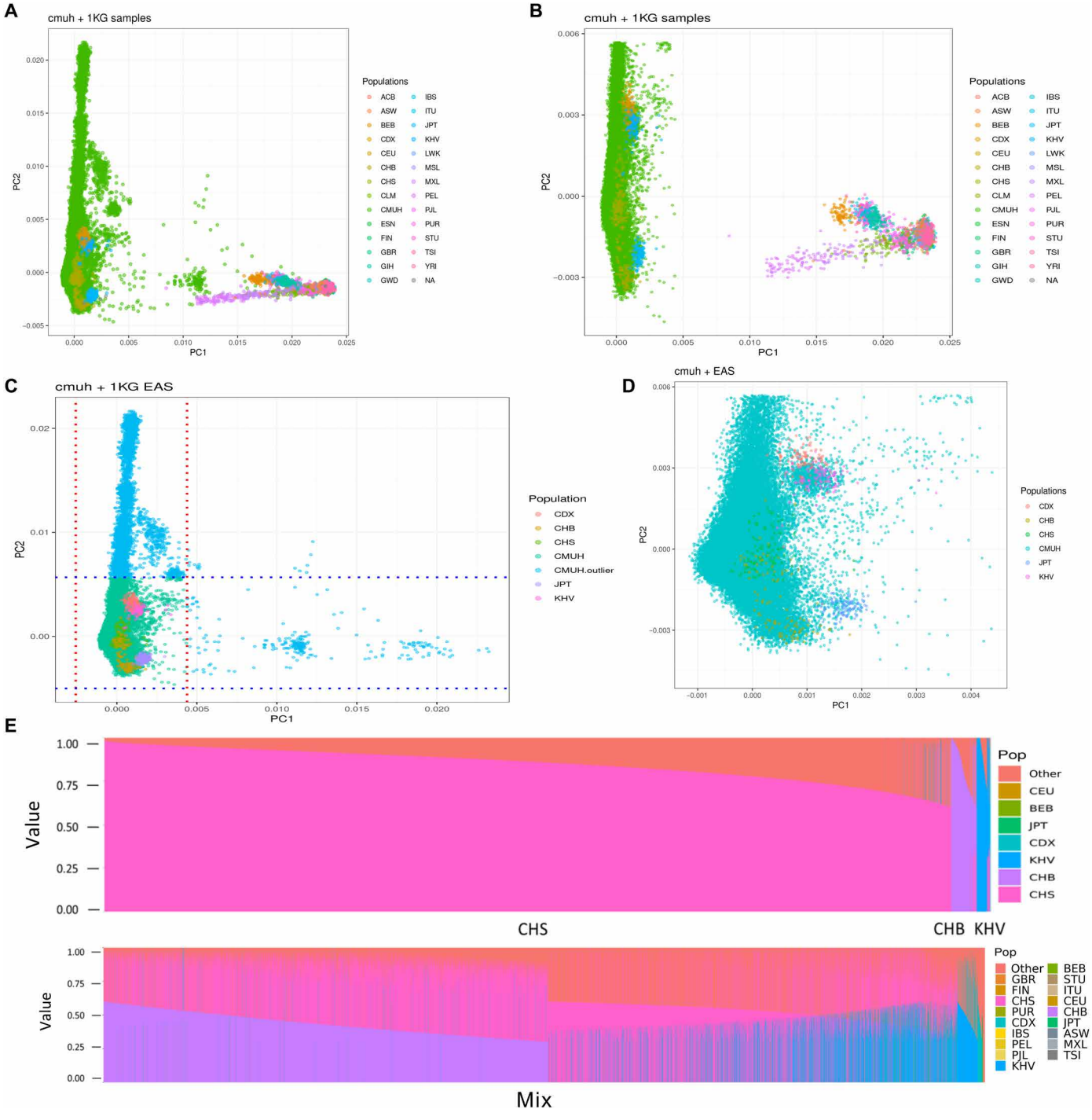


Fig. 3. PCA and ancestral analysis of data from the HiGenome cohort and 1000 Genomes Project. (**A**) Scatterplot depicting the PCA results for principal components 1 and 2. This analysis was conducted using data from both the HiGenome cohort and the 1000 Genomes Project. Most patients in the HiGenome cohort were clustered within the EAS cohort of the 1000 Genomes Project. (**B**) Visualization after the exclusion (from the EAS cohort) of data points with deviations exceeding an IQR of 3. (**C**) Focused view of the EAS region, with data points excluded because of deviation. (**D**) Subset of the HiGenome cohort juxtaposed with the EAS cohort. (**E**) Primary ancestral components for each individual. Most individuals in the HiGenome cohort belonged to the EAS population, which comprised Southern Han Chinese individuals, Han Chinese individuals from Beijing, and Kinh individuals from Ho Chi Minh, Vietnam. Postvisualization of ancestry for all participants; those with >50% from a singular ancestry are depicted in the upper section, whereas those without >50% from a single ancestry are depicted in the lower section, with different colors indicating varying origins.

Analysis of drug metabolism genes revealed that the intermediate metabolizers for *CYP2C19* (49.72%) and *CYP3A5* (43.10%) were the most common genes, with the least common genes being the ultrarapid metabolizers for *CYP2C19* and *MT-RNR1*:m.1494C>T, both at a rate of 0.003% (fig. S6 and table S5). In many genetic databases, results pertaining to drug metabolism genes are rarely presented, primarily because questionnaire-based clinical data do not cover the extensive use of various drugs. Overall, our genetic database is capable of tracking changes in drug dosages throughout the course of treatment with drugs such as warfarin (*21*) and aminoglycosides (*22*), making it particularly valuable.

Atlas of GWASs conducted using PheCodes available in HiGenome

To examine the genetic bases of common diseases, we analyzed our patients' diagnostic data. The case group comprised patients with diseases confirmed by three or more diagnostic instances conforming to the PheCode definition, whereas the control group comprised patients with at least a single diagnosis not conforming to the PheCode definition. After the extracted phenotypes were stratified (with a distribution of 8:2) into base and target datasets, the base dataset was used for GWASs on disease associations, whereas the target dataset was used for PRS calculation. For the case group, a patient threshold of more than 150 was exceeded; therefore, we examined the disease-gene associations for all 1085 traits (table S6), depicted in a circular Manhattan plot (Fig. 4, outer ring). In this plot, the 10 most significant genes for each trait are highlighted. Among these associations, the most significant pertained to traits associated both with the musculoskeletal, hematopoietic, circulatory, endocrine, or metabolic systems and with mental disorders ($P < 1 \times 10^{-70}$). Of the 187 traits with a P value of $<1 \times 10^{-10}$, the most prevalent disease classifications were related to the circulatory system, neoplasms, and the endocrine/metabolic. The predominant disease groups for which the number of patients exceeded 10,000 (across 13 traits) were found to be those affecting the circulatory, endocrine, metabolic, digestive, or genitourinary systems (Fig. 4, middle ring). Traits with a significant gene count of 10 or more were predominantly related to the endocrine, metabolic, circulatory, or integumentary (dermatology) systems (Fig. 4, inner ring). Table S2 presents detailed information on each trait. Each of 101 traits was associated with four or more significant genes. Diseases for which the sample was larger were associated with a larger number of significant genes. Given these findings, we stratified significant gene loci ($P < 5 \times 10^{-8}$) into three categories on the basis of the number of genes: strong association (more than eight loci), moderate association (four to eight loci), and weak association (fewer than four loci). Predominant associations were observed for traits related to neoplasms and the endocrine, metabolic, circulatory, integumentary, or genitourinary systems (fig. S7). However, weak associations were observed for most traits included in our database.

GWASs of diseases prevalent in the HiGenome cohort

The HiGenome database contains data on the following four traits: type 2 diabetes (T2D, 250-2), chronic renal failure or chronic kidney disease (CKD, 585-3), gout (274-1), and alcoholic liver damage (ALD, 317-11; Fig. 5). These traits are prevalent in Taiwan. For T2D, 57 significant gene loci were identified ($P < 5 \times 10^{-8}$; Fig. 5A, top). The most significant variant was rs2237897 (*KCNQ1*, $P = 2.9 \times 10^{-93}$), with an MAF of 0.34. As indicated by the region plot, a strong association was observed between rs2237897 and its adjacent variants in the EAS population ($r^2 > 0.6$; Fig. 5A, middle). This variant was primarily associated with diseases affecting the endocrine or metabolic systems, such as diabetes mellitus and hyperlipidemia (Fig. 5A, bottom). For CKD, nine significant gene loci were identified ($P < 5 \times 10^{-8}$; Fig. 5B, top). The most significant variant was rs56094641 (*FTO*, $P = 9.3 \times 10^{-12}$), with an MAF of 0.13. As indicated by the region plot, a strong association was observed between rs56094641 and its adjacent variants in the EAS population ($r^2 > 0.8$; Fig. 5B, middle). This variant was primarily associated with diseases affecting the circulatory, endocrine, metabolic, or genitourinary systems, such as hypertension, diabetes mellitus, or CKD, respectively (Fig. 5B, bottom). For gout, 11 significant gene loci were identified ($P < 5 \times 10^{-8}$; Fig. 5C, top). The most significant variant was rs4148155 (*ABCG2*, $P = 9.7 \times 10^{-187}$), with an MAF of 0.32. As indicated by the region plot, a strong association was observed between rs4148155 and its adjacent variants in the EAS population ($r^2 > 0.8$; Fig. 5C, middle). This variant was primarily associated both with diseases affecting the endocrine, metabolic, or genitourinary systems and with various symptoms, such as gout, abnormal blood chemistry, CKD, and calculus (Fig. 5C, bottom). For ALD, four significant gene loci were identified ($P < 5 \times 10^{-8}$; Fig. 5D, top). The most significant variant was rs3782886 (*BRAP*, $P = 1.2 \times 10^{-43}$), with an MAF of 0.32. As indicated by the region plot, a strong association was observed between rs3782886 and its adjacent variants in the EAS population ($r^2 > 0.8$; Fig. 5D, middle). This variant was primarily associated with mental disorders and diseases affecting the endocrine, metabolic, or circulatory systems, such as ALD, hypertension, and gout (Fig. 5D, bottom). Although only four diseases are presented here, we conducted GWASs for all 1085 traits by using the PheCode definition. The results are available on HiGenome for interested researchers.

Comparative genetic analysis of disease variability in the Taiwanese Han and European populations

To explore the differences between Taiwanese Han and European populations, we conducted a meta-analysis of two databases to examine the weighted effects [odds ratio (OR)] on four diseases. Data from the PheWeb database of the UKBB were used. With the same PheCode definitions used, T2D (250-2) and gout (274-1) were found to exhibit similar outcomes, unlike CKD (585-3) and ALD (317-11; fig. S8A). According to our meta-analysis, the significance of the genetic associations for T2D (250-2) and gout (274-1) increased. However, the outcomes of CKD (585-3) and ALD (317-11) were similar to those of CMUH, and they were not rendered significant by the UKBB data (fig. S8B). As shown in table S7, no SNPs were significant in our CKD data, in the UKBB, or after our meta-analysis. Although certain SNPs were significant across all three datasets for ALD, their number was smaller than that associated with gout and T2D. For common diseases such as T2D, the data of CMUH exhibited the same significant associations with the *CDKAL1* and *FTO* genes as in the UKBB. However, the *RSPO3* and *AUTS2* genes exhibited significant associations only in the CMUH database. For more details, please refer to table S8 (A to D). Together, our findings confirm the distinct genetic SNP profiles associated with diseases in our population and in European populations, particularly in those with CKD and ALD.

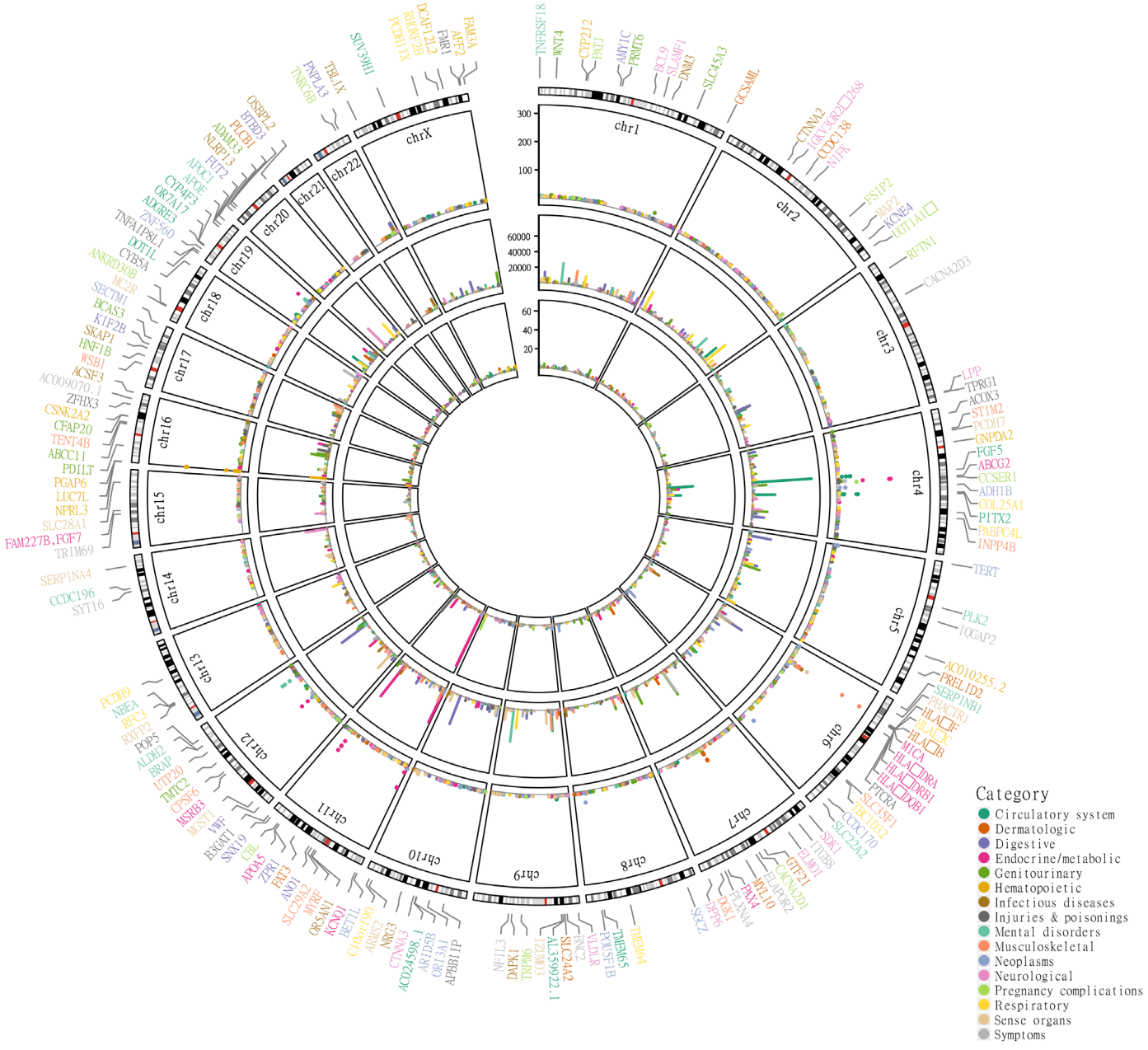


Fig. 4. Atlas of GWASs depicting significant genetic associations across traits. The outermost circle presents the most significant genes for each trait, plotted on the basis of their *P* values on a Manhattan plot. On this circle, the 10 most significant genes for each disease category are marked and color coded in accordance with disease classification. The middle circle depicts the number of individuals affected by each trait. The innermost circle depicts the number of significant genes ($P < 5 \times 10^{-8}$) associated with each trait.

PRSs for diseases prevalent in the HiGenome cohort

To calculate PRSs, we refined our initial group of 1085 traits on the basis of the following criteria. More than 1000 patients had to have a specific trait and at least one significant gene locus ($P < 5 \times 10^{-8}$). This screening procedure yielded 238 traits, whose PRSs were calculated. Area under the curve (AUC) values were calculated for these PRSs. Even when a prediction was exclusively made using a PRS, most traits did not achieve an AUC of >0.6. This threshold was exceeded by only 15 traits, including ankylosing spondylitis, type 1 diabetes, and psoriasis. Notably, when the regression models were adjusted for confounders such as age, sex, and PCA results, the number of traits with an AUC of >0.8 reached 51 (table S9). We identified 31 traits for which neither the PRS alone nor its combination with clinical features achieved an AUC value of >0.6 (Fig. 6, left). By contrast, we identified nine traits for which the PRS alone achieved an AUC value of >0.6, with its combination with clinical features achieving an AUC value of >0.7. These traits were predominantly related to neoplasms and the endocrine, metabolic, circulatory, and musculoskeletal systems (Fig. 6, right).

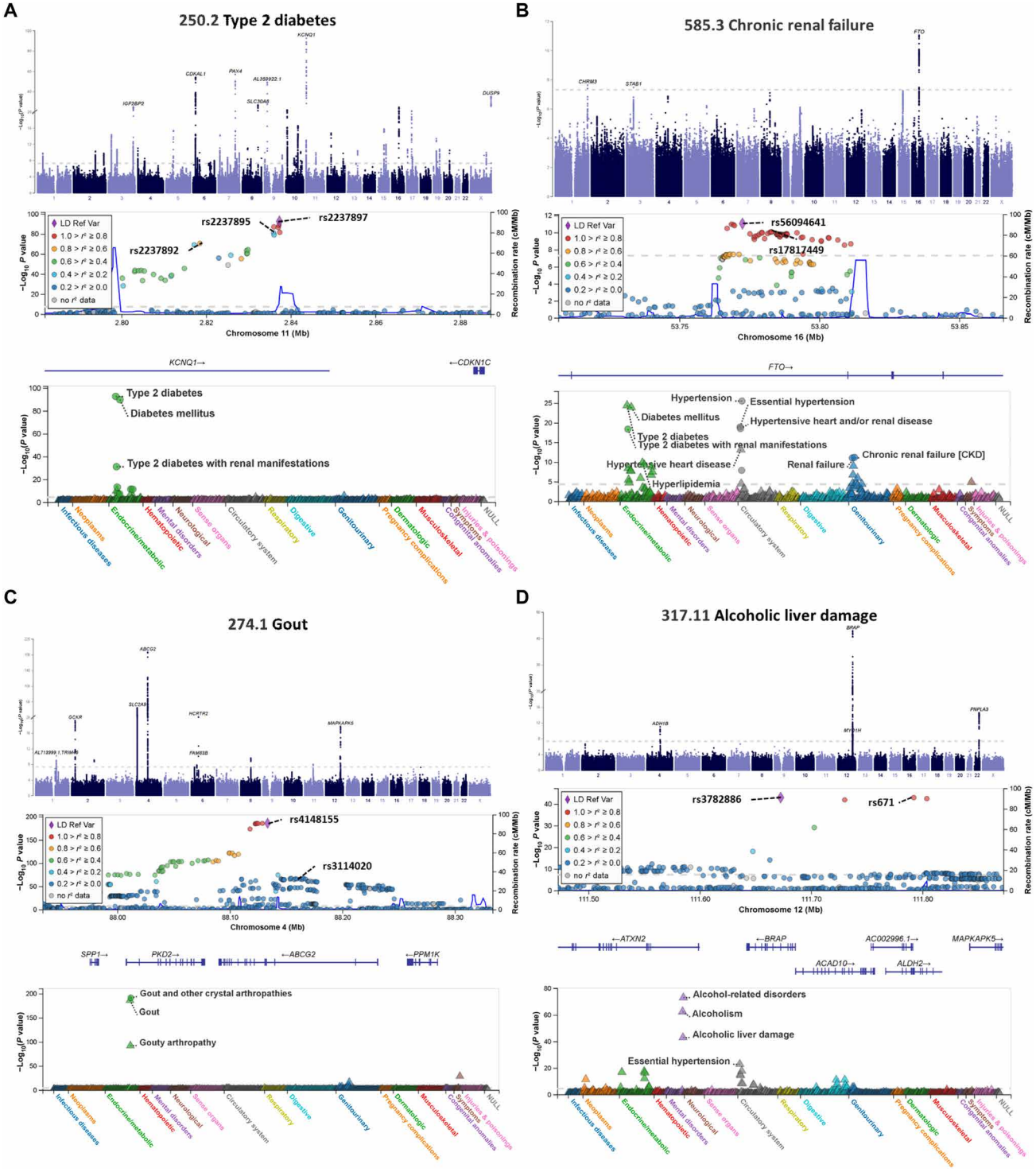


Fig. 5. Comprehensive GWASs of selected diseases. Manhattan plots (top) of significant gene loci associated with various diseases: (**A**) T2D, (**B**) CKD, (**C**) gout, and (**D**) ALD. The *x* axis represents the absolute chromosomal positions of the genes, and the *y* axis represents the corresponding *P* values. Region plots (middle) of the most significant variant loci adjacent to those associated with the EAS population; variant associations are color coded to indicate the degree of correlation. Results of PheWASs (bottom) for the most significant variant loci associated with each disease, color coded in accordance with disease classification.

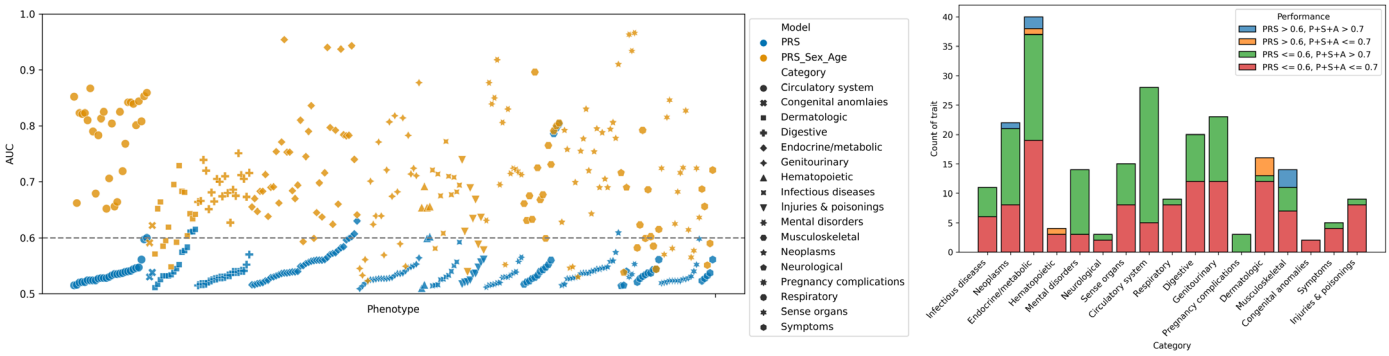
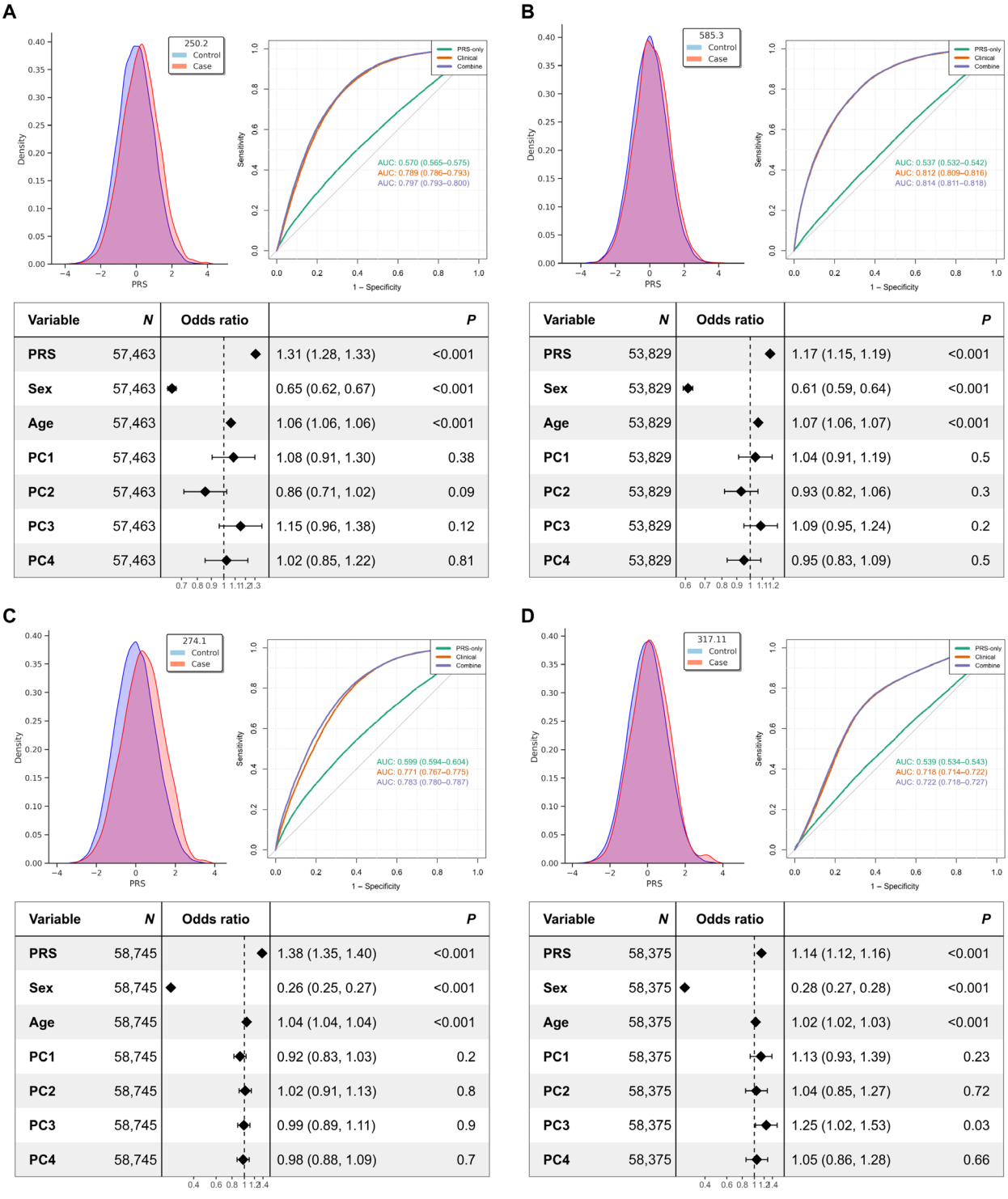


Fig. 6. Statistical analysis of PRS models across traits. (**Left**) AUC value for each trait. Traits in blue indicate AUC values derived exclusively from the PRS, where traits in yellow indicate AUC values derived from the model incorporating both the PRS and clinical features. Each symbol indicates unique disease classification. Most traits initially exhibited an AUC of <0.6; however, with the addition of clinical features, most traits exhibited an AUC of >0.6. (**Right**) High-performance PRS models. Traits with a PRS AUC of >0.6 and PRS+clinical features AUC of >0.7 are highlighted in blue. These traits are predominantly related to endocrinological, musculoskeletal, and other relevant diseases. AUC, area under the curve; PRS, polygenic risk score.

A total of 19,486 variants were selected for the T2D PRS model. Differences were observed between the case and control groups in terms of PRS distribution. Specifically, the median PRS was significantly higher in the case group than in the control group (Fig. 7A, top left). A forest plot revealed that the ORs for PRS, sex, and age were 1.31 [95% confidence interval (CI), 1.28 to 1.33], 0.65 (95% CI, 0.62 to 0.67), and 1.06 (95% CI, 1.06 to 1.06), respectively, all with significant between-group differences ($P < 0.001$). No significant between-group difference was observed in any principal component (principal components 1 to 4; Fig. 7A, bottom). Cross-validation revealed median AUC values of 0.57 [interquartile range (IQR) = 0.565 to 0.575], 0.789 (IQR = 0.786 to 0.793), and 0.793 (IQR = 0.793 to 0.8) for PRS, clinical features, and their combination, respectively (Fig. 7A, top right). A total of 23,191 variants were selected for the CKD PRS model. The median PRS was significantly higher in the case group than in the control group (Fig. 7B, top left). The ORs for PRS, sex, and age were 1.17 (95% CI, 1.15 to 1.19), 0.61 (95% CI, 0.59 to 0.64), and 1.07 (95% CI, 1.06 to 1.07), respectively (all $P < 0.001$). No significant between-group difference was observed in any principal component (Fig. 7B, bottom). The AUC values were 0.537 (IQR = 0.532 to 0.542), 0.812 (IQR = 0.809 to 0.816), and 0.814 (IQR = 0.811 to 0.818) for PRS, clinical features, and their combination, respectively (Fig. 7B, top right). A total of 32 variants were selected for the gout PRS model. Significant differences were observed in the distribution patterns of the PRS between the case and control groups. The median PRS was significantly higher in the case group than in the control group (Fig. 7C, top left). The ORs for PRS, sex, and age were 1.38 (95% CI, 1.35 to 1.4), 0.26 (95% CI, 0.25 to 0.27), and 1.04 (95% CI, 1.04 to 1.04), respectively (Fig. 7C, bottom). The AUC values were 0.599 (IQR = 0.594 to 0.604), 0.771 (IQR = 0.767 to 0.775), and 0.783 (IQR = 0.78 to 0.787) for PRS, clinical features, and their combination, respectively (Fig. 7C, top right). A total of 23 variants were selected for the ALD PRS model. Differences were observed between the case and control groups in terms of PRS distribution. Specifically, the median PRS was significantly higher in the case group than in the control group (Fig. 7D, top left). The ORs for PRS, sex, and age were 1.14 (95% CI, 1.12 to 1.16), 0.26 (95% CI, 0.25 to 0.27), and 1.04 (95% CI, 1.04 to 1.04), respectively (Fig. 7D, bottom). The AUC values were 0.539 (IQR = 0.534 to 0.543), 0.718 (IQR = 0.714 to 0.722), and 0.722 (IQR = 0.718 to 0.727) for PRS, clinical features, and their combination, respectively (Fig. 7D, top right). For the four diseases, the AUC values for the PRS models were ~0.6. Among all clinical features, only age and sex had significant effects; however, no contributions from principal components were observed. Notably, the AUC value for clinical features reached 0.8, and the AUC value for the combination of PRS and clinical features occasionally exceeded this threshold, indicating the strong predictive ability of the combined model. For more information on the PRS results, please refer to HiGenome. Together, our findings underscore the benefits of incorporating clinical features with PRSs to increase the accuracy of predicting diseases prevalent among the Taiwanese Han population. This integrated approach holds promise for precision medicine.

DISCUSSION

Supported by CMUH, the HiGenome database was developed as a specialized resource for the Taiwanese Han population. This database sets a benchmark through its integration of deidentified EMRs and genomic data, providing a comprehensive longitudinal dataset for genetic research. Major international biobanks, such as the UKBB, FinnGen, TWB, and the MVP, incorporate both clinical data and questionnaire-based self-reported information. Although UKBB and MVP include EHRs from hospitals and clinics, much of their phenotype data relies on baseline health assessments and self-reported medical questionnaires, which are subject to recall bias. Questionnaire responses are typically more accurate for recent medical events, but may be less reliable for conditions with variable onset or long latency periods (*26*–*28*). In contrast, HiGenome eliminates reliance on self-reported data by integrating detailed physician-documented EMRs. This approach enhances data accuracy and disease classification, particularly for chronic and progressive diseases where multiple clinical visits refine the diagnosis over time. In addition, HiGenome benefits from up to 19 years of longitudinal follow-up, making it one of the most extensive EAS genetic datasets with deeply integrated clinical records. Another key distinction of the HiGenome cohort is its age distribution as a significant proportion of participants are under 45 years of age. This younger demographic provides early insights into disease manifestation and enhances PRS validation, offering a predictive advantage for early intervention in



Variable	N	Odds ratio	P
PRS	57,463	1.31 (1.28, 1.33)	<0.001
Sex	57,463	0.65 (0.62, 0.67)	<0.001
Age	57,463	1.06 (1.06, 1.06)	<0.001
PC1	57,463	1.08 (0.91, 1.30)	0.38
PC2	57,463	0.86 (0.71, 1.02)	0.09
PC3	57,463	1.15 (0.96, 1.38)	0.12
PC4	57,463	1.02 (0.85, 1.22)	0.81

Variable	N	Odds ratio	P
PRS	53,829	1.17 (1.15, 1.19)	<0.001
Sex	53,829	0.61 (0.59, 0.64)	<0.001
Age	53,829	1.07 (1.06, 1.07)	<0.001
PC1	53,829	1.04 (0.91, 1.19)	0.5
PC2	53,829	0.93 (0.82, 1.06)	0.3
PC3	53,829	1.09 (0.95, 1.24)	0.2
PC4	53,829	0.95 (0.83, 1.09)	0.5

Variable	N	Odds ratio	P
PRS	58,745	1.38 (1.35, 1.40)	<0.001
Sex	58,745	0.26 (0.25, 0.27)	<0.001
Age	58,745	1.04 (1.04, 1.04)	<0.001
PC1	58,745	0.92 (0.83, 1.03)	0.2
PC2	58,745	1.02 (0.91, 1.13)	0.8
PC3	58,745	0.99 (0.89, 1.11)	0.9
PC4	58,745	0.98 (0.88, 1.09)	0.7

Variable	N	Odds ratio	P
PRS	58,375	1.14 (1.12, 1.16)	<0.001
Sex	58,375	0.28 (0.27, 0.28)	<0.001
Age	58,375	1.02 (1.02, 1.03)	<0.001
PC1	58,375	1.13 (0.93, 1.39)	0.23
PC2	58,375	1.04 (0.85, 1.27)	0.72
PC3	58,375	1.25 (1.02, 1.53)	0.03
PC4	58,375	1.05 (0.86, 1.28)	0.66

Fig. 7. Detailed PRS analysis of key diseases. The top-left panel displays the distribution of PRSs in the case and control groups for (**A**) T2D, (**B**) CKD, (**C**) gout, and (**D**) ALD. The *x* axis represents the normalized PRSs. The top-right panel displays the results of 10-fold cross-validation for AUC values; the outcomes of models including PRS, clinical features, or their combination are indicated using different colors. The bottom panel presents a forest plot for each feature, offering insights into patient count, OR, and statistical significance. PRS, polygenic risk score.

precision medicine. By combining longitudinal clinical data with genomic insights, HiGenome serves as a valuable resource for advancing genetic research in the Taiwanese Han population and contributes to global efforts in expanding non-European representation in genetic studies.

In addition to its primary characteristics, the CMUH database exhibits significant growth potential, including longitudinal tracking for up to 19 years (and ongoing). It also seamlessly integrates into the Golden Baby Project, a project run by the children's division of CMUH to start tracking participants since birth. This collaboration enables long-term monitoring of genetic effects on health. In addition, the CMUH database has potential for future integration with Taiwan's National Health Insurance Database, which would provide comprehensive records of medical visits and causes of death. These capabilities position the CMUH database as an exceptionally promising hospital-based database with major potential for advancing medical research.

When comparing our findings with those from the UKBB, several key differences emerge (fig. S8 and table S8). These include variations in case numbers (which affect statistical power), differences in MAFs across ancestries, and variations in effect sizes. Notably, we observed that some associations present only in Taiwanese Han population were absent in UKBB, likely due to the corresponding variants being extremely rare in European cohorts. For example, rs671 in ALDH2, a well-known variant associated with alcohol dependence (ALD), is common in the Taiwanese Han population (MAF = 0.28) but extremely rare in Europeans (MAF < 0.001 not reported in the PheWeb), making it likely to be excluded from UKBB analyses due to low MAF thresholds (table S8D). Similarly, we found significant differences in effect sizes for certain variants between the two populations, emphasizing the need to consider ancestry-specific genetic architectures in PRS models. For example, rs6546932 in the *SELENOI* gene showed a notable discrepancy: In the Taiwanese Han population, it had an OR of 1.58, whereas in the UKBB, the effect size was BETA = 0.189 (corresponding to OR = 1.21) (table S8D). This difference, as demonstrated in our previous study (*29*), highlights the potential impact of population-specific genetic backgrounds on disease associations and underscores the importance of tailoring PRS models to different ancestries.

In this study, many variations were observed in the variants selected for different diseases. PRS models were constructed using variants identified automatically by PRSice-2 (*30*). For certain diseases, only one variant was selected, whereas for others, up to 35,295 variants were selected (table S9). Notably, no correlation was observed between the number of variants and model efficacy. Instead, the predictive power of each model was accurately reflected by the cohort size (*31*, *32*). Generally, to establish PRS models, PRSice-2 is used to identify the most significant disease-associated variant combinations through linkage disequilibrium calculations. Given the varying heritability of different diseases, AUC values are not typically robust for all diseases. In addition, the performance of PRSice-2 is typically limited when the sample size is small (*30*). In this study, when we used the PRS alone in our disease models, we consistently obtained AUC values of <0.7. When we adjusted our models for age and sex, the AUC values rarely exceeded 0.9 (table S3). These findings are consistent with those of other studies demonstrating the effects of multiple factors on polygenic diseases (*15*, *16*). Given that the prevalence of most diseases increases with age, incorporating age invariably increases model accuracy (Fig. 6). In addition to age and sex, other clinical features, such as body mass index, blood pressure, glycated hemoglobin level, various biomarkers, and environmental factors (e.g., exercise, diet, alcohol consumption, and smoking), can be included to achieve further increases in model accuracy, thereby offering a promising direction for future research. Recently, many researchers have started to acknowledge the effect of ancestry on PRS accuracy (*29*). In ethnically diverse countries, a key challenge arises when models primarily designed for European cohorts are applied to other ethnic groups; this challenge may lead to suboptimal outcomes (*5*). Therefore, in future studies, PRS models should be adjusted for ancestry factors to increase their applicability across populations. These adjustments should align with our findings (Fig. 7), which emphasize the importance of incorporating multiple clinical features and ancestry factors to increase the accuracy and applicability of PRS models, particularly in multiethnic contexts. Early application of PRSs may limit unnecessary screenings (*33*).

Overall, the diseases selected in this study largely represent those of significant concern among the Taiwanese Han population. Given the importance of PRS models in research, in a previous study, we identified variants associated T2D (*34*, *35*). In Asian populations, rs2237895, rs2237897, and rs2237892 are identified as three linked SNPs located on *KCNQ1* and significantly associated with T2D (*35*). These SNPs are implicated in the modulation of insulin secretion (*36*). In this study, we found that our findings for gout, a condition characterized by sex-specific prevalence, are consistent with those of previous studies (Fig. 7C), indicating the importance of *ABCG2*, particularly when sex is adjusted for (*37*, *38*). CKD is prevalent in Taiwan, as evidenced by the high rate of dialysis, and has attracted major attention (*39*, *40*). In this study, we found an association between the *FTO* gene and CKD, linking *FTO* with the so-called triad of diabetes, hypertension, and hyperlipidemia. In addition to *FTO*, genes such as *CHRM3*, *STAB1*, *WDR72*, *BHLHE22*, *ABCG2*, *ZMAT4*, *MAT2B*, and *RABGAP1* were found to be associated with CKD, indicating the high predictive power of PRS models for CKD (*41*, *42*). Overall, ALD represents an intriguing case for Taiwan, with the *ALDH2* variant rs671 being highly prevalent among the Taiwanese Han population (*43*). In this study, we identified a significant association between ALD and the *BRAP* variant rs3782886, which was found to be strongly linked to rs671 (Fig. 5D). This association is consistent with that observed between rs671 and alcohol-related diseases in Taiwan.

Several HLA-associated diseases were identified, including ankylosing spondylitis, psoriasis, eye inflammation, chronic sinusitis, acute sinusitis, Graves' disease, asthma, hepatitis B, type 1 diabetes, palindromic rheumatism, systemic lupus erythematosus, hypothyroidism, rheumatoid arthritis, and primary liver malignancy (table S6). These diseases are predominantly related to autoimmunity, immunity, or viral infection. In previous studies, we explored the associations of Graves' disease (*23*) and rheumatoid arthritis (*44*) with HLA. Further comprehensive research is required to explore the associations between various HLA subtypes and these diseases.

This study has several limitations. First, this study relied on EMR data collected from a single center. Second, the study involved unrecorded comorbidities, which may have led to false-negative outcomes in our case and control groups. However, given the generally low prevalence of many diseases in the study population, the rate of false-negative results may have been negligible (*45*). Our GWAS results are largely consistent with the literature (*25*, *38*, *46*–*48*), indirectly confirming the minimal effects of false-negative results. Last,

in Taiwan, diagnostic recording is influenced by the health care system. Many diagnoses depend on physicians' decisions to order specific tests, resulting in the documentation of unconfirmed diagnoses. To minimize this effect, we implemented a criterion of three or more diagnoses when selecting patients for the case group, thereby limiting the inclusion of patients with a single diagnosis. This approach effectively reduced the number of false-positive results, as evidenced by the consistency observed between our findings and those of previous studies (*25*, *46*, *48*–*50*). In future studies, we recommend the implementation of stricter and more comprehensive criteria, with a combination of diagnosis, medication history, and laboratory test results, to yield clearer outcomes, similar to our single-disease GWASs (*25*, *38*, *44*, *46*, *48*). Because HiGenome is a hospital-centric database, a major challenge is the absence of subhealthy individuals, meaning that virtually all participants have at least one documented ailment. In actual scenarios, no individual is entirely free of disease; problems arise from either underinvestigation or a lack of documentation. Therefore, our methodology of excluding conditions strongly associated with the control group can be regarded as a feasible approach for research (*20*, *51*, *52*).

PRS models are developed for clinical applications, such as the prediction and prevention of diseases before their onset. These PRSs have multifaceted clinical applications. In addition to disease prediction and prevention, they can increase the predictive power of medical imaging or electrophysiological reading models within smart health care systems. In addition, integrating PRSs into artificial intelligence models is regarded as a promising approach. Together, our findings underscore the importance of integrating EMRs with genomics to provide valuable insights into disease predisposition among the Taiwanese Han population. Considered a pioneering effort in Taiwan, the HiGenome database not only bridges the gap between clinical practice and genomic research but also lays the foundation for advancements in personalized medicine. Our findings, particularly in the context of multifactorial diseases, emphasize the need for a holistic approach, one that incorporates genetic and nongenetic factors, to increase the accuracy of predictive models. Overall, the strength of our research lies in its integrative approach, extensive follow-up period, and focus on diseases prevalent among the Taiwanese population. This study sets the stage for future targeted interventions.

MATERIALS AND METHODS

Databases

The individuals in this study were divided into two cohorts. One cohort comprised individuals who had participated in a precision medicine project at CMUH. The primary objective of this project, which was launched in 2018, was to explore genetic predisposition to common diseases within the Taiwanese Han population and establish a refined system for predicting and preventing these diseases. The project focused on patients treated at CMUH and was approved by the Institutional Review Board of CMUH. The second cohort comprised patients whose data were extracted from EMRs covering the period from 2003 to 2021. This dataset included patients' anamnesis and laboratory data, such as DNA SNP microarray findings, which are essential for investigating drug-induced side effects. At the Department of Laboratory Medicine of CMUH, which is accredited by the American College of Pathologists, the TPMv1 array is used to identify single-nucleotide variants, such as HLA types, associated with drug-induced side effects. This array is also used to facilitate GWASs of common diseases.

As of the time of this study, a total of 413,210 patients were enrolled. After excluding twin pairs, first-degree relatives, and individuals who were not of EAS ancestry, the final study cohort comprised 323,397 participants. Recruitment is still ongoing.

To ensure patient confidentiality, personal medical details were encrypted, and patient data were used exclusively for research purposes. In our case-control study, EMR data obtained from CMUH were used as the foundational dataset. This dataset included patient demographics, laboratory results, medical procedures, and diagnostic codes as outlined by the *International Classification of Diseases*, *Ninth Revision*, *Clinical Modification* (*ICD-9-CM*) and *International Classification of Diseases*, *Tenth Revision*, *Clinical Modification* (*ICD-10-CM*). CMUH archives disease data by using *ICD-9-CM* and *ICD-10-CM* codes, with the *ICD-9-CM* codes automatically converted into their corresponding *ICD-10-CM* codes. In this study, EMR data corresponding to the period from 2003 to 2021 were analyzed. Medical diagnoses were established in accordance with the PheCode criteria, which were applied on at least three distinct occasions. To establish a control group, we applied the PheCode criteria, including individuals who did not have PheCode-defined diseases (*20*). This study was approved by the Institutional Review Board of CMUH (approval nos. CMUH110-REC3-005 and CMUH111-REC1-176).

Genotyping

Blood samples were collected from all participants. Genomic DNA was extracted from 200-μl blood samples by using a MagCore Genomic DNA Whole Blood Kit (RBC Bioscience, New Taipei City, Taiwan) in accordance with the manufacturer's instructions, followed by elution at a final volume of 60 μl. An Affymetrix Axiom genotyping platform was used to obtain genetic information from the samples; specifically, we used a custom Axiom TPMv1 SNP array (Thermo Fisher Scientific, Santa Clara, CA, USA), which includes 714,457 SNPs across the entire human genome. Genotype analysis was conducted using PLINK 1.9. After SNPs with a call rate of <0.95 (–geno 0.05) were excluded, samples and SNPs with missing rates (–geno 0.02 for SNPs and –mind 0.02 for samples) were removed. Subsequently, monomorphic SNPs with a count of <10 (–mac 10) and multiallelic SNPs were eliminated. Variants with Hardy-Weinberg equilibrium (HWE) P values of $<1 \times 10^{-6}$ and MAF values of $<1 \times 10^{-4}$ were also excluded. We incorporated the following analysis criteria into our study methodology: heterozygous outliers exceeding a standard deviation value of 5, PCA outliers exceeding an IQR of 3, and mismatches between genotypic and actual sex. We also used the KING-robust kinship estimator (PLINK 2.0) (*53*) to remove duplicate samples and first-degree relatives from our cohort and ensure that the genetic data were free from inflationary effects. SHAPEIT4 was used to phase the TPM arrays. Imputation was performed using Beagle 5.2, which is more effective and accurate than other imputation tools. The imputed data were filtered on the basis of the following criteria: an R^2 alternate allele dosage of <0.3 and a genotype posterior probability of <0.9 (*18*). Ultimately, a total of 14,181,206 applicable variants were obtained.

Genome-wide association studies

To identify disease-associated genetic variants, summary statistics were derived from the dataset of Taiwanese Han individuals. These summary statistics were derived using PLINK 2.0 (*54*). After relevant

data were extracted from patient EMRs, the participants were divided into a case group and a control group on the basis of their PheCode classification. For details on participant distribution, please refer to table S1. Logistic regression was used to determine the correlations between various traits. Regression models were adjusted for several confounders, such as age, sex, and PCA results. To minimize the influence of pronounced linkage disequilibrium, which can lead to overestimation, we examined the most significant variant within each genomic region. A stringent P value ($<5 \times 10^{-8}$) was adopted to identify significant associations between the case and control groups. The results were visualized using Manhattan, region, and quantile-quantile plots. A regional visualization highlighting the variants of interest was constructed using the PheWeb platform (*55*).

PCA and ancestry analysis

After integrating our dataset with that of the 1000 Genomes Project, we performed SNP pruning. For this purpose, we used PLINK, adopted a window of 200 SNPs and an r^2 threshold of >0.1, and advanced in steps of 100 SNPs. This process enabled us to retain 66,728 variants that were common between our dataset and that of the 1000 Genomes Project. Ancestry analysis was conducted using ADMIXTURE in a supervised manner (*56*).

HLA imputation

To impute HLA alleles at a four-digit resolution, HLA genotype imputation was conducted using HiBAG (*57*) with attribute bagging analysis. After combining the results of the TWB2 chip from the TWB (approval no. TWBR11208-04) with targeted sequencing data for the HLA region, we used a cohort of 861 individuals to train our model. During the training phase, quality control measures were applied to the TWB2 chip, and 41,857 SNPs located in the HLA gene region on chromosome 6 were extracted. These SNPs were subsequently input into the HiBAG software to generate an HLA training model. After quality control, the SNP results of the TPMv1 chip were also input into the model to obtain corresponding HLA predictions. Default settings were used for analysis. HLA genes including HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1, and HLA-DPB1 were imputed, with an imputation posterior probability of >0.9 regarded as reliable (*23*).

Pharmacogene allele construction

After imputation, variants with an INFO R^2 of ≥0.3, an MAF of >0.0001, and an HWE P value of $>1 \times 10^{-7}$ in variant call format were included for analysis. With the default settings for PGxPOP used (*58*), pharmacogenomic phenotypes were generated from the imputed variant call format file, with GRCh38 as the reference genome. These tools are typically used to extract gene variants and infer star allele haplotypes for each individual in accordance with Clinical Pharmacogenetics Implementation Consortium (CPIC) allele definitions. A total of 10 genes, namely, CYP2B6, CYP2C19, CYP2C9, CYP3A5, CYP4F2, DPYD, NUDT15, SLCO1B1, TPMT, and VKORC1, were included for analysis. In cases where multiple pharmacogenomic phenotypes were inferred by PGxPOP for a single individual, the phenotype was recorded as "NA." To enhance the results of drug metabolism for HLA genes, we used the results of HiBAG analysis.

Meta-analysis using UKBB data

A meta-analysis was conducted using data from TOPMed-imputed PheWeb (*59*) (https://pheweb.org/UKB-TOPMed/), with corresponding PheCode matching. This analysis was conducted using METAL software (*60*) (https://genome.sph.umich.edu/wiki/METAL_Documentation; version: generic-metal-2011-03-25).

PRS calculation and model construction

To calculate patient PRSs, data from the CMUH cohort was randomly divided into base and target datasets. The base dataset was used to investigate the associations between study variables and diseases delineated by PheCodes; this investigation was conducted using PLINK 1.9. After variants with MAF values of >0.01 were filtered, PRSs were compiled from the target dataset by using PRSice2, with the dataset of the 1000 Genomes Project (Phase 3), which is specific to the EAS population, selected as the reference standard. PRSs were calculated using z-score normalization. These scores, along with their clinical features, were used to construct logistic regression models, which were subsequently adjusted for confounding factors such as age, sex, and PCA results.

Phenome-wide association studies

A total of 58,257,251 *ICD-9-CM* or *ICD-10-CM* diagnostic codes were combined into 1791 PheCodes (*61*). Because of the limited variation in data and the insufficient number of participants in certain categories, the final categorization was narrowed down to 1085 PheCodes for subsequent analyses. Logistic regression was used to analyze the associations between variants and each PheCode by using the "PheWAS" package in R software (R Foundation for Statistical Computing, Vienna, Austria) (*62*). All summary statistics from PheWASs are publicly available on the HiGenome website, which provides interactive Manhattan plots, region plots, and PheWAS visualizations from the PheWeb platform (*55*).

Statistical analysis

Baseline continuous and categorical variables were analyzed using Wilcoxon rank sum test and Pearson's chi-square test, respectively. Between-group comparisons were conducted using one-way analysis of variance (ANOVA) followed by Tukey's post hoc test. A two-sided P value of <0.05 was considered statistically significant. The false discovery rate was adjusted using the Benjamini-Hochberg method. All statistical analyses were conducted using IBM SPSS Statistics version 22 (IBM, Armonk, NY, USA) and R software version 4.1.0.

Ethics statement

The study protocol was approved by the Institutional Review Board of China Medical University Hospital (CMUH111-REC1-176 and CMUH110-REC3-005). Deidentified genetic and clinical data were collected after informed consent was obtained from the patients.

Supplementary Materials

This PDF file includes:

Figs. S1 to S8

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S9

REFERENCES AND NOTES

1. S. Fatumo, T. Chikowore, A. Choudhury, M. Ayub, A. R. Martin, K. Kuchenbaecker, A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
2. L. A. Hindorff, V. L. Bonham, L. C. Brody, M. E. C. Ginoza, C. M. Hutter, T. A. Manolio, E. D. Green, Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
3. C. Huntley, B. Torr, A. Sud, C. F. Rowlands, R. Way, K. Snape, H. Hanson, C. Swanton, J. Broggio, A. Lucassen, M. M. Cartney, R. S. Houlston, A. D. Hingorani, M. E. Jones, C. Turnbull, Utility of polygenic risk scores in UK cancer screening: A modelling analysis. *Lancet Oncol.* **24**, 658–668 (2023).
4. S. A. Thomas, C. J. Browning, F. J. Charchar, B. Klein, M. G. Ory, H. Bowden-Jones, S. R. Chamberlain, Transforming global approaches to chronic disease prevention and management across the lifespan: Integrating genomics, behavior change, and digital health solutions. *Front. Public Health* **11**, 1248254 (2023).
5. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
6. A. Abdellaoui, L. Yengo, K. J. H. Verweij, P. M. Visscher, 15 years of GWAS discovery: Realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
7. E. Wong, N. Bertin, M. Hebrard, R. Tirado-Magallanes, C. Bellis, W. K. Lim, C. Y. Chua, P. M. L. Tong, R. Chua, K. Mak, T. M. Lim, W. Y. Cheong, K. E. Thien, K. T. Goh, J.-F. Chai, J. Lee, J. J.-Y. Sung, T. Y. Wong, C. W. L. Chin, P. D. Gluckman, L. L. Goh, K. H. K. Ban, T. W. Tan, SG10K_Health Consortium, X. Sim, C.-Y. Cheng, S. Davila, N. Karnani, K. P. Leong, J. Liu, S. Prabhakar, S. Maurer-Stroh, C. S. Verma, P. Krishnaswamy, R. S. M. Goh, I. Chia, C. Ho, D. Low, S. Virabhak, J. Yong, W. Zheng, S. W. Seow, Y. K. Seck, M. Koh, J. C. Chambers, E. S. Tai, P. Tan, The Singapore National Precision Medicine Strategy. *Nat. Genet.* **55**, 178–186 (2023).
8. S. Moon, Y. J. Kim, S. Han, M. Y. Hwang, D. M. Shin, M. Y. Park, Y. Lu, K. Yoon, H.-M. Jang, Y. K. Kim, T.-J. Park, D. S. Song, J. K. Park, J.-E. Lee, B. J. Kim, The Korea Biobank Array: Design and identification of coding variants associated with blood biochemical traits. *Sci. Rep.* **9**, 1382 (2019).
9. K. Nam, J. Kim, S. Lee, Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits. *Cell Genom.* **2**, 100189 (2022).
10. C.-H. Chen, J.-H. Yang, C. W. K. Chiang, C.-N. Hsiung, P. E. Wu, L.-C. Chang, H.-W. Chu, J. Chang, I.-W. Song, S.-L. Yang, Y.-T. Chen, F.-T. Liu, C.-Y. Shen, Population structure of Han Chinese in the modern Taiwanese population based on 10,000 participants in the Taiwan Biobank project. *Hum. Mol. Genet.* **25**, 5321–5331 (2016).
11. C. T. Fan, J.-C. Lin, C.-H. Lee, Taiwan Biobank: A project aiming to aid Taiwan's transition into a biomedical island. *Pharmacogenomics* **9**, 235–246 (2008).
12. C.-Y. Wei, J.-H. Yang, E.-C. Yeh, M.-F. Tsai, H.-J. Kao, C.-Z. Lo, L.-P. Chang, W.-J. Lin, F.-J. Hsieh, S. Belsare, A. Bhaskar, M.-W. Su, T.-C. Lee, Y.-L. Lin, F.-T. Liu, C.-Y. Shen, L.-H. Li, C.-H. Chen, J. D. Wall, J.-Y. Wu, P.-Y. Kwok, Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. *NPJ Genom. Med.* **6**, 10 (2021).
13. S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T. Konuma, K. Yamamoto, M. Akiyama, K. Ishigaki, A. Suzuki, K. Suzuki, W. Obara, K. Yamaji, K. Takahashi, S. Asai, Y. Takahashi, T. Suzuki, N. Shinozaki, H. Yamaguchi, S. Minami, S. Murayama, K. Yoshimori, S. Nagayama, D. Obata, M. Higashiyama, A. Masumoto, Y. Koretsune, FinnGen, K. Ito, C. Terao, T. Yamauchi, I. Komuro, T. Kadowaki, G. Tamiya, M. Yamamoto, Y. Nakamura, M. Kubo, Y. Murakami, K. Yamamoto, Y. Kamatani, A. Palotie, M. A. Rivas, M. J. Daly, K. Matsuda, Y. Okada, A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
14. G. Sato, Y. Shirai, S. Namba, R. Edahiro, K. Sonehara, T. Hata, M. Uemura, Y. Yamanashi, Y. Furukawa, T. Morisaki, Y. Murakami, Y. Kamatani, K. Muto, A. Nagai, W. Obara, K. Yamaji, K. Takahashi, S. Asai, Y. Takahashi, T. Suzuki, N. Sinozaki, H. Yamaguchi, S. Minami, S. Murayama, K. Yoshimori, S. Nagayama, D. Obata, M. Higashiyama, A. Masumoto, Y. Koretsune, K. Matsuda, Y. Doki, H. Eguchi, Y. Okada, Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis. *Nat. Commun.* **14**, 3671 (2023).
15. T. Ge, C. Y. Chen, B. M. Neale, M. R. Sabuncu, J. W. Smoller, Correction: Phenome-wide heritability analysis of the UK Biobank. *PLOS Genet.* **14**, e1007228 (2018).
16. Y.-C. A. Feng, T. Ge, M. Cordioli, FinnGen, A. Ganna, J. W. Smoller, B. M. Neale, Findings and insights from the genetic investigation of age of first reported occurrence for complex disorders in the UK Biobank and FinnGen. medRxiv 2020.11.20.20234302 [Preprint] (2020). https://doi.org/10.1101/2020.11.20.20234302.
17. H. Hunter-Zinck, Y. Shi, M. Li, B. R. Gorman, S. G. Ji, N. Sun, T. Webster, A. Liem, P. Hsieh, P. Devineni, P. Karnam, X. Gong, L. Radhakrishnan, J. Schmidt, T. L. Assimes, J. Huang, C. Pan, D. Humphries, M. Brophy, J. Moser, S. Muralidhar, G. D. Huang, R. Przygodzki, J. Concato, J. M. Gaziano, J. Gelernter, C. J. O'Donnell, E. R. Hauser, H. Zhao, T. J. O'Leary, V. A. M. V. Program, P. S. Tsao, S. Pyarajan, Genotyping array design and data quality control in the Million Veteran Program. *Am. J. Hum. Genet.* **106**, 535–548 (2020).
18. T.-Y. Liu, C.-F. Lin, H.-T. Wu, Y.-L. Wu, Y.-C. Chen, C.-C. Liao, Y.-P. Chou, D. Chao, Y.-S. Chang, H.-F. Lu, Comparison of multiple imputation algorithms and verification using whole-genome sequencing in the CMUH genetic biobank. *Biomedicine* **11**, 57–65 (2021).
19. H.-F. Lu, T.-Y. Liu, Y.-B. Chou, S.-S. Chang, Y.-W. Hsieh, J.-G. Chang, F.-J. Tsai, Comprehensive characterization of pharmacogenes in a Taiwanese Han population. *Front. Genet.* **13**, 948616 (2022).
20. P. Wu, A. Gifford, X. Meng, X. Li, H. Campbell, T. Varley, J. Zhao, R. Carroll, L. Bastarache, J. C. Denny, E. Theodoratou, W. Q. Wei, Mapping ICD-10 and ICD-10-CM codes to phecodes: Workflow development and initial evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).
21. T.-Y. Liu, H.-Y. Hsu, Y.-S. You, Y.-W. Hsieh, T.-C. Lin, C.-W. Peng, H.-Y. Huang, S.-S. Chang, F.-J. Tsai, Efficacy of Warfarin Therapy guided by pharmacogenetics: A real-world investigation among Han Taiwanese. *Clin. Ther.* **45**, 662–670 (2023).
22. H.-K. Chen, Y.-W. Hsieh, H.-Y. Hsu, T.-Y. Liu, Y.-T. Zhang, C.-D. Lin, F.-J. Tsai, Increased risk of hearing loss associated with MT-RNR1 gene mutations: A real-world investigation among Han Taiwanese population. *BMC Med. Genomics* **17**, 155 (2024).
23. W.-L. Liao, T.-Y. Liu, C.-F. Cheng, Y.-P. Chou, T.-Y. Wang, Y.-W. Chang, S.-Y. Chen, F.-J. Tsai, Analysis of HLA variants and Graves' disease and its comorbidities using a high resolution imputation system to examine electronic medical health records. *Front. Endocrinol.* **13**, 842673 (2022).
24. Y.-C. Chen, W.-D. Lin, T.-Y. Liu, F.-J. Tsai, Identification of the efficacy of parentage testing based on bi-allelic autosomal single nucleotide polymorphism markers in Taiwanese population. *Forensic Sci. Med. Pathol.* **20**, 801–809 (2024).
25. S.-Y. Chen, Y.-C. Chen, T.-Y. Liu, K.-C. Chang, S.-S. Chang, N. Wu, D. L. Wu, R. K. Dunlap, C.-J. Chan, F.-J. Tsai, Identification of novel genes associated with atrial fibrillation and development of atrial fibrillation predictive models by incorporating polygenic risk scores and PheWAS-derived risk factors. medRxiv 2023.08.14.23294097 [Preprint] (2023). https://doi.org/10.1101/2023.08.14.23294097.
26. R. Woodfield, UK Biobank Stroke Outcomes Group, UK Biobank Follow-up and Outcomes Working Group, C. L. Sudlow, Accuracy of patient self-report of stroke: A systematic review from the UK Biobank Stroke Outcomes Group. *PLOS ONE* **10**, e0137538 (2015).
27. T. Schoeler, J.-B. Pingault, Z. Kutalik, Self-report inaccuracy in the UK Biobank: Impact on inference and interplay with selective participation. medRxiv 2023.10.06.23296652 [Preprint] (2023). https://doi.org/10.1101/2023.10.06.23296652.
28. W. Guo, G. K. Fensom, G. K. Reeves, T. J. Key, Physical activity and breast cancer risk: Results from the UK Biobank prospective cohort. *Br. J. Cancer* **122**, 726–732 (2020).
29. T.-H. Sun, C.-C. Wang, T.-Y. Liu, S.-C. Lo, Y.-X. Huang, S.-Y. Chien, Y.-D. Chu, F.-J. Tsai, K.-C. Hsu, Utility of polygenic scores across diverse diseases in a hospital cohort for predictive modeling. *Nat. Commun.* **15**, 3168 (2024).
30. S. W. Choi, P. F. O'Reilly, PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).
31. D. J. M. Crouch, W. F. Bodmer, Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 18924–18933 (2020).
32. Y. Ding, K. Hou, K. S. Burch, S. Lapinska, F. Prive, B. Vilhjalmsson, S. Sankararaman, B. Pasaniuc, Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nat. Genet.* **54**, 30–39 (2022).
33. C. Gao, E. C. Polley, S. N. Hart, H. Huang, C. Hu, R. Gnanaolivu, J. Lilyquist, N. J. Boddicker, J. Na, C. B. Ambrosone, P. L. Auer, L. Bernstein, E. S. Burnside, A. H. Eliassen, M. M. Gaudet, C. Haiman, D. J. Hunter, E. J. Jacobs, E. M. John, S. Lindström, H. Ma, S. L. Neuhausen, P. A. Newcomb, K. M. O'Brien, J. E. Olson, I. M. Ong, A. V. Patel, J. R. Palmer, D. P. Sandler, R. Tamimi, J. A. Taylor, L. R. Teras, A. Trentham-Dietz, C. M. Vachon, C. R. Weinberg, S. Yao, J. N. Weitzel, D. E. Goldgar, S. M. Domchek, K. L. Nathanson, F. J. Couch, P. Kraft, Risk of breast cancer among carriers of pathogenic variants in breast cancer predisposition genes varies by polygenic risk score. *J. Clin. Oncol.* **39**, 2564–2573 (2021).
34. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium, A. Mahajan, M. J. Go, W. Zhang, J. E. Below, K. J. Gaulton, T. Ferreira, M. Horikoshi, A. D. Johnson, M. C. Y. Ng, I. Prokopenko, D. Saleheen, X. Wang, E. Zeggini, G. R. Abecasis, L. S. Adair, P. Almgren, M. Atalay, T. Aung, D. Baldassarre, B. Balkau, Y. Bao, A. H. Barnett, I. Barroso, A. Basit, L. F. Been, J. Beilby, G. I. Bell, R. Benediktsson, R. N. Bergman, B. O. Boehm, E. Boerwinkle, L. L. Bonnycastle, N. Burtt, Q. Cai, H. Campbell, J. Carey, S. Cauchi, M. Caulfield, J. C. N. Chan, L.-C. Chang, T.-J. Chang, Y.-C. Chang, G. Charpentier, C.-H. Chen, H. Chen, Y.-T. Chen, K.-S. Chia, M. Chidambaram, P. S. Chines, N. H. Cho, Y. M. Cho, L.-M. Chuang, F. S. Collins, M. C. Cornelis, D. J. Couper, A. T. Crenshaw, R. M. van Dam, J. Danesh, D. Das, U. de Faire, G. Dedoussis, P. Deloukas, A. S. Dimas, C. Dina, A. S. Doney, P. J. Donnelly, M. Dorkhan, C. van Duijn, J. Dupuis, S. Edkins, P. Elliott, V. Emilsson, R. Erbel, J. G. Eriksson, J. Escobedo, T. Esko, E. Eury, J. C. Florez, P. Fontanillas, N. G. Forouhi, T. Forsen, C. Fox, R. M. Fraser, T. M. Frayling, P. Froguel, P. Frossard, Y. Gao, K. Gertow, C. Gieger, B. Gigante, H. Grallert,

G. B. Grant, L. C. Grrop, C. J. Groves, E. Grundberg, C. Guiducci, A. Hamsten, B.-G. Han, K. Hara, N. Hassanali, A. T. Hattersley, C. Hayward, A. K. Hedman, C. Herder, A. Hofman, O. L. Holmen, K. Hovingh, A. B. Hreidarsson, C. Hu, F. B. Hu, J. Hui, S. E. Humphries, S. E. Hunt, D. J. Hunter, K. Hveem, Z. I. Hydrie, H. Ikegami, T. Illig, E. Ingelsson, M. Islam, B. Isomaa, A. U. Jackson, T. Jafar, A. James, W. Jia, K.-H. Jöckel, A. Jonsson, J. B. M. Jowett, T. Kadowaki, H. M. Kang, S. Kanoni, W. H. L. Kao, S. Kathiresan, N. Kato, P. Katulanda, K. M. Keinanen-Kiukaanniemi, A. M. Kelly, H. Khan, K.-T. Khaw, C.-C. Khor, H.-L. Kim, S. Kim, Y. J. Kim, L. Kinnunen, N. Klopp, A. Kong, E. Korpi-Hyövälti, S. Kowlessur, P. Kraft, J. Kravic, M. M. Kristensen, S. Krithika, A. Kumar, J. Kumate, J. Kuusisto, S. H. Kwak, M. Laakso, V. Lagou, T. A. Lakka, C. Langenberg, C. Langford, R. Lawrence, K. Leander, J.-M. Lee, N. R. Lee, M. Li, X. Li, Y. Li, J. Liang, S. Liju, W.-Y. Lim, L. Lind, C. M. Lindgren, E. Lindholm, C.-T. Liu, J. J. Liu, S. Lobbens, J. Long, R. J. F. Loos, W. Lu, J. Luan, V. Lyssenko, R. C. W. Ma, S. Maeda, R. Mägi, S. Männisto, D. R. Matthews, J. B. Meigs, O. Melander, A. Metspalu, J. Meyer, G. Mirza, E. Mihailov, S. Moebus, V. Mohan, K. L. Mohlke, A. D. Morris, T. W. Mühleisen, M. Müller-Nurasyid, B. Musk, J. Nakamura, E. Nakashima, P. Navarro, P.-K. Ng, A. C. Nica, P. M. Nilsson, I. Njølstad, M. M. Nöthen, K. Ohnaka, T. H. Ong, K. R. Owen, C. N. A. Palmer, J. S. Pankow, K. S. Park, M. Parkin, S. Pechlivanis, N. L. Pedersen, L. Peltonen, J. R. B. Perry, A. Peters, J. M. Pinidiyapathirage, C. G. Platou, S. Potter, J. F. Price, L. Qi, V. Radha, L. Rallidis, A. Rasheed, W. Rathman, R. Rauramaa, S. Raychaudhuri, N. W. Rayner, S. D. Rees, E. Rehnberg, S. Ripatti, N. Robertson, M. Roden, E. J. Rossin, I. Rudan, D. Rybin, T. E. Saaristo, V. Salomaa, J. Saltevo, M. Samuel, D. K. Sanghera, J. Saramies, J. Scott, L. J. Scott, R. A. Scott, A. V. Segrè, J. Sehmi, B. Sennblad, N. Shah, S. Shah, A. S. Shera, X. O. Shu, A. R. Shuldiner, G. Sigurđsson, E. Sijbrands, A. Silveira, X. Sim, S. Sivapalaratnam, K. S. Small, W. Y. So, A. Stančáková, K. Stefansson, G. Steinbach, V. Steinthorsdottir, K. Stirrups, R. J. Strawbridge, H. M. Stringham, Q. Sun, C. Suo, A.-C. Syvänen, R. Takayanagi, F. Takeuchi, W. T. Tay, T. M. Teslovich, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, E. Tikkanen, J. Trakalo, E. Tremoli, M. D. Trip, F. J. Tsai, T. Tuomi, J. Tuomilehto, A. G. Uitterlinden, A. Valladares-Salgado, S. Vedantam, F. Veglia, B. F. Voight, C. Wang, N. J. Wareham, R. Wennauer, A. R. Wickremasinghe, T. Wilsgaard, J. F. Wilson, S. Wiltshire, W. Winckler, T. Y. Wong, A. R. Wood, J.-Y. Wu, Y. Wu, K. Yamamoto, T. Yamauchi, M. Yang, L. Yengo, M. Yokota, R. Young, D. Zabaneh, F. Zhang, R. Zhang, W. Zheng, P. Z. Zimmet, D. Altshuler, D. W. Bowden, Y. S. Cho, N. J. Cox, M. Cruz, C. L. Hanis, J. Kooner, J.-Y. Lee, M. Seielstad, Y. Y. Teo, M. Boehnke, E. J. Parra, J. C. Chambers, E. S. Tai, M. I. M. Carthy, A. P. Morris, Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).

35. F.-J. Tsai, C.-F. Yang, C.-C. Chen, L.-M. Chuang, C.-H. Lu, C.-T. Chang, T.-Y. Wang, R.-H. Chen, C.-F. Shiu, Y.-M. Liu, C.-C. Chang, P. Chen, C.-H. Chen, C. S. J. Fann, Y.-T. Chen, J.-Y. Wu, A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLOS Genet.* **6**, e1000847 (2010).
36. Y.-C. Chang, Y.-F. Chiu, P.-H. Liu, K.-C. Shih, M.-W. Lin, W. H.-H. Sheu, T. Quertermous, J. D. Curb, C. A. Hsiung, W.-J. Lee, P.-C. Lee, Y.-T. Chen, L.-M. Chuang, Replication of genome-wide association signals of type 2 diabetes in Han Chinese in a prospective cohort. *Clin. Endocrinol. (Oxf)* **76**, 365–372 (2012).
37. Y.-S. Chang, C.-Y. Lin, T.-Y. Liu, C.-M. Huang, C.-C. Chung, Y.-C. Chen, F.-J. Tsai, J.-G. Chang, S.-J. Chang, Polygenic risk score trend and new variants on chromosome 1 are associated with male gout in genome-wide association study. *Arthritis Res. Ther.* **24**, 229 (2022).
38. C.-Y. Lin, Y.-S. Chang, T.-Y. Liu, C.-M. Huang, C.-C. Chung, Y.-C. Chen, F.-J. Tsai, J.-G. Chang, S.-J. Chang, Genetic contributions to female gout and hyperuricaemia using genome-wide association study and polygenic risk score analyses. *Rheumatology* **62**, 638–646 (2023).
39. M.-H. Tsai, C.-Y. Hsu, M.-Y. Lin, M.-F. Yen, H.-H. Chen, Y.-H. Chiu, S.-J. J. N. Hwang, Incidence, prevalence, and duration of chronic kidney disease in Taiwan: Results from a community-based screening program of 106,094 individuals. *Nephron* **140**, 175–184 (2018).
40. Y.-T. Lin, Y.-C. Lo, H.-Y. Chiang, C.-R. Jung, C.-M. Wang, T.-C. Chan, C.-C. Kuo, B.-F. Hwang, Particulate air pollution and progression to kidney failure with replacement therapy: An advanced CKD registry–based cohort study in Taiwan. *Am. J. Kidney Dis.* **76**, 645–657.e1 (2020).
41. S. Weber, H. Thiele, S. Mir, M. R. Toliat, B. Sozeri, H. Reutter, M. Draaken, M. Ludwig, J. Altmüller, P. Frommolt, H. M. Stuart, P. Ranjzad, N. A. Hanley, R. Jennings, W. G. Newman, D. T. Wilcox, U. Thiel, K. P. Schlingmann, R. Beetz, P. F. Hoyer, M. Konrad, F. Schaefer, P. Nürnberg, A. S. Woolf, Muscarinic acetylcholine receptor M3 mutation causes urinary bladder disease and a prune-belly-like syndrome. *Am. J. Hum. Genet.* **89**, 668–674 (2011).
42. C. Robinson-Cohen, J. L. Triozzi, B. Rowan, J. He, H. C. Chen, N. S. Zheng, W. Q. Wei, O. D. Wilson, J. N. Hellwege, P. S. Tsao, J. M. Gaziano, A. Bick, M. E. Matheny, C. P. Chung, L. Lipworth, E. D. Siew, T. A. Ikizler, R. Tao, A. M. Hung, Genome-wide association study of CKD progression. *J. Am. Soc. Nephrol.* **34**, 1547–1559 (2023).
43. T.-G. Chang, T.-T. Yen, C.-Y. Wei, T.-H. Hsiao, I.-C. Chen, Impacts of ADH1B rs1229984 and ALDH2 rs671 polymorphisms on risks of alcohol-related disorder and cancer. *Cancer Med.* **12**, 747–759 (2023).
44. Y.-C. Chen, C.-M. Huang, T.-Y. Liu, N. Wu, C.-J. Chan, P.-Y. Shih, H.-H. Chen, S.-Y. Chen, F.-J. Tsai, Effects of human leukocyte antigen DRB1 genetic polymorphism on anti-cyclic citrullinated peptide (ANTI-CCP) and rheumatoid factor (RF) expression in rheumatoid arthritis (RA) patients. *Int. J. Mol. Sci.* **24**, 12036 (2023).
45. Y. Suzuki, H. Ménager, B. Brancotte, R. Vernet, C. Nerin, C. Boetto, A. Auvergne, C. Linhard, R. Torchet, P. Lechat, L. Troubat, M. H. Cho, E. Bouzigon, H. Aschard, H. Julienne, Trait selection strategy in multi-trait GWAS: Boosting SNPs discoverability. bioRxiv 2023.10.27.564319 [Preprint] (2023). https://doi.org/10.1101/2023.10.27.564319.
46. Y.-C. Hsu, H.-L. Chen, C.-F. Cheng, A. Chattopadhyay, P.-S. Chen, C.-C. Lin, H.-Y. Chiang, T.-Y. Liu, C.-H. Huang, C.-C. Kuo, E. Y. Chuang, T.-P. Lu, F.-J. Tsai, The largest genome-wide association study for breast cancer in Taiwanese Han population. *Breast Cancer Res. Treat.* **203**, 291–306 (2024).
47. W.-L. Liao, Y.-N. Huang, Y.-W. Chang, T.-Y. Liu, H.-F. Lu, Z.-Y. Tiao, P.-H. Su, C.-H. Wang, F.-J. Tsai, Combining polygenic risk scores and human leukocyte antigen variants for personalized risk assessment of type 1 diabetes in the Taiwanese population. *Diabetes Obes. Metab.* **25**, 2928–2936 (2023).
48. J.-S. Yang, T.-Y. Liu, Y.-C. Chen, S.-C. Tsai, Y.-J. Chiu, C.-C. Liao, F.-J. Tsai, Genome-wide association study of Alopecia areata in Taiwan: The conflict between individuals and hair follicles. *Clin. Cosmet. Investig. Dermatol.* **16**, 2597–2612 (2023).
49. S.-Y. Chen, T.-Y. Liu, W.-L. Liao, T.-Y. Wang, C.-J. Chan, J.-G. Chang, Y.-C. Chen, H.-F. Lu, H.-H. Yang, F.-J. Tsai, Genome-wide association study of hyperthyroidism based on electronic medical record from Taiwan. *Front. Med.* **9**, 830621 (2022).
50. D.-T. Bau, T.-Y. Liu, C.-W. Tsai, W.-S. Chang, J. Gu, J.-S. Yang, L.-C. Shih, F.-J. Tsai, A genome-wide association study identified novel genetic susceptibility loci for oral cancer in Taiwan. *Int. J. Mol. Sci.* **24**, 2789 (2023).
51. N. S. Zheng, Q. Feng, V. E. Kerchberger, J. Zhao, T. L. Edwards, N. J. Cox, C. M. Stein, D. M. Roden, J. C. Denny, W.-Q. Wei, PheMap: A multi-resource knowledge base for high-throughput phenotyping within electronic health records. *J. Am. Med. Inform. Assoc.* **27**, 1675–1687 (2020).
52. W. W. Stead, A. Lewis, N. B. Giuse, T. Y. Koonce, L. Bastarache, Knowledgebase strategies to aid interpretation of clinical correlation research. *J. Am. Med. Inform. Assoc.* **30**, 1257–1265 (2023).
53. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W.-M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
54. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
55. S. A. Gagliano Taliun, P. VandeHaar, A. P. Boughton, R. P. Welch, D. Taliun, E. M. Schmidt, W. Zhou, J. B. Nielsen, C. J. Willer, S. Lee, L. G. Fritsche, M. Boehnke, G. R. Abecasis, Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).
56. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
57. X. Zheng, J. Shen, C. Cox, J. C. Wakefield, M. G. Ehm, M. R. Nelson, B. S. Weir, HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
58. G. McInnes, A. Lavertu, K. Sangkuhl, T. E. Klein, M. Whirl-Carrillo, R. B. Altman, Pharmacogenetics at scale: An analysis of the UK Biobank. *Clin. Pharmacol. Ther.* **109**, 1528–1537 (2021).
59. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. L. Faive, S.-B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y.-D. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. De Meo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Köttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. M. Manus, S. T. M. Garvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman,

P. Qasba, W. Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zöllner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

60. C. J. Willer, Y. Li, G. R. Abecasis, METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
61. J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, M. A. Basford, D. S. Carrell, P. L. Peissig, A. N. Kho, J. A. Pacheco, L. V. Rasmussen, D. R. Crosslin, P. K. Crane, J. Pathak, S. J. Bielinski, S. A. Pendergrass, H. Xu, L. A. Hindorff, R. Li, T. A. Manolio, C. G. Chute, R. L. Chisholm, E. B. Larson, G. P. Jarvik, M. H. Brilliant, C. A. McCarty, I. J. Kullo, J. L. Haines, D. C. Crawford, D. R. Masys, D. M. Roden, Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
62. R. J. Carroll, L. Bastarache, J. C. Denny, R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).

Acknowledgments: We would like to thank the National Core Facility for Biopharmaceuticals (111-2740-B-492-001) and National Center for High-Performance Computing (National Applied Research Laboratories, Taiwan) for providing computational and storage resources. We would also like to acknowledge the invaluable assistance of all physicians at CMUH and its affiliated hospitals for their cooperation in patient recruitment. Sincere thanks are due to the Taiwan Precision Medicine Initiative team at Academia Sinica. We would thank T.-H. Sun, S.-C. Lo, and C.-C. Wang for assistance in establishing the cloud computing resources and monitoring the computational progress. This manuscript was edited by Wallace Academic Editing. **Funding:** This study was supported by the China Medical University Hospital (grant nos. 1JA6 and DMR-113-184) and National Science and Technology Council (113-2321-B-039-006). **Author contributions:** Conceptualization: T.-Y.L., H.-F.L., C.-C.L., J.-S.Y., S.-Y.C., K.-C.H., S.-S.C., C.-M.H., P.-Y.W., J.-Y.W., D.-Y.C., C.-H.T., and F.-J.T. Data curation: T.-Y.L., Y.-C.C., C.-C.L., Y.-J.L., Y.-P.C., C.-C.Y., P.-Y.W., C.-C.K., and F.-J.T. Formal analysis: T.-Y.L., H.-F.L., Y.-C.C., C.-C.L., W.-L.L., S.-Y.C., Y.-H.L., C.-C.Y., R.-H.C., and F.-J.T. Funding acquisition: F.-J.T. Investigation: T.-Y.L., J.-S.Y., W.-D.L., Y.-C.H., W.-Y.L., Y.-H.L., J.-G.C., C.-H.W., C.-T.C., C.-M.H., T.-Y.W., C.-C.Y., J.-H.C., C.-P.H., H.-C.L., R.-H.C., H.-J.L., P.-Y.W., and F.-J.T. Methodology: T.-Y.L., H.-F.L., C.-C.L., Y.-J.L., J.-S.Y., K.-C.H., S.-S.C., C.-M.H., and F.-J.T. Project administration: T.-Y.L., J.-S.Y., H.-C.L., and F.-J.T. Resources: T.-Y.L., S.-S.C., C.-H.W., C.-T.C., C.-M.H., K.-J.Y., T.-Y.W., C.-C.Y., J.-H.C., C.-P.H., H.-C.L., R.-H.C., H.-J.L., P.-Y.W., J.-Y.W., C.-C.K., D.-Y.C., C.-H.T., and F.-J.T. Software: T.-Y.L., H.-F.L., C.-C.L., K.-C.H., H.-D.C., and Y.-P.C. Supervision: D.-Y.C., C.-H.T., and F.-J.T. Validation: T.-Y.L., Y.-C.C., Y.-J.L., W.-L.L., Y.-H.L., J.-G.C., and F.-J.T. Visualization: T.-Y.L., H.-F.L., Y.-C.C., C.-C.L., W.-D.L., H.-D.C., C.-M.H., and F.-J.T. Writing—original draft: T.-Y.L., Y.-C.C., and C.-M.H. Writing—review and editing: T.-Y.L., H.-F.L., J.-S.Y., W.-L.L., W.-D.L., S.-Y.C., S.-S.C., C.-H.W., C.-T.C., T.-Y.W., C.-C.Y., J.-H.C., H.-J.L., J.-Y.W., and F.-J.T. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Deidentified genomic and clinical data used in this study are available to qualified researchers upon application to the China Medical University Genetic Biobank. All data access requests are independently reviewed by the Human Biobank Ethics Committee and the Medical Data Governance Committee at China Medical University Hospital. Data access is not managed by the authors. For application instructions and more information, please visit the HiGenome website: http://cmuh-biobank.eastasia.cloudapp.azure.com:6001. The original contributions presented here are included in the Supplementary Materials.

Submitted 9 September 2024
Accepted 24 April 2025
Published 4 June 2025
10.1126/sciadv.adt0539