Population-Specific Polygenic Risk Scores Developed for the Han Chinese

Hung-Hsin Chen[1,2,*], Chien-Hsiun Chen[1,*], Ming-Chih Hou[3,*], Yun-Ching Fu[4,5,6,*], Ling-Hui Li[1], Che-Yu Chou[1], Erh-Chan Yeh[1], Ming-Fang Tsai[1], Chun-houh Chen[7], Hsin-Chou Yang[7,8], Yen-Tsung Huang[7,9,10], Yi-Min Liu[1], Chun-yu Wei[1,11], Jen-Ping Su[1], Wan-Jia Lin[1], Elin H.F. Wang[1], Chi-Lu Chiang[12,13], Jeng-Kai Jiang[14,13], I-Hui Lee[15,16,13], Kung-Hao Liang[17,18,19], Wei-Sheng Chen[20,21], Hung-Cheng Tsai[20,21], Shih-Yao Lin[22,13], Fu-Pang Chang[22], Hsiang-Ling Ho[22,23], Yi-Chen Yeh[22,13], Wei-Cheng Tseng[24,25], Ming-Hwai Lin[26], Hsiao-Ting Chang[26,25], Ling-Ming Tseng[27,28,13], Wen-Yih Liang[22], Paul Chih-Hsueh Chen[22], Yu-Cheng Hsieh[29,30,31], Yi-Ming Chen[29,30], Tzu-Hung Hsiao[29], Ching_Heng Lin[29], Yen-Ju Chen[29], I-Chieh Chen[29], Chien-Lin Mao[29], Shu-Jung Chang[29], Yen-Lin Chang[32], Yi_Ju Liao[32], Chih-Hung Lai[33], Wei-Ju Lee[34,30], Hsin Tung[34,30], Ting-Ting Yen[35], Hsin-Chien Yen[36], Ming-Yao Chen[37,38,39], Ying-Chin Lin[40,41,42], Yung-Ta Kao[43,44,45], Bi-Zhen Kao[39], Jing-Er Lee[46], Chi-Li Chung[47,48], Ju-Chi Liu[49,43,45], Paul Chan[50], Chang-Hsien Lin[41], Chia-Hsin, Chen[51,52], I-Chen Wu[53,54], Lung-Chang Lin[55,56], Jiunn-Wei Wang[53,57,58], Shen-liang Shih[59,60], Sun-Wung Hsieh[61,62,63], Chih-Hsing Hung[64,65,56], Wei-Ming Li[66,67,68], Chih-Jen Yang[69,70], Cheng-Shin Yang[1], Ru-Hui Weng[1], Yu-Chi Chen[1], Chun-Ping Chang[1], Tai-Hsun Wu[1], Yu-Chang Lin[1], Yi-Jing Sheen[71,72,30], Shi-Heng Wang[73], Sye-Pu Chen[1], Timothy Raben[74], Erik Widen[74,75], Stephen Hsu[74,75], Feng-Jen Hsieh[1,76], Dong-Ru Ho[77,78,79], Yu-Huei Huang[80,81], Chung-Han Yang[82], Yu-Shu Huang[83,84], Yen-Fu Chen[82], Hsien-Ming Wu[85], Ping-Han Tsai[86,82], Kuan-Gen Huang[85], Chih-Yen Chien[87,88], Yi-Lwun Ho[89,90], Ming-Shiang Wu[89,90], Jia-Horng Kao[89,91,92], Yen-Bin Liu[89,90], Jyh-Ming Jimmy Juang[93,90], Mao-Hsin Lin[89,90], Yen-Hung Lin[89,90,94], Ji-Yuh Lee[95], Hsueh-Ju Lu[96,97], Chieh-Hua Lu[98], An-Chieh Feng[99], Jhih-Syuan Liu[98], Chien-Ping Chiang[100,101], Nain-Feng Chu[98], Jung-Chun Lin[102], Yi-Wei Yeh[103], En Meng[104], Chih-Yang Huang[105,106], Chi-Cheng Li[107,108,109], Tso-Fu Wang[109,110,111], Kuei-Ying Su[112,108], Jia-Kang Wang[113,114,115], Mei-Hsiu Chen[116,117,114], Hua-Fen Chen[118,119,120], Gwo-Chin Ma[121], Ting-Yu Chang[121], Fu-Tien Chiang[122,119], Hsing-Jung Chang[123,124], Kuo-Jang Kao[125], Chen-Fang Hung[125], Ching-Yao Tsai[126,127,128], Po-Yueh Chen[129,130], Kochung Tsui[131,132,133], Pui-Yan Kwok[1], Wayne Huey-Herng Sheu[134,71,135,†], Shun-Fa Yang[136,137,†], Jyh-Ming Liou[138,90,139,†], Jaw-Yuan Wang[140,58,†], Jeng-Fong Chiou[141,142,†], Jer-Yuarn Wu[1,†], Cathy S.-J. Fann[1,†]

1 Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
2 Vanderbilt Genetics Institute, Vanderbilt University Medical Center, TN, USA
3 Division of Gastroenterology and Hepatology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
4 Department of Pediatric Cardiology, Taichung Veterans General Hospital, Taichung, Taiwan
5 Children's Medical Center, Taichung Veterans General Hospital, Taichung, Taiwan
6 Department of Pediatrics, School of Medicine, National Chung-Hsing University, Taichung, Taiwan
7 Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

8 Biomedical Translation Research Center, Academia Sinica, Taipei, Taiwan
9 Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan
10 Department of Mathematics, Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan
11 Department of Clinical Pharmacy, School of Pharmacy, Taipei Medical University, Taipei, Taiwan
12 Department of Chest Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
13 School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
14 Division of Colon and Rectal Surgery, Department of Surgery, Taipei Veterans General Hospital, Taipei, Taiwan
15 Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan
16 Institute of Brain Science, National Yang Ming Chiao Tung University, Taipei, Taiwan
17 Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan
18 Institute of Food Safety and Health Risk Assessment, National Yang Ming Chiao Tung University, Taipei, Taiwan
19 Institute of Biomedical Informatics, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
20 Division of Allergy, Immunology & Rheumatology, Taipei Veterans General Hospital, Taipei, Taiwan
21 Faculty of Medicine, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
22 Department of Pathology and Laboratory Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
23 Department of Biotechnology and Laboratory Science in Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
24 Division of Nephrology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
25 School of Medicine, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
26 Department of Family Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
27 Division of General Surgery, Department of Surgery, Taipei Veterans General Hospital, Taipei, Taiwan
28 Comprehensive Breast Health Center, Taipei Veterans General Hospital, Taipei, Taiwan
29 Department of Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan
30 Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung, Taiwan
31 Department of Medical Research, Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
32 Department of Pharmacy, Taichung Veterans General Hospital, Taichung, Taiwan
33 Department of Medicine and Cardiovascular Center, Taichung Veterans

General Hospital, Taichung, Taiwan
34 Neurological Institute, Taichung Veterans General Hospital, Taichung, Taiwan
35 Department of Otolaryngology, Taichung Veterans General Hospital, Taichung, Taiwan
36 Division of Pediatric Genetics and Metabolism, Children's Medical Center, Taichung Veterans General Hospital, Taichung, Taiwan
37 Division of Gastroenterology and Hepatology, Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
38 TMU Research Center for Digestive Medicine, Taipei Medical University, Taipei, Taiwan
39 Division of Gastroenterology, Department of Internal Medicine, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan
40 Department of Family Medicine, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
41 Department of Family Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
42 Department of Occupational Medicine, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
43 Division of Cardiology, Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
44 Division of Cardiology, Department of Internal Medicine, Taipei Medical University Hospital, Taipei, Taiwan
45 Taipei Heart Institute, Taipei Medical University, Taipei, Taiwan
46 Department of Neurology, Wan Fang Hospital and Taipei Medical University, Taipei, Taiwan
47 Division of Pulmonary Medicine, Department of Internal Medicine, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
48 School of Respiratory Therapy, College of Medicine, Taipei Medical University, Taipei, Taiwan
49 Division of Cardiology, Department of Internal Medicine, Shuang Ho Hospital, Taipei Medical University, New Taipei City, Taiwan
50 Division of Cardiovascular Medicine, Department of Internal Medicine, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan
51 Department of Physical Medicine and Rehabilitation, School of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
52 Regenerative Medicine and Cell Therapy Research Center, Kaohsiung Medical University, Kaohsiung, Taiwan
53 Division of Gastroenterology, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
54 Center for Cancer Research, Kaohsiung Medical University, Kaohsiung, Taiwan
55 Departments of Pediatrics, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
56 Department of Pediatrics, School of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
57 Department of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

58 Graduate Institute of Clinical Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
59 Division of Breast Oncology and Surgery, Department of Surgery, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
60 Center for Medical Education and Humanizing Health Professional Education, Kaohsiung Medical University, Kaohsiung, Taiwan
61 Department of Neurology, Kaohsiung Municipal Siaogang Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
62 Department of Neurology, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
63 Department of Neurology, Faculty of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
64 Research Center for Precision Environmental Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
65 Department of Pediatrics, Kaohsiung Municipal Siaogang Hospital, Kaohsiung, Taiwan
66 Department of Urology, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
67 Department of Urology, School of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
68 Department of Urology, Kaohsiung Medical University Gangshan Hospital, Kaohsiung, Taiwan
69 Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
70 School of Post-Baccalaureate Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
71 Division of Endocrinology and Metabolism, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan
72 Department of Medicine, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
73 National Center for Geriatrics and Welfare Research, National Health Research Institutes, Miaoli, Taiwan
74 University of Michigan State, MI, USA
75 Genomic Prediction, Inc.
76 Khoury College of Computer Sciences, Northeastern University, MA, USA
77 Division of Urology, Department of Surgery, Chang Gung Memorial Hospital, Chiayi, Taiwan;
78 Graduate Institute of Clinical Medical Sciences, College of Medicine, Chang Gung University, Taoyuan, Taiwan
79 School of Medicine, National Tsing Hua University, Hsinchu, Taiwan
80 Department of Dermatology, Chang Gung Memorial Hopistal, Linkou, Taiwan
81 School of Medicine, College of Medicine, Chang-Gung University, Taoyuan, Taiwan
82 Division of Rheumatology, Allergy and Immunology, Department of Internal Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan
83 Department of Psychiatry and Sleep center, Chang Gung Memorial Hospital, Taoyuan, Taiwan

84 College of Medicine, Chang Gung University, Taoyuan, Taiwan
85 Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital, Linkou Medical Center and Chang Gung University College of Medicine, Taoyuan, Taiwan
86 Division of Rheumatology, Allergy and Immunology, Department of Internal Medicine, New Taipei City Municipal TuCheng Hospital, New Taipei City, Taiwan
87 Department of Otolaryngology Head & Neck Surgery, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan
88 Doctoral Program of Clinical and Experimental Medicine, National Sun Yat-sen University, Kaohsiung, Taiwan
89 Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
90 Department of Internal Medicine, National Taiwan University College of Medicine, Taipei, Taiwan
91 Hepatitis Research Center, National Taiwan University Hospital, Taipei, Taiwan
92 Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, Taipei, Taiwan
93 Cardiovascular Center and Heart Failure Center, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
94 Cardiovascular Center, National Taiwan University Hospital, Taipei, Taiwan
95 Department of Internal Medicine, National Taiwan University Hospital, Yunlin branch, Yunlin, Taiwan
96 Division of Hematology and Oncology, Department of Internal Medicine, Chung Shan Medical University Hospital, Taichung, Taiwan
97 School of Medicine, Chung Shan Medical University, Taichung, Taiwan
98 Endocrinology and Metabolism, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
99 General Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
100 Dermatology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
101 Department And Graduate Institute of biochemistry, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
102 Gastroenterology and Hepatology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
103 Psychiatry, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
104 Urology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan
105 Cardiovascular and Mitochondria Related Disease Research Center, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan
106 Center of General Education, Buddhist Tzu Chi Medical Foundation, Tzu Chi University of Science and Technology, Hualien, Taiwan
107 Center of Stem Cell and Precision Medicine, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan
108 School of Medicine, Tzu Chi University, Hualien, Taiwan

109 Department of Hematology and Oncology, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan
110 Department of Medicine, College of Medicine, Tzu Chi University, Hualien, Taiwan
111 Buddhist Tzu Chi Stem Cells Center, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan
112 Division of Allergy, Immunology and Rheumatology, Hualien Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Hualien, Taiwan
113 Department of Ophthalmology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
114 Department of electrical engineering, Yuan Ze University, Taoyuan, Taiwan
115 Department of Medicine, National Taiwan University, Taipei, Taiwan
116 Department of Internal Medicine, Far Eastern Memorial Hospital, New Taipei City, Taiwan
117 Department of Biomedical Engineering, Ming Chuan University, Taoyuan, Taiwan
118 Division of Endocrinology, Department of Internal Medicine, Far-Eastern Memorial Hospital, Taipei, Taiwan
119 School of Medicine, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan
120 Department of Public Health, College of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan
121 Department of Genomic Medicine and Center for Medical Genetics, Changhua Christian Hospital, Changhua, Taiwan
122 Department of Cardiology, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan
123 Precision Medicine Center, Fu Jen Catholic University Hospital , Fu Jen Catholic University, New Taipei City, Taiwan
124 Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei City , Taiwan
125 Koo Foundation Sun Yat-Sen Cancer Center, Taipei, Taiwan
126 Department of Ophthalmology, Taipei City Hospital, Taipei, Taiwan
127 Institute of Public Health, National Yang Ming Chiao Tung University, Taipei, Taiwan
128 Department of Health and Welfare, University of Taipei, Taipei, Taiwan
129 Division of Gastroenterology and Hepatology, Department of Internal Medicine, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi, Taiwan
130 Clinical Trial Center, Department of Medical Research, Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi City, Taiwan
131 Fu-Jen Catholic University School of Medicine, New Taipei City, Taiwan
132 Cathay General Hospital Department of Internal Medicine, Taipei, Taiwan
133 Cathay General Hospital Department of Clinical Pathology, Taipei, Taiwan
134 National Health Research Institutes, Miaoli, Taiwan
135 Division of Endocrinology and Metabolism, Department of Internal Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
136 Institute of Medicine, Chung Shan Medical University, Taichung, Taiwan
137 Department of Medical Research, Chung Shan Medical University

Hospital, Taichung, Taiwan
138 Division of Gastroenterology and Hepatology, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
139 Department of Internal Medicine, National Taiwan University Cancer Center, Taipei, Taiwan
140 Division of Colorectal Surgery, Department of Surgery, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan
141 Department of Radiology, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
142 Division of Radiation Oncology , Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan
*,† These authors contributed equally

Corresponding author:

Hung-Hsin Chen, PhD
Assistant Research Fellow
Institute of Biomedical Sciences
Academia Sinica
128 Sec. 2, Academia Rd., IBMS N121
Taipei 115, Taiwan
hunghsinchen@ibms.sinica.edu.tw

Cathy S. J. Fann, PhD
Research Fellow
Institute of Biomedical Sciences
Academia Sinica
128 Sec. 2, Academia Rd., IBMS 100
Taipei 115, Taiwan
csjfann@ibms.sinica.edu.tw

Abstract

Predicting complex disease risks based on individual genomic profiles is an advancing field in human genetics.[1,2] However, most genetic studies have focused on European populations, creating a global imbalance in precision medicine and underscoring the need for genomic research in non-European groups[3,4]. The Taiwan Precision Medicine Initiative (TPMI) recruited over half a million Taiwanese residents, providing the largest dataset of genetic profiles and electronic medical record data for the Han Chinese. Using extensive phenotypic data, we conducted the largest genomic analyses of Han Chinese across the medical phenome. These analyses identified population-specific genetic risk variants and novel findings on the genetic architecture of complex traits. We developed polygenic risk scores, demonstrating strong predictive performance for conditions such as cardiometabolic diseases, autoimmune disorders, cancers, and infectious diseases. We observed consistent findings in an independent sample from our Biobank and among East Asians in the UK Biobank and the All of Us Project. The identified genetic risks accounted for up to 9.1% of the disease burden variance in Taiwan. Our approach of characterizing the phenome-wide genomic landscape, developing population-specific risk prediction models, assessing their performance, and identifying the genetic impact on health, serves as a model for similar studies in other diverse study populations.

Introduction

A major promise of modern genetics is the ability to predict complex disease risk based on a person's genetic profile. If successful, health management strategies can be developed to mitigate the risk (disease prevention) and to optimize care (early diagnosis and effective treatment). Large-scale studies in Europeans, such as those conducted with data from the UK Biobank (UKB) and Electronic Medical Records and Genomics (eMERGE) Network, show that risk prediction based on genetics holds promise and several countries are exploring ways to implement risk-based management in clinical practice[1,2]. Using polygenic risk scores (PRS) to predict disease risk and identify individuals at high risk is an emerging "precision medicine" approach to leverage individual genetic findings in clinical practice. However, a significant limitation of current PRS models is that they are predominantly based on genome-wide association studies (GWAS) comprising particpants from European populations[4,5], often leading to reduced predictive performance in non-European groups[6,7]. To fully realize the potential of precision medicine for diverse global populations, population specific phenome wide genomic discovery must be perfomed at scale and clinically applicable polygenic risk models must be optimized within and across populations. To fill this major research gap in an East Asian population, we characterized the complex genetic architecture of the Han Chinese phenome wide, developed population-specific PRS, and assesed the external validity of the models across populations with varying degrees of genetic similarity.

Populations with East Asian (EAS) ancestry represent nearly a quarter of the global population, but they account for only 3.95% of the participants in previously published GWAS[3]. Although several biobanks have been built to recruit subjects from East Asia, they have moderate sample size (72K - 212K) and many focus on specific conditions[8-12]. In contrast, biobanks with predominantly European ancestry participants[13-16], have significantly larger sample sizes (224K – 635K) and access to more comprehensive clinical data. The moderate sample size and limited phenotypes in existing EAS biobanks hamper discovery of unique genetic effcts and preclude the development of

robust and clinically useful PRS models for EAS.

We have assembled the largest non European cohort to date, the Taiwan Precision Medicine Initiative (TPMI). From 2019 to 2023, TPMI enrolled and genotyped over half a million participants across sixteen medical centers in Taiwan. All the participants, who are overwhelmingly of Han Chinese ancestry, contributed DNA samples for genetic profiling with a custom-designed genotyping array and consented to provide their longitudinal electronic medical record (EMR) data from 5 years prior to enrollment and into the future. The EMR dataset includes rich and accurate health-related phenotypes, including medical diagnoses and biochemical examinations[17]. In this paper, we present the results of comprehensive genomic analyses of extensive genetic and medical data derived from the Han Chinese population, including phenome-wide GWAS and PRS model development. We identified numerous population-specific risk variants/genes, observed evidence of genetic pleiotropy, and pinpointed clusters of traits that shared similar genetic etiology. We developed and validated PRS prediction models for numerous conditions against external datasets including those from the Taiwan Biobank (TWB), the UKB, and the All of Us Project. Our results reveal the benefits of leveraging large cohort from understudied population to identify unique genetic underpinnings of the human phenome, interpret causal effects via fine mapping and colocalization, and improve the performance of population-specific PRS models, which together, better illuminate the clinical implications of genetic risk. The lack of representation of diverse populations in genetic research will result in inequitable access to precision medicine for those with the highest burden of disease. Thus, continued, large scale genome-wide efforts in diverse populations will be required to maximize genetic discovery and reduce health disparities.

Results

Dichotomized Phenotype (Phecodes) and Quantitative Traits in TPMI

We performed comprehensive genomic analyses, including GWAS, heritability estimation, and PRS model building and evaluation, across a wide range of diseases and quantitative traits using data from TPMI. We examined 700

dichotomized phenotypes (phecode, case n > 2,000) and 24 quantitative traits (sample size > 100,000), spanning numerous disease categories, such as neoplasms, metabolic disorders, circulatory conditions, autoimmune diseases, and more (Fig. 1). The phecodes, derived from International Classification of Diseases (ICD) codes[18,19], alongside quantitative traits such as blood pressure, BMI, liver enzymes, and lipid levels, provide a robust dataset for exploring genetic contributions to human health (Table S1 and S2). The log-transformed case proportion identified from EMR showed a significant correlation with the log-transformed 5-year disease prevalence from National Health Insurance Research Database in Taiwan[20] ($r = 0.64$, p-value = 4.4×10^{-33}) (Fig. 1A), which suggests that TPMI represented the Taiwanese population well. Fig. 1B displays the sample sizes for 24 quantitative traits in the TPMI cohort and highlights sample size variation across traits, which is crucial for determining the power and precision of association analyses within the cohort.

Genome-Wide Association Studies, Fine-Mapping, and Novel Finding

Our GWAS identified at least one significant locus (p-value < 5×10^{-8}) for 265 phecodes and 24 quantitative traits. The high replication rate of 74.4% for reported disease loci from EAS GWAS highlights the robustness of the TPMI data, particularly for endocrine/metabolic and hematopoietic diseases (88.68% and 84.62%, respectively) (Extended Data Fig. 1 and Table S3). Lower replication rates for respiratory and psychiatric disorders (27.78% and 19.81%) may reflect limited case numbers, other untyped genetic variants, such as rare variant, copy number variation and structure variants, or recruitment bias.

Our GWAS revealed 1,139 independent association signals for phecodes and 1,305 signals for quantitative traits via fine-mapping. Notably, 77 novel associations were identified across 44 phecodes and 7 quantitative traits that had not been previously reported in nearby regions (±1Mb), and 307 novel hits from reported regions (Table S4 and S5). Some of our novel findings are biologically relevant to the corresponding phenotypes. For example, the SNP rs17089782, a missense variant in the *PIBF1* (p.R405Q) gene on chromosome 13 is significantly associated with thyroid cancer ($p = 2.8 \times 10^{-9}$). *PIBF1* is essential for immune regulation, especially during pregnancy, and is relevant

to autoimmune diseases and cancer[21]. Another novel variant from the known region associated with BMI found in *PHOX2B* (rs761018157, p-value = 7.6×10^{-9}). This gene, highly expressed in the nervous system, had previously been linked to obesity hypoventilation syndrome in a small study (n = 30)[22] and associated with bone mineral density[23]. Moreover, among the 22 identified independent loci for hepatitis B, 16 fine-mappped loci have not been previously linked to hepatitis B in the GWAS Catalog (Extended Data Fig. 2). Notably, 15 of these 16 loci were found to be associated with liver function or diseases (Table S5). These novel associations highlight the uniqueness of certain disease loci in the TPMI cohort, presenting opportunities for developing population-specific therapeutic interventions and advancing precision medicine. We summarized these identified independent associations in Fig. 2. The identification of the major histocompatibility complex (MHC) region as a significant hotspot on chromosome 6 underscores its extensive involvement in immune-related diseases across multiple categories. Similarly, the short arm of chromosome 11 also affect various traits, including metabolic, endocrine, and genitourinary diseases. These hotspots of trait-relevant variants implied the shared genetic mechanism among diseases and genes' pleiotropy.

Genome- and Gene-level Heritability, and Colocalization

Linkage disequilibrium score regression analysis[24] showed strong liability-scaled SNP-heritability for conditions such as alcoholism ($h^2 = 0.242$), open-angle glaucoma ($h^2 = 0.171$), and retention of urine ($h^2 = 0.173$). In terms of quantitative traits, body height ($h^2 = 0.323$), BMI ($h^2 = 0.218$), and high-density lipoprotein cholesterol ($h^2 = 0.191$) exhibited the highest heritability estimates (Table S6), highlighting the significant role of genetics in these traits. These results have far-reaching implications for precision medicine, as higher heritability signals suggest the potential for more accurate genetic risk prediction models that could improve personalized disease risk assessments.

We then partitioned the heritability to gene level and identified 368 unique genes contributing significantly to phenotypic variation ($h^2 > 0.1\%$ and z-score > 1.64), and 51 of them affected more than one category, including key genes such as *APOE, APOC1, TOMM40, ABCG2,* and *KCNQ1* (Fig. 3 and Table S7).

We also conducted a colocalization analysis to elucidate the potential molecular function of identified GWAS signals (Fig. 3 and Table S8). Our results identified 335 unique genes that might mediate the outcome through their expression (posterior probability > 0.9), including *GBAP1* which colocalized with five different traits (uric acid, serum creatinine, hematocrit, hypertension, and gout). These findings highlight the pleiotropic effects of these genes and present potential avenues for cross-disease therapeutic strategies, where targeting one gene could influence multiple related disorders. Understanding these shared genetic effects is crucial for devising broader precision medicine approaches, particularly for managing comorbidities.

Genetic Correlation and Clusters

Pairwise genetic correlation analysis revealed three major phenotype clusters: cardiometabolic traits, autoimmune and infectious diseases, and kidney-related traits (Fig. 4 and Extended Data Fig. 3). The cardiometabolic cluster, which includes type 2 diabetes, hypertension, and BMI, reinforces the interconnected genetic architecture of cardiovascular and metabolic diseases. The cluster of autoimmune and infectious diseases, which includes viral hepatitis B, psoriasis, and systemic lupus erythematosus, highlights shared immune system pathways and potential gene-pathogen interaction. The kidney-related cluster involved gout, chronic kidney disease, calculus of kidney and ureter, ankylosing spondylitis, and measures of urea nitrogen, creatinine, and uric acid. These findings are significant for clinical applications, suggesting that shared genetic risks across diseases could enable earlier detection of comorbidities and prevention strategies based on an individual's genetic profile. The shared genetic architecture also implied that we may leverage the genetic risk of correlated traits while developing the PRS model.

Cross-Population Comparison Based on GWAS

Cross-population comparisons[25] with the Europeans from UKB showed varying degrees of genetic correlation, with strong correlations for traits like cholelithiasis (ρ_{ge} > 0.999), type 2 diabetes (ρ_{ge}= 0.829), and ischemic heart disease (ρ_{ge}= 0.756), but moderate correlations for gout (ρ_{ge}= 0.616) and psoriasis (ρ_{ge}= 0.418) (Table S6). These findings demonstrate the importance

of population-specific genetic studies, as differences in genetic architectures between populations can significantly affect the accuracy of PRS models.

Polygenic Risk Score (PRS) Development

Building on these insights, we developed and validated PRS models that demonstrated strong predictive performance for a wide range of diseases. Although we used five PRS tools, including , LDpred2[26], Lassosum2[27], PRS-CS[28], SBayesR[29], and MegaPRS[30] (Tables S9-S13), we found that LDpred2 outperformed the others for most traits. Therefore, we took the results of LDpred2 for further comparisons. Out of the 289 PRS models, AUC values surpassed 0.55 for 106 dichotomized phecodes, while all 24 models for quantitative traits accounted for more than 3% of the phenotypic variance. (Table S9 and Extended Data Fig. 4). The top-performing PRS models included well-known heritable traits such as ankylosing spondylitis (AUC = 0.812±0.016), psoriasis (0.710±0.016), atrial fibrillation (0.702±0.014), prostate cancer (0.696±0.018), systemic lupus erythematosus (0.695±0.015), rheumatoid arthritis (0.649±0.011), type 2 diabetes (0.642±0.005), female breast cancer (0.610±0.010), and hypertension (0.610±0.004). Interestingly, the PRS for hepatitis B also demonstrated genetic predictability (0.655±0.008). Comparing these results with heritability, which serves as the theoretical upper bound of PRS explained variance (r^2), we found that 27 phecodes and 10 quantitative traits explained over 60% of their heritability, including prostate cancer, type 2 diabetes, female breast cancer, HDLC, triglycerides, and red blood cell count (Fig. S2).

Leveraging the identified clusters, we performed a multi-trait PRS training, PRSmix+[31], for the traits within each cluster (Fig 5 and Table S14). Notably, multi-trait PRS models improved prediction accuracy for the cardiometabolic cluster with a 3.9% increase in AUC and a 1.70-fold improvement in phenotypic variance explained (r^2). The performances of autoimmune and kidney-related clusters were also enhanced, with AUC improvements of 2.1% and 0.8%, respectively, and 1.44- and 1.32-fold improvements in variance explained (r^2). The significant enhancement of multi-trait PRS (comparing r^2 of LDpred2 and PRSmix+ with paired t-test, $p=1.1x10^{-9}$) highlights the potential of leveraging

shared genetic architecture to enhance disease prediction. Fig. 5 demonstrated the performance of single and multi-trait PRS across three disease clusters, and the differing effectiveness of PRS in predicting genetic risk across various disease categories.

External Validation and Comparison of PRS Models

To evaluate the robustness of the clinical implications of these findings, we performed the external validation of these PRS models (hypertension, type 2 diabetes, viral hepatitis B, gout, calculus of kidney from PRSmix+ and others from LDpred2) in Taiwan Biobank (TWB), UKB, and All of Us and confirmed their robustness and generalizability in EAS across different biobanks (Fig. 6). Although the TWB questionnaire did not contain specific details on hepatitis B status, we used anti-hepatitis B core total antibodies (Anti-HBc) as an indicator of infection or past infection and hepatitis B e-antigen (HBeAg) as a marker of active viral replication. Intriguingly, the AUC for HBeAg was 0.678±0.012, and for Anti-HBc, 0.531±0.002. These results suggest that the PRS for hepatitis B is predicting for the symptoms or severity of the disease.

TPMI-derived PRS models also consistently outperformed UKB European-derived models when applied to EAS (Fig. 6). These results suggest that population-specific PRS models could play a pivotal role in precision medicine, allowing for more accurate risk stratification and enabling personalized healthcare interventions. Additionally, we assessed the performance of TPMI-derived PRS across various ancestry groups, including European, African, Admixed American, and South Asian populations from the UK Biobank and All of Us cohorts (Extended Data Fig. 5). The performance varied by diseases, but consistent results were observed for female breast cancer and glaucoma across populations.

Impact of Genetic Risks on Overall Disease Burden

Although overall health is hard to define with a few metrics, herein, we used the count of clinical visits and duration of hospitalization to roughly describe individuals' overall disease burden. We found the 128 well-performed PRS models (PRSmix+ and LDpred2 model with AUC > 0.55 for phecodes and r^2 >

0.03 for continuous traits) shown to influence overall health indices, explaining 8.06% of the variation in clinical visit frequency (p-value = 5.7×10^{-13}) and 9.07% of the variation in hospitalization duration (p-value = 1.4×10^{-23}, Table 1). Among the identified clusters, cardiometabolic cluster contributed the most to the indices, accounting for 1.14% of clinical visits and 3.22% of hospitalizations. In short, we quantified the proportion of the developed polygenic risk of various diseases and traits on human health, and the results underlined the importance of genomic impact and necessary for precision health development.

Discussion

This study represents the largest GWAS conducted to date in the Han Chinese population, utilizing data of around 500,000 individuals recruited from 16 medical centers across Taiwan. We investigated the genetic architecture of 700 dichotomized phecodes and 24 quantitative traits, identifying 2,444 independent variant-trait associations and showed that population-specific genetic risk-prediction PRS models for a wide range of diseases performed well in the population. Indeed, for the 128 traits where there is sufficient sample size in the cohort, the PRS performance (AUC > 0.55 for dichotomized phecodes and $r^2 > 0.03$ for quantitative traits) rival those developed for Europeans using UKB data. These findings show that population-specific PRS models can be developed successfully for non-European populations and our project serves as a model for large-scale genetic studies in non-European populations.

Recent large-scale projects that emphasize ancestral diversity in human genetic studies have discovered new findings with the inclusion of non-European subjects. MVP conducted multi-ancestry GWAS on 635,000 participants, identifying over 2,000 signals unique to non-European populations[16]. With the TPMI dataset, we performed the largest-ever GWAS in the Han Chinese for several traits. For instance, the previous largest meta-analysis for type 2 diabetes included 77,418 cases from EAS populations, of which 20,573 were Han Chinese[32]. In contrast, our GWAS included 59,289 cases of type 2 diabetes, almost tripling the number of cases ever tested, and identified eight unreported T2D SNPs from known regions, demonstrating the power of TPMI sample size. Identification of new and population-specific risk

variants may lead to further understanding of their molecular mechanism and underline the need for population-specific weightings in PRS models. Moreover, population-specific findings also explain better performance of population-specific PRS model in the population in question. In short, our Han Chinese specific genomic profiles for comprehensive phenotypes provide a solid foundation for population-specific PRS development and precision medicine implementation.

Our understanding of how the genetic factors influencing Hepatitis B, an epidemic infectious disease in Taiwan with an estimated hepatitis B virus carrier rate of 15-20%, also benefited from the large dataset. With 23,618 cases, a significant increase from prior studies of only a few thousand cases, we identified novel loci and demonstrated that host genome may determine the severity and symptoms of this infectious disease. This is similar to that previously reported in COVID-19 and pneumonia, where genetic factors have been shown to influence disease outcomes[33-36]. Our unexpected success of GWAS and PRS for hepatitis B not only demonstrate the power of the large sample size of TPMI, but also reveal the necessity of population-specific genetic study for population-unique diseases.

In addition, the comprehensive phenotypic data allows us to leverage the genetic correlation of multiple traits to improve the performance of PRS models. Three large clusters of correlated traits were identified from pairwise genetic correlation analysis. Including the correlated traits in PRS model development improved performance, resulting in an average 1.53-fold increase in the explained percentage of phenotypic variation. Although previous studies have proven the utility of multi-traits on target diseases[31,37,38], we are first to use this approach at a phenome-wide level and demonstrate the improvement across different types of traits. As a result, we produced well-performed PRS for various categories of diseases, including cardiometabolic diseases, autoimmune disorders, and infectious diseases.

We evaluated our PRS models across several large cohorts, including the TWB, UK Biobank, and All of Us. The TPMI-derived PRS models consistently outperformed those developed from European populations when applied to

diseases in EAS from the three large cohorts. When comparing with European-derived PRS models, we also observed better performance across several traits in EAS, particularly for cardiometabolic and autoimmune diseases. By integrating these well-developed PRS models, we estimate that genetics account for 9.1% of the disease burden in Taiwan. As Taiwan has an aging (and soon to be super-aging) population, implementing genetic risk-based health management strategies may decrease the disease burden while extending health-span significantly.

As with other large-scale studies, our study has several commonly found limitations. First, the TPMI cohort size is not sufficiently large to study some of the severe subtypes of many (common or rare) diseases, such as diabetes insipidus and neurofibromatosis. Second, we attempted to use eQTLs to elucidate the molecular mechanism of diseases, but the underrepresentation of EAS in current eQTL datasets, such as GTEx, poses challenges[39]. Gene expression regulation varies across ancestries[40,41], and differences in LD structures further complicate colocalization analyses. Therefore, ancestral diversity is an urgent need not only in genomic data but also in transcriptomic, proteomic, metabolomic, and epigenomic datasets. Third, the EMR of the participants are incomplete, as some participants receive care from multiple health providers, but the TPMI only has access to EMRs from their enrollment hospitals. Furthermore, the current project retrieved EMRs from an average of 5 years prior to enrollment, so some important data such as age of disease onset for the older participants are not available. Incomplete EMR leads to less precise case definition of some participants. Fourth, some of the younger participants have high-risk genetic profiles but are disease free for those diseases. The duration of the project is too short to determine whether they will eventually develop those diseases.

Effort is underway to gain access to the complete EMRs of the TPMI participants and to recruit additional participants with severe subtypes of common diseases. The high-risk participants who are symptom-free are being followed to monitor disease development. Future studies are being planned to study the high-risk individuals who escape disease development to identify

genetic and non-genetic factors that mitigate their disease risk.

This study formally validates the common belief that population-specific risk-prediction models perform well in that population, pointing to the need to build these models in the major populations across the world. As the PRSs we developed perform well in EAS, the implementation of genetic findings is now available to an additional 25% of the world's population. It is hopeful that if all EAS obtain their genetic profiles and determine their risk for major diseases, many diseases can be prevented or their onset can be delayed significantly, thereby fulfilling the promise of modern genetics.

We also found notable evidence that, for certain diseases, e.g. female breast cancer and glaucoma, PRS models developed from one population can perform equally well in other populations, suggesting a shared genetic etiology for those conditions. This raises the possibility of creating universal PRS models that could be effective across diverse populations for at least some diseases.

In conclusion, we used a large-scale Han Chinese dataset produced by the TPMI to conduct pheno-wide genetic analyses and leverage these genetic findings to train risk prediction models for multiple diseases and traits. The developed models are validated in EAS of different biobanks and demonstrate a consistent performance that bodes well for their use in the general Han Chinese/EAS population. Our approach can serve as a template for developing PRS models in populations currently without such resources, anticipating the time when all populations around the world can benefit from risk-based health management as part of the precision health movement.

Methods

Study Population and Phenotyping

We utilized the Taiwan Precision Medicine Initiative (TPMI) dataset, which links extensive electronic medical records (EMR) with genotypic data for 486,956 individuals. Dichotomized disease status was defined by phecodes, which were based on information extracted from the EMR using International Classification of Diseases (ICD) codes[18,19]. To ensure robustness, cases were defined by

having the diagnosis of the relevant condition on two or more clinical visits. We also extracted quantitative traits from the EMR, including anthropometric, vital sign and laboratory measurements, and we excluded the extreme outliers and removed or adjusted the treated and/or medicated measures based on previous research, and the median value was kept if the participant had multiple qualified measures[42]. (Supplementary methods) In this study, we focused on 702 dichotomized phenotypes (phecodes) that had at least 2,000 cases and 24 quantitative traits that were measured in at least 100,000 individuals. These phecodes spanned 17 disease categories, including but not limited to infectious diseases, neoplasms, endocrine/metabolic disorders, and circulatory system diseases. The 24 quantitative traits were categorized into anthropometric, circulatory, hematological, kidney-related, liver-related, and metabolic measurements.

Genotyping and Quality Control

Genotyping was performed using two customized high-density Axiom SNP arrays produced by Thermo Fisher (Waltham, MA, USA), TPM1 and TPM2. The genotyping experiments were conducted in six genotyping centers in Taiwan[17]. The raw genotypic data underwent quality control measures, and the genetic variants were excluded when they had call rate <0.02, minor allele frequency (MAF) <0.01, Hardy-Weinberg equilibrium test p-value $<1x10^{-6}$. We also excluded individuals with overall call rate <0.95, failed heterozygosity check, or inconsistent documented versus genetically determined sex. For this study, we only included the genetic variants found on both genotyping arrays and excluded variants with a significant batch effect in GWAS. The proportion of genetic ancestry was determined by ADMIXTURE[43], and the projected principal component scores with 1000 Genome as reference panel were applied to determine individuals' ancestry[44]. As a result, 401,710 genetic variants and 463,447 Han Chinese participants passed all quality control measures and were used in the subsequent studies. Details are found in supplementary methods and GitHub (https://github.com/TPMI-Taiwan/tpmi-qc).

Phasing and Imputation

Phasing was conducted on QC-passed genotype data with SHAPEIT5[45].

Genome imputation was carried out with IMPUTE5 using a reference panel of 1,498 whole genome sequenced Taiwan Biobank subjects[12,46]. We also conducted post-imputation quality control with exclusion criteria INFO score ≤ 0.7 and MAF ≤ 0.01. In addition, we also performed a chip-GWAS for minimizing the bias from different chips, resulting in a dataset of 8,046,864 well-imputed common genetic variants.

Population Structure and Relatedness Estimation

We performed a principal component analysis (PCA) based on genotyped variants to capture the effect of population structure. To diminish the effect of close relatives, the main PCA was conducted in a genetically unrelated subset, and other subjects were projected with the calculated PC weightings. And then, these PCA scores were leveraged to accurately quantify the proportion of identity-by-decent (IBD) and degree of relatedness. The maximum unrelated set was determined based on these estimated degrees of relatedness. PC-AiR and PC-Relate were used for PCA and relatedness estimation and PRIMUS was used for identifying the maximum unrelated set with the third degree as threshold[47-49].

Genome-Wide Association Study (GWAS)

To train and validate the proposed PRS models, we left 100,000 unrelated subjects out of GWAS. To maximize the statistical power, we used a mixed-effect regression model to examine the association between genotype and outcome of interest, logistic regression for dichotomized phecode, and linear regression for quantitative traits. The quantile-normalization was applied to quantitative traits to ensure the normal distribution. The mixed-effect model accounted for relatedness among individuals by including a random effect for pairwise kinship. The model also adjusted for key covariates, including age, sex, age^2, interactions between age/age^2 and sex, genotyping chip, enrollment hospital, and 10 genetic principal components to control for population stratification. SAIGE was applied for the mixed effect model GWAS[50]. In addition, we also performed a generalized linear model from PLINK2 for GWAS among the unrelated subset (n = 248,754)[51], and these statistics were used for heritability and genetic correlation estimation.

Replication Evaluation

To systematically evaluate the performance of our GWAS, we leveraged a pre-summarized phenotype-genotype reference map[52], which collected 5,879 genetic associations for 149 unique phecodes from 523 published GWAS and 1,215 associations from EAS. We calculate the overall and power-adjusted replication rates and actual over expected ratio for each available phecode and categories respectively. For measuring the quality of biobank data through replication an R package PGRM was used[52].

Fine-mapping

We performed fine-mapping to identify the independent GWAS signals in all genomic regions containing any variant with a p-value < $5x10^{-8}$ and ±1.5 Mb of the regional lead variant[14], except the major histocompatibility complex region (MHC region, chr6: 25,391,792-33,424,245) due to its complex LD structure. We used the reported 95% credible set to determine the independent signals, and up to ten signals were allowed for each region. The genome-wide significant threshold was applied for defining a credible set as an independent hit, and an additional requirement of log Bayes factor (BF) > 2 was applied for the second hit. For the failed fine-mapping regions and MHC region, we used the lead SNP as the hit of each significant region. SuSiE was conducted for this summary statistics-based fine-mapping with LD derived from our imputation reference panel[53].

Novel Association Identification

We comprehensively compared our GWAS results with reported significant signals on the NHGRI-EBI GWAS Catalog[54], download at 2024-03-11. The mapping of phecodes and quantitative traits to GWAS catalog phenotypes is summarized in Table S15. We classified a gene or region as novel if the fine-mapped independent signal was not located within 1 Mb of any reported genome-wide significant association (p-value < $5x10^{-8}$) for the corresponding phenotype[55]. Additionally, a variant was considered a novel hit if the highest linkage disequilibrium (LD) r^2 was less than 0.1 with any reported significant association. Associations derived from uncertain and umbrella phecodes were

excluded, and for duplicated genetic variants or regions, we only reported the association with the smaller p-value or from the phecode with the more specific definition. Finally, we used ANNOVAR to annotate the novel variants with data from the RefSeq Gene database (2020-08-17 updated)[56,57].

Heritability, Genetic Correlation, and Clustering

To quantify the genomic contribution of the specific traits, we applied linkage disequilibrium score regression to estimate the SNP-based heritability with LDSC[24]. The GWAS summary statistics and the pre-calculated LD score from EAS superpopulation of 1000 Genome were used[44]. For the dichotomized traits, we performed a liability-scaled transformation on the observed heritability using the 5-years population prevalence from the National Health Insurance dataset[20,58]. Additionally, we conducted LDSC to obtain pairwise genetic correlations to assess the similarity of genetic mechanisms between traits[59]. Based on the genetic correlation matrix, we used a hierarchical cluster analysis to identify groups of traits that share genetic mechanisms. We employed the weighted pair group method with arithmetic mean (WPGMA) for clustering, and the resulting cluster tree was used for group identification. Moreover, we estimate the genetic correlation across populations, TPMI and UK biobank, to demonstrate varied genetic architecture in different ancestry populations. For the UK biobank GWAS, we applied a generalized linear model from PLINK2 with the predefined phecode, https://github.com/umich-cphds/createUKBphenome, and corresponding baseline quantitative measures among the identified unrelated set (n = 378,544). Popcorn was performed for the cross-population genetic correlation, and two correlation coefficients were calculated, the transethnic genetic-effect correlation (ρ_{ge}) and transethnic genetic-impact correlation (ρ_{gi})[25].

Gene-Level Heritability and Colocalization

We used both gene-level heritability estimation and colocalization analysis to map our GWAS findings to functional units, specifically genes. We conducted h2gene analysis to partition SNP-based heritability to the gene level[60]. We estimated heritability for genes that overlapped with fine-mapped regions, where gene regions were defined as the gene body ±10 kb for gene-level

heritability. Additionally, to illustrate the molecular functions of genes of interest, we used colocalization analysis to examine whether there are shared common genetic causal variants between tissue-specific gene expression and traits of interest. We utilized expression quantitative traits locus (eQTL) resources from 49 tissues in GTEx v8[39], testing any gene with genome-wide significant signals in the cis-regulation region (±1 Mb). The posterior probabilities were used to evaluate colocalization between gene expression and the trait of interest. The R package, coloc, was used with SuSiE relaxing the single causal variant assumption[61,62].

Single and Multi-Trait Polygenic Risk Score (PRS)

The preserved dataset of 100,000 unrelated TPMI subjects was split into two subsets, training (n = 80,000) and validation (n = 20,000) for PRS model building. Five popular PRS tools were used, LDpred2[26], Lassosum2[27], PRS-CS[28], SBayesR[29], and MegaPRS[30], and the training subset was applied for parameter selection and model optimization if needed. LDpred2, PRS-CS, and SBayesR assumed the effect of genetic variants following a mixture distribution with different pre-defined parameters and applied a Bayesian framework for distribution estimation. Lassosum2 utilized a penalized regression (LASSO) for weight estimating, and MegaPRS leveraged minor allele frequency and linkage disequilibrium for model building. We then used the validation subset to evaluate the performance of PRS models. Individual score was calculated with PLINK2[51]. The explained variance (r^2) was used to evaluate the performance of PRS for quantitative traits[63,64], and two indices, area under the receiver operating characteristic curve (AUC) and liability-scaled $r^{2,}$ were used for PRS of dichotomized phenotypes. The likelihood ratio test was used to obtain the significance for r^2 with R package, lmtest, and standard error for AUC was calculated with R package, auctestr. To further leverage the gene's pleiotropy and shared genetic mechanism among traits, we conducted a multi-trait PRS model building for the traits in the same genetic cluster based on pairwise genetic correlation identified in the previous step. We pooled all PRS models from five tools for those identified traits and applied an elastic net regression to combine their weighting and find the most optimized model for the target trait.

PRSmix+ was performed for the multiple traits PRS model building[31].

External Validation and Comparison

We conducted an external validation of our developed PRS using data from the Taiwan Biobank, EAS from UK Biobank and All of Us. TWB is a community-based biobank, and it has recruited over 200,000 participants in Taiwan. Herein, we used 120,460 independent subjects, who were genotyped with the Axiom customized chip TWB2 (equivalent to TPM1), and their genotyping QC, phasing, and imputation followed the same protocol as described above. The self-reported disease condition was queried from their baseline questionnaire, except for cancer. Since the study design of TWB excluded cancer patients at recruitment, we used both baseline and follow-up self-reporting data to define cancer cases and controls. UKB has enrolled ~500,000 participants since 2006 and linked their genetic data with enriched phenotypic data. For external validation, we only used self-reported Chinese (more diverse than Han Chinese) as East Asian and their inpatient record for case definition. All of Us intends to enroll over 1 million participants in the United States and has released whole genome genotyping data for ~312,000 participants as of the first quarter of 2024. The genetically confirmed EAS as well as other superpopulations and their linked EMR were used for validating our PRS models. Moreover, we compared the TPMI-derived PRS model with UKB-derived models to investigate the performance of population-specific PRS. The UKB-derived models were based on published UKB European GWAS (https://pheweb.org/UKB-TOPMed/), and LDpred2-auto was applied for model building.

Overall Disease Burden Evaluation

We evaluated the genetic impact on overall disease burden. We used the number of clinical visits and the aggregate duration of hospitalization as disease burden indices. Due to collinearity among PRS for different traits, we utilized a partial least square-generalized linear model (PLS-GLM) to extract components from the PRS of qualified traits with R package, plsRglm[65]. The number of extracted components was determined by the Akaike Information Criterion (AIC). We then estimated the covariate-adjusted proportion of genetic contribution (r^2) by comparing the full model with the null model, which included

only covariates such as sex, age, and hospital. Likelihood ratio test was used to obtain the significances of regression models. For each index, we employed three models to compare the top and bottom 5%, 10%, and 20%. We selected covariate-matched controls from subjects without hospitalization records as the bottom group for hospitalization models.

Code availability

Code for genotyping quality control process and analysis is available at our Github (https://github.com/TPMI-Taiwan/).

Data availability

The genotyping and electronic medical record (EMR) data analyzed in this study are from the Taiwan Precision Medicine Initiative (TPMI) with proper approval from the TPMI Data Access Committee. In compliance with the confidentiality laws governing genetic and health data in Taiwan, the de-identified TPMI data are kept in a secure server at the Academia Sinica and not released to the public. All summary statistics, polygenic risk score (PRS) models, and GWAS results are freely available from the TPMI website (https://tpmi.ibms.sinica.edu.tw). Researchers requesting access to the individual genotyping and EMR data can do so on a collaborative basis. Instructions on requesting access to the data can be found on the TPMI's official website (https://tpmi.ibms.sinica.edu.tw).

Ethics

This study was approved by the Institutional Review Boards of Taipei Veterans General Hospital (2020-08-014A), National Taiwan University Hospital (201912110RINC), Tri-Service General Hospital (2-108-05-038), Chang Gung Memorial Hospital (201901731A3), Taipei Medical University Healthcare System (N202001037), Chung Shan Medical University Hospital (CS19035), Taichung Veterans General Hospital (SF19153A), Changhua Christian Hospital (190713), Kaohsiung Medical University Chung-Ho Memorial Hospital (KMUHIRB-SV(II)-20190059), Hualien Tzu Chi Hospital (IRB108-123-A),Far Eastern Memorial Hospital (110073-F), Ditmanson Medical Foundation Chia-Yi

Christian Hospital (IRB2021128), Taipei City Hospital (TCHIRB-10912016), Koo Foundation Sun Yat-Sen Cancer Center (20190823A), Cathay General Hospital (CGH-P110041), Fu Jen Catholic University Hospital (FJUH109001) and Academia Sinica (AS-IRB01-18079), Taiwan. Written informed consent was obtained from the subjects in accordance with institutional requirements and the Declaration of Helsinki principles. All collected information was de-identified before statistical data analysis. The analysis with Health and Welfare Data Science Center (HWDC) was approved by Institutional Review Boards of Academia Sinica (AS-IRB-BM-23056). This research has been conducted using the UK Biobank Resource under UK Biobank Main Application 15326. Work with All of Us data was performed using the All of Us Researcher Workbench under the workspace “Duplicate of Prediction of Polygenic Traits'”.

Acknowledgements

We thank all the participants and researchers of the Taiwan Precision Medicine Initiative and the Taiwan Biobank. This study was funded in part by the Academia Sinica (40-05-GMM, AS-GC-110-MD02, and 236e-1100202 to P.-Y.K. and J.-Y.W.) and the National Development Fund, Executive Yuan (NSTC 111-3114-Y-001-001 to P.-Y.K.). Analysis using UK Biobank data used computational resources hosted by the Michigan State University High-Performance Computing Center. Data from the UK Biobank includes data provided by patients and collected by the National Health Service (NHS) England as part of their care and support. UK Biobank data also includes data assets made available by National Safe Haven as part of the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (research which commenced between 1st October 2020--31st March 2021 grant ref MC_PC_20029; 1st April 2021--30th September 2022 grant ref MC_PC_20058). Also, we are grateful and acknowledge the contributions of the All of Us participants who make this project possible and the work of the National Institutes of Health's All of Us Research Program for making this data available.

References

1 Lennon, N. J. *et al.* Selection, optimization and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse US populations. *Nat Med* **30**, 480-487 (2024). https://doi.org/10.1038/s41591-024-02796-z

2 Thompson, D. J. *et al.* A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. *PLoS One* **19**, e0307270 (2024). https://doi.org/10.1371/journal.pone.0307270

3 Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet* **52**, 242-243 (2020). https://doi.org/10.1038/s41588-020-0580-y

4 Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* **10**, 3328 (2019). https://doi.org/10.1038/s41467-019-11112-0

5 Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635-649 (2017). https://doi.org/10.1016/j.ajhg.2017.03.004

6 Bentley, A. R., Callier, S. & Rotimi, C. N. Diversity and inclusion in genomic research: why the uneven progress? *J Community Genet* **8**, 255-266 (2017). https://doi.org/10.1007/s12687-017-0316-6

7 Smith, J. L. *et al.* Multi-Ancestry Polygenic Risk Score for Coronary Heart Disease Based on an Ancestrally Diverse Genome-Wide Association Study and Population-Specific Optimization. *Circulation: Genomic and Precision Medicine* **17**, e004272 (2024). https://doi.org/10.1161/CIRCGEN.123.004272

8 Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat Genet* **52**, 669-679 (2020). https://doi.org/10.1038/s41588-020-0640-3

9 Nam, K., Kim, J. & Lee, S. Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits. *Cell Genom* **2**, 100189 (2022). https://doi.org/10.1016/j.xgen.2022.100189

10 Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom* **3**, 100361 (2023). https://doi.org/10.1016/j.xgen.2023.100361

11 Feng, Y. A. *et al.* Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genom* **2**, 100197 (2022). https://doi.org/10.1016/j.xgen.2022.100197

12 Wei, C. Y. *et al.* 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). https://doi.org/10.1038/s41525-021-00178-9

13 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015). https://doi.org/10.1371/journal.pmed.1001779

14 Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023). https://doi.org/10.1038/s41586-022-05473-8

15 All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340-346 (2024).

https://doi.org/10.1038/s41586-023-06957-x
16 Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024). https://doi.org/10.1126/science.adj1182
17 Yang, H.-C. *et al.* The Taiwan Precision Medicine Initiative: A Cohort for Large-Scale Studies. *BIORXIV* (2024). https://doi.org/BIORXIV/2024/616932
18 Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* **4**, 1-19 (2021). https://doi.org/10.1146/annurev-biodatasci-122320-112352
19 Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1110 (2013). https://doi.org/10.1038/nbt.2749
20 Lin, L. Y., Warren-Gash, C., Smeeth, L. & Chen, P. C. Data resource profile: the National Health Insurance Research Database (NHIRD). *Epidemiol Health* **40**, e2018062 (2018). https://doi.org/10.4178/epih.e2018062
21 Zhou, Y. *et al.* Performance of multigene testing in cytologically indeterminate thyroid nodules and molecular risk stratification. *PeerJ* **11**, e16054 (2023). https://doi.org/10.7717/peerj.16054
22 Tyagi, A., Goyal, A., Chaware, P. & Rathinam, B. A. D. Mutations of PHOX2B Gene in Patients of Obesity Hypoventilation Syndrome in Central India. *J Lab Physicians* **14**, 164-168 (2022). https://doi.org/10.1055/s-0041-1735582
23 He, D. *et al.* A longitudinal genome-wide association study of bone mineral density mean and variability in the UK Biobank. *Osteoporos Int* **34**, 1907-1916 (2023). https://doi.org/10.1007/s00198-023-06852-1
24 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015). https://doi.org/10.1038/ng.3211
25 Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes, C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am J Hum Genet* **99**, 76-88 (2016). https://doi.org/10.1016/j.ajhg.2016.05.001
26 Prive, F., Arbel, J. & Vilhjalmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424-5431 (2021). https://doi.org/10.1093/bioinformatics/btaa1029
27 Prive, F., Arbel, J., Aschard, H. & Vilhjalmsson, B. J. Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *HGG Adv* **3**, 100136 (2022). https://doi.org/10.1016/j.xhgg.2022.100136
28 Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019). https://doi.org/10.1038/s41467-019-09718-5
29 Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun* **10**, 5086 (2019). https://doi.org/10.1038/s41467-019-12653-0
30 Zhang, Q., Prive, F., Vilhjalmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat Commun* **12**, 4192 (2021). https://doi.org/10.1038/s41467-021-24485-y
31 Truong, B. *et al.* Integrative polygenic risk score improves the prediction

accuracy of complex traits and diseases. *Cell Genom* **4**, 100523 (2024). https://doi.org/10.1016/j.xgen.2024.100523

32 Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240-245 (2020). https://doi.org/10.1038/s41586-020-2263-3

33 Chen, H. H. *et al.* Host genetic effects in pneumonia. *Am J Hum Genet* **108**, 194-201 (2021). https://doi.org/10.1016/j.ajhg.2020.12.010

34 Covid- Host Genetics Initiative. A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7-E26 (2023). https://doi.org/10.1038/s41586-023-06355-3

35 Covid- Host Genetics Initiative. A first update on mapping the human genetic architecture of COVID-19. *Nature* **608**, E1-E10 (2022). https://doi.org/10.1038/s41586-022-04826-7

36 Asgari, S. & Pousaz, L. A. Human genetic variants identified that affect COVID susceptibility and severity. *Nature* **600**, 390-391 (2021). https://doi.org/10.1038/d41586-021-01773-7

37 Kelemen, M., Vigorito, E., Fachal, L., Anderson, C. A. & Wallace, C. shaPRS: Leveraging shared genetic effects across traits or ancestries improves accuracy of polygenic scores. *Am J Hum Genet* **111**, 1006-1017 (2024). https://doi.org/10.1016/j.ajhg.2024.04.009

38 Zhai, S., Guo, B., Wu, B., Mehrotra, D. V. & Shen, J. Integrating multiple traits for improving polygenic risk prediction in disease and pharmacogenomics GWAS. *Brief Bioinform* **24** (2023). https://doi.org/10.1093/bib/bbad181

39 GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020). https://doi.org/10.1126/science.aaz1776

40 Zhong, Y., Perera, M. A. & Gamazon, E. R. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am J Hum Genet* **104**, 1097-1115 (2019). https://doi.org/10.1016/j.ajhg.2019.04.009

41 Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol* **21**, 233 (2020). https://doi.org/10.1186/s13059-020-02113-0

42 Kirby, J. C. *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* **23**, 1046-1052 (2016). https://doi.org/10.1093/jamia/ocv202

43 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009). https://doi.org/10.1101/gr.094052.109

44 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015). https://doi.org/10.1038/nature15393

45 Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* **55**, 1243-1249 (2023). https://doi.org/10.1038/s41588-023-01415-w

46 Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet* **16**, e1009049 (2020).

https://doi.org/10.1371/journal.pgen.1009049
47 Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**, 127-148 (2016). https://doi.org/10.1016/j.ajhg.2015.11.022
48 Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**, 276-293 (2015). https://doi.org/10.1002/gepi.21896
49 Staples, J. *et al.* PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am J Hum Genet* **95**, 553-564 (2014). https://doi.org/10.1016/j.ajhg.2014.10.005
50 Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341 (2018). https://doi.org/10.1038/s41588-018-0184-y
51 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). https://doi.org/10.1186/s13742-015-0047-8
52 Bastarache, L. *et al.* The phenotype-genotype reference map: Improving biobank data science through replication. *Am J Hum Genet* **110**, 1522-1533 (2023). https://doi.org/10.1016/j.ajhg.2023.07.012
53 Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet* **18**, e1010299 (2022). https://doi.org/10.1371/journal.pgen.1010299
54 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023). https://doi.org/10.1093/nar/gkac1010
55 Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112-1118 (2010). https://doi.org/10.1093/bioinformatics/btq099
56 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010). https://doi.org/10.1093/nar/gkq603
57 Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**, D942-D949 (2023). https://doi.org/10.1093/nar/gkac1071
58 Ojavee, S. E., Kutalik, Z. & Robinson, M. R. Liability-scale heritability estimation for biobank studies of low-prevalence disease. *Am J Hum Genet* **109**, 2009-2017 (2022). https://doi.org/10.1016/j.ajhg.2022.09.011
59 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-1241 (2015). https://doi.org/10.1038/ng.3406
60 Burch, K. S. *et al.* Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. *Am J Hum Genet* **109**, 692-709 (2022). https://doi.org/10.1016/j.ajhg.2022.02.012
61 Wallace, C. Statistical testing of shared genetic control for potentially related traits. *Genet Epidemiol* **37**, 802-813 (2013). https://doi.org/10.1002/gepi.21765
62 Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to

variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Series B Stat Methodol* **82**, 1273-1300 (2020). https://doi.org/10.1111/rssb.12388

63 Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* **36**, 214-224 (2012). https://doi.org/10.1002/gepi.21614

64 Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol Psychiatry* **90**, 611-620 (2021). https://doi.org/10.1016/j.biopsych.2021.04.018

65 Bertrand, F. & Maumy-Bertrand, M. plsRglm: Partial least squares linear and generalized linear regression for processing incomplete datasets by cross-validation and bootstrap techniques with R. arXiv:1810.01005 (2018). <https://ui.adsabs.harvard.edu/abs/2018arXiv181001005B>.

Figure legend

Fig. 1. Scatter plots of the case proportion for dichotomized phenotypes and sample size for quantitative traits in TPMI dataset. (A) The case proportion in TPMI is compared to the 5-year prevalence in the National Health Insurance Research Database (NHIRD) for 702 dichotomized phenotypes (phecodes). Each dot represents a specific phecode, with the x-axis showing the prevalence in NHIRD and the y-axis showing the case proportion in TPMI. (B) Scatter plot shows the sample sizes for 24 quantitative traits in the TPMI cohort. Each point represents a trait, with the x-axis indicating the different category of quantitative traits and the y-axis representing the corresponding sample sizes.

Fig. 2. Pheno-wide independent variant-trait associations. Vertical bars show the accumulated number of independent variant-trait associations for dichotomized phecodes (top panel) and quantitative traits (bottom panel). Each category of diseases and traits is represented by a corresponding color. The X-axis is chromosome number, and the Y-axis represents the accumulated number of associations, highlighting the uneven distribution of trait-associated variants across phenotypes.

Fig. 3. Gene-level heritability and colocalization with gene expression. Circle plot showing gene-level heritability and colocalization with gene expression for (A) dichotomized phenotypes, summarized in parent (integer) phecodes, and (B) quantitative traits. Dots represent gene-level heritability (h^2) > 10^{-3}, squares indicate colocalization posterior probability > 0.9, and triangles show both. Inner circle indicates the number associated traits for each identified gene. The bar chart shows the number of identified genes by category and grouped by type of pleiotropy.

Fig. 4. Genetic correlation among three identified trait clusters. Heatmap displays genetic correlations between trait clusters: cardiometabolic, autoimmune/infectious diseases, and kidney-related traits. Genetic correlation was estimated using LDSC, with colors representing the correlation coefficients between traits.

Fig. 5. PRS performance for the three identified trait clusters. Bar plot shows SNP-heritability and PRS explained variance (r^2) for (A) cardiometabolic trait cluster, (B) autoimmune trait cluster, and (C) kidney-related trait cluster. Gray bars indicate SNP-heritability, and the colored bar chart presents the r^2 values,

indicating the proportion of variance explained by the PRS from single-trait PRS (LDpred2, red bar) or multi-trait PRS (PRSmix+, blue bar), while the dot and whisker plot showcases predictive accuracy using AUC. The area under the receiver operating characteristic curve (AUC) presented with 95% confidence interval for dichotomized traits.

Fig. 6. External validation of PRS models in Taiwan Biobank and other cohorts. PRS performance is presented as area under the receiver operating characteristic curve (AUC) ± 95% CI in TPMI (orange), Taiwan Biobank (green), East Asians in UK Biobank (blue), and East Asians in All of Us (purple). Circles represent TPMI-derived PRS, and triangles indicate UKB (European)-derived PRS models. Only the estimates with case size > 40 were showed on the figure.

Extended Data Fig. 1. Comparison of TPMI GWAS-identified loci to the previously published GWAS. The replication rates of TPMI GWAS-identified loci when compared to previously reported loci from the GWAS catalog are presented in this bar chart. Red bars indicate the comparison of TPMI findings to that of all ancestries and blue bars represent the comparison to East Asian ancestries. The categories of diseases are shown under the bars.

Extended Data Fig. 2. The Manhanttan plot of GWAS for viral hepatitis B in TPMI. The names of nearest mapped gene were labed for the independent GWAS significant loci.

Extended Data Fig. 3. Genetic correlation heatmap for all heritable traits. Heatmap showing genetic correlations among heritable traits. Genetic correlations were estimated using LDSC, with colors representing the correlation coefficients between traits. The weighted pair group method with arithmetic mean (WPGMA) was used for clustering with the correlation coefficient as distance between traits.

Extended Data Fig. 4. The bar chart and dot plot for PRS performance. Bar and dot plot showing PRS explained variance (r^2) and SNP-heritability for dichotomous traits. Gray bars represent SNP-heritability, and dots show AUC. An asterisk (*) indicates estimates considering the MHC region.

Extended Data Fig. 5. External validation of PRS models across populations. PRS validation (AUC ± 95% CI) in East Asian (red), European (olive green), African (green), South Asian (blue), and All of Us (purple) populations from TPMI (circle), UKB (triangle), and All of Us (square) cohorts.

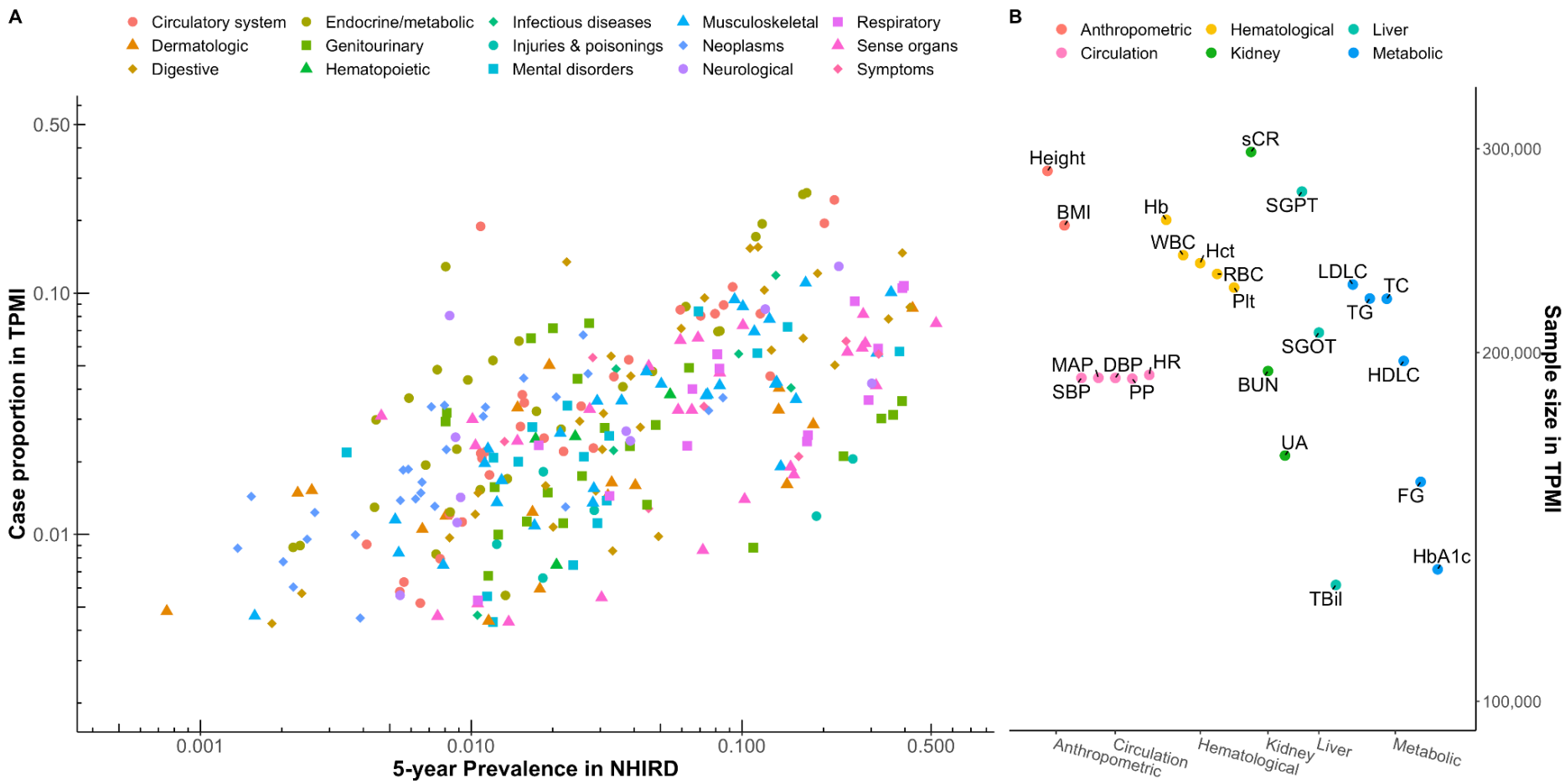


Fig 1. Scatter plots of the case proportion and case number in TPMI dataset. (A) The case proportion in TPMI is compared to the 5-year prevalence in the National Health Insurance Research Database (NHIRD) for 702 dichotomized phenotypes (phecodes). Each dot represents a specific phecode, with the x-axis showing the prevalence in NHIRD and the y-axis showing the case proportion in TPMI. (B) Scatter plot shows the sample sizes for 24 quantitative traits in the TPMI cohort. Each point represents a trait, with the x-axis indicating the different category of quantitative traits and the y-axis representing the corresponding sample sizes.

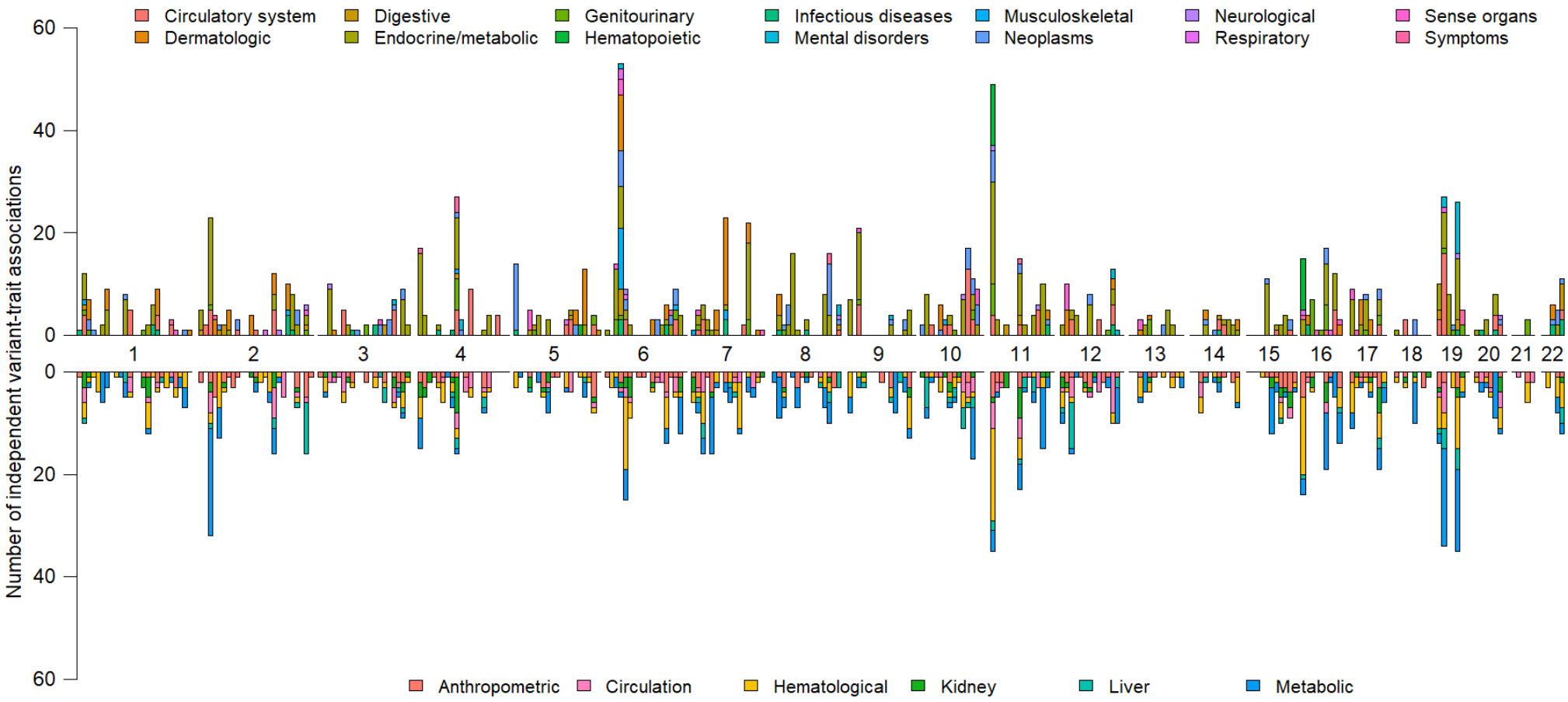


Fig. 2. Pheno-wide independent variant-trait associations. Vertical bars show the accumulated number of independent variant-trait associations for dichotomized phecodes (top panel) and quantitative traits (bottom panel). Each category of diseases and traits is represented by a corresponding color. The X-axis is chromosome number, and the Y-axis represents the accumulated number of associations, highlighting the uneven distribution of trait-associated variants across phenotypes.

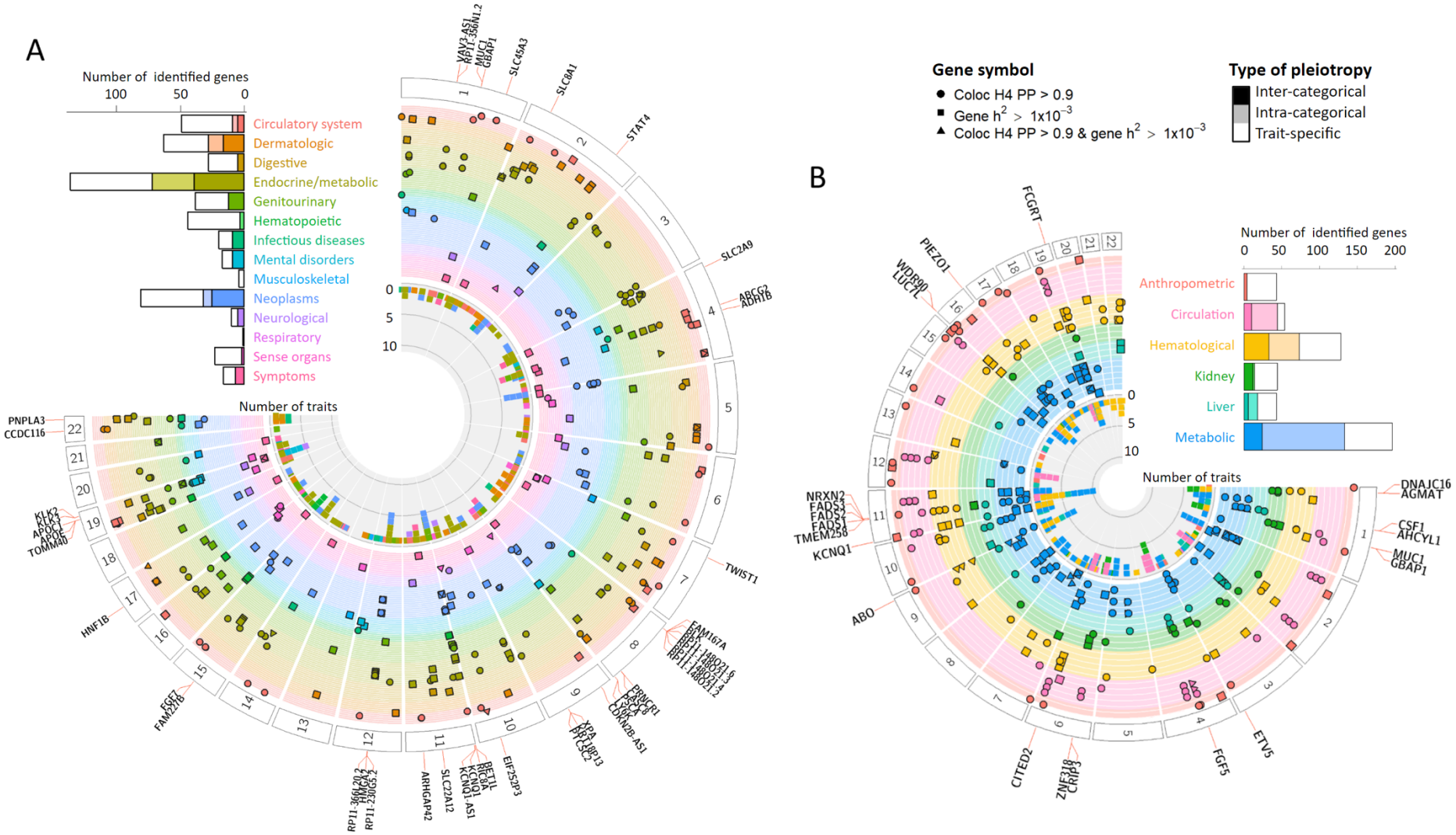


Fig. 3. Gene-level heritability and colocalization with gene expression. Circle plot showing gene-level heritability and colocalization with gene expression for (A) dichotomized phenotypes, summarized in parent (integer) phecodes, and (B) quantitative traits. Dots represent gene-level heritability (h^2) $>10^{-3}$, squares indicate colocalization posterior probability > 0.9, and triangles show both. Inner circle indicates the number associated traits for each identified gene. The bar chart shows the number of identified genes by category and grouped by type of pleiotropy.

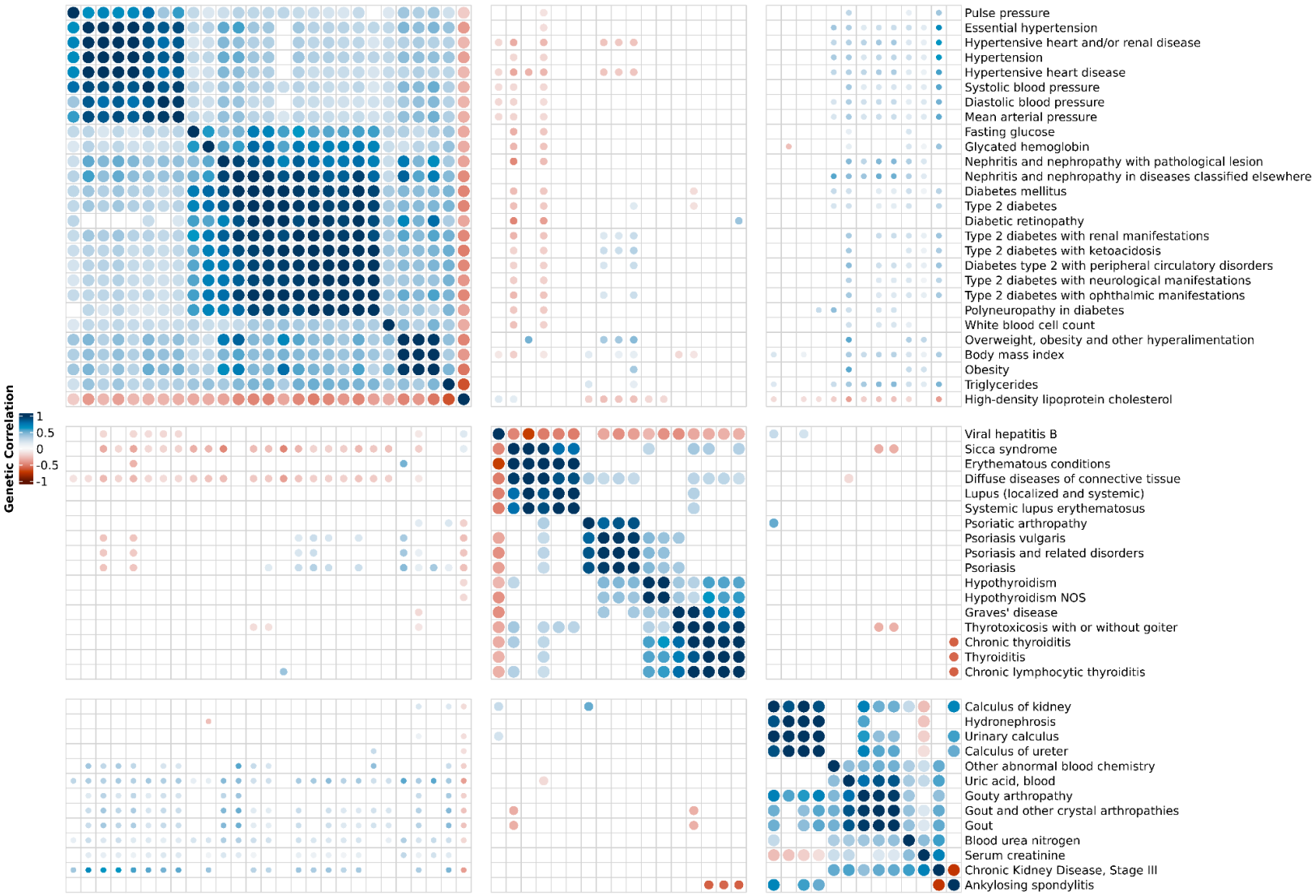


Fig. 4. Genetic correlation among three identified trait clusters. Heatmap displays genetic correlations between trait clusters: cardiometabolic, autoimmune/infectious diseases, and kidney-related traits. Genetic correlation was estimated using LDSC, with colors representing the correlation coefficients between traits.

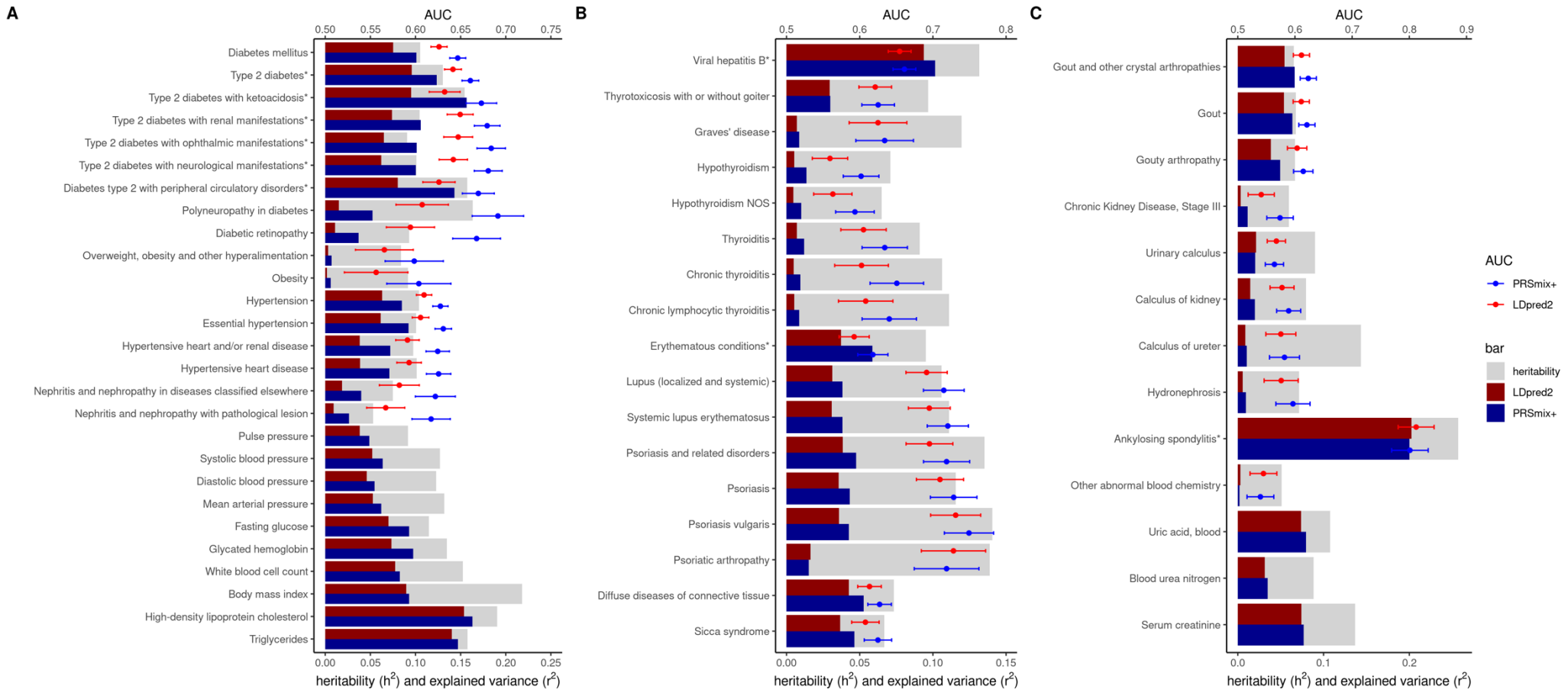


Fig. 5. PRS performance for the three identified trait clusters. Bar plot shows SNP-heritability and PRS explained variance (r^2) for (A) cardiometabolic trait cluster, (B) autoimmune trait cluster, and (C) kidney-related trait cluster. Gray bars indicate SNP-heritability, and the colored bar chart presents the r^2 values, indicating the proportion of variance explained by the PRS from single-trait PRS (LDpred2, red bar) or multi-trait PRS (PRSmix+, blue bar), while the dot and whisker plot showcases predictive accuracy using AUC. The area under the receiver operating characteristic curve (AUC) presented with 95% confidence interval for dichotomized traits.

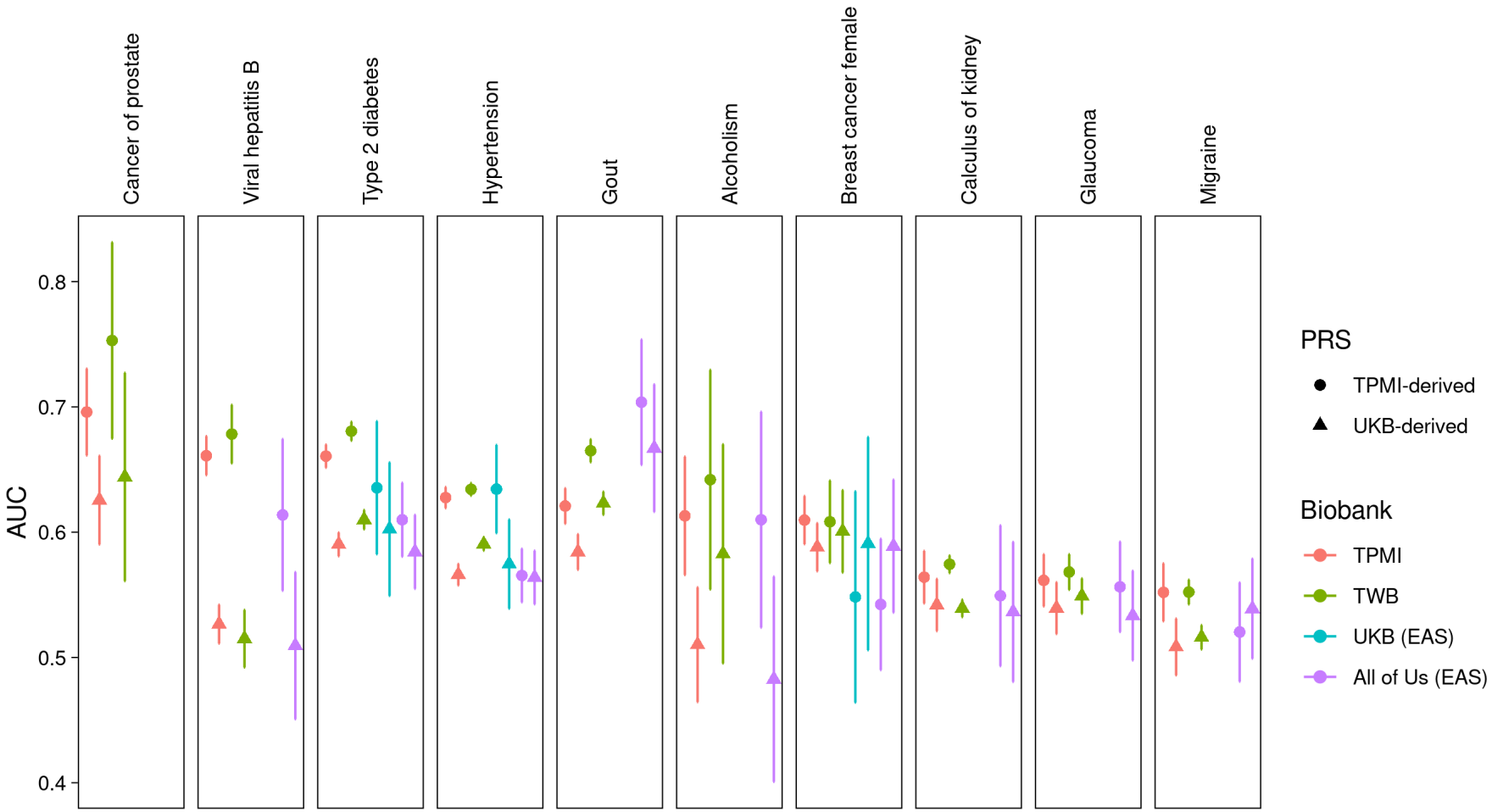
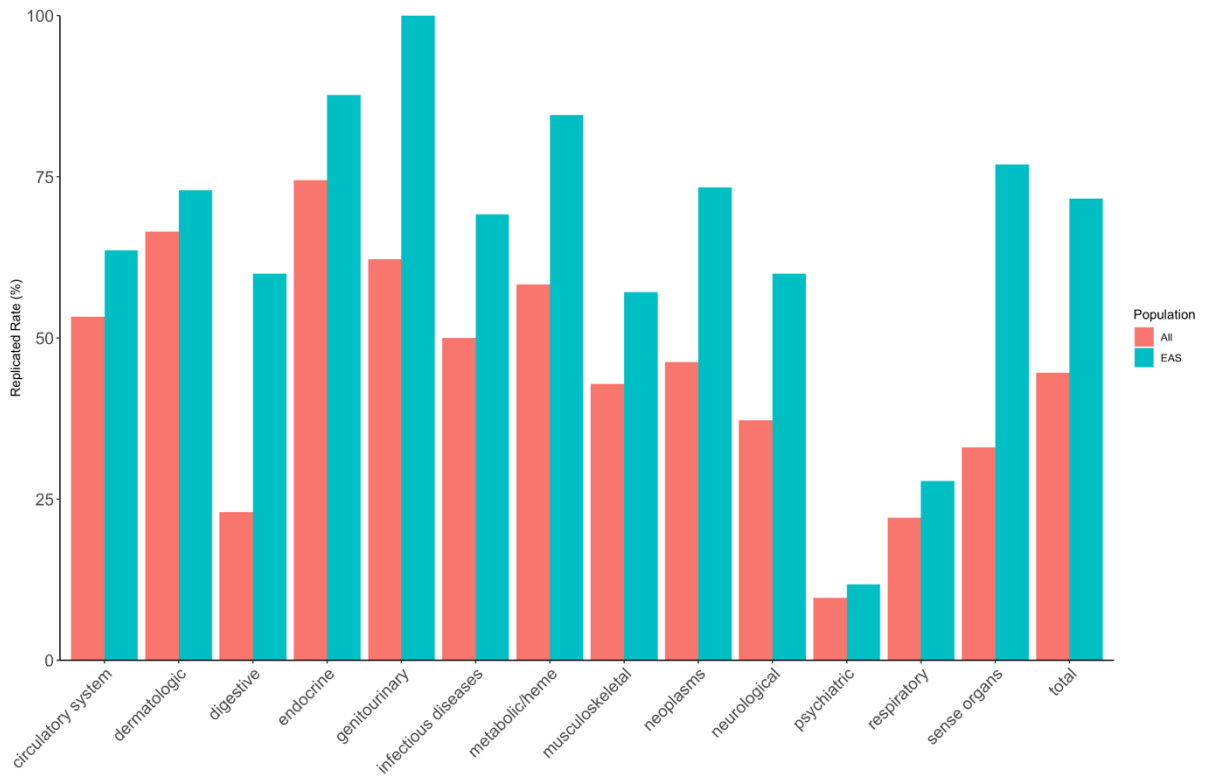
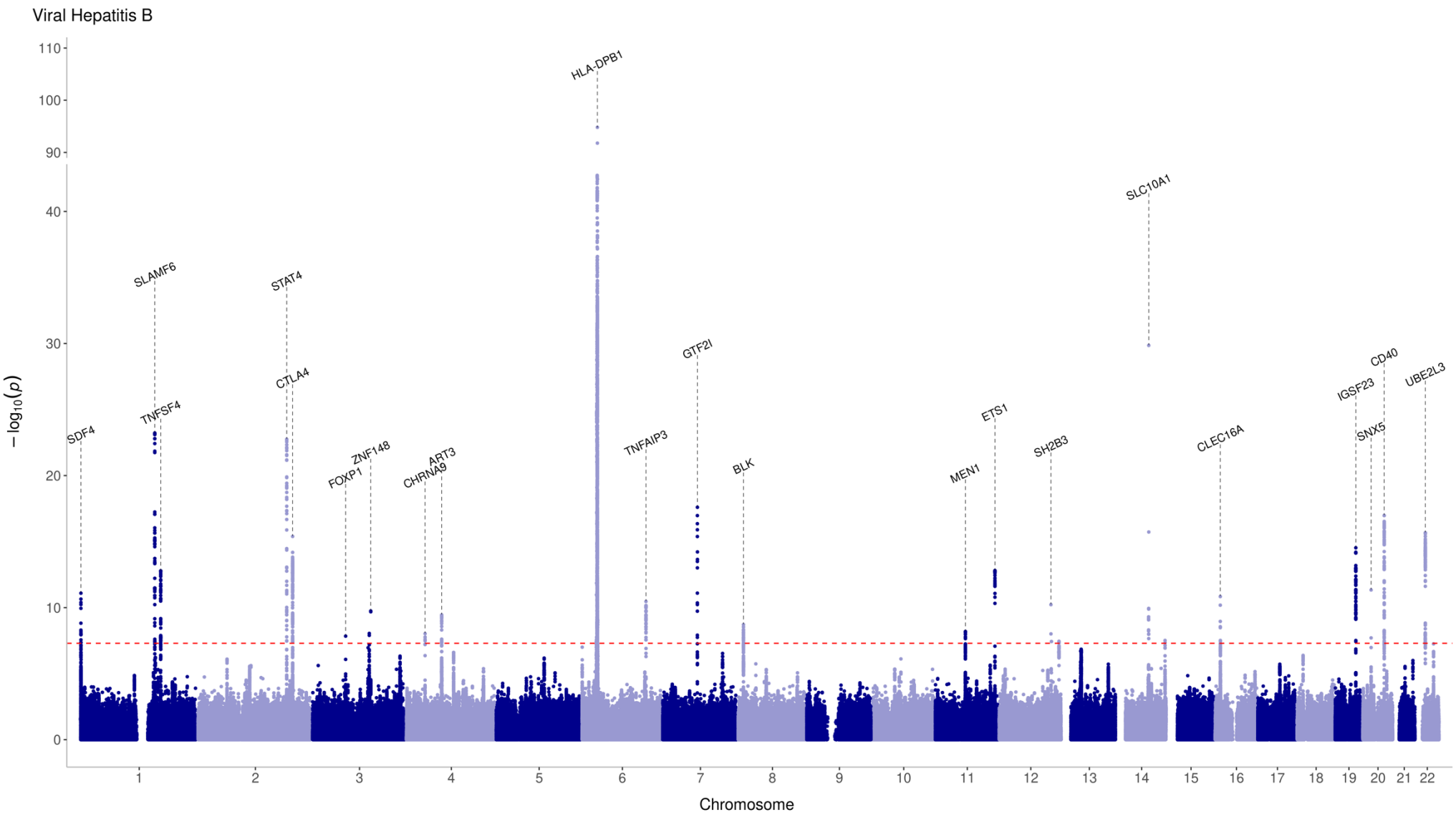


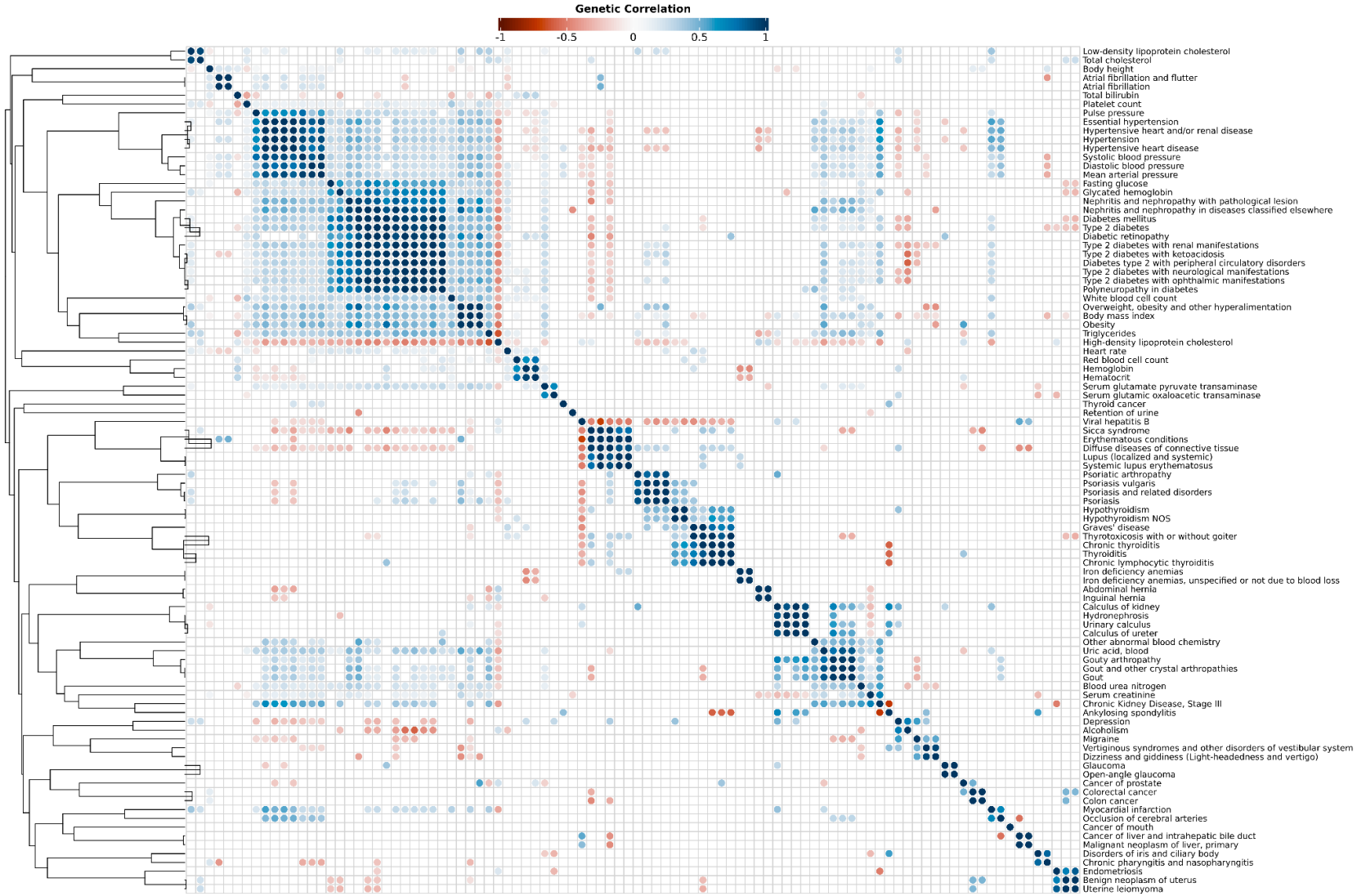
Fig. 6. External validation of PRS Models in Taiwan Biobank and other cohorts. PRS performance is presented as area under the receiver operating characteristic curve (AUC) ± 95% CI in TPMI (orange), Taiwan Biobank (green), East Asians in UK Biobank (blue), and East Asians in All of Us (purple). Circles represent TPMI-derived PRS, and triangles indicate UKB (European)-derived PRS models. Only the estimates with case size > 40 were showed on the figure.



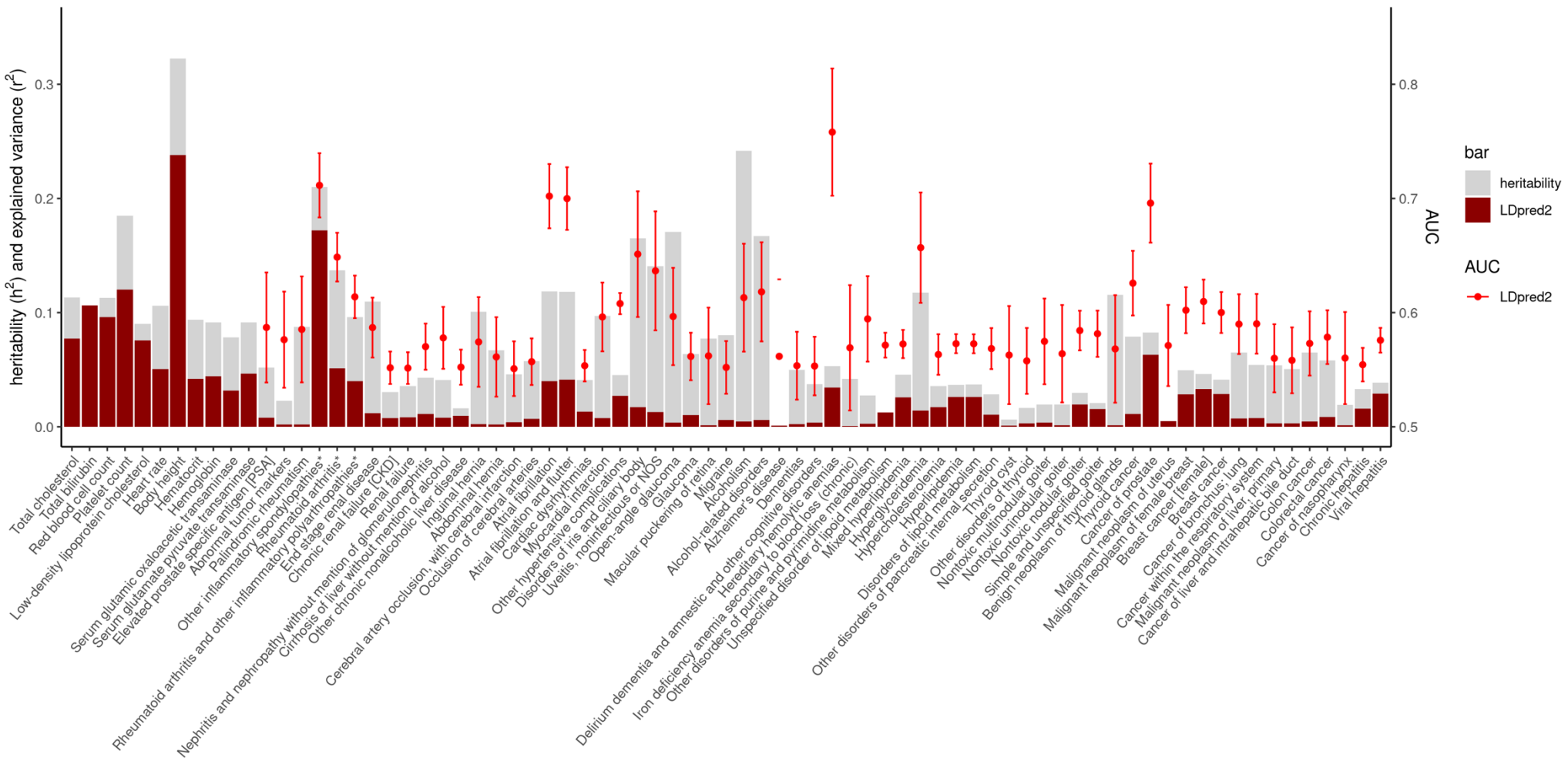
Extended Data Fig. 1. Comparison of TPMI GWAS-identified loci to the previously published GWAS. The replication rates of TPMI GWAS-identified loci when compared to previously reported loci from the GWAS catalog are presented in this bar chart. Red bars indicate the comparison of TPMI findings to that of all ancestries and blue bars represent the comparison to East Asian ancestries. The categories of diseases are shown under the bars.



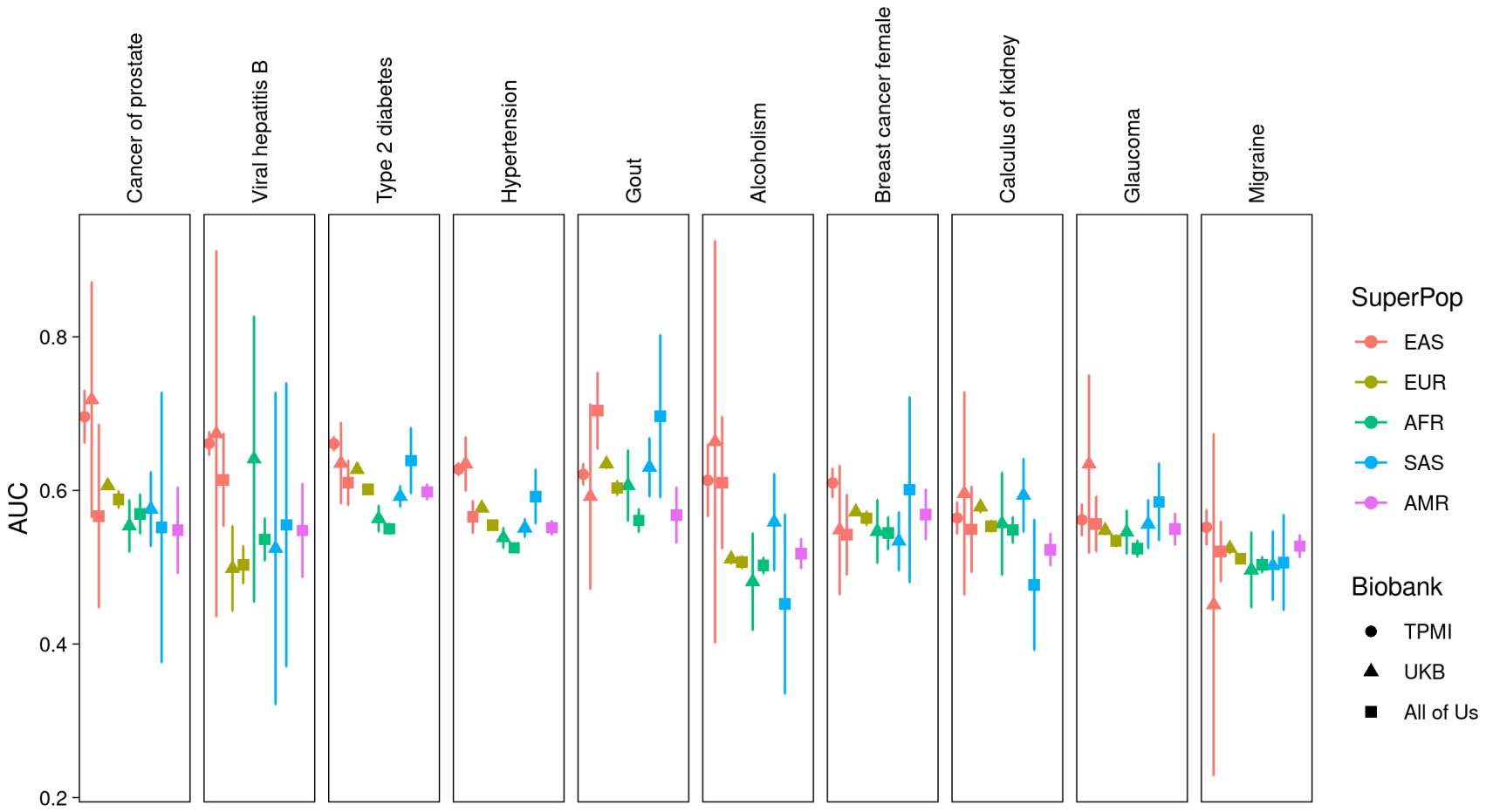
Extended Data Fig. 2. The Manhanttan plot of GWAS for viral hepatitis B in TPMI. The names of nearest mapped gene were labed for the independent GWAS significant loci.



Extended Data Fig. 3. Genetic correlation heatmap for all heritable traits. Heatmap showing genetic correlations among heritable traits. Genetic correlations were estimated using LDSC, with colors representing the correlation coefficients between traits. The clustering three is based on the weighted pair group method with arithmetic mean (WPGMA) with the correlation coefficient as distance between traits.



Extended Data Fig. 4. The bar chart and dot plot for PRS performance. Bar and dot plot showing PRS explained variance (r²) and SNP-heritability for dichotomous traits. Gray bars represent SNP-heritability, and dots show AUC. An asterisk (*) indicates estimates considering the MHC region.



Extended Data Fig. 5. External validation of PRS models across populations. PRS validation (AUC ± 95% CI) in East Asian (red), European (olive green), African (green), South Asian (blue), and All of Us (purple) populations from TPMI (circle), UKB (triangle), and All of Us (square) cohorts.

Table 1

Table 1. Proportion of disease burden explained by genetic risk

		top 5 % vs. bottom 5 %				top 10 % vs. bottom 10 %				top 20% vs. bottom 20%			
		Raw model		Adjusted model[1]		Raw model		Adjusted model[1]		Raw model		Adjusted model[1]	
Index	Genetic risk	R^2	p-value	R^2	p-value	R^2	p-value	R^2	p-value	R^2	p-value	R^2	p-value
Clinical visit	Cardiometabolic traits	0.67%	8.5×10^{-4}	1.14%	4.4×10^{-2}	0.38%	1.9×10^{-4}	0.77%	1.3×10^{-3}	0.27%	1.3×10^{-4}	0.56%	1.7×10^{-4}
	Autoimmune and infectious diseases	1.10%	1.8×10^{-4}	0.93%	6.2×10^{-2}	0.65%	2.0×10^{-5}	0.58%	4.0×10^{-3}	0.37%	4.0×10^{-5}	0.32%	2.4×10^{-2}
	Kidney-related traits	0.88%	1.3×10^{-3}	0.87%	7.8×10^{-2}	0.57%	2.0×10^{-4}	0.37%	9.4×10^{-1}	0.26%	9.4×10^{-5}	0.19%	2.0×10^{-2}
	All predictable traits (128 traits)	7.74%	1.5×10^{-27}	7.99%	5.7×10^{-13}	3.84%	4.0×10^{-27}	3.84%	2.0×10^{-25}	1.79%	2.0×10^{-25}	1.95%	8.0×10^{-14}
Hospitalization	Cardiometabolic traits	2.40%	1.5×10^{-8}	3.22%	2.4×10^{-8}	1.11%	6.5×10^{-8}	1.48%	4.1×10^{-9}	0.65%	4.1×10^{-15}	0.86%	6.1×10^{-9}
	Autoimmune and infectious diseases	1.09%	6.8×10^{-4}	1.42%	3.1×10^{-3}	0.21%	2.6×10^{-3}	0.29%	2.1×10^{-3}	0.15%	2.1×10^{-11}	0.19%	2.0×10^{-2}
	Kidney-related traits	0.39%	5.2×10^{-3}	0.51%	2.2×10^{-2}	0.43%	3.4×10^{-3}	0.58%	2.8×10^{-3}	0.21%	2.8×10^{-11}	0.28%	1.1×10^{-2}
	All predictable traits (128 traits)	6.75%	1.1×10^{-25}	9.09%	1.4×10^{-23}	3.80%	4.9×10^{-27}	5.08%	6.8×10^{-32}	2.36%	6.8×10^{-32}	3.12%	2.2×10^{-31}

[1]Model adjusting for sex, age and enrollment hospital

Supplementary Information

Supplemental tables

Table S1-S15

Supplemental text

Detailed methods about genotyping and phenotyping quality control