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], Yuan-Tsong Chen[1], Pui-Yan Kwok[1,134], Wayne Huey-Herng Sheu[135,71,136,†], Shun-Fa Yang[137,138,†], Jyh-Ming Liou[139,90,140,†], Jaw-Yuan Wang[141,58,†], Jeng-Fong Chiou[142,143,†], 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 Core Laboratory of Neoantigen Analysis for Personalized Cancer Vaccine, Office of R&D, 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 Cardiovascular Research Institute, Institute for Human Genetics, and Department of Dermatology, University of California, San Francisco, USA
135 National Health Research Institutes, Miaoli, Taiwan
136 Division of Endocrinology and Metabolism, Department of Internal Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

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

Correspondence and requests for information should be addressed to Hung-Hsin Chen (hunghsinchen@ibms.sinica.edu.tw) or Cathy S.-J. Fann (csjfann@ibms.sinica.edu.tw)

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 a large dataset of genetic profiles and electronic medical record data for the Han Chinese. Using extensive phenotypic data, we conducted comprehensive genomic analyses of Han Chinese across the medical phenome. These analyses identified population-specific genetic risk variants and novel findings for various 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 dataset, Taiwan Biobank, and among East Asians in the UK Biobank and the All of Us Project. The identified genetic risks accounted for up to 10.3% of the overall health variation in the TPMI cohort. 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.

358 **Introduction**

359 A major promise of modern genetics is the ability to predict complex disease
360 risk based on a person's genetic profile. If successful, health management
361 strategies can be developed to mitigate the risk (disease prevention) and to
362 optimize care (early diagnosis and effective treatment). Large-scale studies by
363 the UK Biobank (UKB) and Electronic Medical Records and Genomics
364 (eMERGE) Network show that risk prediction based on genetics holds promise
365 and several countries are exploring ways to implement risk-based management
366 in clinical practice[1,2]. Using polygenic risk scores (PRS) to predict disease risk
367 and identify individuals at high risk is an emerging "precision medicine"
368 approach to leverage genetic findings in clinical practice. However, a significant
369 limitation is that current PRS models are predominantly based on genome-wide
370 association studies (GWAS) with participants of European ancestries (EUR)[4,5],
371 often leading to reduced predictive performance in groups of other anestry[6,7].
372 To fully realize the potential of precision medicine for diverse global populations,
373 population specific phenome-wide genomic discovery must be perfomed at
374 scale and clinically applicable polygenic risk models must be optimized within
375 and across populations. To fill this major research gap in an population of East
376 Asian ancestry (EAS), we characterized the complex genetic architecture of the
377 population of Han Chinese ancestry phenome-wide, developed population-
378 specific PRS, and assessed the external validity of the models across
379 populations with varying degrees of genetic similarity.

380 Populations of EAS ancestry represent nearly a quarter of the global population,
381 but they account for only 3.95% of the participants in previous GWAS[3].
382 Although several biobanks have been built to recruit subjects from East Asia,
383 they have moderate sample size (72K-212K) and many focus on specific
384 conditions[8-12]. In contrast, biobanks with predominantly European ancestry
385 participants[13-16], have significantly larger sample sizes (224K-635K) and
386 access to more comprehensive clinical data. The moderate sample size and
387 limited phenotypes in existing EAS biobanks hamper discovery of unique
388 genetic effects and preclude the development of robust and clinically useful
389 PRS models for EAS.

390 We assembled a large non-EUR cohort, the Taiwan Precision Medicine

391 Initiative (TPMI), and genotyped over half a million participants across sixteen
392 medical centers in Taiwan from 2019 to 2023. All the participants, who are
393 overwhelmingly of Han Chinese ancestry, contributed DNA samples for genetic
394 profiling with a custom-designed genotyping array and consented to provide
395 their longitudinal electronic medical records (EMR) from 5 years prior to
396 enrollment and into the future. The EMR dataset includes rich and accurate
397 health-related phenotypes, including medical diagnoses and biochemical
398 examinations[17]. Here, we present the results of comprehensive genomic
399 analyses with extensive genetic and medical data derived from the TPMI cohort,
400 including phenome-wide GWAS and PRS model development. We identified
401 numerous population-specific risk variants/genes, observed evidence of
402 genetic pleiotropy, and pinpointed clusters of traits that shared similar genetic
403 etiology. Then, we developed and validated PRS prediction models for
404 numerous conditions against external datasets including those from the Taiwan
405 Biobank (TWB), the UKB, and the All of Us Project. Our results reveal the
406 benefits of leveraging a large cohort from understudied population to identify
407 unique genetic underpinnings of the human phenome, interpret causal effects
408 via fine mapping and colocalization, and improve the performance of
409 population-specific PRS models, which together, better illuminate the clinical
410 implications of genetic risk.
411
412 **Results**
413 **Diseases and Quantitative Traits in TPMI**
414 We performed comprehensive genomic analyses, including GWAS, heritability
415 estimation, and PRS model building and evaluation, across a wide range of
416 diseases and quantitative traits using 463,447 genetically inferred Han Chinese
417 from TPMI. We examined 695 dichotomized phenotypes (phecode, case n
418 >2,000) and 24 quantitative traits (sample size >100,000), spanning numerous
419 disease categories (defined by phecode groupings[18,19]), such as neoplasms,
420 metabolic disorders, circulatory conditions, autoimmune diseases, and more
421 (Fig. 1). The phecodes, derived from International Classification of Diseases
422 (ICD) codes[18,19], alongside quantitative traits such as blood pressure, BMI, liver
423 enzymes, and lipid levels, provide a robust dataset for exploring genetic

424 contributions to human health (Table S1 and S2). The log-transformed case
425 proportion identified from EMR showed a moderate but significant correlation
426 with the log-transformed 5-year disease prevalence from National Health
427 Insurance Research Database (NHIRD) in Taiwan[20] (r = 0.656, p-value =
428 2.69×10^{-84}) (Fig. 1A and Extended Data Fig. 1), suggesting that the TPMI's
429 hospital-based design may not fully capture mild and common illness, which
430 are primarily observed in local and primary care clinics. Fig. 1B displays the
431 sample sizes for 24 quantitative traits in the TPMI and highlights sample size
432 variation across traits, a key measure impacting the power and precision of
433 association analyses within the cohort.

434 **GWAS, Fine-Mapping, and Novel Results**

435 Our GWAS identified at least one significant locus (p-value < 5×10^{-8}) for 265
436 phecodes of the 695 tested and all 24 quantitative traits. Highlighting the
437 robustness of the TPMI data, we observed a high replication rate of reported
438 disease loci from EAS GWAS on GWAS catalog (actual/expected ratio (AER)
439 = 78.17%, considering the statistical power with the published tool, PGRM[21]),
440 particularly for endocrine and metabolic/hematopoietic diseases (AER =
441 88.68% and 84.62%, respectively; Extended Data Fig. 2 and Table S3). Lower
442 replication rates for respiratory (AER = 23.53%) may reflect limited case
443 numbers, untyped genetic variants, such as rare variants, copy number
444 variation, and structural variants, or recruitment bias such as age distribution.
445 We applied the sum of single effects model for fine-mapping to identify the
446 independent variant-trait associations, and reported the genetic variant with
447 highest posterior inclusion probability of identified credible sets as well as the
448 single lead variant for the failed fine-mapping regions and major
449 histocompatibility complex region (MHC region, chr6:25M-33M). Our analyses
450 revealed a total number of 2,656 fine-mapping identified independent
451 association signals, including 1,309 from phecodes GWAS and 1,347 from
452 quantitative traits. Notably, 95 novel associations, defined as having no
453 previously reported results within 1Mb in the NHGRI-EBI GWAS Catalog[21] of
454 relevant GWAS, were identified across 50 phecodes and 7 quantitative traits.
455 In addition, we identified 217 new hits from previously reported regions, defined
456 as having r^2<0.1 with any variant observed in the NHGRI-EBI GWAS Catalog

457 within 1 MB for the same phenotype (Table S4 and S5). After applying multiple
458 test correction, 1,502 fine-mapped associations passed Bonferroni-adjusted
459 threshold ($5x10^{-8}/(695+24) = 6.95x10^{-11}$), as well as 21 novel variants and 115
460 new hits.
461 Of the 95 novel genetic associations, 30 variants are rare (minor allele
462 frequency [MAF]<0.05) in populations with other ancestry (with African ancestry,
463 Admixed American ancestry, South Asian ancestry, and European ancestry
464 (non-Finnish) in gnomAD), and 33 variants less than 0.01 in the EUR, the most
465 extensively studied population, which explains why they were not reported in
466 previous GWAS. For example, the SNP rs17089782, a missense variant in the
467 *PIBF1* (p.R405Q) on chromosome 13 is significantly associated with thyroid
468 cancer ($p = 2.8x10^{-9}$) in the TPMI cohort. This SNP has a MAF of 5.65% in
469 TPMI but 0.01% in EUR, which may explain why this association was only
470 detectable in TPMI. However, *PIBF1* is essential for immune regulation,
471 especially during pregnancy, and is relevant to autoimmune diseases and
472 cancer[22]. Another novel variant identified in our quantitative analysis of BMI
473 (rs761018157 p-value = $4.8x10^{-9}$, MAF in TPMI = 4.34%, MAF in EUR<0.01%,)
474 maps to *PHOX2B*. This gene, highly expressed in the nervous system, had
475 previously been linked to obesity hypoventilation syndrome in a small study (n
476 = 30)[23] and associated with bone mineral density[24]. In addition, when we
477 compare the effect size in TPMI and UKB for the rest of the novel findings, 25
478 exhibit a significant different effect size (p-value<0.05). For instance, a TPMI-
479 identified platelet count-associated variant, rs12955741, located in the
480 intergenic region between *TGIF1* and *DLGAP1* exhibits a different effect size
481 compared to that in UKB ($\beta_{TPMI} = 0.044$, $\beta_{UKB} = -0.005$, p-value = $1.7x10^{-9}$).
482 Moreover, the high hepatitis B virus (HBV) carrier rate in Taiwan[25] contrasts
483 sharply with its rarity in European cohorts, enabling TPMI to identify novel loci
484 associated with viral hepatitis B (case number in TPMI = 23,618 vs. UKB = 132).
485 Among the 26 independent loci identified in our analysis of hepatitis B, 19 fine-
486 mapped loci are novel (Extended Data Fig. 3). Notably, 18 of these 19 loci were
487 found to be associated with liver function or diseases (Table S5). These novel
488 associations highlight the uniqueness of certain disease loci in the TPMI cohort,
489 presenting opportunities for developing population-specific therapeutic

interventions and advancing precision medicine.

All identified independent associations are summarized in Fig. 2. The identification of the 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 (*INS-KCNQ1* region) also affect various traits, including metabolic, endocrine, and genitourinary diseases. These hotspots of trait-relevant variants implied the shared genetic mechanism among diseases and potential of pleiotropic effects.

Heritability and Colocalization

Linkage disequilibrium score regression analysis (LDSC)[26] showed strong liability-scaled SNP-heritability for conditions such as alcoholism ($h^2 = 0.213$), retention of urine ($h^2 = 0.163$), and open-angle glaucoma ($h^2 = 0.160$). Among 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 and Supplmentary Information), 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 at the gene level and identified 329 unique genes contributing significantly to phenotypic variation (h^2 >0.1% and z-score >1.64). Of these, 45 affected more than one phecode and/or quantitative 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 with three QTL datasets, including GTEx[27], MAGE[28], and JCTF[29] (Fig. 3 and Table S8). Our results identified 391 unique genes that potentially mediate the outcome through their expression level (posterior probability > 0.9), including *GBAP1* which colocalized with five different traits (uric acid, serum creatinine, hematocrit, hypertension, and gout). Among the colocalized genes, 75 of them can be identified only in the multi-ancestry lymphoblastoid cell lines eQTL, MAGE (20 genes), and/or Japanese whole blood eQTL, JCTF (59 genes). Our findings demonstrate the effect of these genes (such as *APOE, ABCG2*, and

523 *KCNQ1*) on multiple traits and disorders. By elucidating shared genetic effects,
524 these results offer opportunities to develop precision medicine approaches that
525 address comorbidities, such as treating hyperlipidemia and reducing dementia
526 risk through a single intervention targeting *APOE*, which influences both lipid
527 metabolism and Alzheimer's disease risk. This gene-level understanding
528 emphasizes the potential to optimize therapeutic strategies by leveraging
529 genetic pleiotropy in disease management.

530 **Genetic Correlation and Clusters**

531 Pairwise genetic correlation analyses revealed three major phenotype clusters:
532 cardiometabolic traits, autoimmune and infectious diseases, and kidney-related
533 traits (Fig. 4 and Extended Data Fig. 4). The cardiometabolic cluster, which
534 includes type 2 diabetes, hypertension, and BMI, reinforces the interconnected
535 phenotypic and genetic architectures of cardiovascular and metabolic diseases.
536 The cluster of autoimmune and infectious diseases, which includes viral
537 hepatitis B, psoriasis, and systemic lupus erythematosus, illuminates shared
538 immune system pathways and potential gene-pathogen interaction. The
539 kidney-related cluster involved gout, chronic kidney disease, calculus of kidney
540 and ureter, ankylosing spondylitis, and measures of urea nitrogen, creatinine,
541 and uric acid. The shared genetic architecture provides opportunities to
542 leverage the genetic risk of correlated traits while developing the PRS model.

543 **Cross-Population Comparison**

544 Cross-population comparisons[30] with EUR GWAS from UKB showed varying
545 degrees of genetic correlation, with strong, statistically significant correlations
546 for traits like cholelithiasis ($\rho_{ge} > 0.999$), type 2 diabetes ($\rho_{ge} = 0.829$), and
547 ischemic heart disease ($\rho_{ge} = 0.756$), but moderate correlations for gout ($\rho_{ge} =$
548 0.616) and psoriasis ($\rho_{ge} = 0.418$) (Table S6). The moderate correlations
549 suggest the differentiated genetic mechanism and disease distribution across
550 populations (gout case n = 24,411 in TPMI and 3,179 in UKB; psoriasis case n
551 = 4,166 in TPMI and 2,197 in UKB). Therefore, these findings demonstrate the
552 importance of population-specific genetic studies, as differences in genetic
553 architectures between populations can significantly affect the accuracy of PRS
554 models.

555 **Polygenic Risk Score (PRS) Development**

556 Building on these insights, we developed and validated PRS models that
557 demonstrated strong predictive performance for a wide range of diseases.
558 Although we used five PRS tools, including LDpred2[31], Lassosum2[32], PRS-
559 CS[33], SBayesR[34], and MegaPRS[35] (Tables S9-S13), we found that LDpred2
560 outperformed the others for most traits (Extended Fig. 5). Therefore, we took
561 the results of LDpred2 for further comparisons. Out of the 265 PRS models for
562 phecodes, AUC values exceeded 0.55 for 105 dichotomized phecodes with a
563 significant p-value (p-value<0.05). Additionally, the explained variance
564 of models for 24 quantitative traits ranged from 0.028 (aspartate
565 aminotransferase, AST) to 0.227 (height). (Table S9 and Extended Data Fig. 6
566). The most predictive PRS models included highly heritable traits such as
567 ankylosing spondylitis (AUC = 0.812±0.016), psoriasis (0.709±0.016), atrial
568 fibrillation (0.702±0.014), prostate cancer (0.696±0.018), systemic lupus
569 erythematosus (0.696±0.015), rheumatoid arthritis (0.646±0.011), type 2
570 diabetes (0.640±0.005), female breast cancer (0.611±0.010), and hypertension
571 (0.610±0.004). Interestingly, the PRS for hepatitis B also demonstrated high
572 genetic predictability (0.654±0.008). Since SNP-heritability (h^2) represents the
573 upper bound of variance that can be explained by PRS, we examined the
574 proportion of heritability captured by our models (r^2/h^2). A total of 36 traits,
575 including prostate cancer (r^2/h^2 = 0.054/0.070), type 2 diabetes (0.066/0.126),
576 and HDL cholesterol (0.136/0.191), reached over 50% of their SNP-heritability,
577 suggesting that PRS can achieve near-optimal predictive accuracy for highly
578 heritable traits. However, for complex diseases influenced by both genetic and
579 environmental factors, PRS performance is inherently constrained by the
580 fraction of heritability attributable to common variants. These findings reinforce
581 the importance of SNP-heritability as a reference point for evaluating PRS utility
582 and highlight the need for larger, ancestrally diverse datasets to further
583 enhance genetic prediction models (Extended Data Fig. 6).
584 Leveraging the identified clusters, we performed a multi-trait PRS training,
585 PRSmix+[36], for the traits within each cluster (Fig 5 and Table S14). Notably,
586 multi-trait PRS models improved prediction accuracy for the cardiometabolic
587 disease cluster with a 0.040 increase in AUC (from 0.608 to 0.648) and a 1.770-
588 fold improvement in phenotypic variance explained (r^2). The performances of

589 autoimmune and kidney-related disease clusters were also enhanced, with
590 averaged AUC improvements of 0.018 and 0.009, respectively (from 0.641 to
591 0.659, and 0.601 to 0.610 respectively), and 1.351- and 1.349-fold
592 improvements in variance explained (r^2). The significant enhancement of multi-
593 trait PRS prediction (comparing r^2 of LDpred2 and PRSmix+ with paired t-test,
594 $p = 1.07 \times 10^{-13}$) highlights the potential of leveraging shared genetic architecture
595 to enhance disease risk prediction. Fig. 5 demonstrates the performance of
596 single and multi-trait PRS across three disease clusters, as well as the differing
597 effectiveness of PRS in predicting genetic risk across various disease
598 categories.

599 **PRS External Validation and Comparison**

600 To evaluate the robustness and generalizability of our PRS models, we
601 performed an external validation of the models (hypertension, type 2 diabetes,
602 viral hepatitis B, gout, calculus of kidney from PRSmix+ and others from
603 LDpred2) in Taiwan Biobank (TWB, genetically inferred unrelated Han Chinese,
604 n = 88,628), UKB (self-reported EAS, n = 9,893), and All of Us (genetically
605 inferred EAS, n = 6,895). We found that the prediction accuracy, AUC, of our
606 models ranged from 0.548 (glaucoma) to 0.712 (prostate cancer) in TWB, 0.557
607 (female breast cancer) to 0.634 (hypertension) in UKB, and 0.520 (migraine) to
608 0.709 (gout) in All of Us (Extended Data Fig. 7). Although the TWB
609 questionnaire did not contain specific details on hepatitis B status, we used anti-
610 hepatitis B core total antibodies (Anti-HBc) as an indicator of infection or past
611 infection and hepatitis B surface-antigen (HBsAg) as a marker of acute/chronic
612 infection. Intriguingly, the AUC for the TPMI-derived model of hepatitis B were
613 0.674±0.003 for HBsAg and 0.530±0.002 for Anti-HBc in the TWB. These
614 results demonstrate the high predictive value of the PRS for hepatitis B for
615 predicting symptoms and severity of the disease.

616 TPMI-derived PRS models perform better than the UKB European-derived
617 models when applied to EAS for viral hepatitis B, type 2 diabetes, hypertension,
618 gout and migraine. (Extended Data Fig. 7) For the other traits, TPMI derived
619 models consistently outperform the UKB ones although the confidence intervals
620 overlap. However, the overlapping confidence intervals in UKB and All of Us
621 may be due to their limited sample size of EAS. These results suggest that

622 population-specific PRS models allow for more accurate risk stratification and
623 enable personalized healthcare interventions for EAS. Additionally, we
624 assessed the performance of TPMI-derived PRS and TPMI-included cross-
625 population PRS across various ancestry groups, including European, African,
626 Admixed American, and South Asian populations from the UK Biobank and All
627 of Us cohorts (Extended Data Fig. 8). The performance varied by diseases, but
628 consistent results were observed for female breast cancer and glaucoma
629 across populations, and TPMI-included cross-population PRS slightly but not
630 significantly improved in the populations other than EAS and EUR.

631 **Genetic Risks on Overall Health Measures**

632 Although overall health is hard to define with a few metrics, herein, we used the
633 count of clinical visits and duration of hospitalization to roughly describe
634 individuals' overall health. We found that 131 of the top performing PRS models
635 (LDpred2 models with AUC > 0.55 and all PRSmix+ models for phecodes and
636 all models for quantitative traits) are significantly associated with overall health
637 indices, explaining 8.47% of the variation in clinical visit frequency (p-value =
638 2.69×10^{-14}) and 10.29% of the variation in hospitalization duration (p-value =
639 5.62×10^{-27}, Extended Data Table 1 and S15) in the comparison between top
640 and bottom 5% groups after adjusting for sex, age and recruiting hospital.
641 Among the identified clusters, the cardiometabolic disease cluster contributed
642 the most to the indices, accounting for 1.32% of clinical visits (p-value = 0.02)
643 and 3.55% of hospitalizations (p-value = 7.10×10^{-9}). This may reflect the high
644 prevalence of cardiometabolic diseases in the hospital-based TPMI cohort. In
645 short, quantification of the impact of PRS for various diseases and traits on
646 human health opens up opportunities for developing precision health
647 management strategies.

648 **Discussion**

649 This study represents a lagre-scale GWAS in the population of Han Chinese
650 ancestry, utilizing data of around 500,000 individuals recruited from 16 medical
651 centers across Taiwan. We investigated the genetic architecture of 695
652 dichotomized phecodes and 24 quantitative traits, identifying 2,656
653 independent variant-trait associations and showed that population-specific
654 genetic risk-prediction PRS models for a wide range of diseases performed well

655 in the population. Indeed, for the traits with sufficient sample size in the cohort,
656 the PRS performance rival those developed for EUR using UKB data. These
657 findings show that population-specific PRS models can be developed
658 successfully for populations of non-European ancestry and our project serves
659 as a model for large-scale genetic studies in other populations.
660 Recent large-scale projects that emphasize ancestral diversity in human
661 genetic studies have discovered new findings with the inclusion of subjects of
662 non-European ancestry. MVP conducted multi-ancestry GWAS on 635,000
663 participants, identifying over 2,000 signals unique to populations with non-
664 European ancestry[16]. With the TPMI dataset, we performed larger GWAS in
665 subjects of Han Chinese ancestry for several traits than published studies. For
666 instance, the previous largest meta-analysis for type 2 diabetes included 20,573
667 cases who were of Han Chinese ancestry[37]. In contrast, our GWAS included
668 59,289 cases of type 2 diabetes, almost tripling the number of cases ever tested,
669 and identified five unreported T2D SNPs from known regions, demonstrating
670 the power of TPMI sample size. Identification of new and population-specific
671 risk variants may lead to further understanding of their molecular mechanism
672 and underline the need for population-specific weightings in PRS models.
673 Moreover, population-specific findings also explain better performance of
674 population-specific PRS model in the population in question. In short, our
675 population-specific genomic profiles for comprehensive phenotypes provide a
676 solid foundation for PRS development.
677 Our understanding of how the genetic factors influencing hepatitis B, an
678 endemic infectious disease in Taiwan with an estimated hepatitis B virus carrier
679 rate of 9.78% among the unvaccinated cohort (born before 1984)[25], also
680 benefited from the large dataset. With 23,618 cases, a significant increase from
681 prior studies of only a few thousand cases[38-40], we identified 26 fine-mapped
682 signals, including 19 novel loci, and revealed a significant negative correlation
683 between hepatitis B and other autoimmune diseases, such as Sicca syndrome,
684 psoriasis, and systemic lupus erythematosus. Our well-performed and
685 validated PRS model for hepatitis B demonstrated that host genome may
686 determine the severity and symptoms of this infectious disease. This is similar
687 to that previously reported in COVID-19 and pneumonia, where genetic factors

688 have been shown to influence disease outcomes[41-44]. Our unexpected success
689 of GWAS and PRS for hepatitis B not only demonstrate the power of the large
690 sample size of TPMI, but also reveal the necessity of population-specific
691 genetic study for population-enriched diseases. The benefits extend beyond
692 differences in ancestry to include environmental factors such as pathogen
693 exposure, food intake, and lifestyle influences. Exploring how the human
694 genome interacts with these diverse external and environmental factors can
695 significantly enhance our understanding of how genetic variants contribute to
696 disease susceptibility or severity.
697 In addition, the comprehensive phenotypic data allow us to investigate the
698 genetic correlation among multiple traits, which have substantial implications
699 for clinical applications, and leverage them to improve the performance of PRS
700 models. By identifying shared genetic risks across diseases, at-risk individuals
701 can be alerted to pursue early detection of comorbidities and targeted
702 prevention strategies. For example, the clustering of cardiometabolic traits,
703 such as type 2 diabetes, hypertension, and BMI, highlights their interconnected
704 genetic basis and suggests that individuals with a high genetic risk for one
705 condition may benefit from early screening and intervention for related
706 conditions[45]. Additionally, the shared genetic architecture allows the
707 development of multi-trait PRS models that integrate genetic risks across
708 correlated traits, improving prediction accuracy and enabling precision
709 medicine approaches that address multiple health outcomes simultaneously[46].
710 Including the correlated traits in PRS model development improved
711 performance, resulting in an average 1.55-fold increase in the explained
712 percentage of phenotypic variation (p-value = 1.07×10^{-13}). Although previous
713 studies have proven the utility of multi-traits on target diseases[36,47,48], we have
714 extended the use this approach on a phenome-wide level and demonstrated
715 the improvement across different types of traits. As a result, we produced well-
716 performed PRS for various categories of diseases, including cardiometabolic
717 diseases, autoimmune disorders, and infectious diseases.
718 We evaluated our PRS models across several large cohorts, including the TWB,
719 UK Biobank, and All of Us. The TPMI-derived PRS models consistently
720 outperformed those developed from European populations when applied to

721 diseases in Han Chinese descents or EAS from the three large cohorts. When
722 comparing with EUR-derived PRS models, we also observed better
723 performance across several traits in EAS, particularly for cardiometabolic and
724 autoimmune diseases. Similarly, the TPMI-included cross-population model
725 slightly improves performance in populations of other ancestries. These results
726 highlight the need for population-specific models and emphasize the
727 importance of genetic data from diverse populations to advance cross-
728 population models. By integrating these well-developed PRS models, we
729 estimate that genetics account for 10.3% of variation of hospitalization duration
730 in TPMI. Although the estimates of genetic contributions to health measure may
731 be influenced by disease prevalences and ascertainment biases, our result
732 suggest integrating genetic risk-based health management strategies with
733 traditional risk factors, such as age, sex, smoking, and BMI, may enhance
734 prediction models and refine personalized risk stratification.
735 As with other large-scale epidemiological studies, ascertainment bias is also
736 observed in our study. TPMI's case proportion shows a significant but moderate
737 correlation with the prevalence from NHIRD, implying potential ascertainment
738 bias of TPMI's hospital-based design. Compared to the general population in
739 Taiwan (NHIRD), TPMI participants are overrepresented in the middle-aged
740 group (year of birth 1940-1970, 54.3% vs. 38.3%), include slightly more females
741 (55.1% vs. 50.6%), and have a higher proportion of participants from northern
742 Taiwan (59.5% vs. 47.3%). These demographic differences, along with the
743 volunteer-based recruitment process, likely contribute to the lower case
744 proportions observed in TPMI (Figure 1a). Importantly, a significant portion of
745 disease records in the NHIRD originate from local clinics and primary care
746 settings, which are not covered in TPMI. Notably, the EMR of the participants
747 are incomplete, as some participants receive care from multiple health
748 providers, but the TPMI only has access to EMRs from their enrollment
749 hospitals. We acknowledge that ascertainment biases may influence disease
750 prevalence and heritability estimates. Thus, we accounted for case-control
751 ascertainment by applying liability-scale transformations using population
752 prevalence data from NHIRD and utilized independent validation for PRS (TWB,
753 UKB, All of Us) to mitigate the impact of ascertainment biases, while

754 acknowledging that residual biases may persist. Methods like inverse
755 probability weighting could mitigate such biases[49], but these require detailed
756 external reference data that are currently unavailable. Additionally, we
757 observed a relatively low estimated heritability for body height and BMI in TPMI,
758 compared to values reported in the literature.[50,51] These estimates may be
759 affected by factors such as inconsistencies in assessment across EMR,
760 variations in statistical approaches, and reduced bias of assortative mating in
761 TPMI population.[52-54] These limitations also emphasise the need for future
762 adjustments to enhance generalizability.
763 In addition to the ascertainment bias, our study has other commonly found
764 limitations. First, the TPMI cohort size is not sufficiently large to study some of
765 the severe subtypes of many diseases, such as diabetes insipidus and
766 neurofibromatosis. Second, we attempted to use eQTLs to elucidate the
767 molecular mechanism of diseases, but the underrepresentation of EAS in
768 current eQTL datasets, such as GTEx, poses challenges[27]. Gene expression
769 regulation varies across ancestries[55,56], and differences in LD structures further
770 complicate colocalization analyses. Comparing to GTEx whole blood eQTL, the
771 multi-ancestry lymphoblastoid eQTL and Japanese whole blood eQTL revealed
772 309 additional gene-trait pairs. Therefore, ancestral diversity is an urgent need
773 not only in genomic data but also in transcriptomic, proteomic, metabolomic,
774 and epigenomic datasets. Third, the current project retrieved EMRs from an
775 average of 5 years prior to enrollment, so some important data such as age of
776 disease onset for the older participants are not available. Incomplete EMR
777 leads to less precise case definition of some participants. Fourth, some of the
778 younger participants have high-risk genetic profiles but are disease free for
779 those diseases. The duration of the project is too short to determine whether
780 they will eventually develop those diseases.
781 Effort is underway to gain access to the complete EMRs of the TPMI
782 participants and to recruit additional participants with severe subtypes of
783 common diseases. The high-risk participants who are symptom-free are being
784 followed to monitor disease development. Future studies are being planned to
785 study the high-risk individuals who escape disease development to identify
786 genetic and non-genetic factors that mitigate their disease risk. Furthermore,

787 the meta-analysis integrating TPMI with other large-scale EAS biobanks, such
788 as TWB[11,12], Korean Genome and Epidemiology Study[9], China Kadoorie
789 Biobank[10], and Biobank Japan[8], may further enhance our understanding of the
790 genetic etiology in EAS and improve prediction models.
791 This study demonstrates that population-specific risk-prediction models, such
792 as those developed for EAS in this work, can achieve strong predictive
793 performance for traits with high relevance in that population. The PRSs we
794 developed for EAS performed well for multiple traits, including diseases with
795 significant public health implications, such as type 2 diabetes and systemic
796 lupus erythematosus. However, for certain traits, such as female breast cancer
797 and glaucoma, PRS derived from both UK Biobank and TPMI performed
798 comparably, suggesting that the genetic architecture of some traits allows for
799 generalizable models. These findings emphasize the importance of developing
800 and validating PRS models in diverse global populations to maximize their utility
801 and equity in genetic risk prediction. While our results emphasize the utility of
802 developing population-specific PRS, further research is needed to directly
803 compare their performance with multi-population models and assess their
804 generalizability and to assess their impact on disease prevention and
805 management. In particular, longitudinal studies and real-world implementations
806 will be critical to determine the extent to which PRS-guided interventions can
807 delay disease onset or improve health outcomes. Furthermore, it is hopeful that
808 if all can obtain their genetic profiles and determine their risk for major diseases,
809 many diseases can be prevented or their onset can be delayed significantly,
810 thereby fulfilling the promise of modern genetics.
811 In conclusion, we used a large-scale Han Chinese dataset produced by the
812 TPMI to conduct pheno-wide genetic analyses and leverage these genetic
813 findings to train risk prediction models for multiple diseases and traits. The
814 developed models are validated in EAS of different biobanks and demonstrate
815 a consistent performance that bodes well for their use in the general Han
816 Chinese/EAS population. Our approach can serve as a template for developing
817 PRS models in populations currently without such resources, anticipating the
818 time when all populations around the world can benefit from risk-based health
819 management as part of the precision health movement.

Figure legend

Fig. 1. Scatter plot of the case proportion for dichotomized phenotypes and barchart of 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 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. The correlation coefficient and significance were calculated using the two-sided Pearson correlation test. (B) Scatter plot shows the sample sizes for 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. The gene-level heritability was estimated with h2gene[57], and the colocalization analysis was performed with coloc[58].

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 (h^2) 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 (estimated from TPMI GWAS unrelated set [n=248,754] with LDSC), and the colored bar chart presents the r^2 values, indicating the proportion of variance explained by the PRS among TPMI validation set (n=20,000) from single-trait PRS (LDpred2, red bar) or multi-trait PRS (PRSmix+, blue bar), while the dot and whisker plot showcases predictive accuracy using Area Under the receiver operating characteristic Curve (AUC) with 95% confidence interval for dichotomized traits.

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*, 2024.2010.2014.616932 (2024). https://doi.org/10.1101/2024.10.14.616932

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 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

22 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

23 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

24 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

25 Chang, K. C. *et al.* Survey of hepatitis B virus infection status after 35 years of universal vaccination implementation in Taiwan. *Liver Int* **44**, 2054-2062 (2024). https://doi.org/10.1111/liv.15959

26 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

27 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

28 Taylor, D. J. *et al.* Sources of gene expression variation in a globally diverse human cohort. *Nature* **632**, 122-130 (2024). https://doi.org/10.1038/s41586-024-07708-2

29 Wang, Q. S. *et al.* The whole blood transcriptional regulation landscape in 465 COVID-19 infected samples from Japan COVID-19 Task Force. *Nat Commun* **13**, 4830 (2022). https://doi.org/10.1038/s41467-022-32276-2

30 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

31 Prive, F., Arbel, J. & Vilhjalmsson, B. J. LDpred2: better, faster, stronger.

Bioinformatics **36**, 5424-5431 (2021). https://doi.org/10.1093/bioinformatics/btaa1029

32 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

33 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

34 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

35 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

36 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

37 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

38 Chang, S. W. *et al.* A genome-wide association study on chronic HBV infection and its clinical progression in male Han-Taiwanese. *PLoS One* **9**, e99724 (2014). https://doi.org/10.1371/journal.pone.0099724

39 Zeng, Z. *et al.* Genome-wide association study identifies new loci associated with risk of HBV infection and disease progression. *BMC Med Genomics* **14**, 84 (2021). https://doi.org/10.1186/s12920-021-00907-0

40 Li, Y. *et al.* Genome-wide association study identifies 8p21.3 associated with persistent hepatitis B virus infection among Chinese. *Nat Commun* **7**, 11664 (2016). https://doi.org/10.1038/ncomms11664

41 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

42 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

43 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

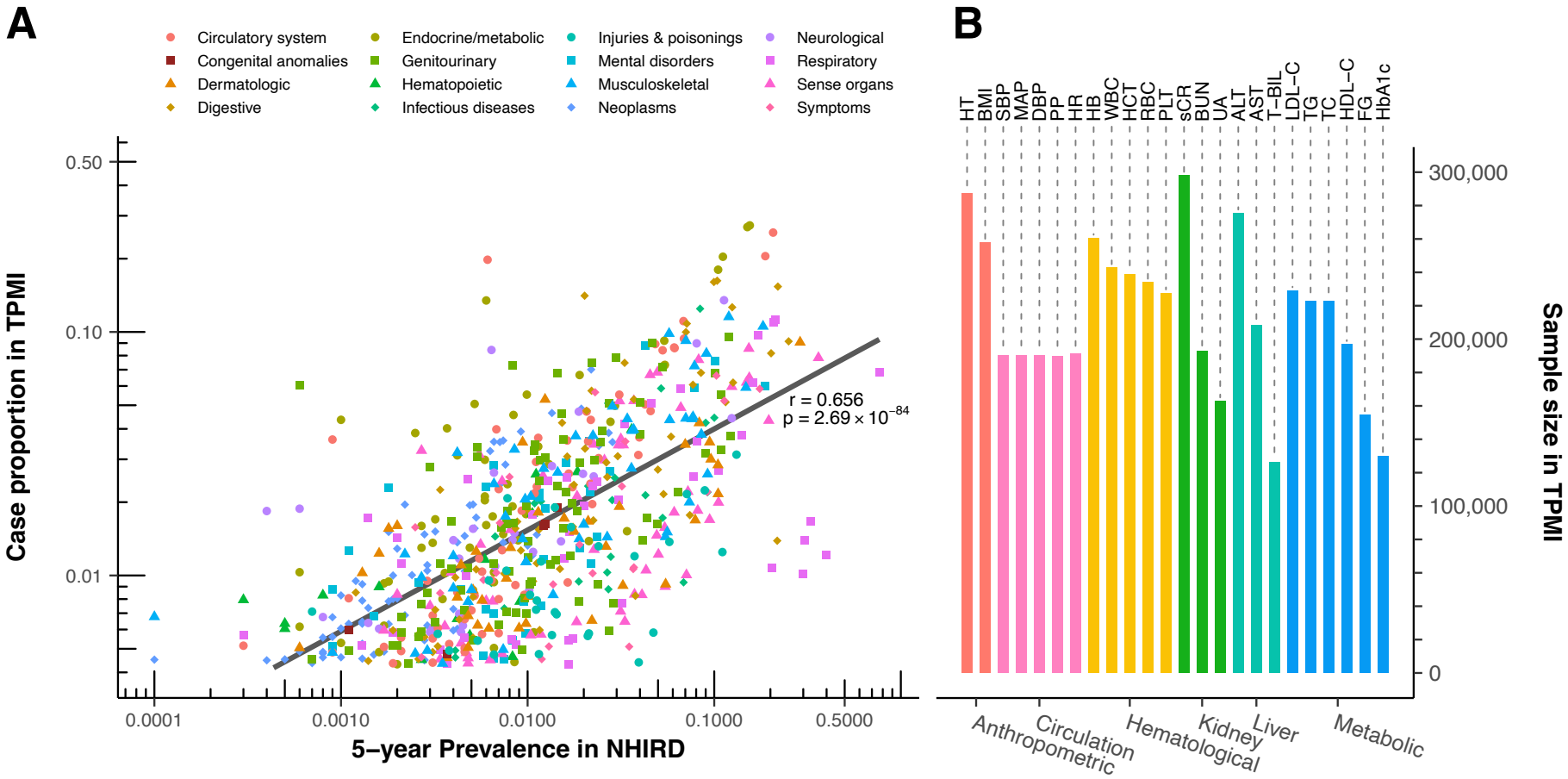
44 Asgari, S. & Pousaz, L. A. Human genetic variants identified that affect COVID susceptibility and severity. *Nature* **600**, 390-391 (2021). https://doi.org/10.1038/d41586-021-01773-7

45 Kember, R. L. *et al.* Polygenic risk scores for cardiometabolic traits demonstrate importance of ancestry for predictive precision medicine. *Pac Symp Biocomput* **29**, 611-626 (2024).

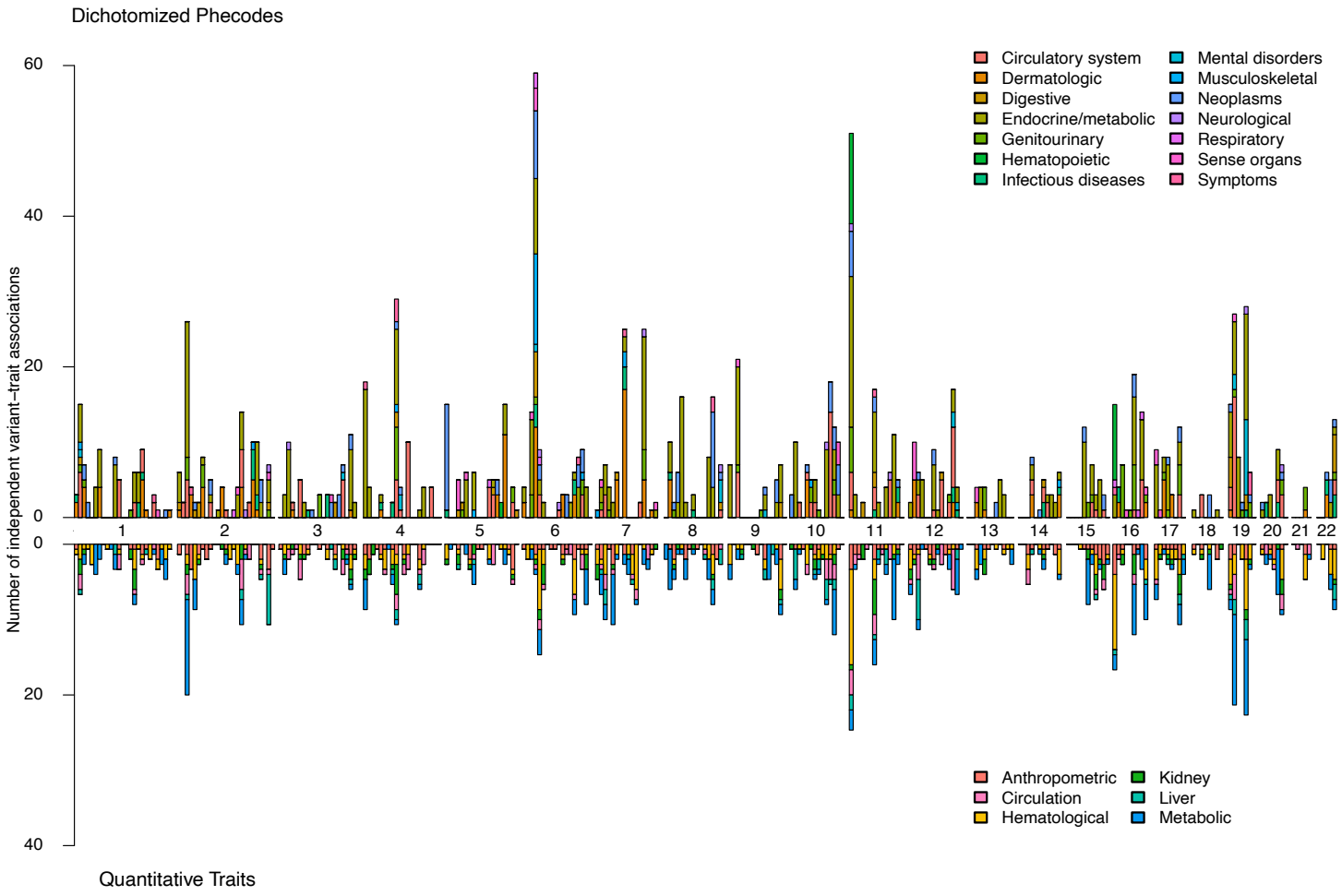
46 Khunsriraksakul, C. *et al.* Multi-ancestry and multi-trait genome-wide association meta-analyses inform clinical risk prediction for systemic lupus erythematosus. *Nat Commun* **14**, 668 (2023). https://doi.org/10.1038/s41467-023-36306-5

1005 47 Kelemen, M., Vigorito, E., Fachal, L., Anderson, C. A. & Wallace, C. shaPRS:
1006 Leveraging shared genetic effects across traits or ancestries improves accuracy
1007 of polygenic scores. *Am J Hum Genet* **111**, 1006-1017 (2024).
1008 https://doi.org/10.1016/j.ajhg.2024.04.009
1009 48 Zhai, S., Guo, B., Wu, B., Mehrotra, D. V. & Shen, J. Integrating multiple traits
1010 for improving polygenic risk prediction in disease and pharmacogenomics
1011 GWAS. *Brief Bioinform* **24** (2023). https://doi.org/10.1093/bib/bbad181
1012 49 Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic
1013 associations and downstream analyses. *Nat Hum Behav* **7**, 1216-1227 (2023).
1014 https://doi.org/10.1038/s41562-023-01579-9
1015 50 Chen, C.-Y. *et al.* Analysis across Taiwan Biobank, Biobank Japan, and UK
1016 Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell*
1017 *Genomics* **3**, 100436 (2023).
1018 https://doi.org/https://doi.org/10.1016/j.xgen.2023.100436
1019 51 Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex
1020 trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263-273
1021 (2022). https://doi.org/10.1038/s41588-021-00997-7
1022 52 Yengo, L. *et al.* Imprint of assortative mating on the human genome. *Nat Hum*
1023 *Behav* **2**, 948-954 (2018). https://doi.org/10.1038/s41562-018-0476-3
1024 53 Border, R. *et al.* Assortative mating biases marker-based heritability estimators.
1025 *Nat Commun* **13**, 660 (2022). https://doi.org/10.1038/s41467-022-28294-9
1026 54 Li, M. X. *et al.* A major gene model of adult height is suggested in Chinese. *J*
1027 *Hum Genet* **49**, 148-153 (2004). https://doi.org/10.1007/s10038-004-0125-8
1028 55 Zhong, Y., Perera, M. A. & Gamazon, E. R. On Using Local Ancestry to
1029 Characterize the Genetic Architecture of Human Traits: Genetic Regulation of
1030 Gene Expression in Multiethnic or Admixed Populations. *Am J Hum Genet* **104**,
1031 1097-1115 (2019). https://doi.org/10.1016/j.ajhg.2019.04.009
1032 56 Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS
1033 colocalization in GTEx. *Genome Biol* **21**, 233 (2020).
1034 https://doi.org/10.1186/s13059-020-02113-0
1035 57 Burch, K. S. *et al.* Partitioning gene-level contributions to complex-trait
1036 heritability by allele frequency identifies disease-relevant genes. *Am J Hum*
1037 *Genet* **109**, 692-709 (2022). https://doi.org/10.1016/j.ajhg.2022.02.012
1038 58 Wallace, C. A more accurate method for colocalisation analysis allowing for
1039 multiple causal variants. *PLoS Genet* **17**, e1009440 (2021).
1040 https://doi.org/10.1371/journal.pgen.1009440
1041

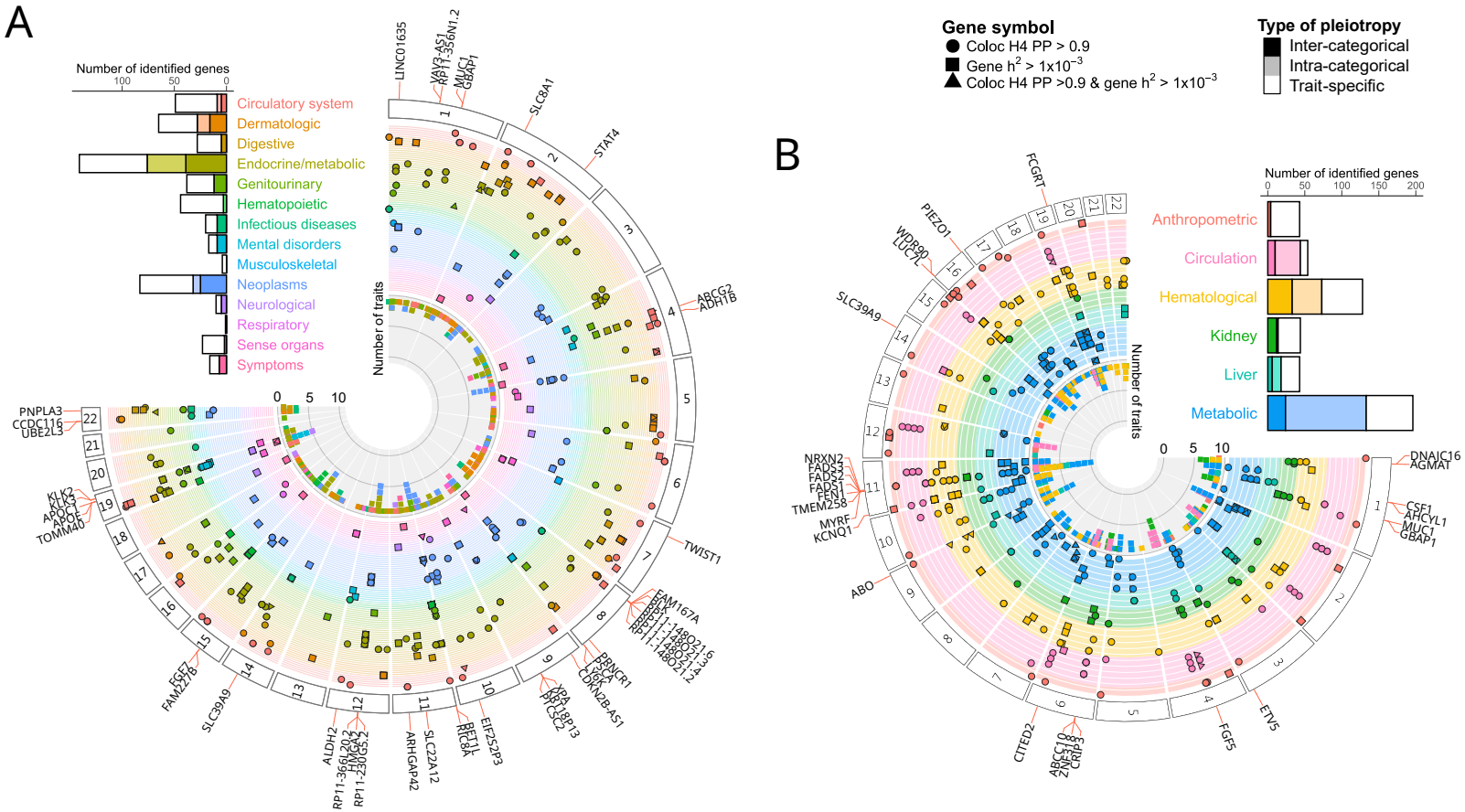
1042 **Figure 1**



1043 **Figure 2**



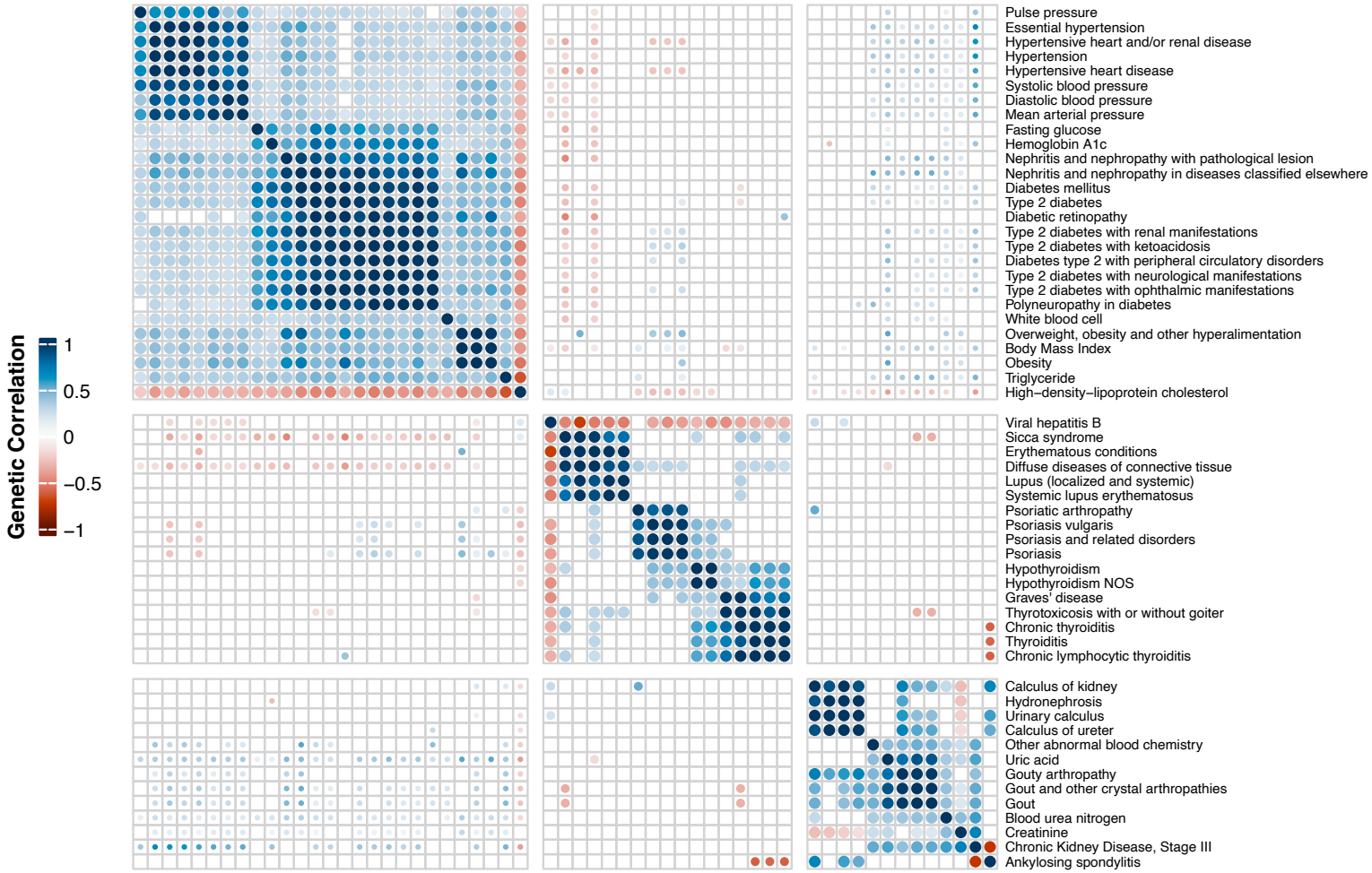
1044 **Figure 3**



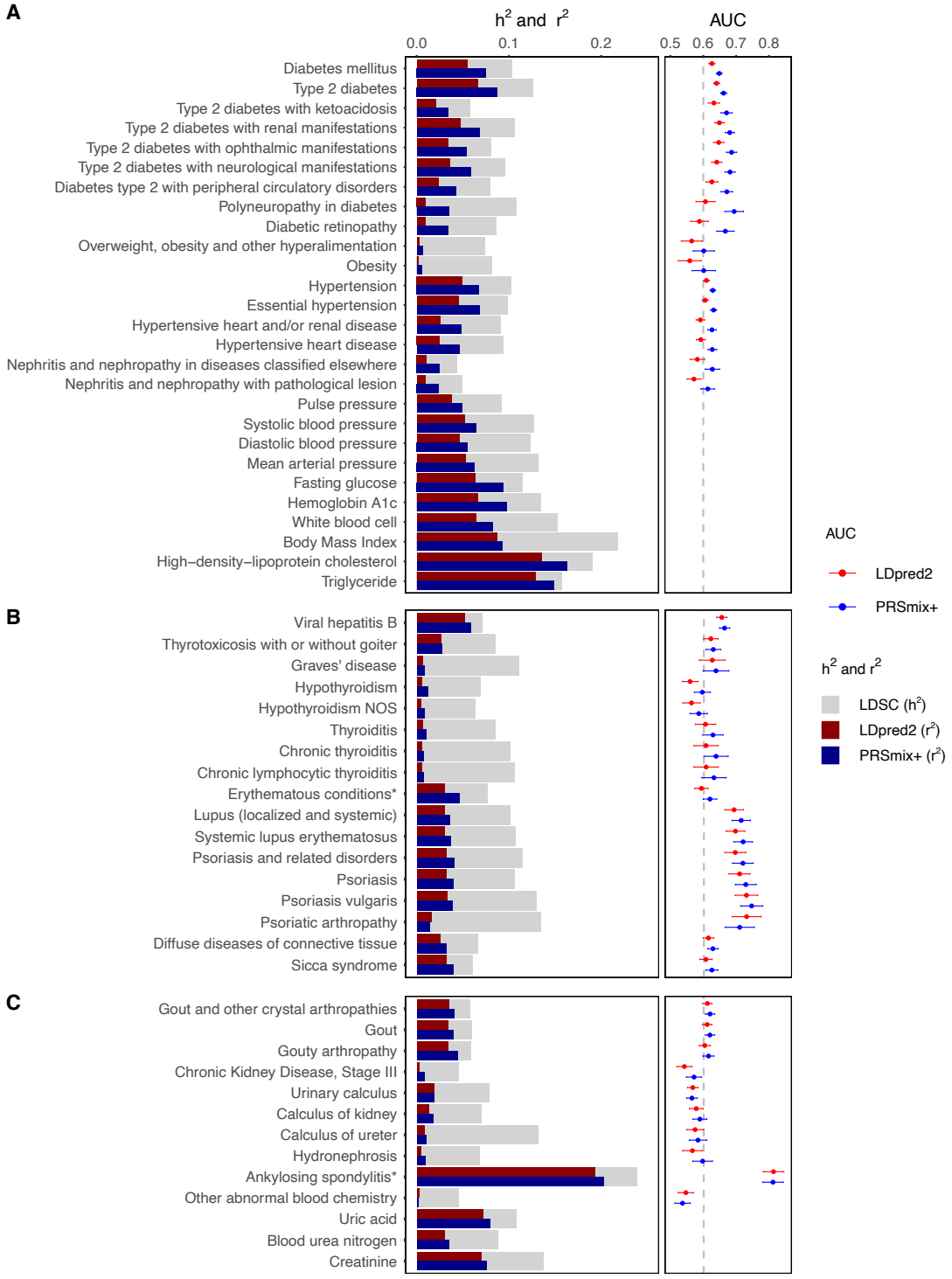
Dichotomized Phecodes

Quantitative Traits

1045 **Figure 4**



1046 **Figure 5**



1047 **Methods**

1048 **Study Population and Phenotyping**

1049 We utilized the Taiwan Precision Medicine Initiative (TPMI) dataset, which links
1050 extensive electronic medical records (EMR) with genotypic data for 486,956
1051 individuals. Dichotomized disease status was defined by phecodes, which were
1052 based on information extracted from the EMR using International Classification
1053 of Diseases (ICD) codes[18,19]. To ensure robustness, cases were defined by
1054 having the diagnosis of the relevant condition on two or more clinical visits. We
1055 also extracted quantitative traits from the EMR, including anthropometric, vital
1056 sign and laboratory measurements, and we excluded the extreme outliers and
1057 removed or adjusted the treated and/or medicated measures based on previous
1058 research, and the median value was kept if the participant had multiple qualified
1059 measures[59]. (Supplementary methods) In this study, we focused on 695
1060 dichotomized phenotypes (phecodes) that had at least 2,000 cases and 24
1061 quantitative traits that were measured in at least 100,000 individuals. These
1062 phecodes spanned 17 disease categories, including but not limited to infectious
1063 diseases, neoplasms, endocrine/metabolic disorders, and circulatory system
1064 diseases. The 24 quantitative traits were categorized into anthropometric,
1065 circulatory, hematological, kidney-related, liver-related, and metabolic
1066 measurements.

1067 **Genotyping and Quality Control**

1068 Genotyping was performed using two customized high-density Axiom SNP
1069 arrays produced by Thermo Fisher (Waltham, MA, USA), TPMv1 and TPMv2.
1070 The genotyping experiments were conducted in six genotyping centers in
1071 Taiwan[60]. The raw genotypic data underwent quality control measures, and the
1072 genetic variants were excluded when they had call rate <0.98 minor allele
1073 frequency (MAF) <0.01, or Hardy-Weinberg equilibrium test p-value $<1x10^{-6}$.
1074 We also excluded individuals with overall call rate <0.95, failed heterozygosity
1075 check, or inconsistent documented versus genetically determined sex. For this
1076 study, we only included the genetic variants found on both genotyping arrays
1077 and excluded variants with a significant batch effect in GWAS. The proportion
1078 of genetic ancestry was determined by ADMIXTURE[61], and the projected
1079 principal component scores with 1000 Genome as reference panel were

1080 applied to determine individuals' ancestry[62]. As a result, 401,710 genetic
1081 variants and 463,447 Han Chinese participants passed all quality control
1082 measures and were used in the subsequent studies. Details are found in
1083 supplementary methods and GitHub (https://github.com/TPMI-Taiwan/tpmi-qc).

1084 **Phasing and Imputation**

1085 Phasing was conducted on QC-passed genotype data with SHAPEIT5[63].
1086 Genome imputation was carried out with IMPUTE5 using a reference panel of
1087 1,498 whole genome sequenced Taiwan Biobank subjects[12,64]. We also
1088 conducted post-imputation quality control with exclusion criteria INFO score ≤
1089 0.7 and MAF ≤ 0.01. In addition, we also performed a chip-GWAS for minimizing
1090 the bias from different chips, resulting in a dataset of 8,046,864 well-imputed
1091 common genetic variants.

1092 **Population Structure and Relatedness Estimation**

1093 We performed a principal component analysis (PCA) based on genotyped
1094 variants to capture the effect of population structure. To diminish the effect of
1095 close relatives, the main PCA was conducted in a genetically unrelated subset,
1096 and other subjects were projected with the calculated PC weightings. And then,
1097 these PCA scores were leveraged to accurately quantify the proportion of
1098 identity-by-decent (IBD) and degree of relatedness. The maximum unrelated
1099 set was determined based on these estimated degrees of relatedness. PC-AiR
1100 and PC-Relate were used for PCA and relatedness estimation and PRIMUS
1101 was used for identifying the maximum unrelated set with the third degree as
1102 threshold[65-67].

1103 **Genome-Wide Association Study (GWAS)**

1104 The entire dataset was divided into three subgroups: the GWAS set (N =
1105 363,447), the training set (N = 80,000), and the testing set (N = 20,000). To
1106 maximize the statistical power, we used a mixed-effect regression model to
1107 examine the association between genotype and outcome of interest, logistic
1108 regression for dichotomized phecode, and linear regression for quantitative
1109 traits. The quantile-normalization was applied to quantitative traits to ensure the
1110 normal distribution. The mixed-effect model accounted for relatedness among
1111 individuals by including a random effect for pairwise kinship. The model also
1112 adjusted for key covariates, including age, sex, age^2, interactions between

1113 age/age^2 and sex, genotyping chip, enrollment hospital, and 10 genetic
1114 principal components to control for population stratification. SAIGE was applied
1115 for the mixed effect model GWAS[68]. Within the GWAS set, we selected an
1116 unrelated subset (N = 248,754) to perform GWAS using a generalized linear
1117 model with PLINK2, and we conducted 1:10 age, sex-matching for the traits
1118 with imbalanced case/control ratio (<1/20). These GWAS statistics were then
1119 used for heritability and genetic correlation estimation[69].

1120 **Replication Evaluation**

1121 To systematically evaluate the performance of our GWAS, we leveraged a pre-
1122 summarized phenotype-genotype reference map[70], which collected 5,879
1123 genetic associations for 149 unique phecodes from 523 published GWAS and
1124 1,215 associations from EAS. We calculate the overall and power-adjusted
1125 replication rates and actual over expected ratio for each available phecode and
1126 categories respectively. For measuring the quality of biobank data through
1127 replication an R package PGRM was used[70].

1128 **Fine-mapping**

1129 We performed fine-mapping to identify the independent GWAS signals in all
1130 genomic regions containing any variant with a p-value<5×10^{-8} and ±1.5 Mb of
1131 the regional lead variant[14], except the major histocompatibility complex region
1132 (MHC region, chr6: 25,391,792-33,424,245) due to its complex LD structure.
1133 We used the reported 95% credible set to determine the independent signals,
1134 and up to ten signals were allowed for each region. The genome-wide
1135 significant threshold was applied for defining a credible set as an independent
1136 hit, and an additional requirement of log Bayes factor (BF) > 2 was applied for
1137 the second hit. For the failed fine-mapping regions and MHC region, we used
1138 the lead SNP as the hit of each significant region. SuSiE was conducted for this
1139 summary statistics-based fine-mapping with LD derived from our imputation
1140 reference panel[71], which reflects our study population's genetic architecture.
1141 While using LD from the GWAS sample might improve accuracy, we used the
1142 imputation panel due to the computation efficiency.

1143 **Novel Association Identification**

1144 We comprehensively compared our GWAS results with reported significant
1145 signals on the NHGRI-EBI GWAS Catalog[21], download at 2024/03/11. The

1146 mapping of phecodes and quantitative traits to GWAS catalog phenotypes is
1147 summarized in Table S16. We classified a variant as novel if the fine mapped
1148 independent signal was not located within 1 Mb of any reported genome-wide
1149 significant association (p-value<$5x10^{-8}$) for the corresponding phenotype.
1150 Additionally, a variant was considered a new hit if the highest linkage
1151 disequilibrium (LD) r^2 was less than 0.1 with any reported significant association
1152 within 1Mb. Associations derived from uncertain and umbrella phecodes were
1153 excluded, and for duplicated genetic variants or regions, we only reported the
1154 association with the smaller p-value or from the phecode with the more specific
1155 definition. Finally, we used ANNOVAR to annotate the novel variants with data
1156 from the RefSeq Gene database (2020-08-17 updated)[72,73]. For the novel
1157 variants, we explored their allele frequencies in EUR (non-Finnish), AFR, SAS,
1158 AMR from gnomAD[74]. We also compared the the effect size between TPMI and
1159 UKB with t-test for investigating the ancestry-specific effect.

1160 **Heritability, Genetic Correlation, and Clustering**

1161 To quantify the genomic contribution of the specific traits, we applied linkage
1162 disequilibrium score regression to estimate the SNP-based heritability with
1163 LDSC[26]. The GWAS summary statistics and the pre-calculated LD score from
1164 EAS superpopulation of 1000 Genome were used[62]. For the dichotomized traits,
1165 we performed a liability-scaled transformation on the observed heritability using
1166 the 5-years population prevalence from the National Health Insurance dataset
1167 (NHIRD) of the Health and Welfare Data Science Center[20,75]. For the traits with
1168 a higher prevalence in our dataset (TPMI) than the population (NHIRD), we
1169 applied the equation derived by Lee et al[76]. For other traits, we used the
1170 adapted equation from Ojavee et al[75]. Additionally, we conducted LDSC to
1171 obtain pairwise genetic correlations to assess the similarity of genetic
1172 mechanisms between traits[77]. Based on the genetic correlation matrix, we used
1173 a hierarchical cluster analysis to identify groups of traits that share genetic
1174 mechanisms. We employed the weighted pair group method with arithmetic
1175 mean (WPGMA) for clustering, and the resulting cluster tree was used for group
1176 identification. Moreover, we estimate the genetic correlation across populations,
1177 TPMI and UK biobank, to demonstrate varied genetic architecture in different
1178 ancestry populations. For the UK biobank GWAS, we applied a generalized

1179 linear model from PLINK2 with the predefined phecode,
1180 https://github.com/umich-cphds/createUKBphenome, and corresponding
1181 baseline quantitative measures among the identified unrelated set (n =
1182 378,544). Popcorn was performed for the cross-population genetic correlation,
1183 and two correlation coefficients were calculated, the transethnic genetic-effect
1184 correlation (ρ_{ge}) and transethnic genetic-impact correlation (ρ_{gi})[30].

1185 **Gene-Level Heritability and Colocalization**

1186 We used both gene-level heritability estimation and colocalization analysis to
1187 map our GWAS findings to functional units, specifically genes. We conducted
1188 h2gene analysis to partition SNP-based heritability to the gene level[57]. We
1189 estimated heritability for genes that overlapped with fine-mapped regions,
1190 where gene regions were defined as the gene body ±10 kb for gene-level
1191 heritability. Additionally, to illustrate the molecular functions of genes of interest,
1192 we used colocalization analysis to examine whether there are shared common
1193 genetic causal variants between tissue-specific gene expression and traits of
1194 interest. We utilized expression quantitative traits locus (eQTL) resources from
1195 49 tissues in GTEx v8[27], lymphoblastoid in MAGE[28], and whole blood in JCTF[29] ,
1196 testing any gene with genome-wide significant signals in the cis-regulation
1197 region (±1 Mb). The posterior probabilities were used to evaluate colocalization
1198 between gene expression and the trait of interest. The R package, coloc, was
1199 used with SuSiE relaxing the single causal variant assumption[58,78,79].

1200 **Single and Multi-Trait Polygenic Risk Score (PRS)**

1201 The preserved dataset of 100,000 unrelated TPMI subjects was split into two
1202 subsets, training (n = 80,000) and validation (n = 20,000) for PRS model
1203 building. Five popular PRS tools were used, LDpred2[31], Lassosum2[32], PRS-
1204 CS[33], SBayesR[34], and MegaPRS[35], and the training subset was applied for
1205 parameter selection and model optimization if needed. LDpred2, PRS-CS, and
1206 SBayesR assumed the effect of genetic variants following a mixture distribution
1207 with different pre-defined parameters and applied a Bayesian framework for
1208 distribution estimation. Lassosum2 utilized a penalized regression (LASSO) for
1209 weight estimating, and MegaPRS leveraged minor allele frequency and linkage
1210 disequilibrium for model building. We then used the validation subset to
1211 evaluate the performance of PRS models. Individual score was calculated with

1212 PLINK2[69]. The explained variance (r^2) was used to evaluate the performance
1213 of PRS for quantitative traits[76,80], and two indices, area under the receiver
1214 operating characteristic curve (AUC) and liability-scaled $r^{2,}$ were used for PRS
1215 of dichotomized phenotypes. We followed the approach by Ni et al.[80] and report
1216 both raw r^2 and r^2 adjusted for covariates (sex, age, and PCs). Additionally, we
1217 include partial r^2 estimates, calculated using the R package rsq. To account for
1218 population stratification in cross-cohort predictions, we also report r^2 with PCs
1219 as covariates in the supplemental tables. For AUC comparisons, we include a
1220 baseline model incorporating standard covariates (sex, age, and PCs) to better
1221 assess the added predictive power of PRS. The likelihood ratio test was used
1222 to obtain the significance for r^2 with R package, lmtest, and standard error for
1223 AUC was calculated with R package, auctestr. To further leverage the gene's
1224 pleiotropy and shared genetic mechanism among traits, we conducted a multi-
1225 trait PRS model building for the traits in the same genetic cluster based on
1226 pairwise genetic correlation identified in the previous step. We pooled all PRS
1227 models from five tools for those identified traits and applied an elastic net
1228 regression to combine their weighting and find the most optimized model for the
1229 target trait. PRSmix+ was performed for the multiple traits PRS model building[36].
1230 The cross-population PRS models was based on both TPMI and UKB
1231 European GWAS, and PRS-CSx was applied[81].

1232 **External Validation and Comparison**

1233 We conducted an external validation of our developed PRS using data from the
1234 Taiwan Biobank, EAS from UK Biobank and All of Us. TWB is a community-
1235 based biobank, and it has recruited over 200,000 participants in Taiwan. Herein,
1236 we used 88,628 unrelated subjects (> third degree and removed 5,242
1237 overlapped individuals), who were genotyped with the Axiom customized chip
1238 TWB2 (equivalent to TPMv1), and their genotyping QC, phasing, and
1239 imputation followed the same protocol as described above. The self-reported
1240 disease condition was queried from their baseline questionnaire, except for
1241 cancer. Since the study design of TWB excluded cancer patients at recruitment,
1242 we used both baseline and follow-up self-reporting data to define cancer cases
1243 and controls. UKB has enrolled ~500,000 participants since 2006 and linked
1244 their genetic data with enriched phenotypic data. For UKB validation, we used

their inpatient record for case definition. Their ancestral population was determined by self-reported ethnic background, such as self-reported Chinese as EAS (n = 1,572), White, British, Irish, and any other white background as EUR (n = 472,869), Black or Black British, Caribbean, African, and any other Black background as AFR (n = 8,074), and Asian or Asian British, Indian, Pakistani, Bangladeshi, and any other Asian background as SAS (n = 9,893). 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. We applied ADMIXTURE with 1000 Genomes as reference panel for assigning the genetically inferred ancestral populations, including EAS (n = 6,895), EUR (n = 152,754), AFR (n = 60,964), AMR (n = 32,394), and SAS (n = 2,334). 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 Health Measures Evaluation

We evaluated the genetic impact on overall health meaures. We used the number of clinical visits and the aggregate duration of hospitalization as overall health 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[82]. 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

All polygenic risk score (PRS) models and GWAS results (summary statistics from SAIGE and PLINK) are available from the TPMI website. The UKB phecode GWAS was obtatined from UKBiobank TOPMed-imputed PheWeb (https://pheweb.org/UKB-TOPMed/). The eQTL resources are download from websites of GTEx (https://www.gtexportal.org/), MAGE (https://doi.org/10.5281/zenodo.10535719), and JCTF (https://humandbs.dbcls.jp/en/hum0343-v4). The detailed describtion of data availability and application process of TWB, NHIRD, UKB and All of Us can be found on their websites (TWB: https://www.biobank.org.tw/english.php; NHIRD: https://www.apre.mohw.gov.tw/; UKB: https://www.ukbiobank.ac.uk/; All of Us: https://www.researchallofus.org/).

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 Taiwan Biobank was apporaved by Institutional Review Boards of Academia Sinica (AS-IRB-BM-

19014), and the NHIRD 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. We gratefully acknowledge the use of data from the National Health Insurance Research Database, provided by the Ministry of Health and Welfare of Taiwan. 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.). This work used ASGC (Academia Sinica Grid-computing Center) Distributed Cloud resources, which is supported by Academia Sinica. 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.

Author contributions

Supervision: P.-Y.K., W.H.S., S.-F.Y., J.-M.L., J.-Y.W., J.-F.C., J.-Y.W., ,C.S.F., M.-C.H., Y.-C.F., and C.-H.C.; Formal analysis: H.-H.C., C.-Y.C., J.-P.S., W.-

J.L., F.-J.H., E.H.W., E.-C.Y., S.-H.W., and S.-P.C.; Resources: P.-Y.K., C.-L.C., J.-K.J., I.-H.L., K.-H.L., W.-S.C., H.-C.T., S.-Y.L., F.-P.C., H.-L.H., Y.-C.Y., W.-C.T., M.-H.L., H.-T.C., L.-M.T., W.-Y.L., P.C.C., Y.-C.H., Y.-M.C., T.-H.H., C.-H.L., Y.-J.C., I.-C.C., C.-L.M., S.-J.C., Y.-L.C., Y.-J.L., C.-H.L., W.-J.L., H.T., T.-T.Y., H.-C.Y., M.-Y.C., Y.-C.L., Y.-T.K., B.-Z.K., J.-E.L., C.-L.C., J.-C.L., P.C., C.-H.L., C.-H.C., I.-C.W., L.-C.L., J.-W.W., S.-I.S., S.-W.H., C.-H.H., W.-M.L., C.-J.Y., Y.-T.C., D.-R.H., Y.-H.H, C.-H.Y., Y.-S.H., Y.-F.C., H.-M.W., P.-H.T., K.-G.H., C.-Y.C., Y.-L.H., M.-S.W., J.-H.K., Y.-B.L., J.J.J., Y.-H.L., J.-Y.L., H.-J.L., A.-C.F., J.-S.L., N.-F.C., Y.-W.Y., E.M., C.-Y.H., C.-C.L., T.-F.W., K.-Y.S., J.-K.W., M.-H.C., H.-F.C., G.-C.M., T.-Y.C., F.-T.C., H.-J.C., K.-J.K., C.-F.H., C.-Y.T., P.-Y.C., K.T. and Y.-J.S.; Validation (TWB): H.-H.C., C.-Y.C., J.-P.S., W.-J.L., F.-J.H., E.H.W., E.-C.Y., and J.-Y.W.;Validation (UKB and All of Us): T.R., E.W., S.H.; Data Curation:M.-F.T., T.-H.W., Y.-C.L.; Writing - Original Draft: H.-H.C., C.S.F., L.-H.L., C.-Y.C., J.-P.S., W.-J.L., E.H.W., E.-C.Y.; Writing - Review & Editing: P.-Y.K., E.-C.Y., M.-F.T., C.-H.C., H.-C.Y., Y.-T.H., C.-H.C., C.-Y.W., H.-H.C., C.S.F., and L.-H.L.; Investigation: L.-H.L., R.-H.W., Y.-C.C., and C.-P.C.; Project administration: Y.-M.L. and C.-S.Y. All authors reviewed and approved the final manuscript.

Competing interests

SDHH is a founder, shareholder, and serves on the Board of Directors of Genomic Prediction, Inc. (GP), and EW is an employee and shareholder of GP. All remaining authors declare no competing interests relevant to this work.

References

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

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

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 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

26 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

27 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

28 Taylor, D. J. *et al.* Sources of gene expression variation in a globally diverse human cohort. *Nature* **632**, 122-130 (2024). https://doi.org/10.1038/s41586-024-07708-2

29 Wang, Q. S. *et al.* The whole blood transcriptional regulation landscape in 465 COVID-19 infected samples from Japan COVID-19 Task Force. *Nat Commun* **13**, 4830 (2022). https://doi.org/10.1038/s41467-022-32276-2

30 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

31 Prive, F., Arbel, J. & Vilhjalmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424-5431 (2021). https://doi.org/10.1093/bioinformatics/btaa1029

32 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

33 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

34 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

35 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

36 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

57 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

58 Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet* **17**, e1009440 (2021). https://doi.org/10.1371/journal.pgen.1009440

59 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

60 Yang, H.-C. *et al.* The Taiwan Precision Medicine Initiative: A Cohort for Large-Scale Studies. *BIORXIV* (2024). https://doi.org/10.1101/2024.10.14.616932v1

61 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

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

63 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

64 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

65 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

66 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

67 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

68 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

69 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

70 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

71 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

72 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

73 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

74 Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92-100 (2024). https://doi.org/10.1038/s41586-023-06045-0

75 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

76 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

77 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

78 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

79 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

80 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

81 Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics* **54**, 573-580 (2022). https://doi.org/10.1038/s41588-022-01054-7

82 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>.

Extended Data

Extended Data Fig. 1. Scatter plots of the case proportion for dichotomized phenotypes by disease categories. The significantly positive correlations are observed across disease categories between the case proportion in TPMI (x-axis) and the 5-year prevalence in NHIRD (y-axis), except congenital anomalies.

Extended Data Fig. 2. 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. 3. The Manhattan plot of GWAS for viral hepatitis B in TPMI. The names of nearest mapped gene were labeled for the independent GWAS significant loci.

Extended Data Fig. 4. 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. 5. The scatter plot of performance for PRS developed by different tools. The color represents the phecode category, and the shape indicates the PRS development tool: Lassosum2 (circle), LDpred2 (triangle), MegaPRS (square), PRS-CS (cross), and SbayesR (square cross). Each phecode is positioned at a unique x-coordinate, with the tool that has the highest AUC highlighted.

Extended Data Fig. 6. 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 indicate SNP-heritability (estimated from TPMI GWAS unrelated set [n=248,754] with LDSC), and the colored bar chart presents the r^2 values, indicating the proportion of variance explained by the PRS among TPMI validation set (n=20,000), and dots and error bars show Area Under the receiver operating characteristic Curve (AUC) with 95% confidence interval. An asterisk (*) indicates estimates considering the MHC region.

Extended Data Fig. 7. External validation of PRS models in Taiwan Biobank and other cohorts. PRS performance is presented as Area Under the receiver operating characteristic Curve (AUC) with 95% confidence inverval in TPMI (orange, TPMI validation set, n = 20,000), Taiwan Biobank (green, n = 88,628), East Asians in UK Biobank (blue, n = 1,572), and East Asians in All of Us (purple, n = 6,895). Circles represent TPMI-derived PRS, and triangles indicate UKB (European)-derived PRS models. Only the estimates with case size > 30 were showed on the figure.

Extended Data Fig. 8. External validation of TPMI-derived PRS model and TPMI-UKB cross-population PRS models across populations. The plots show the Area Under the receiver operating characteristic Curve (AUC) with 95% confidence inverval for PRS validation in East Asian (red, n = 88,628 in TWB; 1,572 in UKB; 6,895 in All of Us), European (olive green, n = 472,869 in UKB and 152,754 in All of Us), African (green, n = 8,074 in UKB and 60,964 in All of Us), South Asian (blue, n = 9,893 in UKB and 2,334 in All of Us), and Admixed American (purple, n = 32,394 in All of Us) populations

from TPMI (circle), UKB (triangle), and All of Us (square) cohorts.

Extended Data Table 1. Proportion of overall health indices explained by genetic risk

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

Likelihood ratio test was applied to obtain the explained variation (r2) and significance by comparing the full model and null model.

Supplementary Information

Supplemental tables

Table S1-S16

Supplemental text

Detailed methods of genotyping and phenotyping quality control, and SNP-based heritability for quantitative traits