Global biobank analyses provide lessons for computing polygenic risk scores across diverse cohorts

Ying Wang[1,2,3,*], Shinichi Namba[4], Esteban Lopera[5], Sini Kerminen[6], Kristin Tsuo[1,2,3], Kristi Läll[7], Masahiro Kanai[1,2,3,8,9], Wei Zhou[1,2,3], Kuan-Han Wu[10], Marie-Julie Favé[11], Laxmi Bhatta[12], Philip Awadalla[11,13], Ben Brumpton[12,14,15], Patrick Deelen[5,16], Kristian Hveem[12,14], Valeria Lo Faro[17,18,19], Reedik Mägi[7], Yoshinori Murakami[20], Serena Sanna[5,21], Jordan W. Smoller[22], Jasmina Uzunovic[11], Brooke N. Wolford[10,12], Global Biobank Meta-analysis Initiative, Cristen Willer[12,23,24,25], Eric R. Gamazon[26,27,28], Nancy J. Cox[26,28], Ida Surakka[23], Yukinori Okada[4,29,30,31,32], Alicia R. Martin[1,2,3,‡,*], Jibril Hirbo[26,28,‡,*]

Affiliations

1. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
2. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
3. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
4. Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
5. University of Groningen, UMCG, Department of Genetics, Groningen, the Netherlands
6. Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland
7. Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia
8. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
9. Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan
10. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor MI, 48103, USA
11. Ontario Institute for Cancer Research, Toronto, Ontario, Canada
12. K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim, 7030, Norway
13. Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
14. HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, 7600, Norway
15. Clinic of Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, 7030, Norway
16. Oncode Institute, Utrecht, The Netherlands
17. Department of Ophthalmology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
18. Department of Clinical Genetics, Amsterdam University Medical Center (AMC), Amsterdam, The Netherlands
19. Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
20. Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan
21. Institute for Genetics and Biomedical Research (IRGB), National Research Council (CNR), Cagliari 09100, Italy
22. Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
23. Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA
24. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA
25. Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
26. Department of Medicine, Division of Genetic Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA
27. MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
28. Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA
29. Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
30. Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan
31. Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan
32. Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita 565-0871, Japan

[1‡]These authors contributed equally.
Lead Contact: Ying Wang (yiwang@broadinstitute.org)
*Correspondence: yiwang@broadinstitute.org, armartin@broadinstitute.org, jibril.hirbo@vumc.org.

Summary

With the increasing availability of biobank-scale datasets that incorporate both genomic data and electronic health records, many associations between genetic variants and phenotypes of interest have been discovered. Polygenic risk scores (PRS), which are being widely explored in precision medicine, use the results of association studies to predict the genetic component of disease risk by accumulating risk alleles weighted by their effect sizes. However, limited studies have thoroughly investigated best practices for PRS in global populations across different diseases. In this study, we utilize data from the Global-Biobank Meta-analysis Initiative (GBMI), which consists of individuals from diverse ancestries and across continents, to explore methodological considerations and PRS prediction performance in 9 different biobanks for 14 disease endpoints. Specifically, we constructed PRS using heuristic (pruning and thresholding, P+T) and Bayesian (PRS-CS) methods. We found that the genetic architecture, such as SNP-based heritability and polygenicity, varied greatly among endpoints. For both PRS construction methods, using a European ancestry LD reference panel resulted in comparable or higher prediction accuracy compared to several other non-European based panels; this is largely attributable to European descent populations still comprising the majority of GBMI participants. PRS-CS overall outperformed the classic P+T method, especially for endpoints with higher SNP-based heritability. For example, substantial improvements are observed in East-Asian ancestry (EAS) using PRS-CS compared to P+T for heart failure (HF) and chronic obstructive pulmonary disease (COPD). Notably, prediction accuracy is heterogeneous across endpoints, biobanks, and ancestries, especially for asthma which has known variation in disease prevalence across global populations. Overall, we provide lessons for PRS construction, evaluation, and interpretation using the GBMI and highlight the importance of best practices for PRS in the biobank-scale genomics era.

Keywords

Global-biobank meta-analysis initiative (GBMI), polygenic risk scores (PRS), multi-ancestry genetic prediction, accuracy heterogeneity

Introduction

Population- and hospital-based biobanks are increasingly coupling genomic and electronic health record data at sufficient scale to evaluate the promise of personalized medicine[1]. The growth of these paired datasets enables genome-wide association studies (GWAS) to estimate increasingly precise genetic effect sizes contributing to disease risk. In turn, GWAS summary statistics can be used to aggregate the effects of many SNPs to estimate individuals' genetic predispositions for complex diseases via polygenic risk scores (PRS). As GWAS power has increased, PRS accuracy has also improved, with PRS for some traits having comparable accuracies to independent biomarkers already routinely used in clinical risk models[2]. Consequently, several areas of medicine have already begun investigating the potential for integrating PRS alongside other biomarkers and information currently used in clinical risk models[3–5]. However, evidence of clinical utility for PRS across disease areas is currently limited or inconsistent[2,6–8]. Furthermore, many methods have been developed to compute PRS, each with different strengths and weaknesses[9–11]. Thus, guidelines that delineate best practices while considering a range of real-world healthcare settings and disease areas are critically needed.

Best practices for PRS are critical but lacking for a range of considerations that have been shown to contribute to variability in accuracy and interpretation. These include guidance for variable phenotype definitions and precision for both discovery GWAS and target populations, which varies with cohort ascertainment strategy, geography, environmental exposures and other common covariates[12–14]. Other considerations include varying genetic architectures, statistical power of the discovery GWAS, and PRS methods, which vary in which variants (generally in the form of SNPs) are included and how weights are calculated[9,15]. A particularly pernicious issue requiring best practices is regarding maximizing generalizability of PRS accuracy among ancestry groups[16,17]. Developing best practices for PRS therefore requires harmonized genetic data spanning diverse phenotypes, participants, and ascertainment strategies.

To facilitate the development of best practices, we evaluate several considerations for PRS in the Global Biobank Meta-analysis Initiative (GBMI). GBMI brings together population- and hospital-based biobanks developed in twelve countries spanning four different continents: North America (USA, Canada), East Asia (Japan and China), Europe (Iceland, UK, Estonian, Finland, Scotland, Norway and Netherlands) and Oceania (Australia). GBMI aggregates paired genetic and phenotypic data from >2.1 million individuals across diverse ancestries, including: ~1.4 million Europeans (EUR), ~18,000 Admixed Americans (AMR), ~1,600 Middle Eastern (MID),

~31,000 Central and South Asians (CSA), ~341,000 East Asians (EAS) and ~33,000 Africans (AFR). Biobanks have collated phenotype information through different sources including electronic health records, self-report data from epidemiological survey questionnaires, billing codes, doctors' narrative notes, and death registries. Detailed description of each biobank is found in Zhou et al.[18].

Here we outline a framework for PRS analyses of multi-ancestry GWAS across multiple biobanks, as shown in **Figure 1**. By evaluating PRS across 14 endpoints (**Table S1**) and 9 biobanks, we review and explore practical considerations for three steps: genetic architecture estimation, PRS method optimization and selection, and evaluation of PRS accuracy. The endpoints examined are: asthma, chronic obstructive pulmonary disease (COPD), heart failure (HF), stroke, acute appendicitis (AcApp), venous thromboembolism (VTE), gout, appendectomy, primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC) and hypertrophic or obstructive cardiomyopathy (HCM), for which the phenotype definitions can be found in Zhou et al.[18]. Our framework applies to biobank-scale resources with both homogenous and diverse ancestries.

Results

Genetic architecture of 14 endpoints in GBMI

We first estimated the genetic architecture of 14 endpoints based on HapMap3 SNPs (see STAR Methods). Different prediction methods vary in which SNPs are selected and which effect sizes are assigned to them. Thus, understanding the genetic architecture of complex traits along with sample size and ancestry composition of the discovery GWAS is critical for choosing optimal prediction methods. For example, the SNP-based heritability (h^2_{SNP}) bounds PRS accuracy. We used SBayesS[19] to estimate h^2_{SNP}, polygenicity (the proportion of SNPs with nonzero effects), and the relationship between minor allele frequency (MAF) and SNP effects (i.e., a metric of negative selection, hereafter denoted as S) for the 14 endpoints in GBMI. Meta-analyses in GBMI were performed across 19 different biobanks on 14 endpoints using an inverse-variance weighted method as described in Zhou et al.[18], including individuals from diverse ancestries.

Most diseases analyzed here had low but significant h^2_{SNP} and a range of polygenicity estimates (**Figure 2**). Note that here we reported the h^2_{SNP} on the liability scale (see STAR Methods). The SBayesS model failed to converge for HCM, likely because of known predisposing monogenic mutations, so this endpoint was dropped from downstream analyses. In addition to presenting results using meta-analysis from all ancestries (multi-ancestry GWAS), we also reported estimates using EUR only GWAS summary statistics (EUR GWAS). We observed that the estimates were overall higher using multi-ancestry GWAS compared to EUR GWAS (**Figure 2**). Overall, the median estimates of SNPs with nonzero effects across 13 endpoints were 0.34% for multi-ancestry GWAS and 0.14% for EUR GWAS, respectively. The corresponding median estimates for h^2_{SNP} were 0.051 for multi-ancestry GWAS and 0.043 for EUR GWAS, respectively.

Polygenicity and h^2_{SNP} estimates varied greatly among different endpoints. Specifically, the h^2_{SNP} estimates were highest for asthma and gout using multi-ancestry GWAS (h^2_{SNP}= 0.085, s.e. = 0.0011 and h^2_{SNP}= 0.111, s.e. = 0.0024, respectively), while asthma was found to be much more polygenic than gout. We caution that the numeric interpretation of polygenicity depends on various factors and cannot be interpreted as the number of causal variants. For example, larger and more powerful GWAS tend to discover more trait-associated variants, thus appear to have higher polygenicity. Because we used the same set of SNPs in SBayesS analyses for all endpoints, we hence used the results as a relative measurement of the degree of polygenicity. We observed that the estimate of polygenicity for UtC using multi-ancestry GWAS was not statistically different from 0 (Wald test, *p*-value > 0.05/13) due to limited power observed as relatively low h^2_{SNP}. Overall, COPD and asthma were estimated to be the most polygenic traits, followed by HF and stroke, whereas AcApp, UtC and ThC were the least polygenic. Lastly, we observed signals of negative selection for traits including asthma (S = -0.56, s.e. = 0.05), COPD (S = -0.40, s.e. = 0.11) and POAG (S = -0.50, s.e. = 0.15) when considering using EUR GWAS, consistent with empirical findings of negative selection explaining extreme polygenicity of complex traits[20].

In summary, we observed largely varied key parameters of genetic architecture among 13 endpoints using multi-ancestry and EUR only GWAS. We found that asthma and COPD had the

highest h^2_{SNP} as well as polygenicity. We excluded HCM in our subsequent prediction analyses due to lower evidence of polygenicity.

Optimal prediction performance using heuristic methods depends on phenotype-specific genetic architecture

We first evaluated the pruning and thresholding (P+T) method for PRS across phenotypes given its widespread use and relative simplicity. Specifically, we fine-tuned P+T optimization using different combinations of LD window sizes and LD r^2 thresholds in the most powered GBMI GWAS, asthma. We also explored the effect of ancestry composition, sample size, and SNP density of the LD reference panel on prediction performance in diverse ancestry groups for asthma (see STAR Methods). We then performed broadly optimized P+T parameters (*p*-value thresholds ranged from 5×10^{-8} to 1) for all 13 endpoints in the UKBB and BBJ.

We found that different LD window sizes maximized the prediction accuracy (referred to as R^2 on the liability scale, $R^2_{liability}$, if not specified) in different settings for asthma (**Figure S1 and Table S2**). PRS accuracy tends to decrease with larger LD r^2 thresholds (e.g., > 0.1) when using more stringent *p*-value thresholds, but tends to increase with more liberal *p*-value thresholds, possibly because more stratified signal is tagged. To balance the computational burden and signal-to-noise ratio, we used an LD window size of 250Kb and LD r^2 of 0.1 to report the following results. We repeated our analyses using genome-wide common SNPs and compared the prediction accuracy with that using HapMap3 SNPs only (**Figure S1 and Table S2**). There were no significant improvements in predictions using a denser SNP set, which suggests that HapMap3 SNP set represents genome-wide common SNPs well. Moreover, we found that the sample sizes of the LD reference panel had little impact on P+T performance (**Figure S2**); but, the parameters described above including LD window sizes and LD r^2 thresholds had a larger impact on accuracy. We also showed that using 1KG-EUR as the LD reference panel performed well compared to using other ancestral populations with similar sample sizes in the 1KG dataset, which can be explained by the overrepresentation of EUR participants in GBMI (**Figure S3**). Because the sample was primarily European and the 1KG-EUR LD panel performed similarly well as diverse ancestry LD panels we evaluated, we therefore used 1KG-EUR as the LD reference panel for all following P+T analyses. But the choice of LD

reference panel for multi-ancestry GWAS remains an open question to further explore especially when the discovery GWAS becomes more diverse.

We found that the optimal *p*-value threshold (the *p*-value threshold with highest prediction accuracy) differed considerably between various endpoints (**Figure S4**). This pattern is found to be related to polygenicity of studied endpoints; but it is also due to a combination of factors such as the GWAS discovery cohort sample size, disease prevalence, trait-specific genetic architecture, and genetic and environmental differences between discovery and target ancestries[21]. For example, when the optimal *p*-value was determined in the UKBB-EUR subset, gout (132 variants) and AcApp (22 variants) showed highest accuracy at a *p*-value threshold of 5×10^{-7} , while stroke and HF achieved the highest accuracy when including SNPs with *p*-value < 0.5 and 1 (99,858 variants and 136,290 variants), respectively. To investigate whether ancestries affect the optimal *p*-value threshold, we replicated our analysis in the BBJ (**Figure S4**). In the BBJ, *p*-value thresholds of 5×10^{-5}, 0.01 and 5×10^{-5} presented best performance for gout, stroke and HF, respectively. Consistent with previous studies, these results suggest that optimal prediction parameters for P+T appear to be dependent on the ancestry of the target data among other factors[22,23]. Further, we found that for more polygenic traits including asthma, COPD, stroke and HF, prediction was more accurate with more variants in the PRS (i.e., a less significant threshold) than using the genome-wide significance threshold (*p*-value < 5×10^{-8}). On the contrary, less polygenic traits showed no or modest improvement with less stringent *p*-value thresholds, especially for traits such as gout which has trait-associated SNPs with large effects. However, these trends were less obvious in the BBJ which might be attributed to the small proportion of EAS included in the discovery GWAS.

Finally, we investigated the impact of per-variant effective sample size heterogeneity. Since GBMI consists of a number of biobanks with diverse ancestries, the number of samples used for meta-analysis was notably heterogeneous among the variants; the majority of the variants in the GWAS meta-analysis had only a limited number of effective samples (N_{eff}) (**Figure 3-A**). Therefore, although sample size heterogeneity is not usually considered for PRS, it may confound the PRS prediction accuracy in the case of global biobank collaborations. By filtering the variants according to N_{eff} per-variant (i.e., N_{eff} larger than 50% or 80% thresholds of the maximum N_{eff} of the trait of interest, see STAR Methods), we observed that the $R^2_{liability}$ increased substantially for less stringent thresholds (*p*-value > 5×10^{-5}) in the UKBB (**Figure S5-A**). As a representative example, the largest

$R^2_{liability}$ (0.034) was obtained for asthma when the *p*-value threshold was 5 × 10^{-3}, whereas the $R^2_{liability}$ was 6.6 × 10^{-3} at the threshold without N_{eff} filtering (**Figure 3-B and Table S3**). Next, we investigated whether N_{eff} filtering could be substituted by other filtering criteria. Although excluding variants with MAF less than 0.1 partially compensated for PRS transferability, the improvement of N_{eff} filtering in $R^2_{liability}$ was still observed (**Figure S5-B**). Heterogeneity in N_{eff} might be confounding especially in multi-ancestry meta-analyses because it can be distorted by heterogeneous allele frequencies and imputation quality spectra among ancestries. Indeed, as rarer variants tend to be more ancestry-specific, variants with low N_{eff} tend to be unique to specific ancestries (**Figure 3-C**). Of note, the dependency of $R^2_{liability}$ on the N_{eff} was, however, largely rectified for most of the traits by using only HapMap3 SNPs (**Figure S5-C**). Given that the $R^2_{liability}$ for HapMap3 SNPs was comparable to that for genome-wide SNPs (**Figure S1**), filtering to HapMap3 SNPs might be suitable for meta-analysis of diverse populations. On the other hand, HapMap3 SNPs generally have good imputation quality, although a recent study shows that relaxing imputation INFO score from 0.9 to 0.3 has negligible impacts on prediction accuracy[9]. We replicated the N_{eff} filtering in BBJ, and confirmed that improved $R^2_{liability}$ attributable to N_{eff} filtering was also observed (**Figure S5-D**). Although the effect of the N_{eff} filtering was diminished by the MAF filtering in relatively stringent thresholds (*p*-value < 5 × 10^{-4}), the effect was still observed in the other thresholds (**Figure S5-E**). Using only HapMap3 SNPs almost completely reduced the dependency of $R^2_{liability}$ on the N_{eff} (**Figure S5-F**).

Overall, we found the prediction performance of P+T to be affected by a combination of factors, with *p*-value thresholds showing larger effects as compared to other parameters, such as LD window sizes, LD r^2 thresholds, and variant filtering by N_{eff} or MAF. Moreover, the optimal *p*-value threshold varied substantially between different endpoints in GBMI. We also demonstrated that restricted use of HapMap3 SNPs showed comparable or better prediction accuracy relative to using genome-wide common SNPs for P+T, particularly for GWAS from diverse cohorts as in GBMI with genetic variants showing considerable heterogeneity in effective sample sizes.

Bayesian approaches for calculating PRS improve accuracy

We also evaluated fully genome-wide polygenic risk scores, by first fine-tuning the parameters in PRS-CS. We ran PRS-CS using both the grid model and automated optimization model

(referred to as auto model), the former of which specifies a global shrinkage parameter (phi, in which smaller values indicate less polygenic architecture and vice versa for larger values), with 1KG-EUR as the LD reference panel. We note that the optimized phi parameter with highest prediction accuracy in the grid model differed among traits (**Figure S6**). Specifically, we found that for more polygenic traits (as estimated using SBayesS) including asthma, COPD and stroke (**Figure 2**), the optimal phi parameter was 1×10^{-3} in EUR (**Figure S6**). There was no significant difference between prediction accuracy using the optimal grid model versus auto model (**Figure S6**), which suggests PRS-CS can learn the phi parameter from discovery GWAS well when its sample size is considerably large. Therefore, we hereafter used the auto model because of its computational efficiency. Across target ancestral populations in the UKBB, PRS from EUR-based LD reference panels showed significantly higher or comparable prediction accuracies compared to PRS using other ancestry-based LD reference panels (**Figure S7-A**). This result suggests that it is reasonable to use a EUR-based LD reference panel in GBMI because EUR ancestry constitutes the largest proportion of GWAS participants (~69%). Note that we also compared the prediction accuracy of LD reference panels derived from UKBB-EUR, which has a much larger sample size, against 1KG-EUR and found no significant difference (**Figure S7-B**). These results suggest that PRS-CS is not sensitive to the sample size of the LD reference panel, which is consistent with previous findings[24].

We then compared the optimal prediction accuracy of P+T versus the PRS-CS auto model in the UKBB and BBJ and found that PRS-CS showed overall better prediction performance for traits with higher h^2_{SNP} but no to slight improvements for traits with lower h^2_{SNP} (**Figure 4**). Specifically, the highest improvement of PRS-CS relative to that of P+T in EUR was observed for HF, of 66.1%, followed by stroke (62.7%) and COPD (55.9%). Substantial increments were observed for HF (105.2%), COPD (102.5%) and Stroke (41.6%) in EAS. A 92.5% and 74.0% improvement was shown for asthma in CSA and AFR, respectively. P+T on the contrary saw better prediction performance over PRS-CS for a few trait-ancestry comparisons, such as the largest relative improvements of 60.5% for IPF and 50.2% for UtC in EAS. Compared with P+T, which requires tuning *p*-value thresholds and is affected by variant-level quality controls such as N_{eff}, there is no need to tune prediction parameters using the PRS-CS auto model, thus reducing the computational burden.

Overall, after examining 13 disease endpoints, these results favor the use of PRS-CS for developing PRS from multi-ancestry GWAS of primarily European-samples, which is also

consistent with previous findings that Bayesian methods generally show better prediction accuracy over P+T across a range of different traits[9,24].

PRS accuracy is heterogeneous across ancestries and biobanks

For each of the participating biobanks, we used leave-one-out meta-analysis as the discovery GWAS to estimate the prediction performance of PRS in each biobank (see STAR Methods). The disease prevalence and effective sample size of each biobank is shown in **Figure S8**. Generally, the PRS prediction accuracy of different traits increased with larger h^2_{SNP} (**Figure 5 and Table S4**). For example, the average R^2 on the liability scale across biobanks (hereafter denoted as $\overline{R^2_{liability}}$, see STAR Methods) in EUR ranged from 1.0% for HF, ~2.2% for COPD and ThC to 3.8% for gout and 4.6% for asthma. Notably, accuracy was sometimes heterogeneous across biobanks within the same ancestry for some traits. Specifically, the $R^2_{liability}$ for asthma in ESTBB and BioVU was significantly lower than $\overline{R^2_{liability}}$, which might be attributable to between-biobank differences such as recruitment strategy, phenotyping, disease prevalence, and environmental factors. The prediction accuracy was generally lower in non-European ancestries compared to European ancestries, especially in African ancestry, which is mostly consistent with previous findings[25–27] with a few exceptions. For example, we observed comparable prediction accuracy for gout in EAS relative to that in EUR, which could be reflected by large effective sample sizes and some gout-associated SNPs with large effects exhibiting higher allele frequencies in EAS (**Figure S9**). For example, the MAFs of gout top-associated SNP, rs4148157, were 0.073 in 1KG-EUR and 0.25 in 1KG-EAS, respectively, and the phenotypic variance explained by that SNP in EAS (8.3%) was more than twice as high as that in EUR (3.0%). The accuracy of PRS to predict asthma risks in AMR was found to be significantly higher than that in EUR, which could be due to the small sample size in AMR (**Table S4**). Thus, further validation is needed in larger AMR population cohorts.

The ability of PRS to stratify individuals with higher disease risks was also found to be heterogeneous across biobanks and ancestries as shown in **Figure 6** and **Table S5**. We showed that the PRS distribution across different biobanks slightly varied. Specifically, we calculated the absolute difference of median PRS in each decile for each endpoint between biobanks for cases and controls, separately, and found that the largest absolute differences were 0.06 and 0.21 for stroke controls and stroke cases, respectively (**Figure S10**). This justifies

the comparison of odds ratios (ORs) in terms of relative risks. The ORs between the top 10% and bottom 10% were more heterogeneous between biobanks and also higher relative to other comparisons (e.g., top 10% vs middle and other strata). This is consistent with previous studies where OR reported between tails of the PRS distribution is generally inflated relative to those between top ranked PRS and general populations[11]. We measured the variation of OR between biobanks using the coefficient of variation of OR ($CoeffVar_{OR}$, see STAR Methods). The largest $CoeffVar_{OR}$ in EUR was observed for ThC of 0.46 between top 10% and bottom 10% as compared to 0.27 and 0.23 for top 10% vs middle and other, respectively. We recapitulated the findings using $R^2_{liability}$ that ORs were overall higher for traits with higher h^2_{SNP} and also higher in EUR than non-EUR ancestries, which is expected as the two accuracy metrics are interrelated. For example, the averaged ORs across biobanks weighted by the inverse variance in EUR (see STAR Methods) for gout were 4.6, 2.4 and 2.2 for the top 10% vs bottom 10%, middle and other strata, separately. The corresponding estimates in EUR for stroke were 1.6, 1.3 and 1.3, respectively. Across ancestries, the average OR of asthma between the top 10% and bottom 10% ranged from 4.1 in EUR to 2.4 in AFR.

Overall, the predictive performance of PRS measured by $R^2_{liability}$ and OR was found to be heterogeneous across ancestries. This heterogeneity was also presented across biobanks for traits such as asthma which is considered as a syndrome comprising heterogeneous diseases[28].

GBMI facilitates improved PRS accuracy compared to previous studies

GBMI resources might be expected to improve prediction accuracy due to large sample sizes and the inclusion of diverse ancestries. To explore this, we compared the prediction accuracy achieved by GBMI versus previously published GWAS using the same pipeline to run PRS-CS. As shown in **Figure 7**, the accuracy improvements were most obvious for traits with larger h^2_{SNP} and no to small for traits with lower h^2_{SNP}. Specifically, we calculated the absolute improvement of GBMI relative to that using previously published GWAS and found that on average across biobanks, the largest improvements in EUR were 0.033 for asthma, 0.031 for gout, 0.019 for ThC and 0.017 for COPD. However, PRS accuracy was higher for published GWAS relative to the current GBMI for POAG in EUR and AFR, and COPD in the specific case of Lifelines biobank. We referred to the datasets included in the public GWAS of POAG and found that individuals from diverse datasets of EUR and AFR populations were also part of the discovery

dataset, thus we cannot rule out the possibility of sample overlapping or relatedness between the discovery and target datasets for these populations. This suggests that the PRS evaluation may be biased upwards from the prior GWAS for POAG. Also, the phenotypes of POAG across different biobanks are likely more heterogeneous in GBMI than targeted case-control studies[18,29]. The meta-analysis of GBMI with International Glaucoma Genetics Consortium (IGGC) did not lead to substantially improved prediction performance[29]. Another concern might be the disproportional case/control ratio of POAG in GBMI, of ~27,000 cases and ~1.4M controls, thus POAG-related phenotypes with shared genetics in the controls or possible uncontrolled ancestry differences between cases and controls might confound the GBMI GWAS. A very high heterogeneity for phenotype definitions is also found for COPD, however this does not explain why one biobank alone presents this pattern; a specific environmental or population effect not considered in the broad analysis might affect this particular observation.

Discussion

Here, we share lessons and methodological considerations and provide potential guidelines for best practices when generating PRS in the multi-ancestry GBMI resource. First, because PRS construction methods have thus far been mostly applied in a homogeneous population, ancestry-matched LD reference panels can provide unbiased estimates of LD structure between SNPs. In this study in which European ancestry individuals still account for the majority of the discovery GWAS, we found that a EUR-based LD reference panel provides comparable or better prediction accuracy relative to using other cosmopolitan LD panels for PRS construction methods based on GWAS summary statistics such as P+T and PRS-CS. We expect that this finding is largely generalizable across similar methods for calculating PRS. However, as GWAS ancestry composition becomes increasingly diverse, it will be important to explore how the choice of LD reference panels affects PRS prediction accuracy when using multi-ancestry GWAS as the discovery dataset. Second, for P+T, the best predictor is often obtained through fine-tuning the *p*-value thresholds in a validation dataset, while other LD related parameters, such as LD r^2 and LD window size, are usually arbitrarily specified. Here, we used asthma as an example and found that the prediction accuracy of P+T was much less sensitive to different LD related parameters compared to various *p*-value thresholds. Moreover, the optimal *p*-value threshold varied across phenotypes, likely because of trait-specific genetic architecture, especially the degree of polygenicity measured by SBayeS. However, differences in discovery

GWAS and target dataset such as sample sizes, phenotype definition, disease population prevalence (**Figure S8**) and population characteristics could also contribute to this variation. For PRS-CS, we validated a previous finding that the auto model, without requiring post-hoc tuning of the proportion of SNPs with non-zero effects (phi), showed similar prediction performance relative to the grid model, which requires determining the optimal phi parameter in an independent tuning cohort. Lastly, we explored additional per-variant quality control for P+T by estimating effective sample size (N_{eff}) using both HapMap3 SNPs and genome-wide SNPs. We found considerable heterogeneity for the N_{eff} of genetic variants included in the GBMI analyses, indicating that a filter for variant-specific N_{eff} may improve PRS accuracy when utilizing large-scale multi-ancestry discovery GWAS for prediction (**Figure 3-A**). We found that filtering out variants with extremely small N_{eff} improves prediction performance for P+T especially when using genome-wide SNPs. The current LD reference panels provided by PRS-CS are based on common HapMap3 SNPs only; thus PRS-CS might be less sensitive to variation in per-variant N_{eff}. However, further exploration with genome-wide SNPs is needed. Other quality control procedures on the variant level such as restricting to SNPs with relatively homogeneous LD structure between reference panels and discovery GWAS may improve PRS performance as well[30].

Applying our standardized framework for constructing PRS with PRS-CS in different biobanks, we found that the prediction performance showed great heterogeneity across biobanks and ancestries. As PRS inherently only capture genetic factors, this suggests that other factors such as environmental exposures and demographic history may impact the predictive power of PRS within and across ancestries, an open question for future research and methods development. For example, we found that the $R^2_{liability}$ in OHS was higher overall than in other biobanks, which may be attributed to the more complex relatedness structure in this founder population. Notably, the phenotype definitions, recruitment strategy and disease prevalence also vary to different extents across the biobanks studied here.

By using the large-scale GWAS in GBMI, we found that leveraging the large-scale meta-analysis of GBMI significantly improved the accuracy of PRS over previous studies with smaller sample sizes and less diversity. Overall, traits with higher SNP-based heritability showed greater improvement compared to those with lower SNP-based heritability. This indicates that PRS performance will continually benefit from larger sample sizes and more diverse populations. However, further research is needed to understand more concretely how

the composition of underrepresented populations, including specific ancestries and varying sample sizes, can be modeled alongside current Eurocentric GWAS to best facilitate PRS accuracy and generalizability.

We note a few limitations in our study. First, we chose 1KG-EUR as the LD reference panel because data security practices often preclude the use of individual-level GWAS data across analytical teams. This results in an unavoidable mismatch of LD information between the discovery and target datasets, which might affect SNP effect size estimates and thus prediction performance. Further efforts are required to provide more appropriate LD reference panels, such as utilizing the large-scale UKBB with individual-level genotypes to construct a panel with matched ancestry proportions to GBMI. Second, we have focused on common SNPs, specifically HapMap3 SNPs for PRS-CS. As a result, information from rarer variants missing in the LD reference panel was not captured in other non-European ancestries, which may explain a small fraction of the loss of accuracy across populations. Third, although a harmonized analysis framework was developed for GBMI, such as phenotype definitions, ancestry assignments, and PRS construction, there remains a multitude of factors that may contribute to heterogeneous accuracy across both biobanks and ancestries. These include, but are not limited to, phenotype precision, cohort-level disease prevalence, and environmental factors. Last, we evaluated PRS predictive performance using multi-ancestry GWAS but comparisons with single-ancestry GWAS at sufficient scale would enable us to better understand the specific contributions of ancestry diversity and increasing sample size especially for under-represented ancestries, which also serves as a future direction.

The GBMI resource constitutes remarkable progress in expanding the number of endpoints and ancestry groups studied, laying the groundwork for several future directions for exploration. For example, PRS construction methods that model GWAS summary statistics alongside LD information corresponding to multiple ancestries have shown promising accuracy improvements for some traits[16,31], but early investigation into one of these methods has yielded marginal improvement in both European and non-European ancestries for asthma in GBMI[32]. Apart from utilizing single-ancestry GWAS as mentioned above, sex-stratified GWAS in GBMI provide us opportunities to explore the role of sex-specific effects as well as the sample size ratio of male/female in prediction performance of PRS across biobanks. In addition to genetic effects, biobank-specific risk factors and environmental exposures provide further opportunities to better understand the heterogeneity in PRS accuracy that we have identified across biobanks and

ancestries[33,34]. Finally, extending these collaboration efforts to more biobanks in the future, particularly those including recently admixed populations, will bring more resolution into those effects that are biobank-specific and ancestry-specific. Studies in recently admixed populations show that GWAS power can be improved by utilizing local ancestry-specific SNP effect estimates and thus have the potential to benefit genetic prediction accuracy and generalizability[35,36,37]. Altogether, these initiatives hold great promise for improving transferability of PRS across biobanks and ancestries by harnessing the phenotypic richness and diversity present in different biobanks.

STAR Methods

Datasets and quality control

Discovery datasets: For each of 14 endpoints, we used GWAS summary statistics from both GBMI and public datasets with summary statistics available in GWAS Catalog if applicable (**Table S1**) as the discovery dataset. We filtered out SNPs with ambiguous variants, tri- and multi-allelic variants and low imputation quality (imputation INFO score < 0.3). For the GBMI discovery datasets, leave-one-biobank-out meta-analysis using the inverse-variance weighted meta-analysis strategy was applied[18].

Target datasets: We used 9 biobanks, i.e., BioBank Japan (BBJ)[38], BioVU[39], Lifelines[40], UK Biobank (UKBB)[41], Ontario Health Study (OHS)[42], Estonian Biobank (ESTBB)[43], FinnGen, Michigan Genomics Initiative (MGI)[44] and Trøndelag Health Study (HUNT)[45], as the target datasets, which were independent from the datasets included in the discovery GWAS. Brief descriptions about these biobanks can be found in Zhou et al.[18]. We removed individuals with genetic relatedness larger than 0.05 and applied the same filters as the discovery GWAS for SNPs. In addition, only common SNPs with MAF > 1% were retained.

Genetic architecture of 14 endpoints in GBMI

SBayesS is a summary-level based method utilizing a Bayesian mixed linear model, which can report key parameters describing the genetic architecture of complex traits[19]. It only requires GWAS summary statistics and LD correlation matrix estimated from a reference panel. We ran SBayesS using the GWAS summary statistics from all 14 endpoints in GBMI, including

meta-analyses on all ancestries and on EUR only in 19 biobanks[18]. We evaluated the SNP-based heritability (h^2_{SNP}), polygenicity (proportion of SNPs with nonzero effects) and the relationship between allele frequency and SNP effects (S). We used the shrunk LD matrix (i.e., a LD matrix ignoring small LD correlations due to sampling variance) on HapMap3 SNPs provided by GCTB software. The LD matrix was constructed based on 50K European individuals from UKBB. Note that we observed inflated SNP-based heritability estimates using effective sample size for each SNP and hence used the total GWAS sample size instead. We used other default settings in the software. We calculated the *p*-value of each parameter using Wald test to evaluate whether it was significantly different from 0. The h^2_{SNP} was further transformed into liability-scale with disease prevalence approximated as the case proportions in the GWAS summary statistics[46].

PRS construction

P+T: P+T is used to clump quasi-independent trait-associated loci within a LD window size using a specific LD r^2 threshold. For fine-tuning the parameters, we first ran P+T in the UKBB with the most powered asthma GWAS using different combinations of LD window sizes (LD_{win} = 250, 500, 1000, and 2000Kb) and LD r^2 thresholds (r^2 = 0.01, 0.02, 0.05, 0.1, 0.2, and 0.05) with the following flags: *--clump-p1 1 --clump-p2 1 --clump-r2* LD_{win} *--clump-kb* r^2 in Plink v1.9[47]. We constructed PRS using *--score* implemented in Plink v1.9 using 13 different *p*-value thresholds (5×10^{-8}, 5×10^{-7}, 1×10^{-6}, 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, 5×10^{-3}, 0.01, 0.05, 0.1, 0.2, 0.5, 1). Both HapMap3 SNPs and genome-wide SNPs were used in the analysis. We optimized the parameters in a subset of 10,000 randomly selected European individuals from UKBB. After that, we generalized the process into other endpoints using LD_{win}=250kb and r^2=0.1 in the UKBB and BBJ with a focus on HapMap3 SNPs only. We further explored how per-variant filtering based on effective sample sizes (N_{eff}) and MAF thresholds would affect the prediction performance. We used three thresholds to retain variants by their N_{eff}: >0%, >50%, and >80% of N_{eff} compared to the total ones and also three MAF filters: 0.01, 0.05 and 0.1.

PRS-CS: PRS-CS[24] is a Bayesian regression framework which enables continuous shrinkage priors on SNP effects to infer their posterior mean effects. We ran PRS-CS using both the grid and auto models in the UKBB. In the grid model, we used a series of global shrinkage parameters (phi = 1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 0.01, 0.1, 1), with lower phi values

suggesting less polygenic genetic architecture and vice versa for more polygenic genetic architecture. For the auto model, PRS-CS will learn the phi parameter from the discovery GWAS without requiring post-hoc tuning. We used both total GWAS sample size and effective sample size as input for PRS-CS and found little difference, suggesting that PRS-CS is insensitive to the input of GWAS sample size. We hence used the effective sample size for subsequent analyses in this study. We used the default settings for other parameters. We generalized the auto model for all endpoints in both UKBB and BBJ.

LD reference panel

Both P+T and PRS-CS are summary-level based PRS prediction methods, utilizing GWAS summary statistics and a LD reference panel. To explore the impact of LD reference panels on prediction performance, we used LD reference panels of different ancestral compositions, varying sample sizes and SNP density. Specifically, we used four global ancestry groups, i.e., European (EUR), South-Asian (SAS), East-Asian (EAS) and African (AFR), from 1000G Phase 3 (1KG) as LD reference panels for P+T. Further, we randomly sampled a subset of individuals with sample sizes of 500, 5000, 10,000 and 50,000 from UKBB-EUR to analyze how the sample sizes of LD reference panel would affect prediction accuracy for P+T. Moreover, we ran P+T on both the HapMap3 SNP set and a denser SNP set with whole genome-wide SNPs. We ran PRS-CS with the LD matrix provided by PRS-CS software[24], which are based on both 1KG and UKBB populations from those four ancestry groups and Admixed American population (AMR). We performed those analyses using most powered asthma GWAS summary statistics in GBMI and evaluated the prediction performance in diverse ancestry groups in the UKBB.

Evaluation of prediction performance

After constructing PRS, we evaluated the prediction performance in the independent target datasets. We used a logistic regression to calculate the Nagelkerke's R^2 and variance on the liability-scale explained by PRS as described previously[46]. Area under the receiver operating characteristic curve (AUC) was also reported for full models with additional covariates and models including PRS only. We used bootstrap with 1000 replicates to estimate their corresponding 95% confidence intervals (CIs). Note that the proportion of cases in each ancestry in the target dataset was approximated as the disease population prevalence. The same covariates (usually age, sex and 20 genotypic principal components, PCs) used in the GWAS analyses were included in the full regression model as phenotype ~ PRS + covariates.

We also calculated the average R^2 on the liability scale across biobanks ($\overline{R^2_{liability}}$) in each ancestry by weighting the effective sample size of each biobank for each endpoint. Further, we divided the target individuals into deciles based on the ranking of PRS distribution. We compared the odds ratio (OR) of the top decile relative to those ranked as the bottom, the middle and the remaining, when using the first decile as the referenced group. For endpoints presented in two or more biobanks, we calculated the averaged OR using the inverse variance weighted method and the coefficient of variation of OR ($CoeffVar_{OR}$) as SD(OR)/mean(OR).

Resource Availability

Data and Code Availability

The all-biobank and ancestry-specific GWAS summary statistics are publicly available for downloading at https://www.globalbiobankmeta.org/resources and browsed at the PheWeb Browser http://results.globalbiobankmeta.org/. The PRS weights re-estimated using PRC-CS-auto for multi-ancestry GWAS including all biobanks and leave-UKBB-out multi-ancestry GWAS will be uploaded to PGS Catalog (https://www.pgscatalog.org/). 1000 Genome Phase 3 data can be accessed at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data. The software used in this study can be found at: Plink (https://www.cog-genomics.org/plink/), PRS-CS (https://github.com/getian107/PRScs), SBayesS/GCTB (https://cnsgenomics.com/software/gctb/). The codes used in this study can be found in the github repository: https://github.com/globalbiobankmeta/PRS.

Acknowledgements

A.R.M is funded by the K99/R00MH117229. E.L. is funded by the Colciencias fellowship ed.783. S.N. was supported by Takeda Science Foundation. Y.O. was supported by JSPS KAKENHI (19H01021, 20K21834), and AMED (JP21km0405211, JP21ek0109413, JP21ek0410075, JP21gm4010006, and JP21km0405217), JST Moonshot R&D (JPMJMS2021, JPMJMS2024), Takeda Science Foundation, and Bioinformatics Initiative of Osaka University Graduate School of Medicine, Osaka University. E.R.G. is supported by the National Institutes of Health (NIH) Awards R35HG010718, R01HG011138, R01GM140287, and NIH/NIA AG068026. V.L.F. was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.675033 (EGRET plus). L. B. and B. B. receive support from the K.G. Jebsen Center for Genetic Epidemiology funded by Stiftelsen Kristian Gerhard Jebsen; Faculty of Medicine and Health Sciences, NTNU; The Liaison Committee for education, research and innovation in Central Norway; and the Joint Research Committee between St Olavs Hospital and the Faculty of Medicine and Health Sciences, NTNU. K.L. and R.M. were supported by the Estonian Research Council grant PUT (PRG687) and by INTERVENE - This

project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016775. W.Z. was supported by the National Human Genome Research Institute of the National Institutes of Health under award number T32HG010464. The work of the contributing biobanks was supported by numerous grants from governmental and charitable bodies. The biobank specific acknowledgements and full author list for GBMI are included in the **Supplementary Notes**.

Author Contributions

Study design: A.M., J.H., Y.O., Y.W.
Data collection/contribution: L.B., P.A., B.B., P.D., K.H., R.M., Y.M., S.S., J.U., C.W., N.J.C., I.S., J.H.
Data analysis:Y.W., S.N., E.L., S.K., K.T., K.L., M.K. W.Z., K.H.W, M.J.F., L.B., V.L.F, J.H.
Writing: Y.W., S.N., E.L., Y.O., A.M., J.H
Revision: Y.W., S.N, E.L., K.T., W.Z., S.S., J.W.S., B.N.W., C.W., E.R.G., N.J.C., Y.O., A.M., J.H.

Declaration of Interests

E.R.G. receives an honorarium from the journal Circulation Research of the American Heart Association as a member of the Editorial Board.

Figures

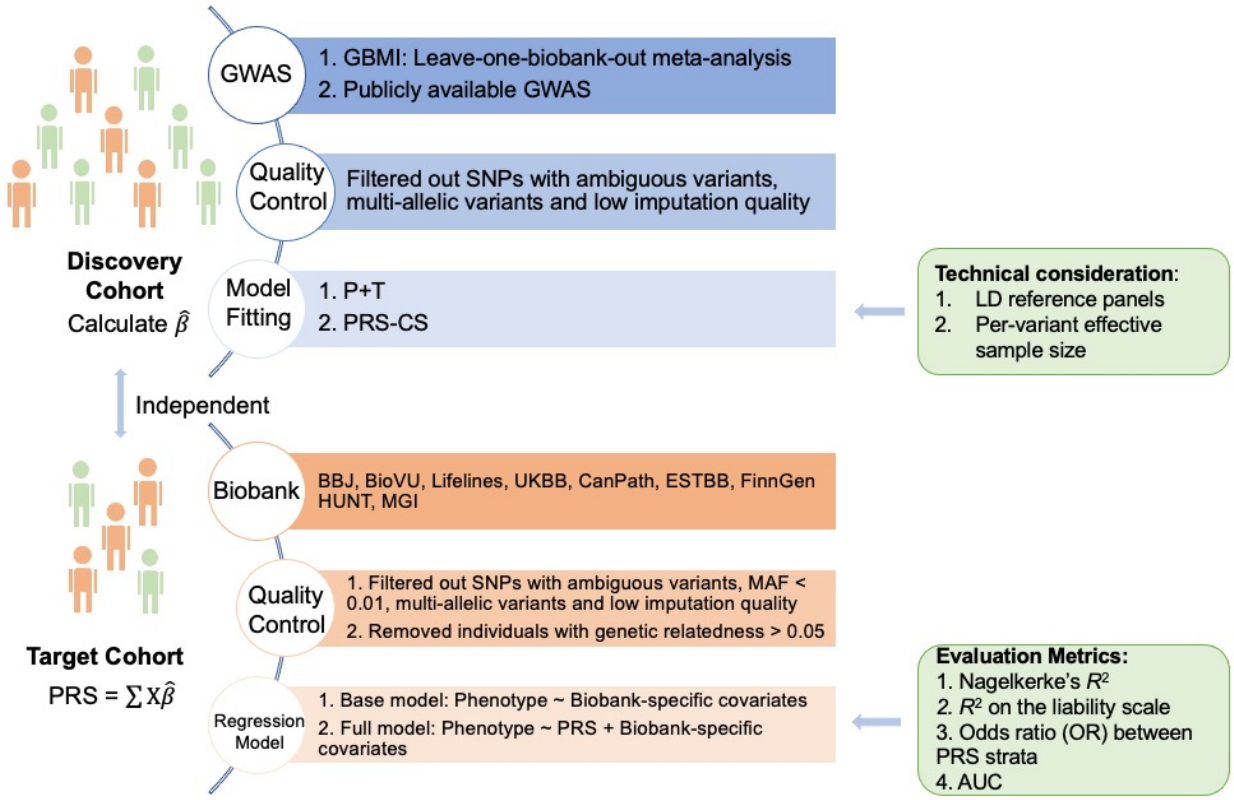

Figure 1. Overview of the study framework.

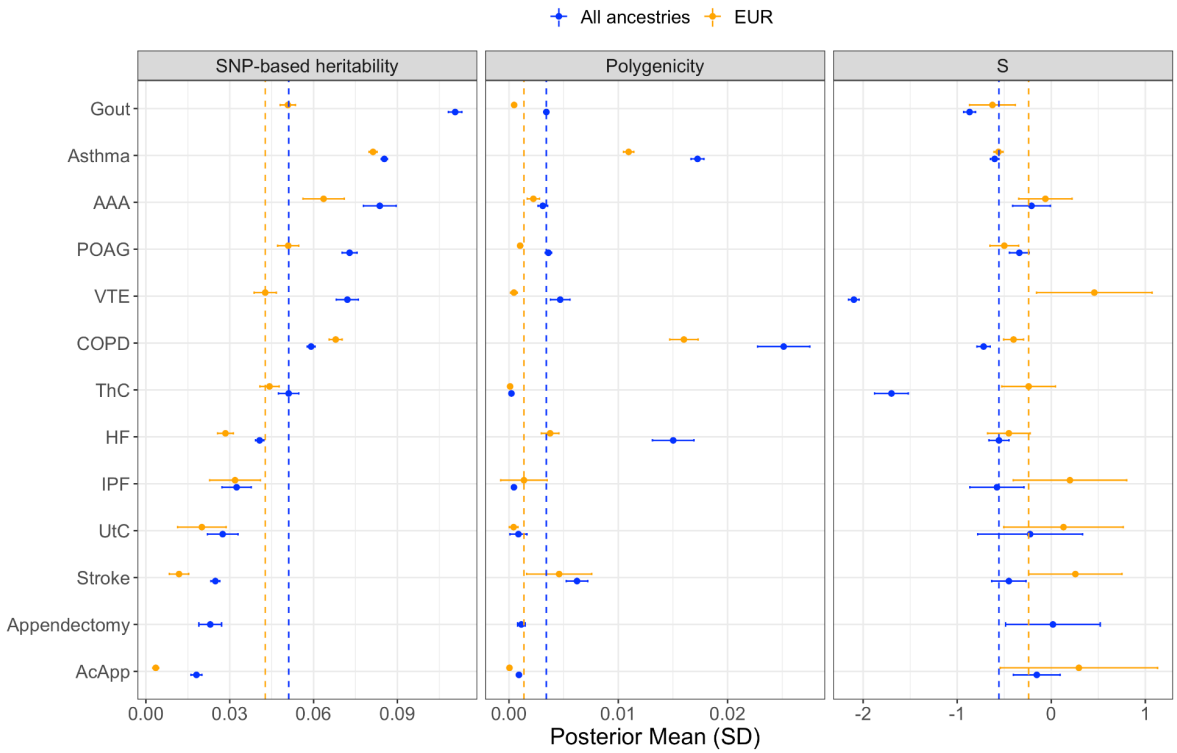

Figure 2. Genetic architecture of endpoints in GBMI. The phenotypes on the y-axis are ranked based on the SNP-based heritability estimates using meta-analysis from all ancestries. Note the SNP-based heritability estimates were transformed on the liability scale. The vertical dashed lines in each panel indicate the corresponding median estimates across 13 endpoints. The results for hypertrophic or obstructive cardiomyopathy (HCM) are not presented as it failed to converge in the SBayesS model.

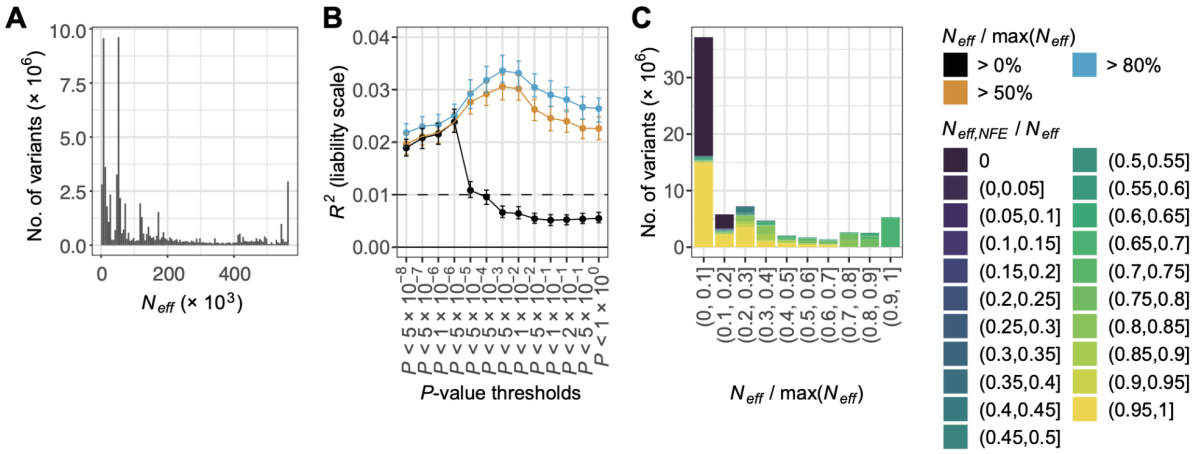

Figure 3. Sample size heterogeneity affects PRS prediction accuracy for P+T. A) the distribution of effective sample sizes (N_{eff}) for asthma as a representative trait. **B**) predictive performance of P+T for European (EUR) samples in the UK Biobank (UKBB). The R^2 for asthma is shown as a representative result. Full results are shown in **Figure S5 and Table S3**. **C**) the ratio of N_{eff} of EUR compared with N_{eff} of all samples for asthma.

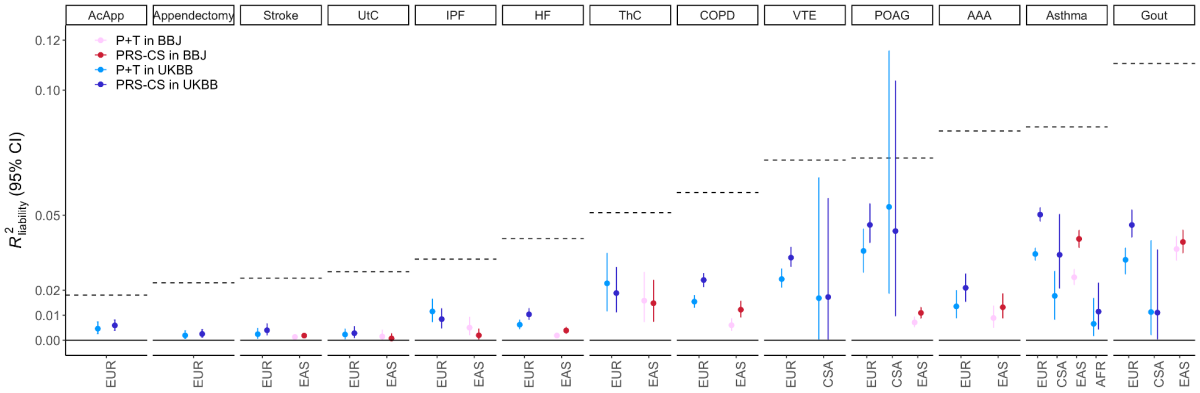

Figure 4. Prediction performance using P+T versus that using PRS-CS. The phenotypes are ranked based on the SNP-based heritability as shown in **Figure 2** (indicated by the dashed line) estimates using all ancestries. Only trait-ancestry pairs with significant accuracies in both P+T and PRS-CS are presented.

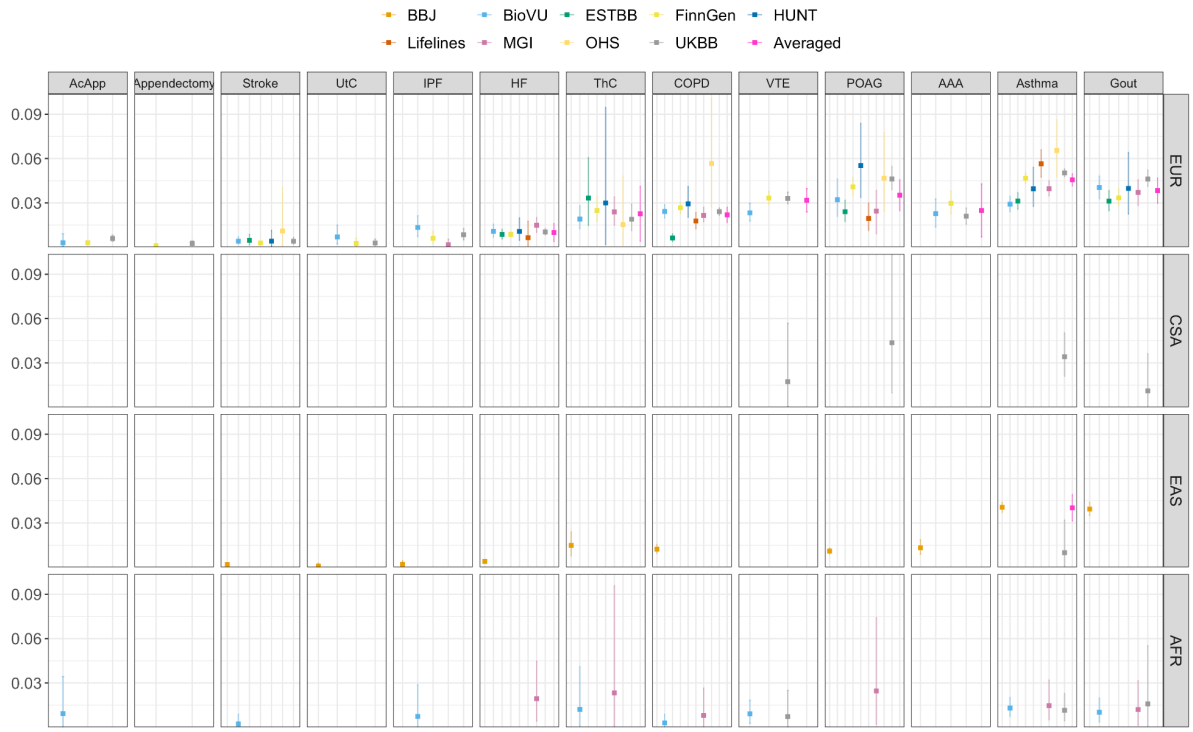

Figure 5. Prediction performance of PRS-CS across biobanks and ancestries. The phenotypes on the y-axis were ranked by the SNP-based heritability using all ancestries as shown in **Figure 2**. Only the significant results were shown. Data for all trait-ancestry pairs in each biobank are provided in **Table S4.** Note that we removed the estimates in AMR and MID due to limited information as a result of small sample sizes.

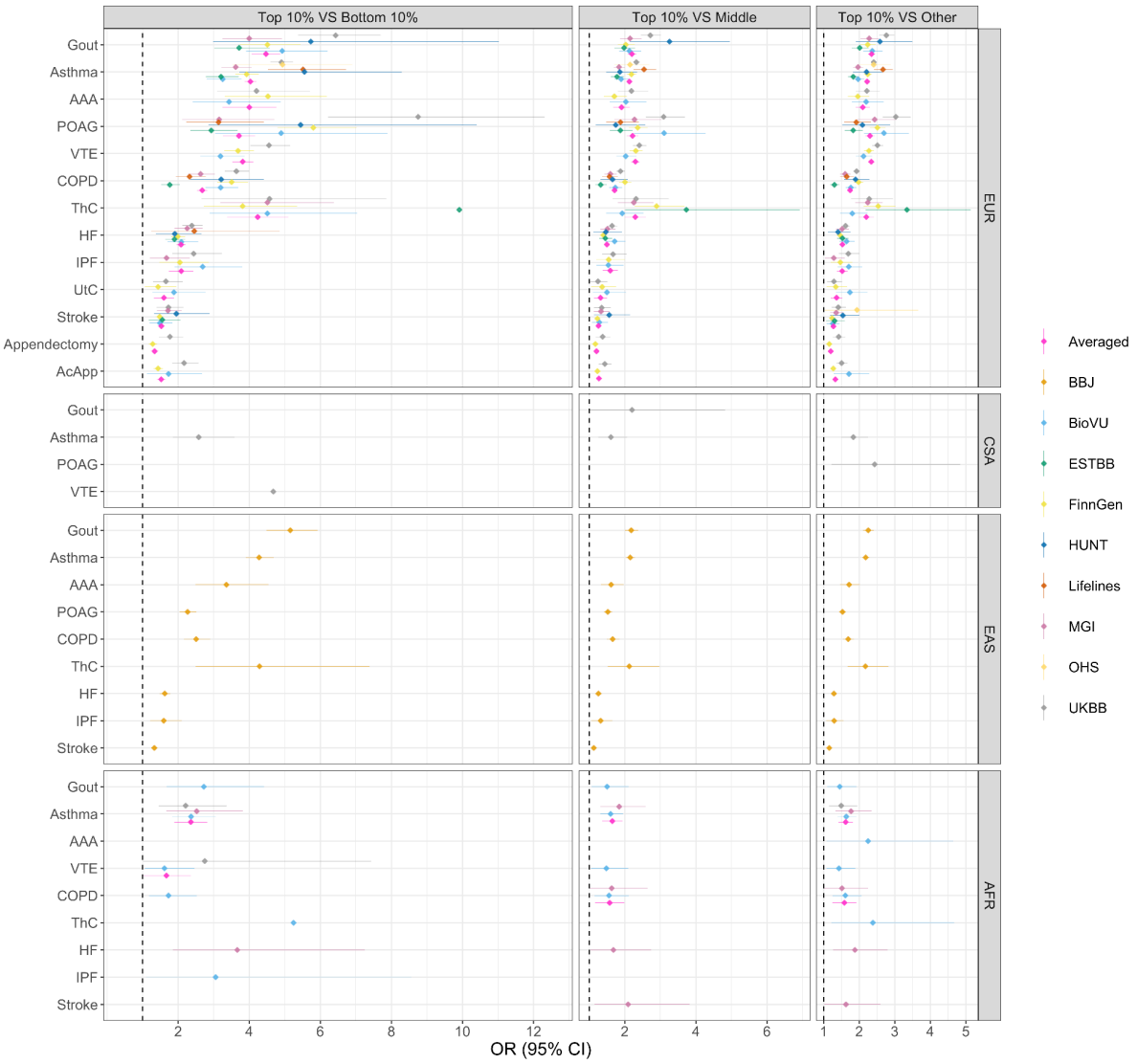

Figure 6. The odds ratio (OR) between different PRS strata for endpoints in GBMI. The dashed line indicates OR=1. Only significant trait-ancestry specific OR was reported, with *p*-value < 0.05. The averaged OR was calculated using the inverse-variance weighted method (see STAR Methods). PRS was stratified into deciles with the first decile (bottom 10%) used as the referenced group. The phenotypes were ranked based on SNP-based heritability estimates using all ancestries (**see Figure 2**).

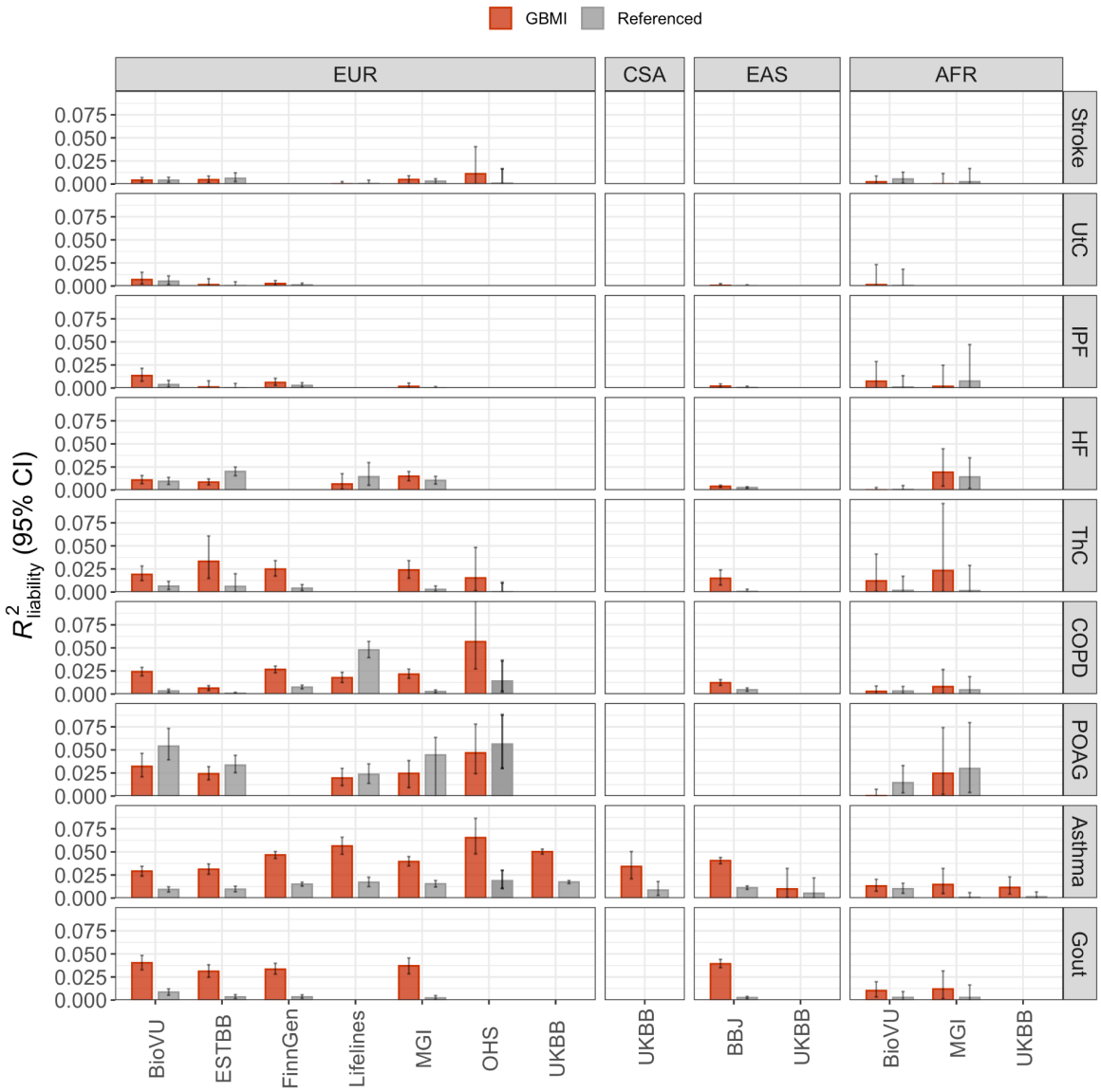

Figure 7. The prediction performance of GBMI versus previously published GWAS. The phenotypes in each ancestry are ranked by the SNP-based heritability estimates from all ancestries. Note that we removed the estimates in AMR and MID due to limited information as a result of small sample sizes.

Supplementary Information

Supplementary Figures

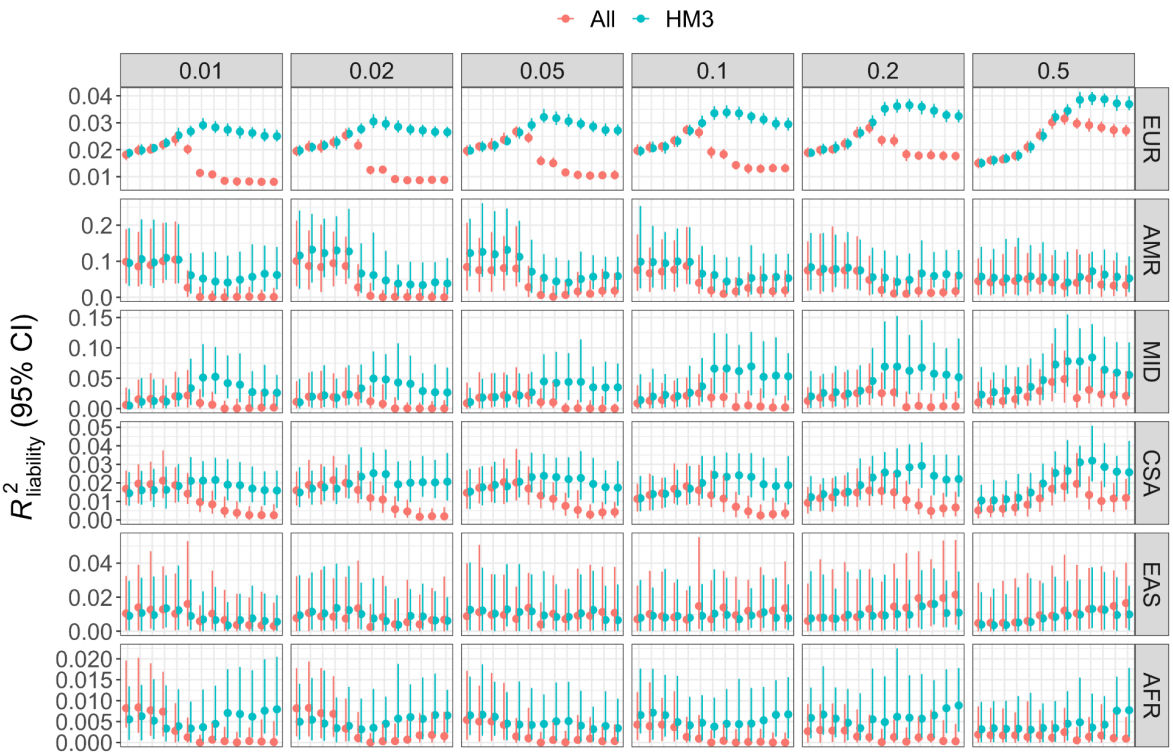

Figure S1. Prediction performance of P+T for asthma using different parameters. We ran P+T with different combinations of *p*-value thresholds (x-axes, 5×10^{-8}, 5×10^{-7}, 1×10^{-6}, 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, 5×10^{-3}, 0.01, 0.05, 0.1, 0.2, 0.5 and 1), LD r^2 thresholds (in columns: r^2=0.01, 0.02, 0.05, 0.1, 0.2, and 0.5) and LD windows (250, 500, 1000, and 2000Kb) for asthma (see STAR Methods). As slight differences between LD parameters were presented, we show the results using an LD window size of 250Kb in the figure and report other results in **Table S2**. Both HapMap3 SNP and a denser whole genome-wide SNP sets were analyzed.1KG-EUR was used as the LD reference panel in all analyses. The accuracies were evaluated in the UKBB.

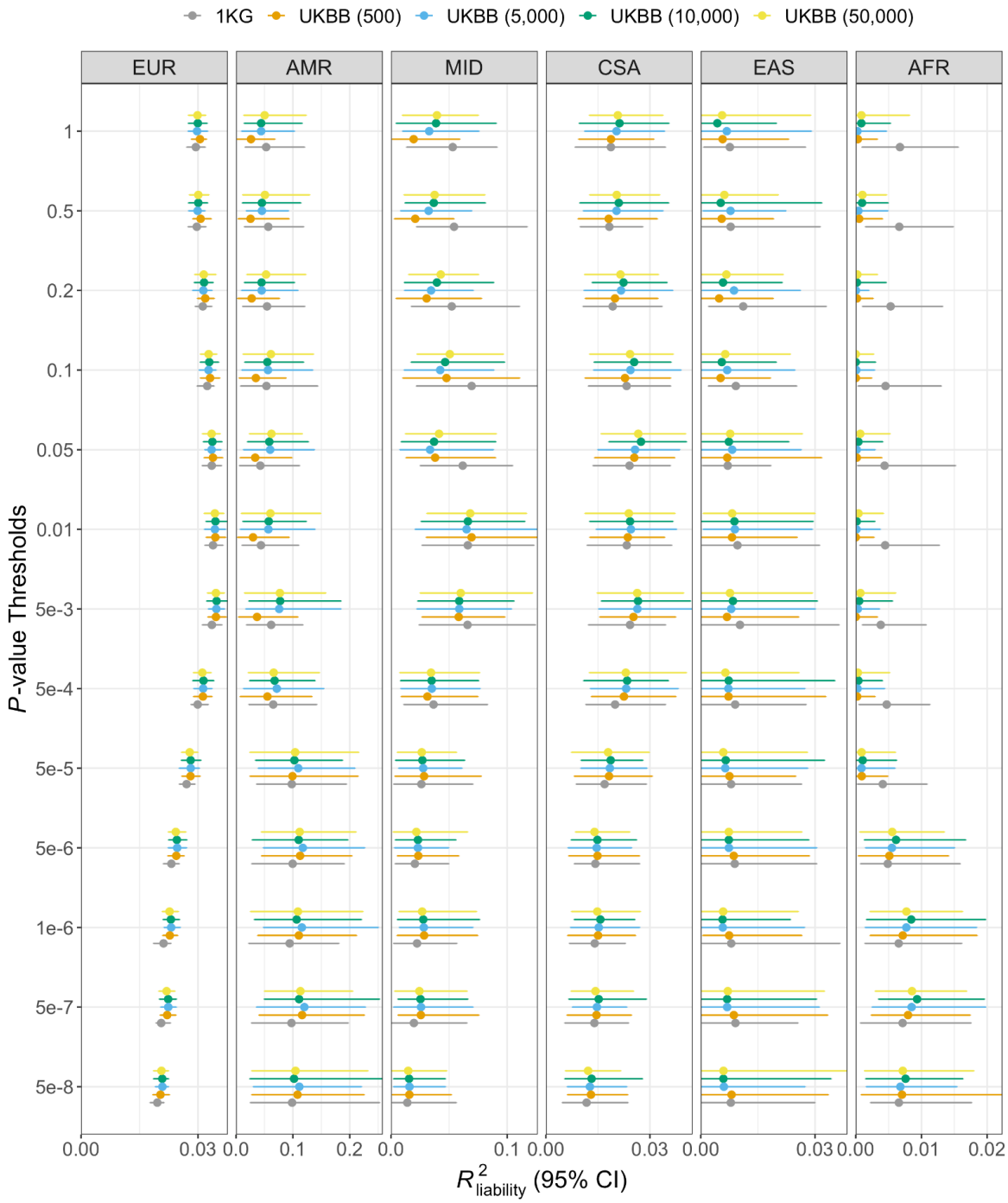

Figure S2. The impact of LD reference panels' sample sizes on P+T prediction performance. We varied the sample sizes of EUR-based LD reference panels from 500 to 50,000. The accuracies were evaluated in the UKBB.

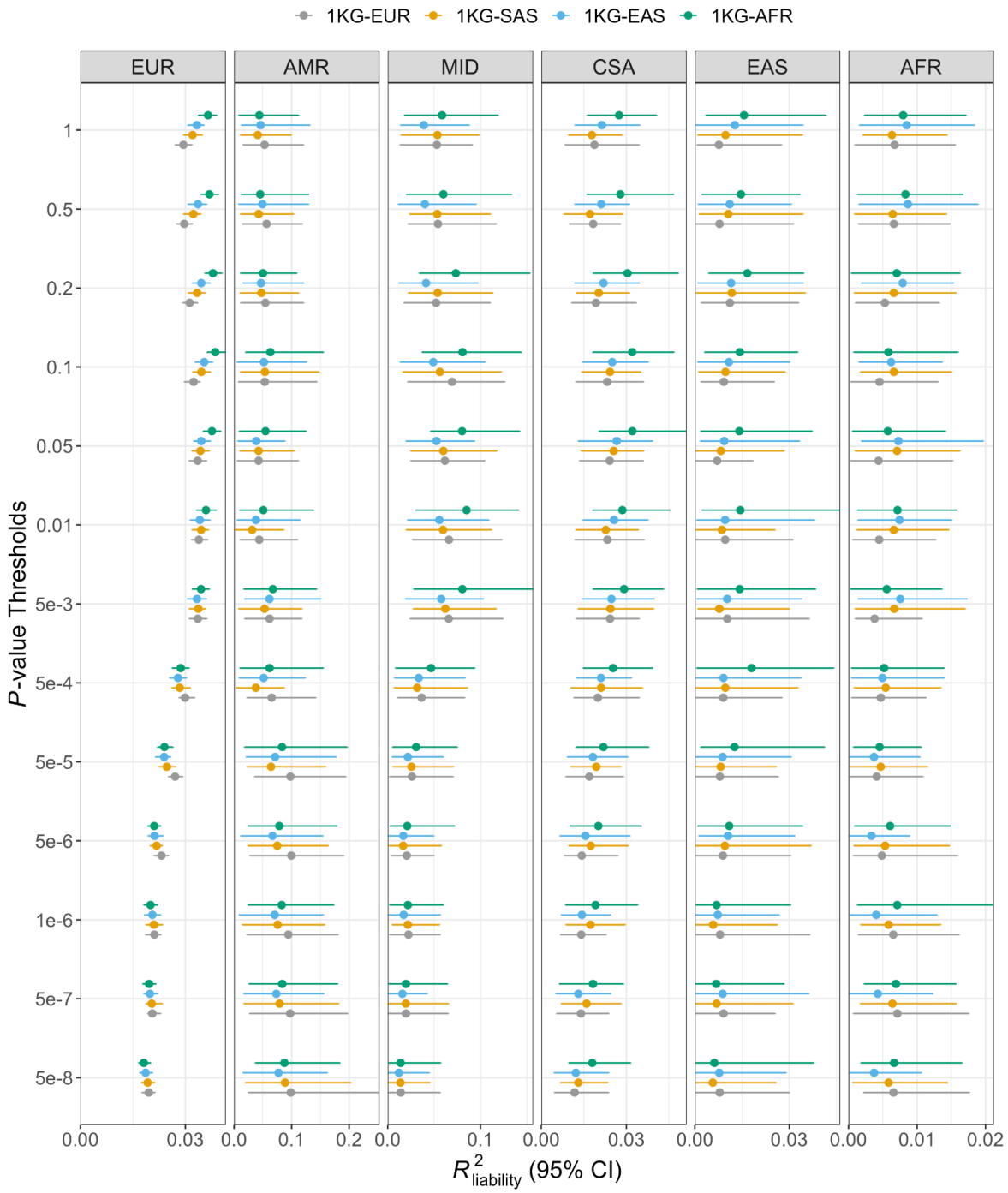

Figure S3. The impact of LD reference panels' ancestral composition on P+T performance. We used different ancestral populations from 1KG as LD reference to run P+T. The accuracies were evaluated in the UKBB.

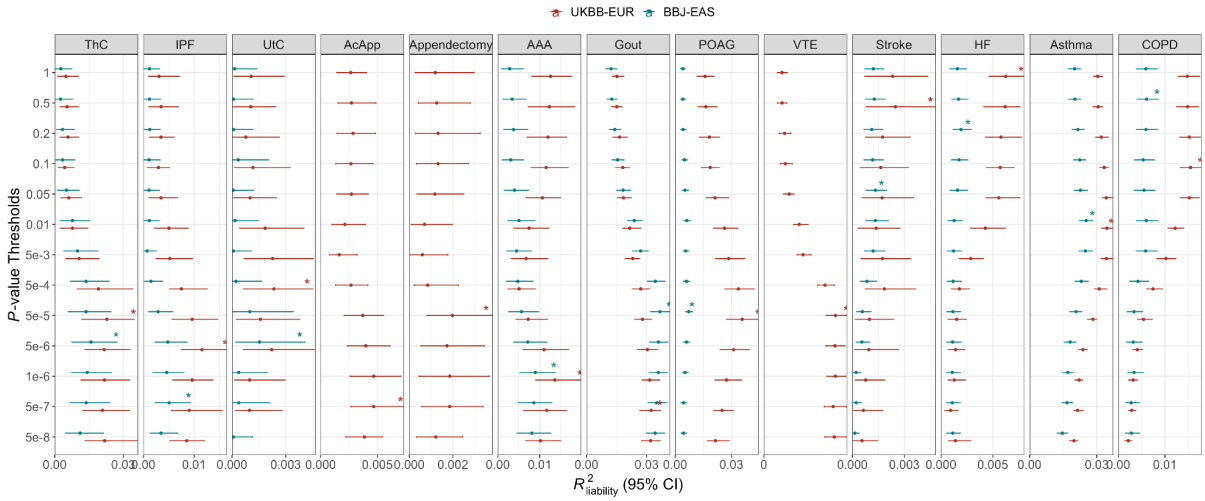

Figure S4. The prediction performance of P+T using different *p*-value thresholds. We evaluated the accuracies in both UKBB-EUR and BBJ-EAS. The asterisks indicate the optimal *p*-value threshold in each endpoint. The phenotypes in columns were ranked based on the polygenicity estimates using all ancestries (**see Figure 2**).

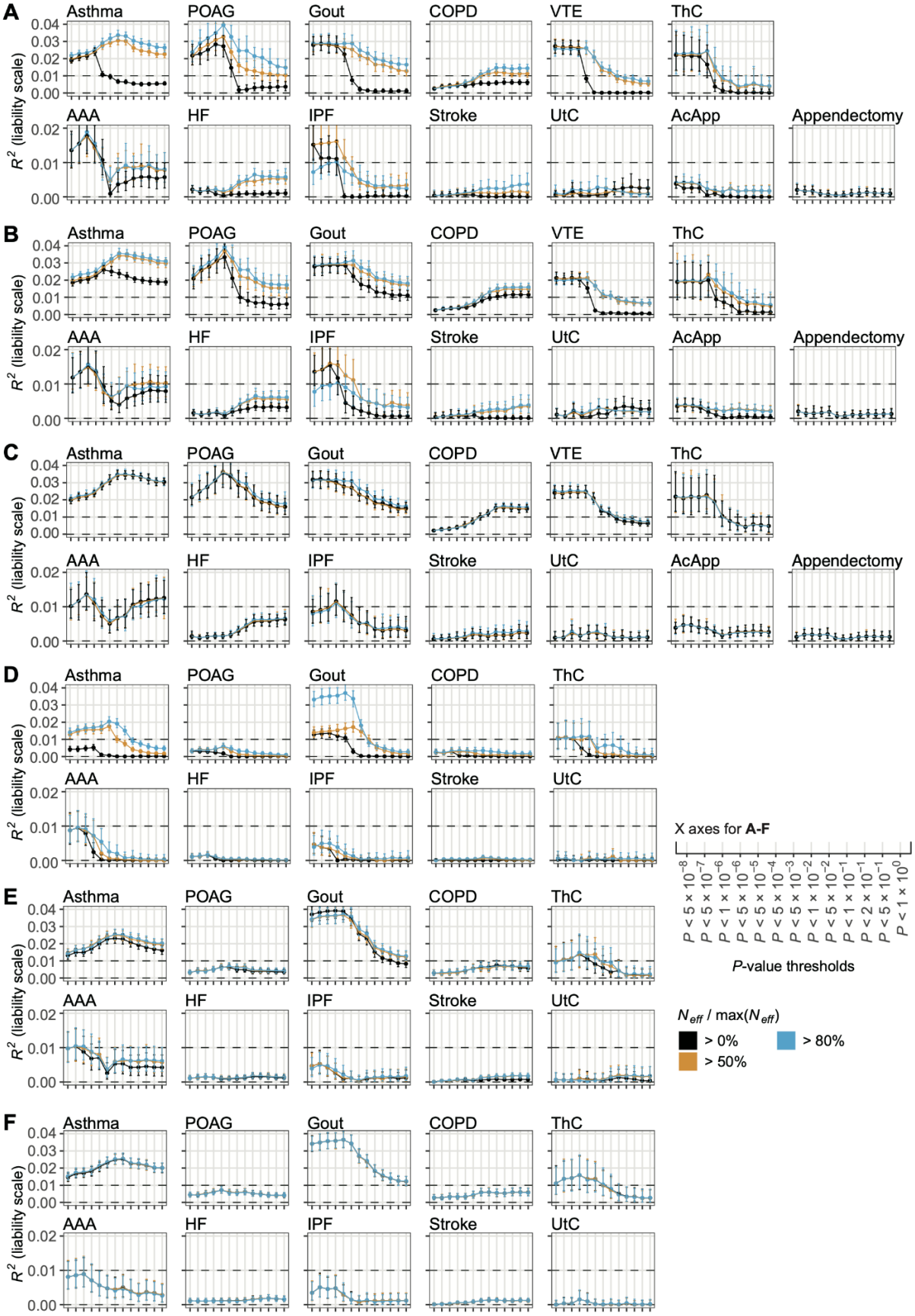
A
B
C
D
E
F
Asthma
POAG
Gout
COPD
VTE
ThC
AAA
HF
IPF
Stroke
UtC
AcApp
Appendectomy
R^2 (liability scale)
X axes for A-F
$P < 5 \times 10^{-8}$
$P < 5 \times 10^{-7}$
$P < 1 \times 10^{-6}$
$P < 5 \times 10^{-6}$
$P < 5 \times 10^{-5}$
$P < 5 \times 10^{-4}$
$P < 5 \times 10^{-3}$
$P < 5 \times 10^{-2}$
$P < 1 \times 10^{-2}$
$P < 5 \times 10^{-2}$
$P < 1 \times 10^{-1}$
$P < 2 \times 10^{-1}$
$P < 5 \times 10^{-1}$
$P < 1 \times 10^{0}$
P-value thresholds
N_{eff} / max(N_{eff})
> 0%
> 50%
> 80%

Figure S5. PRS predictive performance for P+T stratified by sample size heterogeneity. The R^2 of P+T for 13 endpoints for European samples in the UK Biobank (UKBB) (**A**-**C**) and East Asian samples in Biobank Japan (BBJ) (**D**-**F**). Clinical information for VTE, AcApp, and Appendectomy was not collected in BBJ. **A** and **D** show the R^2 of PRS without filtering by minor allele frequency (MAF), while the variants with MAF less than 0.1 were excluded for PRS calculation in **B** and **E**. The HapMap3 variants were used for PRS calculation in **C** and **F**. The full results showing the effect of per-variant N_{eff} and MAF filtering are shown in **Table S3**.

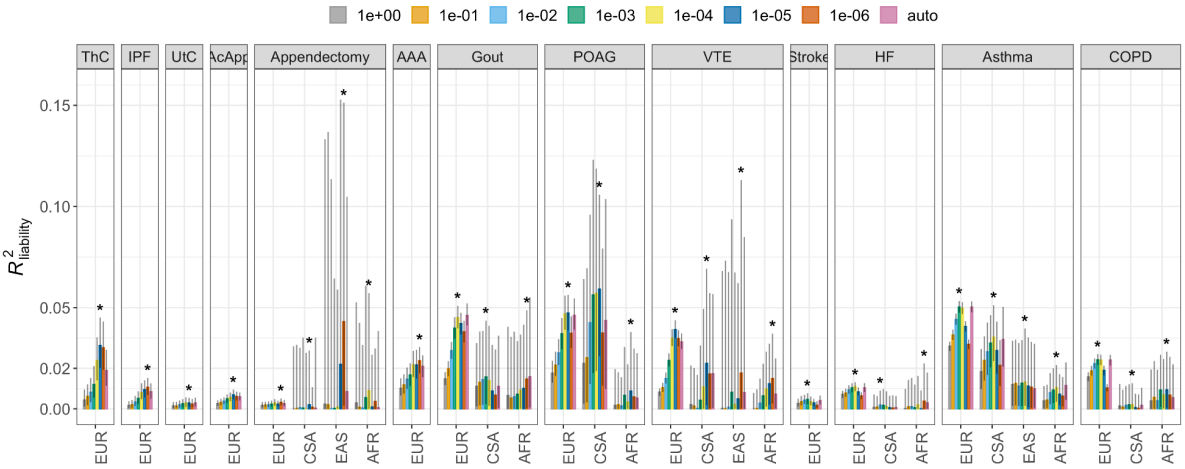

Figure S6. Prediction performance of different PRS-CS models. We ran PRS-CS using both grid model and auto model (see STAR Methods). The asterisks indicate the optimized phi parameter with highest prediction accuracy achieved by grid model in each target ancestry in the UKBB. The phenotypes were ranked by the polygenicity using all ancestries as shown in **Figure 2**. Note that we removed the estimates in AMR and MID due to limited information as a result of small sample sizes.

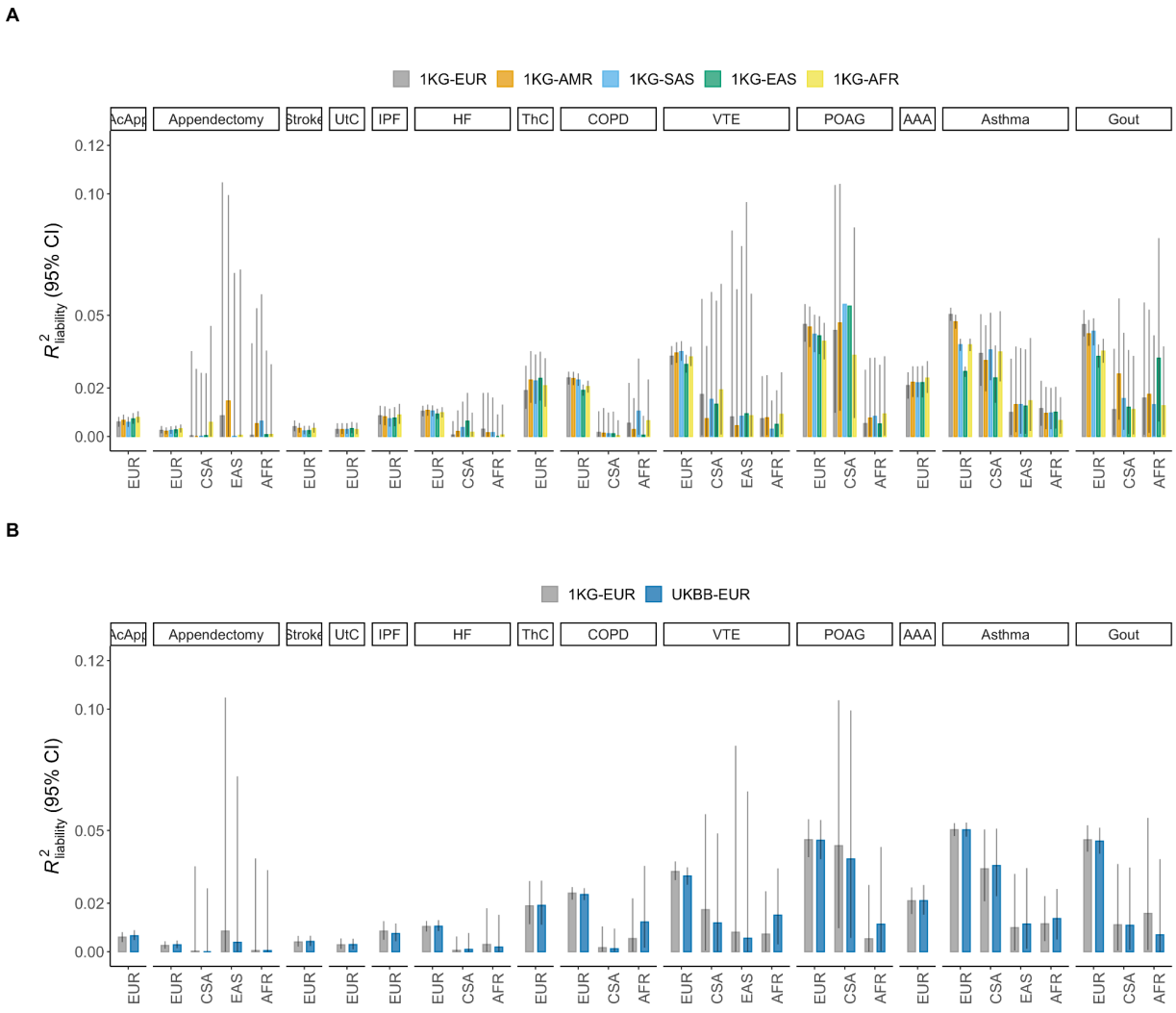

Figure S7. The impact of LD reference panels on prediction accuracy using PRS-CS auto models. A) We used LD references from diverse ancestral populations in 1KG for running PRS-CS auto models. **B)** We used EUR LD reference from both 1KG and UKBB with different sample sizes. The phenotypes were ranked by the SNP-based heritability using all ancestries as shown in **Figure 2**. Note that we removed the estimates in AMR and MID due to limited information as a result of small sample sizes.

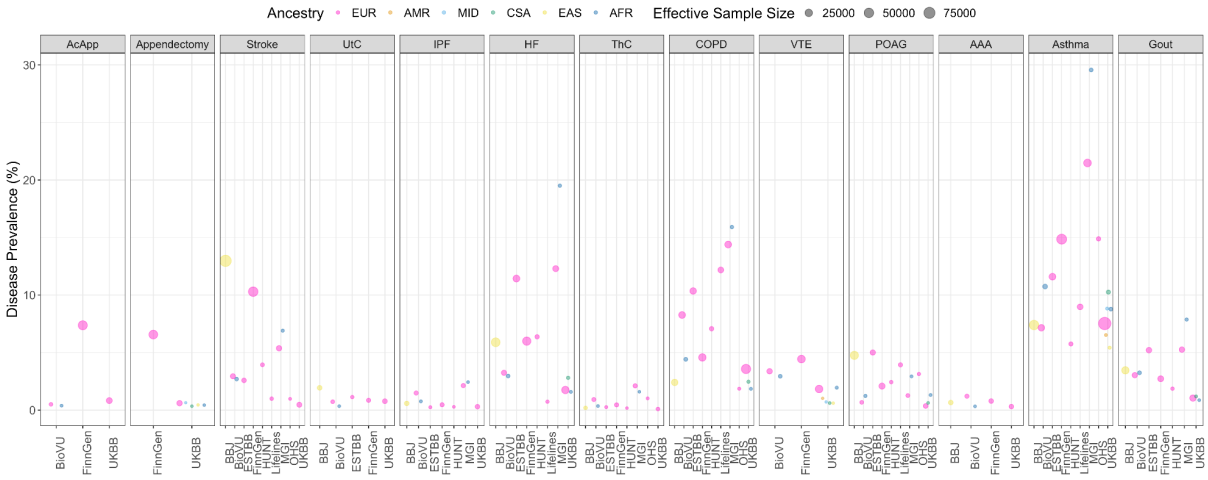

Figure S8. The population prevalence and effective sample size of endpoints in GBMI for each biobank.

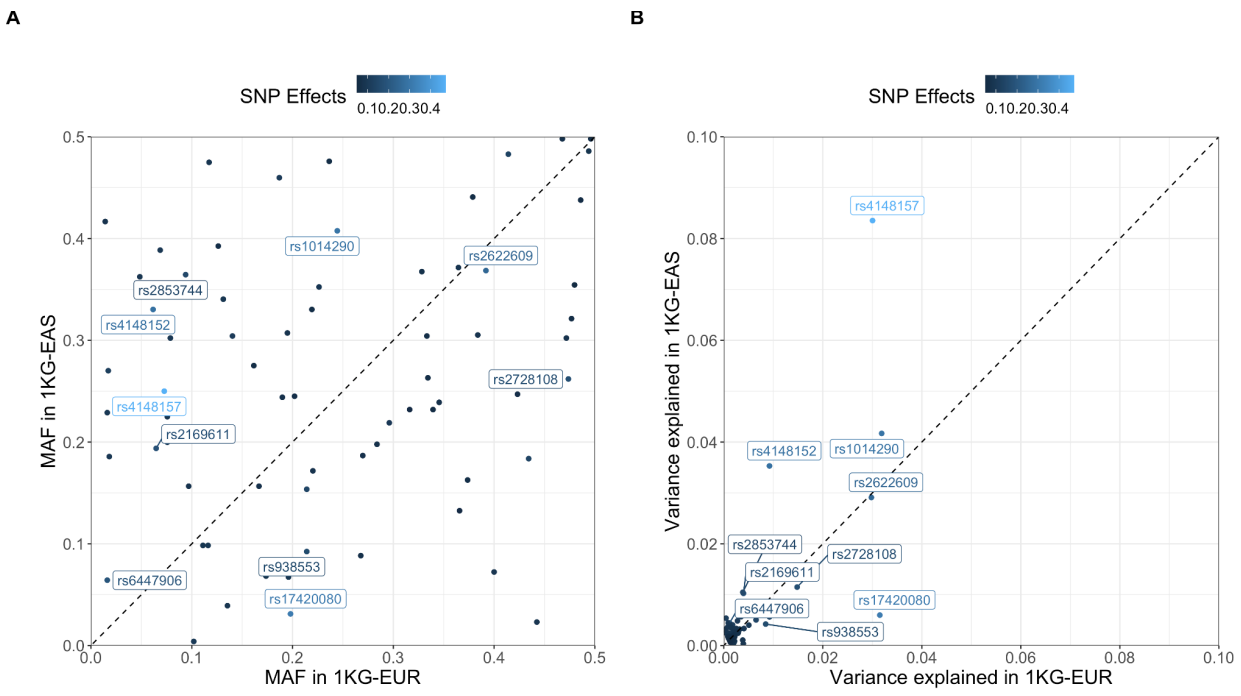

Figure S9. MAF and variance explained for gout-associated genome-wide significant SNPs in 1KG. We showed **A)** The MAF distribution **B)** variance explained for gout-associated GWS SNPs. The genome-wide significant SNPs were identified using P+T (see STAR Methods). The dashed line denotes y = x. The labelled SNPs have top 10 ranked SNP effects.

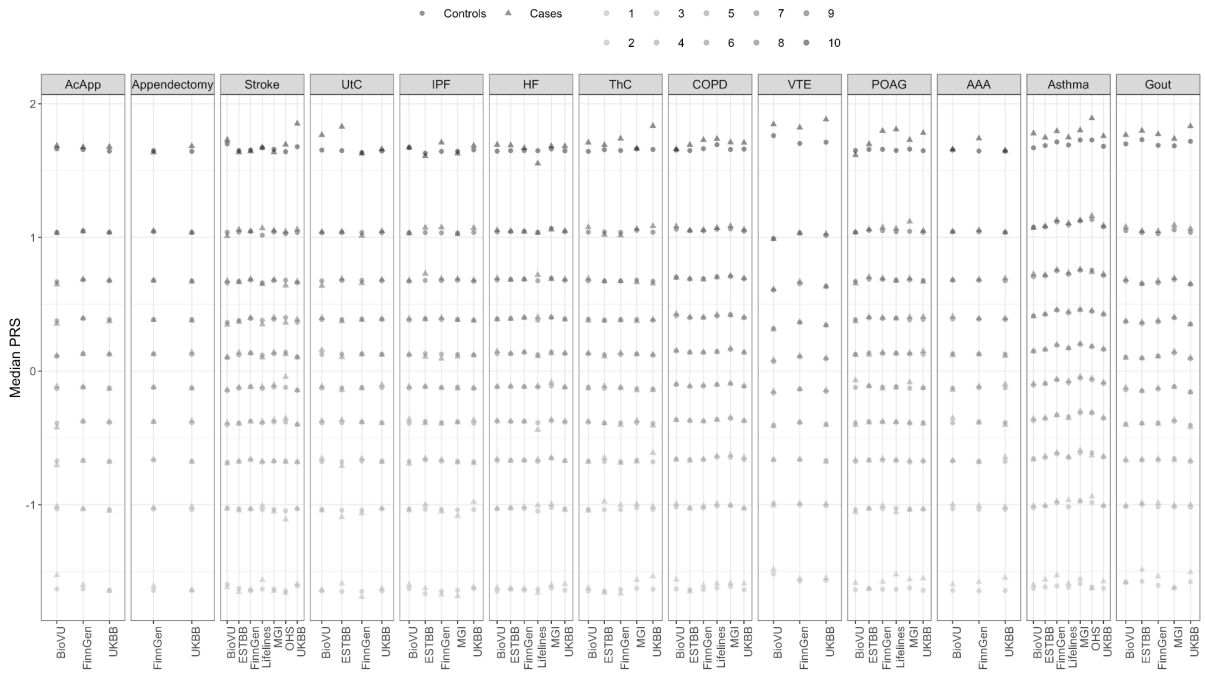

Figure S10. The distribution of median PRS across biobanks in EUR. PRS was splitted into deciles while PRS in controls were normalized with mean of 0 and variance of 1.

Supplementary Tables

Table S1. Public GWAS used in comparison to GBMI GWAS.

Endpoints	Abbreviations	Public GWAS [Reference]
Asthma	Asthma	[48]
Chronic obstructive pulmonary disease	COPD	[49]
Heart Failure	HF	[50]
Stroke	Stroke	[51]
Acute appendicitis	AcApp	
Venous thromboembolism	VTE	
Gout	Gout	[52]
Appendectomy	Appendectomy	
Primary open-angle glaucoma	POAG	[53]
Uterine cancer	UtC	[54]
Abdominal aortic aneurysm	AAA	
Idiopathic pulmonary fibrosis	IPF	[55]
Thyroid cancer	ThC	[54]
Cardiomyopathy (hypertrophic, obstructive)	HCM	

Table S2. Prediction performance for asthma-PRS in the UKBB using different P+T parameters.

Table S3. The impact of per-variant effective sample size on PRS prediction performance.

Table S4. Prediction performance across biobanks in 13 endpoints in GBMI.

Table S5. Odds ratio between PRS distributions across biobanks in 13 endpoints in GBMI.

References

1. Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. Cell. 2019 Mar 21;177(1):58–69.

2. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J Am Coll Cardiol. 2018 Oct 16;72(16):1883–93.

3. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018 Sep;19(9):581–90.

4. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020 May 18;12(1):44.

5. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019 Jan 3;104(1):21–34.

6. Landi I, Kaji DA, Cotter L, Van Vleck T, Belbin G, Preuss M, et al. Prognostic value of polygenic risk scores for adults with psychosis. Nat Med. 2021 Sep 6;1–6.

7. Dudbridge F, Pashayan N, Yang J. Predictive accuracy of combined genetic and environmental risk scores. Genet Epidemiol. 2018 Feb;42(1):4–19.

8. Craig JE, Han X, Qassim A, Hassall M, Cooke Bailey JN, Kinzy TG, et al. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. Nat Genet. 2020 Jan 20;52(2):160–6.

9. Ni G, Zeng J, Revez JA, Wang Y, Zheng Z, Ge T, et al. A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. Biol Psychiatry. 2021 Nov 1;90(9):611–20.

10. Ma Y, Zhou X. Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. 2021 Nov;37(11):995–1011.

11. Kulm S, Marderstein A, Mezey J. A systematic framework for assessing the clinical impact of polygenic risk scores. medRxiv. 2021. Available from: https://www.medrxiv.org/content/10.1101/2020.04.06.20055574v2.full-text

12. Majara L, Kalungi A, Koen N, Zar H, Stein DJ, Kinyanda E, et al. Low generalizability of polygenic scores in African populations due to genetic and environmental diversity. bioRxiv. 2021. p. 2021.01.12.426453. Available from: https://www.biorxiv.org/content/10.1101/2021.01.12.426453v1

13. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584–91.

14. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. Elife. 2020 Jan

30;9:e48376.

15. Martin AR, Daly MJ, Robinson EB, Hyman SE, Neale BM. Predicting Polygenic Risk of Psychiatric Disorders. Biol Psychiatry. 2019 Jul 15;86(2):97–109.

16. Ruan Y, Anne Feng Y-C, Chen C-Y, Lam M, Sawa A, Martin AR, et al. Improving polygenic prediction in ancestrally diverse populations. medRxiv. 2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.12.27.20248738

17. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot W, Khera A, et al. Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores. medRxiv. 2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.01.19.21249483

18. Zhou W, Kanai M, Wu K-HH, Humaira R, Tsuo K, Hirbo JB, et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human diseases. medRxiv. 2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.11.19.21266436

19. Zeng J, Xue A, Jiang L, Lloyd-Jones LR, Wu Y, Wang H, et al. Widespread signatures of natural selection across human complex traits and functional genomic categories. Nat Commun. 2021 Feb 19;12(1):1164.

20. O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. Am J Hum Genet. 2019 Sep 5;105(3):456–76.

21. Zhang Y, Qi G, Park J-H, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat Genet. 2018 Sep;50(9):1318–26.

22. Ware EB, Schmitz LL, Faul J, Gard A, Mitchell C, Smith JA, et al. Heterogeneity in polygenic scores for common human traits. bioRxiv. 2017. p. 106062. Available from: https://www.biorxiv.org/content/10.1101/106062v1

23. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020 Sep;15(9):2759–72.

24. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019 Apr 16;10(1):1776.

25. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2017 Apr 6;100(4):635–49.

26. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019 Jul 25;10(1):1–9.

27. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun. 2020 Jul 31;11(1):3865.

28. Borish L, Culp JA. Asthma: a syndrome composed of heterogeneous diseases. Ann Allergy Asthma Immunol. 2008 Jul;101(1):1–8; quiz 8–11, 50.

29. Lo Faro V, Bhattacharya A, Zhou W, Zhou D, Wang Y, Läll K, et al. Global Biobank Meta-Analysis Initiative: A genome-wide association meta-analysis identifies novel primary open-angle glaucoma loci and shared biology with vascular mechanisms and cell proliferation. In preparation. 2021;

30. Chen W, Wu Y, Zheng Z, Qi T, Visscher PM, Zhu Z, et al. Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. bioRxiv. 2020. p. 2020.07.09.196535. Available from: https://www.biorxiv.org/content/10.1101/2020.07.09.196535v1

31. Márquez-Luna C, Loh P-R, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet Epidemiol. 2017 Dec;41(8):811–23.

32. Tsuo K, Zhou W, Wang Y, Kanai M, Namba S, Gupta R, et al. Multi-ancestry meta-analysis of asthma identifies novel associations and highlights shared genetic architecture across biobanks and traits. In preparation. 2021;

33. Meisner A, Kundu P, Chatterjee N. Case-Only Analysis of Gene-Environment Interactions Using Polygenic Risk Scores. Am J Epidemiol. 2019 Nov 1;188(11):2013–20.

34. Loika Y, Irincheeva I, Culminskaya I, Nazarian A, Kulminski AM. Polygenic risk scores: pleiotropy and the effect of environment. Geroscience. 2020 Dec;42(6):1635–47.

35. Atkinson EG, Maihofer AX, Kanai M, Martin AR, Karczewski KJ, Santoro ML, et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. Nat Genet. 2021 Feb;53(2):195–204.

36. Cavazos TB, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. Human Genetics and Genomics Advances. 2021 Jan 14;2(1):100017.

37. Marnetto D, Pärna K, Läll K, Molinaro L, Montinaro F, Haller T, et al. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. Nat Commun. 2020 Apr 2;11(1):1–9.

38. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol. 2017 Mar;27(3S):S2–8.

39. Bowton EA, Collier SP, Wang X, Sutcliffe CB, Van Driest SL, Couch LJ, et al. Phenotype-Driven Plasma Biobanking Strategies and Methods. J Pers Med. 2015 May 14;5(2):140–52.

40. Scholtens S, Smidt N, Swertz MA, Bakker SJL, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol. 2015 Aug;44(4):1172–80.

41. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct 10;562(7726):203–9.

42. Dummer TJB, Awadalla P, Boileau C, Craig C, Fortier I, Goel V, et al. The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease

prevention. CMAJ. 2018 Jun 11;190(23):E710–7.

43. Leitsalu L, Haller T, Esko T, Tammesoo M-L, Alavere H, Snieder H, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015 Aug;44(4):1137–47.

44. Zawistowski M, Fritsche LG, Pandit A, Vanderwerff B, Patil S, Schmidt EM, et al. The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine patients. In preparation. 2021;

45. Krokstad S, Langhammer A, Hveem K, Holmen TL, Midthjell K, Stene TR, et al. Cohort Profile: the HUNT Study, Norway. Int J Epidemiol. 2013 Aug;42(4):968–77.

46. Lee SH, Goddard ME, Wray NR, Visscher PM. A better coefficient of determination for genetic profile analysis. Genet Epidemiol. 2012 Apr;36(3):214–24.

47. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb 25;4:7.

48. Demenais F, Margaritte-Jeannin P, Barnes KC, Cookson WOC, Altmüller J, Ang W, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat Genet. 2017 Dec 22;50(1):42–53.

49. Kim W, Prokopenko D, Sakornsakolpat P, Hobbs BD, Lutz SM, Hokanson JE, et al. Genome-Wide Gene-by-Smoking Interaction Study of Chronic Obstructive Pulmonary Disease. Am J Epidemiol. 2021 May 4;190(5):875–85.

50. Shah S, Henry A, Roselli C, Lin H, Sveinbjörnsson G, Fatemifar G, et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. Nat Commun. 2020 Jan 9;11(1):1–12.

51. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet. 2018 Mar 12;50(4):524–37.

52. Köttgen A, Albrecht E, Teumer A, Vitart V, Krumsiek J, Hundertmark C, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nat Genet. 2013 Feb;45(2):145–54.

53. Gharahkhani P, Jorgenson E, Hysi P, Khawaja AP, Pendergrass S, Han X, et al. Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. Nat Commun. 2021 Feb 24;12(1):1–16.

54. Rashkin SR, Graff RE, Kachuri L, Thai KK, Alexeeff SE, Blatchins MA, et al. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat Commun. 2020 Sep 4;11(1):1–14.

55. Duckworth A, Gibbons MA, Beaumont R, Wood AR, Almond HP, Lunnon K, et al. A Mendelian randomisation study of smoking causality in IPF compared with COPD. medRxiv. 2020; Available from: https://www.medrxiv.org/content/10.1101/2020.12.04.20243790v1.abstract