#### Global biobank analyses provide lessons for developing polygenic risk scores 1

#### across diverse cohorts 2

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

#### 10 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
- 12345678901234567890123456789012345678901234567 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. Department of Genome Informatics, Graduate School of Medicine, the University of Tokyo, Tokyo 113-0033, Japan.
  - 32. Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita 565-0871, Japan
- 48 49

<sup>&</sup>lt;sup>1</sup><sup>‡</sup>These authors contributed equally.

Lead Contact: Ying Wang (viwang@broadinstitute.org)

<sup>\*</sup>Correspondence: viwang@broadinstitute.org, armartin@broadinstitute.org, jibril.hirbo@vumc.org.

50

# 51 Summary

52 With the increasing availability of biobank-scale datasets that incorporate both genomic data and 53 electronic health records, many associations between genetic variants and phenotypes of interest 54 have been discovered. Polygenic risk scores (PRS), which are being widely explored in precision 55 medicine, use the results of association studies to predict the genetic component of disease risk 56 by accumulating risk alleles weighted by their effect sizes. However, few studies have thoroughly 57 investigated best practices for PRS in global populations across different diseases. In this study, 58 we utilize data from the Global-Biobank Meta-analysis Initiative (GBMI), which consists of 59 individuals from diverse ancestries and across continents, to explore methodological 60 considerations and PRS prediction performance in 9 different biobanks for 14 disease endpoints. 61 Specifically, we constructed PRS using heuristic (pruning and thresholding, P+T) and Bayesian 62 (PRS-CS) methods. We found that the genetic architecture, such as SNP-based heritability and 63 polygenicity, varied greatly among endpoints. For both PRS construction methods, using a 64 European ancestry LD reference panel resulted in comparable or higher prediction accuracy 65 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 66 67 outperformed the classic P+T method, especially for endpoints with higher SNP-based heritability. 68 For example, substantial improvements are observed in East-Asian ancestry (EAS) using PRS-69 CS compared to P+T for heart failure (HF) and chronic obstructive pulmonary disease (COPD). 70 Notably, prediction accuracy is heterogeneous across endpoints, biobanks, and ancestries, 71 especially for asthma which has known variation in disease prevalence across global populations. 72 Overall, we provide lessons for PRS construction, evaluation, and interpretation using the GBMI 73 and highlight the importance of best practices for PRS in the biobank-scale genomics era. 74

# 75 Keywords

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

# 78 Introduction

79 Population- and hospital-based biobanks are increasingly coupling genomic and electronic health 80 record data at sufficient scale to evaluate the potential of personalized medicine<sup>1</sup>. The growth of 81 these paired datasets enables genome-wide association studies (GWAS) to estimate increasingly 82 precise genetic effect sizes contributing to disease risk. In turn, GWAS summary statistics can be used to aggregate the effects of many genetic markers (usually in the form of single-nucleotide 83 84 polymorphisms, SNPs) to estimate individuals' genetic predispositions for complex diseases via 85 polygenic risk scores (PRS). As GWAS power has increased. PRS accuracy has also improved. 86 with PRS for some traits having comparable accuracies to independent biomarkers already 87 routinely used in clinical risk models<sup>2</sup>. Consequently, several areas of medicine have already 88 begun investigating the potential for integrating PRS alongside other biomarkers and information 89 currently used in clinical risk models<sup>3-5</sup>. However, evidence of clinical utility for PRS across disease areas is currently limited or inconsistent<sup>2,6-8</sup>. Furthermore, many methods have been 90 developed to compute PRS, each with different strengths and weaknesses<sup>9–11</sup>. Thus, guidelines 91 92 that delineate best practices while considering a range of real-world healthcare settings and 93 disease areas are critically needed.

94 Best practices for PRS are critical but lacking for a range of considerations that have been shown 95 to contribute to variability in accuracy and interpretation. These include guidance for variable 96 phenotype definitions and precision for both discovery GWAS and target populations, which 97 varies with cohort ascertainment strategy, geography, environmental exposures and other common covariates<sup>12–14</sup>. Other considerations include varying genetic architectures, statistical 98 99 power of the discovery GWAS, and PRS methods, which vary in which variants (generally in the 100 form of SNPs) are included and how weights are calculated<sup>9,15</sup>. A particularly pernicious issue 101 requiring best practices is regarding maximizing generalizability of PRS accuracy among ancestry 102 groups<sup>16,17</sup>. Developing best practices for PRS therefore requires harmonized genetic data 103 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 hospitalbased 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.<sup>18</sup>.

116 Here we outline a framework for PRS analyses of multi-ancestry GWAS across multiple biobanks, 117 as shown in **Figure 1**. The endpoints examined are: asthma, chronic obstructive pulmonary 118 disease (COPD), heart failure (HF), stroke, acute appendicitis (AcApp), venous thromboembolism 119 (VTE), gout, appendectomy, primary open-angle glaucoma (POAG), uterine cancer (UtC). 120 abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC) and 121 hypertrophic or obstructive cardiomyopathy (HCM), for which the phenotype definitions can be 122 found in Zhou et al.<sup>18</sup>. Those 14 endpoints represent the pilot effort of GBMI, which greatly vary 123 in disease prevalence. It ranges from <1% for AAA, IPF, ThC and HCM to ~6% for COPD and 124 ~9% for asthma. Some endpoints (for example, appendectomy which can be extracted from EHR 125 procedure codes) have not been broadly studied in previous GWAS studies. By evaluating PRS 126 across 14 endpoints (Table S1 and Table S2) and 9 biobanks, we review and explore practical 127 considerations for three steps: genetic architecture estimation, PRS method optimization and selection, and evaluation of PRS accuracy. Our framework applies to biobank-scale resources 128 129 with both homogenous and diverse ancestries.

# 130 Results

The diverse ancestries included in GBMI accounted for different proportions ranging from ~76.4% for EUR, 0.1% for MID, 1.0% for AMR, 1.7% for CSA, 18.9% for EAS and 1.8% for AFR. We explored the genetic architecture of 14 endpoints using GWAS summary statistics from all ancestries and EUR only in GBMI<sup>19</sup>. We used leave-one-biobank-out meta-analyzed GWAS in GBMI as our primary discovery datasets for the following PRS analyses. The ancestry compositions of discovery GWAS used in this study can be found in **Table S2**.

# 137 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

140 are assigned to them. Thus, understanding the genetic architecture of complex traits along with 141 sample size and ancestry composition of the discovery GWAS is critical for choosing optimal 142 prediction methods. For example, the SNP-based heritability  $(h_{SNP}^2)$  bounds PRS accuracy. We used SBayesS<sup>20</sup> to estimate  $h_{SNP}^2$ , polygenicity (the proportion of SNPs with nonzero effects), and 143 144 the relationship between minor allele frequency (MAF) and SNP effects (i.e., a metric of negative 145 selection, hereafter denoted as S) for the 14 endpoints in GBMI. Meta-analyses in GBMI were 146 performed across up to 18 different biobanks on 14 endpoints using an inverse-variance weighted 147 method as described in Zhou et al.<sup>18</sup>, including individuals from diverse ancestries. In addition to 148 presenting results using EUR only GWAS summary statistics (EUR GWAS), we also reported 149 estimates using meta-analysis from all ancestries (multi-ancestry GWAS). We explored whether 150 we can reasonably use EUR-based LD reference to approximate the LD of multi-ancestry GWAS 151 in GBMI using the attenuation ratio statistic estimated from LD score regression (LDSC) (see 152 STAR Methods). The attenuation ratio can be used to quantify whether there was a strong LD 153 mismatch, for which the values > 0.2, between GWAS summary statistics and the LD reference 154 panel<sup>21</sup>. We found that the ratio of LDSC using the EUR LD reference panel for GBMI multi-155 ancestry GWAS was not statistically larger than 0.2. Also, the values were not statistically different 156 from those achieved using GBMI EUR GWAS. This is consistent with a previous study which has 157 found that EUR-based LD can reasonably approximate the LD in their multi-ancestry GWAS 158 consisting of ~75% EUR individuals<sup>22</sup>.

159

Most diseases analyzed here had low but significant  $h_{SNP}^2$  and a range of polygenicity estimates 160 161 (Figure 2). Note that here we reported the  $h_{SNP}^2$  on the liability scale (see STAR Methods). The 162 SBayesS model failed to converge for HCM, likely because its estimated  $h_{SNP}^2$  was found to be 163 not significantly different from 0 using LDSC. This could be ascribed to its known predisposing 164 monogenic mutations, the low disease prevalence and heterogeneous subtypes<sup>19</sup>. Therefore, this 165 endpoint was dropped from downstream analyses. We observed that the estimates were overall 166 higher using multi-ancestry GWAS compared to EUR GWAS (Figure 2). Overall, the median 167 estimates of SNPs with nonzero effects across 13 endpoints were 0.34% for multi-ancestry 168 GWAS and 0.14% for EUR GWAS (p-value = 0.002, paired wilcoxon signed rank test), 169 respectively. The corresponding median estimates for  $h_{SNP}^2$  were 0.051 for multi-ancestry GWAS 170 and 0.043 for EUR GWAS (p-value = 0.002, paired wilcoxon signed rank test), respectively. The 171 largest difference of 0.06 was found in gout. This could be due to higher  $h_{SNP}^2$  estimated in non-EUR GWAS. For example, the estimates for  $h_{SNP}^2$  using EUR and EAS GWAS was 0.051 (s.e. = 172 173 0.0027) and 0.088 (s.e. = 0.005), respectively. Moreover, we have also found that the estimated

effect sizes of two gout-associated loci (close to genes *ALDH16A1* and *SLC2A9*) were different
across ancestries<sup>19</sup>. Specifically, we observed that a few top gout-associated variants showed
much higher allele frequencies in EAS as compared to EUR, thus resulting in larger variance
explained (Figure S1).

178

179 Polygenicity and  $h_{SNP}^2$  estimates varied greatly among different endpoints. Specifically, the  $h_{SNP}^2$ estimates were highest for asthma and gout using multi-ancestry GWAS ( $h_{SNP}^2$  = 0.085, s.e. = 180 181 0.0011 and  $h_{SNP}^2$  = 0.111, s.e. = 0.0024, respectively), while asthma was found to be much more 182 polygenic than gout. We caution that the numeric interpretation of polygenicity depends on various 183 factors and cannot be interpreted as the number of causal variants. For example, larger and more 184 powerful GWAS tend to discover more trait-associated variants, thus appear to have higher 185 polygenicity. Because we used the same set of SNPs in SBayesS analyses for all endpoints, we 186 hence used the results as a relative measurement of the degree of polygenicity. We observed 187 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  $h_{SNP}^2$ . Overall, 188 COPD and asthma were estimated to be the most polygenic traits, followed by HF and stroke, 189 190 whereas AcApp, UtC and ThC were the least polygenic. Lastly, we observed signals of negative 191 selection for traits including asthma (S = -0.56, s.e. = 0.05), COPD (S = -0.40, s.e. = 0.11) and 192 POAG (S = -0.50, s.e. = 0.15) when considering using EUR GWAS, consistent with empirical 193 findings of negative selection explaining extreme polygenicity of complex traits<sup>23</sup>. 194

195 In summary, we observed largely varied key parameters of genetic architecture among 13 196 endpoints using multi-ancestry and EUR only GWAS. We found that asthma and COPD had the 197 highest  $h_{SNP}^2$  as well as polygenicity. We excluded HCM in our subsequent prediction analyses 198 due to lower evidence of polygenicity and its non-significant  $h_{SNP}^2$ .

199 Optimal prediction performance using heuristic methods depends on 200 phenotype-specific genetic architecture

We first evaluated the pruning and thresholding (P+T, *p*-value thresholds ranged from  $5 \times 10^{-8}$  to 1) method using the EUR-based LD reference panel for all 13 endpoints in the UKBB and BBJ, respectively, given its widespread use and relative simplicity. Note in this study, we used leaveone-biobank-out meta-analyzed GWAS as the discovery GWAS when evaluating PRS in that specific biobank (**Table S2**). We further explored how different factors impact the prediction

performance of P+T in diverse ancestry groups, including LD parameters (LD window sizes and LD  $r^2$  thresholds), LD reference panels (ancestry composition, sample size, and SNP density) and per-variant effective sample size ( $N_{eff}$ ) and MAF (see STAR Methods).

209

210 First of all, we selected the optimal p-value threshold (the p-value threshold with highest prediction accuracy, as measured by  $R^2$  on the liability scale,  $R^2_{liability}$ , if not specified) in the tuning cohorts 211 212 and evaluated the accuracies in the test cohorts (see STAR Methods). Specifically, we found that 213 for UKBB with diverse ancestries, using ancestry-specific tuning cohorts provided better 214 prediction performance as compared to that using EUR-based tuning cohorts (Figure S2). We 215 found that the optimal p-value threshold differed considerably between various endpoints (Figure 216 S3 and Table S3). This pattern is found to be related to polygenicity of studied endpoints; but it 217 is also due to a combination of factors such as the GWAS discovery cohort sample size, disease 218 prevalence, trait-specific genetic architecture, and genetic and environmental differences 219 between discovery and target ancestries<sup>24</sup>. For example, when the optimal p-value was 220 determined in the UKBB-EUR subset, the less polygenic traits of ThC (106 variants) and AcApp (17 variants) showed highest accuracy at p-value thresholds of  $5 \times 10^{-5}$  and  $5 \times 10^{-7}$ , respectively, 221 222 while for the more polygenic traits of stroke (115,609 variants), HF (115,741 variants), asthma 223 (7,858 variants) and COPD (29,751 variants) achieved the highest accuracy when including SNPs 224 with *p*-value less than 1, 1, 0.01 and 0.1, respectively. To investigate whether ancestries affect 225 the optimal p-value threshold, we replicated our analysis in the BBJ (Figure S3). In the BBJ, p-226 value thresholds of 5 x 10<sup>-5</sup>, 0.01 and 5 x 10<sup>-5</sup> presented best performance for gout, stroke and 227 HF, respectively. Consistent with previous studies, these results suggest that optimal prediction 228 parameters (here p-value threshold specifically) for P+T appear to be dependent on the ancestry 229 of the target data among other factors<sup>25,26</sup>. Further, we found that for more polygenic traits 230 including asthma, COPD, stroke and HF, prediction was more accurate with more variants in the 231 PRS (i.e., a less significant threshold) than using the genome-wide significance threshold (p-value  $< 5 \times 10^{-8}$ ). On the contrary, less polygenic traits showed no or modest improvement with less 232 233 stringent *p*-value thresholds, especially for traits such as gout which has trait-associated SNPs 234 with large effects. However, these trends were less obvious in the BBJ which might be attributed 235 to the small proportion of EAS included in the discovery GWAS. One caveat we noted was that 236 fixed LD parameters of P+T were used, thus the results might be impacted by additional 237 optimization of those parameters, which we will further explore below.

We found that further optimizing LD parameters, including LD window size and LD  $r^2$  thresholds. 239 240 of P+T did not contribute to significant improvement of accuracy across endpoints. Specifically, 241 we observed that the median accuracies with versus without LD parameter optimization were of 242 0.018 and 0.015, respectively (Figure S4). However, there was slight but statistically significant 243 accuracy improvement in EUR for asthma (~0.006). This might be due to more stratified signals 244 being tagged, which results in noise reduction of the predictor. As compared to using fixed LD 245 parameters, we found similar relationships between polygenicity and optimal p-value thresholds 246 when optimizing LD parameters in the UKBB. Specifically, the optimal p-value thresholds were 247 overall less stringent for more polygenic traits and more stringent for less polygenic traits. For 248 example, the accuracy using LD parameter optimization in the UKBB-EUR was highest with the 249 p-value thresholds of 0.5, 1, 0.1 and 0.2 for the highly polygenic traits of stroke, HF, asthma and COPD, respectively. In contrast, the optimal *p*-value thresholds of  $5 \times 10^{-5}$  and  $5 \times 10^{-7}$  were 250 251 observed for less polygenic traits of ThC and AcApp, respectively. To balance the computational 252 burden and signal-to-noise ratio, we used an LD window size of 250Kb and LD  $r^2$  of 0.1 as before. 253 We repeated our analyses using genome-wide common SNPs and compared the prediction 254 accuracy with that using HapMap3 SNPs only (Figure S4 and Table S3). There were no 255 significant improvements in prediction accuracies using a denser SNP set, which suggests that 256 HapMap3 SNP set represents genome-wide common SNPs well. Specifically, we found the 257 accuracies in EUR for the most polygenic traits, asthma (~0.006), COPD (~0.005) and HF 258 (~0.004), to be slightly improved using HapMap3 SNPs. Moreover, we found that the sample sizes 259 of the LD reference panel had little impact on P+T performance (Figure S5); but the parameters 260 described above including LD window sizes and LD  $l^2$  thresholds had a larger impact on accuracy. 261 We also showed that using 1KG-EUR as the LD reference panel performed well compared to 262 using other ancestral populations with similar sample sizes in the 1KG dataset, which could be 263 explained by the overrepresentation of EUR participants (~76.4%) in GBMI (Figure S6 and Table 264 **S3**). We further ran LDSC using the EUR-based LD reference panel on leave-specific-biobank-265 out GWAS in GBMI to estimate the attenuation ratio statistic (see STAR Methods). Similar to 266 previous findings, we found that even in leave-UKBB-out GWAS with the lowest EUR proportion 267 (Table S2), its LD information can be well approximated using the EUR reference panel, which 268 was reflected by the values of ratio not statistically larger than 0.2 and not statistically different 269 from EUR GWAS in GBMI. We therefore used 1KG-EUR as the LD reference panel for all 270 subsequent P+T analyses. But the choice of external LD reference panel for multi-ancestry GWAS 271 needs further exploration especially when the discovery GWAS becomes more diverse. 272

273 Finally, we investigated the impact of per-variant effective sample size heterogeneity. Since GBMI 274 consists of a number of biobanks with diverse ancestries, the number of samples used for meta-275 analysis was notably heterogeneous among the variants; the majority of the variants in the GWAS 276 meta-analysis had only a limited number of effective samples ( $N_{eff}$ ) (**Figure 3-A**). Therefore, although 277 sample size heterogeneity is not usually considered for PRS, it may confound the PRS prediction 278 accuracy in the case of global biobank collaborations. By filtering the variants according to  $N_{eff}$  per-279 variant (i.e., N<sub>eff</sub> larger than 50% or 80% thresholds of the maximum N<sub>eff</sub> of the trait of interest, see 280 STAR Methods), we observed that the  $R_{liability}^2$  increased substantially for less stringent thresholds (*p*-value > 5 × 10<sup>-5</sup>) in the UKBB (**Figure S7-A**). As a representative example, the largest  $R_{liability}^2$ 281 (0.034) was obtained for asthma when the *p*-value threshold was  $5 \times 10^{-3}$ , whereas the  $R_{liability}^2$  was 282 6.6 × 10<sup>-3</sup> at the threshold without  $N_{eff}$  filtering (Figure 3-B and Table S4). Next, we investigated 283 284 whether  $N_{eff}$  filtering could be substituted by other filtering criteria. Although excluding variants with 285 MAF less than 0.1 partially compensated for PRS transferability, the improvement of N<sub>eff</sub> filtering in  $R_{liability}^2$  was still observed (Figure S7-B). Heterogeneity in  $N_{eff}$  might be confounding especially in 286 multi-ancestry meta-analyses because it can be distorted by heterogeneous allele frequencies and 287 288 imputation quality spectra among ancestries. Indeed, as rarer variants tend to be more ancestry-289 specific, variants with low N<sub>eff</sub> tend to be unique to specific ancestries (Figure 3-C). Of note, the 290 dependency of  $R_{lightlity}^2$  on the N<sub>eff</sub> was, however, largely rectified for most of the traits by using only HapMap3 SNPs (Figure S7-C). Given that the  $R_{liability}^2$  for HapMap3 SNPs was comparable to that 291 292 for genome-wide SNPs (Figure S4), filtering to HapMap3 SNPs might be suitable for meta-analysis 293 of diverse populations. On the other hand, HapMap3 SNPs generally have good imputation quality, 294 although a recent study shows that relaxing imputation INFO score from 0.9 to 0.3 has negligible 295 impacts on prediction accuracy<sup>9</sup>. We replicated the  $N_{eff}$  filtering in BBJ and confirmed that improved  $R_{lighility}^2$  attributable to  $N_{eff}$  filtering was also observed (Figure S7-D). Although the effect of the  $N_{eff}$ 296 filtering was diminished by the MAF filtering in relatively stringent thresholds (p-value < 5 x 10<sup>-4</sup>), the 297 298 effect was still observed in the other thresholds (Figure S7-E). Using only HapMap3 SNPs almost completely reduced the dependency of  $R_{liability}^2$  on the  $N_{eff}$  (Figure S7-F). 299

300

301 Overall, we found the prediction performance of P+T to be affected by a combination of factors, with 302 *p*-value thresholds showing larger effects as compared to other parameters, such as LD window sizes, 303 LD  $r^2$  thresholds, and variant filtering by  $N_{\text{eff}}$  or MAF. Moreover, the optimal *p*-value threshold varied 304 substantially between different endpoints in GBMI. We also demonstrated that restricted use of 305 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.

### 308 Bayesian approaches for calculating PRS improve accuracy

309 We also evaluated fully genome-wide polygenic risk scores, by first fine-tuning the parameters in 310 PRS-CS. We ran PRS-CS using both the grid model and automated optimization model (referred 311 to as auto model), the former of which specifies a global shrinkage parameter (phi, in which 312 smaller values indicate less polygenic architecture and vice versa for larger values), with 1KG-313 EUR as the LD reference panel. We note that the optimized phi parameter with highest prediction 314 accuracy in the grid model differed among traits (Figure S8). Specifically, we found that for more 315 polygenic traits (as estimated using SBayesS) including asthma, COPD and stroke (Figure 2), 316 the optimal phi parameter was  $1 \times 10^{-3}$  in EUR (Figure S8). There was no significant difference 317 between prediction accuracy using the optimal grid model versus auto model (Figure S8), which 318 suggests PRS-CS can learn the phi parameter from discovery GWAS well when its sample size 319 is considerably large. Therefore, we hereafter used the auto model because of its computational 320 efficiency. Across target ancestral populations in the UKBB, PRS from EUR-based LD reference 321 panels showed significantly higher or comparable prediction accuracies compared to PRS using 322 other ancestry-based LD reference panels (Figure S9-A). This result suggests that it is 323 reasonable to use a EUR-based LD reference panel in GBMI largely because EUR ancestry 324 constitutes the largest proportion of GWAS participants (~76.4%). Note that we also compared 325 the prediction accuracy of LD reference panels derived from UKBB-EUR, which has a much larger 326 sample size, against 1KG-EUR and found no significant difference (Figure S9-B). These results 327 suggest that PRS-CS is not sensitive to the sample size of the LD reference panel, which is 328 consistent with previous findings<sup>27</sup>.

329

330 We then compared the optimal prediction accuracy of P+T versus the PRS-CS auto model in the 331 UKBB and BBJ and found that PRS-CS showed overall better prediction performance for traits with higher  $h_{SNP}^2$  but no or slight improvements for traits with lower  $h_{SNP}^2$  (Figure 4). Specifically, 332 333 the highest significant improvement of PRS-CS relative to that of P+T in EUR was observed for 334 HF, of 60.9%, followed by COPD (53.2%) and asthma (48.8%). Substantial increments were 335 observed for HF (105.2%), COPD (102.5%) and asthma (60.9%) in EAS. 45.8% and 48.1% 336 improvements were shown for asthma in CSA and AFR, respectively. P+T saw better prediction 337 performance over PRS-CS for a few trait-ancestry comparisons, however, such improvement was 338 not statistically significant. 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.

341

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<sup>9,27</sup>. The practical considerations about the two models, PRS-CS and P+T, used in this study, are shown in **Table S5**.

### <sup>347</sup> PRS accuracy is heterogeneous across ancestries and biobanks

348 For each of the participating biobanks, we used leave-one-out meta-analysis as the discovery 349 GWAS to estimate the prediction performance of PRS in each biobank (see STAR Methods). The 350 disease prevalence and effective sample size of each biobank is shown in Figure S10. Generally, 351 the PRS prediction accuracy of different traits increased with larger  $h_{SNP}^2$  (Figure 5 and Table **S6**). For example, the average  $R^2$  on the liability scale across biobanks (hereafter denoted as 352  $\overline{R_{lightlity}^2}$ , see STAR Methods) in EUR ranged from <1.0% for AcApp, appendectomy, stroke, UtC 353 and IPF, 1.0% for HF, ~2.2% for COPD and ThC to 3.8% for gout and 4.6% for asthma. Notably, 354 355 accuracy was sometimes heterogeneous across biobanks within the same ancestry for some traits. Specifically, the  $R_{liability}^2$  for asthma in ESTBB and BioVU was significantly lower than 356 357  $R_{hability}^2$ , which might be attributable to between-biobank differences such as recruitment strategy, phenotyping, disease prevalence, and environmental factors. The prediction accuracy 358 359 was generally lower in non-European ancestries compared to European ancestries, especially in African ancestry, which is mostly consistent with previous findings<sup>28–30</sup> with a few exceptions. For 360 361 example, we observed comparable prediction accuracy for gout in EAS relative to that in EUR, 362 which could be reflected by large effective sample sizes and some gout-associated SNPs with 363 large effects exhibiting higher allele frequencies in EAS (Figure S1). For example, the MAFs of 364 gout top-associated SNP, rs4148157, were 0.073 in 1KG-EUR and 0.25 in 1KG-EAS, 365 respectively, and the phenotypic variance explained by that SNP in EAS (8.3%) was more than 366 twice as high as that in EUR (3.0%). The accuracy of PRS to predict asthma risks in AMR was 367 found to be significantly higher than that in EUR, which could be due to the small sample size in 368 AMR (Table S6). Thus, further validation is needed in larger AMR population cohorts.

370 The ability of PRS to stratify individuals with higher disease risks was also found to be 371 heterogeneous across biobanks and ancestries as shown in Figure 6 and Table S7. We showed 372 that the PRS distribution across different biobanks slightly varied. Specifically, we calculated the 373 absolute difference of median PRS in each decile for each endpoint between biobanks for cases 374 and controls, separately, and found that the largest absolute differences were 0.06 and 0.21 for 375 stroke controls and stroke cases, respectively (Figure S11). This justifies the comparison of odds 376 ratios (ORs) in terms of relative risks. The ORs between the top 10% and bottom 10% were more 377 heterogeneous between biobanks and also higher relative to other comparisons (e.g., top 10% 378 vs middle and other strata). This is consistent with previous studies where OR reported between 379 tails of the PRS distribution is generally inflated relative to those between top ranked PRS and 380 general populations<sup>11</sup>. We measured the variation of OR between biobanks using the coefficient 381 of variation of OR (CoeffVar<sub>OR</sub>, see STAR Methods). The largest CoeffVar<sub>OR</sub> in EUR was observed 382 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  $R_{liability}^2$  that ORs were overall 383 higher for traits with higher  $h_{SNP}^2$  and also higher in EUR than non-EUR ancestries, which is 384 385 expected as the two accuracy metrics are interrelated. For example, the averaged ORs across 386 biobanks weighted by the inverse variance in EUR (see STAR Methods) for gout were 4.6, 2.4 387 and 2.2 for the top 10% vs bottom 10%, middle and other strata, separately. The corresponding 388 estimates in EUR for stroke were 1.6, 1.3 and 1.3, respectively. Across ancestries, the average 389 OR of asthma between the top 10% and bottom 10% ranged from 4.1 in EUR to 2.4 in AFR.

390

Overall, the predictive performance of PRS measured by  $R_{liability}^2$  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<sup>31</sup>.

## 394 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 and Figure S12**, the accuracy improvements were most obvious for traits with larger  $h_{SNP}^2$  but there was no or slight improvement for traits with lower  $h_{SNP}^2$ . 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 of  $\overline{R_{hability}^2}$  in EUR were

402 0.033 for asthma, 0.031 for gout, 0.019 for ThC and 0.017 for COPD, whilst the corresponding 403 improvements of AUC on average ( $\overline{AUC}$ ) were 0.051, 0.078, 0.078 and 0.041, respectively. Substantial improvements were also observed for gout in EAS ( $\overline{R_{liability}^2}$ : 0.037,  $\overline{AUC}$ : 0.090), for 404 asthma in CSA ( $\overline{R_{liability}^2}$ : 0.026,  $\overline{AUC}$ : 0.060), EAS ( $\overline{R_{liability}^2}$ : 0.017,  $\overline{AUC}$ : 0.047) and AFR 405  $(\overline{R_{liability}^2} 0.009, \overline{AUC}: 0.034)$ , and for ThC in EAS  $(\overline{R_{liability}^2}: 0.014, \overline{AUC}: 0.080)$  and AFR  $(\overline{R_{liability}^2}: 0.014, \overline{R_{liability}^2}: 0.0014, \overline{R_{liability}^2}: 0.014, \overline{R_{liability}^2}:$ 406 407 0.016, AUC: 0.108). However, PRS accuracy was significantly higher for published GWAS relative 408 to the current GBMI for POAG in EUR and AFR, and COPD in the specific case of Lifelines 409 biobank. We referred to the datasets included in the public GWAS of POAG and found that 410 individuals from diverse datasets of EUR and AFR populations were also part of the discovery 411 dataset, thus we cannot rule out the possibility of sample overlapping or relatedness between the 412 discovery and target datasets for these populations. This suggests that the PRS evaluation may 413 be biased upwards from the prior GWAS for POAG. Also, the phenotypes of POAG across 414 different biobanks are likely more heterogeneous in GBMI than targeted case-control studies<sup>18,32</sup>. 415 The meta-analysis of GBMI with International Glaucoma Genetics Consortium (IGGC) did not 416 lead to substantially improved prediction performance<sup>32</sup>. Another concern might be the 417 disproportional case/control ratio of POAG in GBMI, of ~27,000 cases and ~1.4M controls, thus 418 POAG-related phenotypes with shared genetics in the controls or possible uncontrolled ancestry 419 differences between cases and controls might confound the GBMI GWAS. A very high 420 heterogeneity for phenotype definitions is also found for COPD, however this does not explain 421 why one biobank alone presents this pattern; a specific environmental or population effect not 422 considered in the broad analysis might affect this particular observation.

423

To boost statistical power, we can meta-analyze GBMI GWAS with other non-overlapping cohorts as shown in other GBMI working groups<sup>33–35</sup>. However, we should note that more heterogeneity might be introduced from different resources such as population structure and phenotype definitions, which we cannot control with summary statistics data and that could exacerbate the heterogeneous performance of PRS across target populations. On the other hand, GBMI is open to more cohorts and has been continuously working on integrating more datasets.

# 432 Discussion

The GBMI resource is notable in its collection of phenotypes studied and range of participating 433 434 cohorts from multiple ancestry groups; it has therefore offered a unique opportunity to 435 comprehensively evaluate and develop guidelines regarding the effects of multi-ancestry and 436 heterogeneous GWAS discovery data, polygenicity, and PRS methods on prediction performance 437 in diverse target cohorts. In this study, we have used the unique GBMI resource consisting of 438 multi-ancestry GWAS for multiple disease endpoints with varying genetic architectures and 439 prevalences across diverse populations to develop and evaluate PRS. Indeed, we found overall 440 across a range of phenotypes and ancestries that using the large-scale meta-analysis from GBMI 441 significantly improved PRS accuracy compared to previous studies with smaller sample sizes and 442 less diverse cohorts. While some previous studies have benchmarked PRS methods and 443 accuracies, most have been based on relatively homogeneous GWAS discovery cohorts or 444 evaluated for specific phenotypes<sup>3,9,26,36</sup>. Even when assessing the portability of PRS across 445 ancestries, most evaluations have included ancestrally diverse target cohorts but still relatively homogeneous discovery cohorts<sup>12,13,37</sup>. Thus, based on the results of our analyses using GBMI, 446 447 we have provided additional lessons and guidelines for developing PRS with multi-ancestry 448 discovery data for different endpoints (Figure S13). We have organized these best practices 449 according to 1) characteristics of the discovery GWAS, 2) PRS model fitting, and 3) the target 450 cohort.

451

452 First, the GWAS discovery cohort provides the prerequisite input for polygenic score calculations 453 and interpretation, namely how phenotypes are ascertained and in which populations, which 454 SNPs to include, and which effect sizes will be used. We recommend that standard quality 455 controls should be performed with more caution when considering multi-ancestry discovery 456 GWAS. Specifically, we suggest filtering variants based on the per-variant effective sample size 457  $(N_{\rm eff})$  and MAF as they show considerable heterogeneity across datasets and ancestries in our 458 discovery GWAS. When we filtered out variants with extremely small  $N_{eff}$  in our P+T analyses, 459 and in particular when using HapMap3 SNPs, PRS prediction performance improved. As noted 460 in Zhou et al.<sup>19</sup>, the allele frequencies of variants in GBMI meta-analyzed GWAS were compared 461 with those in gnomAD using Mahalanobis distance and flagged if they were three standard 462 deviations away from the mean. We recommend computing such statistics and filtering with this 463 information, or if infeasible, restricting to using only HapMap3 variants.

465 Given the significant improvements in PRS accuracy with GBMI discovery GWAS over previous 466 studies with smaller sample sizes and less diversity, we recommend using the largest and most 467 diverse GWAS discovery cohort available when constructing PRS, even if it matches the ancestry 468 composition of the target cohort slightly less well than a smaller GWAS. Overall, traits with higher 469 SNP-based heritability showed greater improvement compared to those with lower SNP-based 470 heritability. This indicates that PRS performance will continually benefit from larger sample sizes 471 and more diverse populations. However, further research is needed to understand more 472 concretely how the composition of underrepresented populations, including specific ancestries 473 and varying sample sizes, can be modeled alongside current Eurocentric GWAS to best facilitate 474 PRS accuracy and generalizability.

475

476 Second, when fitting PRS models, important choices include which PRS construction methods to 477 use, how to fine-tune hyperparameters, and which LD reference panels to use. So far, PRS 478 models that use GWAS summary statistics have been favored over those that use individual-level 479 data due to their computational efficiency and data access restrictions. These models have been comprehensively reviewed recently<sup>10,38</sup>. In this study, we therefore explored the prediction 480 481 performance of two widely used PRS construction methods, P+T and PRS-CS. We paired the 482 results of these methods with prior knowledge of trait-specific genetic architecture estimates from 483 SBayesS. The best predictor for P+T is often obtained by fine-tuning the p-value thresholds in a 484 validation dataset, while other LD related parameters, such as LD  $r^2$  and LD window size, are 485 usually arbitrarily specified. Here, we found that the prediction accuracy of P+T was much less 486 sensitive to different LD-related parameters compared to various p-value thresholds. Moreover, 487 the optimal p-value threshold varied across phenotypes. likely because of trait-specific genetic 488 architecture, especially the degree of polygenicity measured by SBayesS. However, differences 489 in discovery GWAS and target dataset such as sample sizes, phenotype definition, disease 490 population prevalence and population characteristics could also contribute to this variation. When 491 analyzing PRS-CS results, we validated a previous finding that the auto model, that does not 492 require post-hoc tuning of the proportion of SNPs with non-zero effects (phi), showed similar 493 prediction performance relative to the more computationally intensive grid model, which requires 494 determining the optimal phi parameter in an independent tuning cohort<sup>27</sup>.

495

We also recommend using prior knowledge and empirical measurements of the genetic architecture of studied phenotypes to choose specific types of PRS models. In this study, we evaluated the effects of trait-specific genetic architecture on PRS performance using estimates

499 from SBayesS. Generally, traits with higher SNP-based heritability, such as asthma and gout, 500 showed greater improvement with the GBMI discovery data compared to those with lower SNP-501 based heritability, such as acute appendicitis (AcApp). Trait-specific architecture affected both the 502 choice of method and optimal hyper-parameters. For example, extremely polygenic traits are 503 more suitable for an infinitesimal model or Bayesian models that are adaptive to the trait genetic 504 architecture. The specific model hyper-parameters are also affected by trait genetic architecture. 505 For example, the optimal *p*-value threshold of P+T might be more stringent for less polygenic 506 traits but less stringent for highly polygenic traits.

507

508 Another decision point in fitting PRS models is regarding which LD reference panel to use when 509 multi-ancestry GWAS discovery and target populations are available. An in-sample LD reference 510 panel that spans the full discovery cohort is optimal but rarely available. Here, we have shown 511 that EUR-based LD reference panels can reasonably approximate the LD of GBMI GWAS. 512 However, choosing LD reference panels that mirrors the ancestry composition of the discovery 513 GWAS when in-sample LD reference panels are not available is ideal. For convenience, if one 514 ancestry is dominant in the multi-ancestry GWAS, we suggest using that ancestry-matched 515 reference panel. The attenuation ratio statistic estimated from LDSC can further be used as a 516 measure to quantify the degree of LD mismatch between discovery GWAS and LD reference 517 panels<sup>22</sup>. When ancestry proportions are relatively evenly distributed, we and others have found 518 that using LD reference panels with ancestry proportions that match the discovery GWAS could 519 provide better prediction performance especially for less polygenic traits with large effect variants (unpublished work), such as lipid traits<sup>39</sup>. We also found that prediction performance can be 520 521 improved when using ancestry-matched tuning cohorts for PRS construction to fine-tune hyper-522 parameters and avoid overfitting, such as P+T and the PRS-CS grid models explored in this study. 523 While other studies have also explored options such as pseudo-validation when no additional tuning cohort is available<sup>40,41</sup> 524

525

Third, the practical considerations for target populations involved in PRS analyses are quite consistent between using homogenous GWAS and multi-ancestry GWAS. In this study, we used biobanks with various ancestry compositions and recruitment strategies as the target cohorts<sup>19</sup>. For example, BBJ, BioVU and MGI are hospital-based biobanks whereas others are populationbased or have mixed enrollment strategies, which can impact phenotype precision or ascertainment bias and therefore heritability. UKBB, MGI and BioVU have diverse ancestries while others primarily consist of one ancestry (either European or East-Asian participants). The

performance of PRS in different target populations can also be affected by the ancestry proportions in the discovery GWAS and precision of phenotype definition aside from biobankspecific factors (e.g., environmental factors), which warrants further exploration. We therefore recommend considering those factors and reporting PRS distribution statistics (e.g., median PRS) and accuracy metrics when benchmarking the prediction performance between different PRS predictors. More reporting standards about PRS models have been well-documented in PGS Catalog<sup>36</sup>.

540

541 Related to the target cohorts, we also found that the prediction performance showed great 542 heterogeneity across biobanks and ancestries. Because PRS are only intended to capture genetic 543 factors, other considerations such as environmental exposures and demographic history may 544 impact the predictive power of PRS within and across ancestries, with recommendations for how 545 to model these alongside PRS an open question for future research and methods development. 546 For example, we found that the  $R_{lighility}^2$  in OHS was overall higher than in other biobanks, which 547 may be attributed to the more complex relatedness structure in this founder population. Notably, 548 the phenotype definitions, recruitment strategy and disease prevalence also vary to different 549 extents across the biobanks studied here.

550

551 We note a few limitations in our study. First, we chose 1KG-EUR as the LD reference panel 552 because data security practices often preclude the use of individual-level GWAS data across 553 analytical teams. Although we have shown that the EUR-based LD reference panels can 554 reasonably approximate the LD of GBMI GWAS studied here, it still could affect SNP effect size 555 estimates and thus prediction performance. Further efforts are required to provide more 556 appropriate LD reference panels. For example, utilizing the large-scale UKBB with individual-level 557 genotypes to construct a panel with matched ancestry proportions to the discovery GWAS has 558 been used in a recent study<sup>39</sup>. However, early explorations have shown that using proportional 559 LD reference panels generally achieves similar prediction performance as using EUR-based 560 reference panels when EUR is primarily dominant in the multi-ancestry GWAS (unpublished 561 work). Also, sharing LD matrices from participating biobanks without accessing individual-level 562 data would be another alternative to construct an in-sample LD matrix. On the other hand, 563 individual-level based PRS construction methods across large-scale biobanks without relying on 564 LD reference panels are also promising. Such methods could potentially benefit from secure 565 large-scale GWAS across multiple datasets. For example, Blatt et al.<sup>42</sup> have used homomorphic 566 encryption to establish a privacy-preserving framework to perform GWAS and decrypt the results

567 for sharing through a project coordinator. Second, we have focused on common SNPs, 568 specifically HapMap3 SNPs for PRS-CS. As a result, information from rarer variants missing in 569 the LD reference panel was not captured in other non-European ancestries, which may explain a 570 small fraction of the loss of accuracy across populations. Third, although a harmonized analysis 571 framework was developed for GBMI, such as phenotype definitions, ancestry assignments, and 572 PRS construction, there remains a multitude of factors that may contribute to heterogeneous 573 accuracy across both biobanks and ancestries. These include, but are not limited to, phenotype 574 precision, cohort-level disease prevalence, and environmental factors. Last, we evaluated PRS 575 predictive performance using multi-ancestry GWAS but comparisons with single-ancestry GWAS 576 at sufficient scale would enable us to better understand the specific contributions of ancestry 577 diversity and increasing sample size especially for under-represented ancestries, which also 578 serves as a future direction.

579

580 The GBMI resource constitutes remarkable progress in expanding the number of endpoints and 581 ancestry groups studied, laying the groundwork for several future directions for exploration. For 582 example, PRS construction methods that model GWAS summary statistics alongside LD 583 information from multiple ancestries have shown promising accuracy improvements for some 584 traits<sup>16,43</sup>, but statistical methods are insufficient for equitable accuracy without simultaneous 585 progress in generating large-scale diverse data, as early investigation into one of these methods 586 has yielded marginal improvement in both European and non-European ancestries for asthma in 587 GBMI<sup>44</sup>. In addition to multi-ancestry GWAS, sex-stratified GWAS in GBMI also provides 588 opportunities to explore the role of sex-specific effects as well as impacts from the sample size 589 ratio of males/females on prediction performance of PRS across biobanks. Beyond genetic 590 effects, biobank-specific risk factors and environmental exposures provide further opportunities 591 to better understand the heterogeneity in PRS accuracy that we have identified across biobanks and ancestries<sup>45,46</sup>. This will be extremely important as previous work has shown that prediction 592 593 performance differences between target cohorts are not likely to be reduced using various PRS 594 construction methods<sup>9</sup>. Finally, extending these collaboration efforts to more biobanks in the 595 future, particularly those including recently admixed populations, will bring more resolution into 596 those effects that are biobank-specific and ancestry-specific. Studies in recently admixed 597 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 598 599 generalizability, particularly for less polygenic traits<sup>47,48,49</sup>. Altogether, these initiatives hold great

promise for improving transferability of PRS across biobanks and ancestries by harnessing thephenotypic richness and diversity present in different biobanks.

602

# 603 Acknowledgements

A.R.M is funded by the K99/R00MH117229. E.L. is funded by the Colciencias fellowship ed.783.

605 S.N. was supported by Takeda Science Foundation. Y.O. was supported by JSPS KAKENHI

606 (22H00476), and AMED (JP21gm4010006, JP22km0405211, JP22ek0410075,

- 607 JP22km0405217, JP22ek0109594), 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)
- 610 Awards R35HG010718, R01HG011138, R01GM140287, and NIH/NIA AG068026. V.L.F. was
- 611 supported by the European Union's Horizon 2020 research and innovation programme under
- 612 the Marie Skłodowska-Curie grant agreement No.675033 (EGRET plus). L. B. and B. B. receive
- 613 support from the K.G. Jebsen Center for Genetic Epidemiology funded by Stiftelsen Kristian
- 614 Gerhard Jebsen; Faculty of Medicine and Health Sciences, NTNU; The Liaison Committee for
- 615 education, research and innovation in Central Norway; and the Joint Research Committee
- 616 between St Olavs Hospital and the Faculty of Medicine and Health Sciences, NTNU. K.L. and
- 617 R.M. were supported by the Estonian Research Council grant PUT (PRG687) and by
- 618 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
- 620 supported by the National Human Genome Research Institute of the National Institutes of
- 621 Health under award number T32HG010464. The work of the contributing biobanks was
- 622 supported by numerous grants from governmental and charitable bodies. The biobank specific
- 623 acknowledgements and full author list for GBMI are included in the **Supplementary Notes**.

# 624 Author Contributions

- 625 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.,
- 627 I.S., J.H.
- 628 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.
- 629 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.

# 632 Declaration of Interests

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

# 635 Figure Legends

### 636 **Figure 1. Overview of the study framework.**

637 Figure 2. Genetic architecture of endpoints in GBMI. We reported the estimates from 638 using meta-analyzed GWAS from all ancestries (labeled as All ancestries) and European 639 only (labeled as EUR), respectively. The phenotypes on the y-axis are ranked based on 640 the SNP-based heritability estimates using meta-analysis from all ancestries. Note the 641 SNP-based heritability estimates were transformed on the liability scale. The vertical 642 dashed lines in each panel indicate the corresponding median estimates across 13 endpoints. The results for hypertrophic or obstructive cardiomyopathy (HCM) are not 643 644 presented. Abbreviations: Europeans (EUR), chronic obstructive pulmonary disease 645 (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), 646 primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm 647 (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC).

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 S7 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 653 654 phenotypes are ranked based on the SNP-based heritability as shown in Figure 2 655 (indicated by the dashed line) estimates using all ancestries. Only trait-ancestry pairs with 656 significant accuracies in both P+T and PRS-CS are presented. The prediction accuracy 657 in P+T estimated in the test cohort based on the optimal p-value thresholds fine-tuned in the validation cohort. The auto model was used for PRS-CS. Abbreviations: Europeans 658 659 (EUR), Admixed Americans (AMR), Middle Eastern (MID), Central and South Asians 660 (CSA), East Asians (EAS) and Africans (AFR), chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), 661 662 primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm 663 (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC).

### 664

Figure 5. Prediction performance of PRS-CS across biobanks and ancestries. The 665 666 phenotypes on the y-axis were ranked by the SNP-based heritability using all ancestries 667 as shown in **Figure 2**. Only the significant results were shown. Data for all trait-ancestry 668 pairs in each biobank are provided in **Table S6.** Note that we removed the estimates in 669 AMR and MID due to limited information as a result of small sample sizes. Abbreviations: 670 Europeans (EUR), Admixed Americans (AMR), Middle Eastern (MID), Central and South 671 Asians (CSA), East Asians (EAS) and Africans (AFR), chronic obstructive pulmonary 672 disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous 673 thromboembolism (VTE), primary open-angle glaucoma (POAG), uterine cancer (UtC), 674 abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer 675 (ThC).

Figure 6. The odds ratio (OR) between different PRS strata for endpoints in GBMI. 676 677 The dashed line indicates OR=1. Only significant trait-ancestry specific OR was reported, with *p*-value < 0.05. The full results are shown in **Table S7.** The averaged OR was 678 679 calculated using the inverse-variance weighted method (see STAR Methods). PRS was 680 stratified into deciles with the first decile (bottom 10%) used as the referenced group. The 681 phenotypes were ranked based on SNP-based heritability estimates using all ancestries (see Figure 2). Abbreviations: Europeans (EUR), Admixed Americans (AMR), Middle 682 683 Eastern (MID), Central and South Asians (CSA), East Asians (EAS) and Africans (AFR), 684 chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis 685 (AcApp), venous thromboembolism (VTE), primary open-angle glaucoma (POAG), 686 uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis 687 (IPF), thyroid cancer (ThC).

#### 688 Figure 7. The prediction performance and ancestry compositions of GBMI versus 689 previously published GWAS. A) The ancestry compositions of GBMI and referenced 690 GWAS. The label for biobanks in the x-axis indicated the leave-one-out-biobank meta-691 analyzed GWAS in GBMI. The previously published GWAS was labeled as Referenced. 692 B) The comparison of AUC between GBMI and referenced GWAS. The AUC was 693 calculated by fitting PRS only. The phenotypes in A) were ranked based on the effective sample sizes from all ancestries. The phenotypes in B) were ranked by the SNP-based 694 695 heritability estimates from all ancestries. Note that we removed the estimates in AMR and 696 MID due to limited information as a result of small sample sizes. The full results are shown 697 in Table S4. Abbreviations: Europeans (EUR), Admixed Americans (AMR), Middle Eastern (MID), Central and South Asians (CSA), East Asians (EAS) and Africans (AFR), 698 699 chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), primary open-angle glaucoma (POAG), 700 701 uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis 702 (IPF), thyroid cancer (ThC).

703

704

# 705 STAR Methods

### 706 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 and Table S2**) 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</li>
discovery datasets, leave-one-biobank-out meta-analysis using the inverse-variance weighted
meta-analysis strategy was applied<sup>18</sup>.

713

Target datasets: We used 9 biobanks, i.e., BioBank Japan (BBJ)<sup>50</sup>, BioVU<sup>51</sup>, Lifelines<sup>52</sup>, UK Biobank
 (UKBB)<sup>53</sup>, Ontario Health Study (OHS)<sup>54</sup>, Estonian Biobank (ESTBB)<sup>55</sup>, FinnGen, Michigan
 Genomics Initiative (MGI)<sup>56</sup> and Trøndelag Health Study (HUNT)<sup>57</sup>, 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.<sup>18</sup>. 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.

### 721 Genetic architecture of 14 endpoints in GBMI

722 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<sup>20</sup>. It only requires 723 724 GWAS summary statistics and LD correlation matrix estimated from a reference panel. We ran 725 SBayesS using the GWAS summary statistics from all 14 endpoints in GBMI, including meta-726 analyses on all ancestries and on EUR only in 19 biobanks<sup>18</sup>. We evaluated the SNP-based heritability  $(h_{SNP}^2)$ , polygenicity (proportion of SNPs with nonzero effects) and the relationship 727 between allele frequency and SNP effects (S). We used the shrunk LD matrix (i.e., a LD matrix 728 729 ignoring small LD correlations due to sampling variance) on HapMap3 SNPs provided by GCTB 730 software. The LD matrix was constructed based on 50K European individuals from UKBB. Note 731 that we observed inflated SNP-based heritability estimates using effective sample size for each 732 SNP and hence used the total GWAS sample size instead. We used other default settings in the 733 software. We calculated the p-value of each parameter using Wald test to evaluate whether it was significantly different from 0. The  $h_{SNP}^2$  was further transformed into liability-scale with disease 734 735 prevalence approximated as the case proportions in the GWAS summary statistics<sup>58</sup>.

### 736 PRS construction

737 P+T: P+T is used to clump guasi-independent trait-associated loci within a LD window size using a specific LD  $r^2$  threshold. We first ran P+T in the UKBB and BBJ using a LD  $r^2$  threshold of 0.1 738 and a LD window (LDwin) of 250Kb. We performed the analysis on both HapMap3 SNPs and 739 genome-wide SNPs. We constructed PRS using --score implemented in Plink v1.9<sup>59</sup> using 13 740 different *p*-value thresholds  $(5 \times 10^{-8}, 5 \times 10^{-7}, 1 \times 10^{-6}, 5 \times 10^{-6}, 5 \times 10^{-5}, 5 \times 10^{-4}, 5 \times 10^{-3}, 0.01, 10^{-6}, 5 \times 10^{$ 741 742 0.05, 0.1, 0.2, 0.5, 1). We further explored how per-variant filtering based on effective sample 743 sizes ( $N_{eff}$ ) and MAF thresholds would affect the prediction performance. We used three 744 thresholds to retain variants by their  $N_{eff}$ , >0%, >50%, and >80% of  $N_{eff}$  compared to the total ones 745 and also three MAF filters: 0.01, 0.05 and 0.1. In the UKBB, we also explored the impact of 746 optimizing LD parameters on prediction performance by using different combinations of LD<sub>win</sub> (250, 500, 1000, and 2000Kb) and LD *r*<sup>2</sup> thresholds (0.01, 0.02, 0.05, 0.1, 0.2, and 0.05) with the 747 following flags: --clump-p1 1 --clump-p2 1 --clump-r2 LD<sub>win</sub> --clump-kb r<sup>2</sup> in Plink v1.9. For each 748

population in the specific biobank, we randomly split the individuals into two even parts. One part was used as a validation cohort to fine-tune the parameters and the other part was used as the test cohort to evaluate the performance of PRS. To explore the impact of tuning cohorts on target populations with diverse ancestries such as UKBB in this study, we also used 10,000 EUR samples, not included in the discovery GWAS and independent from the test cohort, as the tuning cohort.

755

756 PRS-CS: PRS-CS<sup>27</sup> is a Bayesian regression framework which enables continuous shrinkage 757 priors on SNP effects to infer their posterior mean effects. We ran PRS-CS using both the grid 758 and auto models in the UKBB. In the grid model, we used a series of global shrinkage parameters 759  $(phi = 1 \times 10^{-6}, 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 0.01, 0.1, 1)$ , with lower phi values suggesting less 760 polygenic genetic architecture and vice versa for more polygenic genetic architecture. For the 761 auto model, PRS-CS will learn the phi parameter from the discovery GWAS without requiring 762 post-hoc tuning. We used both total GWAS sample size and effective sample size as input for 763 PRS-CS and found little difference, suggesting that PRS-CS is insensitive to the input of GWAS 764 sample size. We hence used the effective sample size for subsequent analyses in this study. We 765 used the default settings for other parameters. We generalized the auto model for all endpoints 766 in both UKBB and BBJ. When comparing the two models, we selected the optimal phi parameter 767 from the grid model based on the highest prediction accuracy in the target population.

### 768 LD reference panel

769 Both P+T and PRS-CS are summary-level based PRS prediction methods, utilizing GWAS 770 summary statistics and an LD reference panel. To explore the impact of LD reference panels on 771 prediction performance, we used LD reference panels of different ancestral compositions, varying 772 sample sizes and SNP density. Specifically, we used four global ancestry groups, i.e., European 773 (EUR), South-Asian (SAS), East-Asian (EAS) and African (AFR), from 1000G Phase 3 (1KG) as 774 LD reference panels for P+T. Further, we randomly sampled a subset of individuals with sample 775 sizes of 500, 5000, 10,000 and 50,000 from UKBB-EUR to analyze how the sample sizes of LD 776 reference panel would affect prediction accuracy for P+T. Moreover, we ran P+T on both the 777 HapMap3 SNP set and a denser SNP set with genome-wide SNPs. We ran PRS-CS with the LD 778 matrix provided by PRS-CS software<sup>27</sup>, which are based on both 1KG and UKBB populations from 779 those four ancestry groups and Admixed American population (AMR). We performed those 780 analyses using leave-UKBB-out GWAS in GBMI and evaluated the prediction performance in 781 diverse ancestry groups in the UKBB.

### 782

To explore how well EUR-based LD reference approximated the LD of multi-ancestry GWAS in GBMI, we ran LD score regression (LDSC) to estimate the attenuation ratio statistic<sup>21</sup>. The values of attenuation ratio larger than 0.2 suggest a strong LD mismatch between GWAS summary statistics and LD reference panel. We performed LDSC analyses on different GWAS, including GBMI GWAS from meta-analyses on all ancestries, EUR only and leave-one-biobank-out.

# 788 Evaluation of prediction performance

789 After constructing PRS, we evaluated the prediction performance in the independent target 790 datasets. We used a logistic regression to calculate the Nagelkerke's  $R^2$  and variance on the 791 liability-scale explained by PRS as described previously<sup>58</sup>. Area under the receiver operating 792 characteristic curve (AUC) was also reported for full models with additional covariates and models 793 including PRS only. We used bootstrap with 1000 replicates to estimate their corresponding 95% 794 confidence intervals (CIs). Note that the proportion of cases in each ancestry in the target dataset 795 was approximated as the disease population prevalence. The same covariates (usually age, sex 796 and 20 genotypic principal components. PCs) used in the GWAS analyses were included in the 797 full regression model as phenotype ~ PRS + covariates. We also calculated the average  $R^2$  on the liability scale and AUC across biobanks (denoted as  $\overline{R_{liability}^2}$  and  $\overline{AUC}$ , respectively) in each 798 799 ancestry by weighting the effective sample size of each biobank for each endpoint. Further, we 800 divided the target individuals into deciles based on the ranking of PRS distribution. We compared 801 the odds ratio (OR) of the top decile relative to those ranked as the bottom, the middle and the 802 remaining, when using the first decile as the referenced group. For endpoints presented in two or 803 more biobanks, we calculated the averaged OR using the inverse variance weighted method and 804 the coefficient of variation of OR (CoeffVar<sub>OR</sub>) as SD(OR)/mean(OR).

### 805 Resource Availability

### 806 Data and Code Availability

807 The all-biobank and ancestry-specific GWAS summary statistics are publicly available for

808 downloading at <u>https://www.globalbiobankmeta.org/resources</u> and browsed at the PheWeb

809 Browser <u>http://results.globalbiobankmeta.org/</u>. The PRS weights re-estimated using PRC-CS-

810 auto for multi-ancestry GWAS including all biobanks and leave-UKBB-out multi-ancestry GWAS

811 have been uploaded to PGS Catalog (<u>https://www.pgscatalog.org/</u>) under the study ID

812 PGP000262. 1000 Genome Phase 3 data can be accessed at

813 ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\_collections/1000\_genomes\_project/data. We used

- UKB data via application 31063. The software used in this study can be found at: Plink
- 815 (https://www.cog-genomics.org/plink/), PRS-CS (https://github.com/getian107/PRScs),
- 816 SBayesS/GCTB (<u>https://cnsgenomics.com/software/gctb/</u>). The codes used in this study can be
- found in the github repository: <u>https://github.com/globalbiobankmeta/PRS</u>.
- 818
- 819

# 820 References

- Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. Cell. 2019 Mar 21;177(1):58–69.
- 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.
- Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk
   scores. Nat Rev Genet. 2018 Sep;19(9):581–90.
- Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments.
   Genome Med. 2020 May 18;12(1):44.
- 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.
- 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.
- Budbridge 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.
- 840 9. Ni G, Zeng J, Revez JA, Wang Y, Zheng Z, Ge T, et al. A Comparison of Ten Polygenic
  841 Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. Biol Psychiatry.
  842 2021 Nov 1;90(9):611–20.
- Ma Y, Zhou X. Genetic prediction of complex traits with polygenic scores: a statistical
   review. Trends Genet. 2021 Nov;37(11):995–1011.
- Kulm S, Marderstein A, Mezey J. A systematic framework for assessing the clinical impact
  of polygenic risk scores [Internet]. MedRxiv. 2021. Available from:
  https://www.medrxiv.org/content/10.1101/2020.04.06.20055574v2.full-text
- 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
  [Internet]. Cold Spring Harbor Laboratory. 2021 [cited 2021 Jan 28]. p. 2021.01.12.426453.

- 851 Available from: https://www.biorxiv.org/content/10.1101/2021.01.12.426453v1.abstract
- 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.
- 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.
- 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.
- Ruan Y, Anne Feng YC, Chen CY, Lam M, Sawa A, Martin AR, et al. Improving polygenic
  prediction in ancestrally diverse populations [Internet]. medRxiv. 2021. Available from:
  http://medrxiv.org/lookup/doi/10.1101/2020.12.27.20248738
- 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 [Internet].
  medRxiv. 2021. Available from:
- 866 http://medrxiv.org/lookup/doi/10.1101/2021.01.19.21249483
- 18. Zhou W, Kanai M, Wu KHH, Humaira R, Tsuo K, Hirbo JB, et al. Global Biobank Metaanalysis Initiative: powering genetic discovery across human diseases [Internet]. medRxiv.
  2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.11.19.21266436
- 870 19. Zhou W, Kanai M, Wu KHH, Humaira R, Tsuo K, Hirbo JB, et al. Global Biobank Meta871 analysis Initiative: powering genetic discovery across human diseases [Internet]. MedRxiv.
  872 2021. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.11.19.21266436
- 20. 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.
- 876 21. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group
  877 of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes
  878 confounding from polygenicity in genome-wide association studies. Nat Genet. 2015
  879 Mar;47(3):291–5.
- Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A Saturated Map of Common Genetic Variants Associated with Human Height from 5.4 Million Individuals of Diverse Ancestries [Internet]. bioRxiv. 2022 [cited 2022 Jan 11]. p. 2022.01.07.475305.
  Available from: https://www.biorxiv.org/content/10.1101/2022.01.07.475305v1?rss=1
- 23. 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.
- 24. Zhang Y, Qi G, Park JH, 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.

- 890 25. Ware EB, Schmitz LL, Faul J, Gard A, Mitchell C, Smith JA, et al. Heterogeneity in
  891 polygenic scores for common human traits [Internet]. bioRxiv. 2017. p. 106062. Available
  892 from: https://www.biorxiv.org/content/10.1101/106062v1
- 26. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score
  analyses. Nat Protoc [Internet]. 2020 Jul 24; Available from:
  http://dx.doi.org/10.1038/s41596-020-0353-1
- 896 27. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression
   897 and continuous shrinkage priors. Nat Commun. 2019 Apr 16;10(1):1776.
- 898 28. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human
  899 Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J
  900 Hum Genet. 2017 Apr 6;100(4):635–49.
- 29. 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.
- 30. 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.
- 31. 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.
- 32. Lo Faro, Bhattacharya, Zhou, Zhou, Wang, Läll, 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;
- 33. Faro VL, Bhattacharya A, Zhou W, Zhou D, Wang Y, Läll K, et al. Genome-wide association
  meta-analysis identifies novel ancestry-specific primary open-angle glaucoma loci and
  shared biology with vascular mechanisms and cell proliferation [Internet]. medRxiv. 2021.
  Available from: https://www.medrxiv.org/content/10.1101/2021.12.16.21267891.abstract
- Surakka I, Wu KH, Hornsby W, Wolford BN, Shen F, Zhou W, et al. Multi-ancestry metaanalysis identifies 2 novel loci associated with ischemic stroke and reveals heterogeneity of effects between sexes and ancestries [Internet]. bioRxiv. 2022. Available from: https://www.medrxiv.org/content/10.1101/2022.02.28.22271647.abstract
- 921 35. Partanen JJ, Häppölä P, Zhou W, Lehisto AA, Ainola M, Sutinen E, et al. Leveraging global
  922 multi-ancestry meta-analysis in the study of Idiopathic Pulmonary Fibrosis genetics
  923 [Internet]. bioRxiv. 2021. Available from:
  924 https://www.medrxiv.org/content/10.1101/2021.12.29.21268310.abstract
- 36. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, et al. Improving
  reporting standards for polygenic scores in risk prediction studies. Nature. 2021
  Mar;591(7849):211–9.
- 928 37. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, et al. Portability of 245
   929 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from

- 930 the same cohort. Am J Hum Genet. 2022 Feb 3;109(2):373.
- 38. Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR. Challenges and Opportunities for
  Developing More Generalizable Polygenic Risk Scores. Annu Rev Biomed Data Sci
  [Internet]. 2022 May 16; Available from: http://dx.doi.org/10.1146/annurev-biodatasci111721-074830
- 935 39. Graham SE, Clarke SL, Wu KH, Lin K, Millwood IY, Mahajan A, et al. The power of genetic diversity in genome-wide association studies of lipids. Nature [Internet]. 2021 [cited 2021
  937 Dec 10]; Available from: https://ora.ox.ac.uk/objects/uuid:5d0c9801-0dbf-4d5d-8d19938 95606c30a2c0
- Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized
   regression on summary statistics. Genet Epidemiol. 2017 Sep;41(6):469–80.
- 941 41. Miao J, Guo H, Song G, Zhao Z, Hou L, Lu Q. Quantifying portable genetic effects and 942 improving cross-ancestry genetic prediction with GWAS summary statistics [Internet].
  943 bioRxiv. 2022 [cited 2022 Jun 17]. p. 2022.05.26.493528. Available from: 944 https://www.biorxiv.org/content/10.1101/2022.05.26.493528v1
- 945 42. Blatt M, Gusev A, Polyakov Y, Goldwasser S. Secure large-scale genome-wide association studies using homomorphic encryption. Proc Natl Acad Sci U S A. 2020 May 26;117(21):11608–13.
- 948 43. Márquez-Luna C, Loh PR, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA
  949 Type 2 Diabetes Consortium, Price AL. Multiethnic polygenic risk scores improve risk
  950 prediction in diverse populations. Genet Epidemiol. 2017 Dec;41(8):811–23.
- 44. Tsuo, Zhou, Wang, Kanai, Namba, Gupta, et al. Multi-ancestry meta-analysis of asthma
  identifies novel associations and highlights shared genetic architecture across biobanks
  and traits. In preparation. 2021;
- 45. 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.
- 46. 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.
- 47. 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.
- 48. 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.
- 49. 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.
- 967 50. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the
  968 BioBank Japan Project: Study design and profile. J Epidemiol. 2017 Mar;27(3S):S2–8.

- 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.
- 52. 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.
- 974 53. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank
  975 resource with deep phenotyping and genomic data. Nature. 2018 Oct 10;562(7726):203–9.
- 54. 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.
- 55. Leitsalu L, Haller T, Esko T, Tammesoo ML, 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.
- 56. Zawistowski, Fritsche, Pandit, Vanderwerff, Patil, Schmidt, et al. The Michigan Genomics
  Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine
  patients. In preparation. 2021;
- 57. 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.
- 58. 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.
- 59. 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.

# **Figure list**

Figure 1. Overview of the study framework2
Figure 2. Genetic architecture of endpoints in GBMI
Figure 3. Sample size heterogeneity affects PRS prediction accuracy for P+T4
Figure 4. Prediction performance using P+T versus that using PRS-CS5
Figure 5. Prediction performance of PRS-CS across biobanks and ancestries
Figure 6. The odds ratio (OR) between different PRS strata for endpoints in GBMI7
Figure 7. The prediction performance and ancestry compositions of GBMI versus previously published
GWAS



Figure 1. Overview of the study framework.



### Figure 2. Genetic architecture of endpoints in GBMI.

We reported the estimates from using meta-analyzed GWAS from all ancestries (labeled as All ancestries) and European only (labeled as EUR), respectively. 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. Abbreviations: Europeans (EUR), chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC).



# 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 S7 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. The prediction accuracy in P+T was estimated in the test cohort based on the optimal p-value thresholds finetuned in the validation cohort. The auto model was used for PRS-CS. Abbreviations: Europeans (EUR), Admixed Americans (AMR), Middle Eastern (MID), Central and South Asians (CSA), East Asians (EAS) and Africans (AFR), chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC).



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 traitancestry pairs in each biobank are provided in **Table S6**. Note that we removed the estimates in AMR and MID due to limited information as a result of small sample sizes. Abbreviations: Europeans (EUR), Admixed Americans (AMR), Middle Eastern (MID), Central and South Asians (CSA), East Asians (EAS) and Africans (AFR), chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC).



### 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 full results are shown in Table S7. 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). Abbreviations: Europeans (EUR), Admixed Americans (AMR), Middle Eastern (MID), Central and South Asians (CSA), East Asians (EAS) and Africans (AFR), chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC).



# Figure 7. The prediction performance and ancestry compositions of GBMI versus previously published GWAS.

A) The ancestry compositions of GBMI and referenced GWAS. The label for biobanks in the x-axis indicated the leave-one-out-biobank meta-analyzed GWAS in GBMI. The previously published GWAS was labeled as Referenced. B) The comparison of AUC between GBMI and referenced GWAS. The AUC was calculated by fitting PRS only. The phenotypes in A) were ranked based on the effective sample sizes from all ancestries. The phenotypes in B) were 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. The full results are shown in **Table S4**. Abbreviations: Europeans (EUR), Admixed Americans (AMR), Middle Eastern (MID), Central and South Asians (CSA), East Asians (EAS) and Africans (AFR), chronic obstructive pulmonary disease (COPD), heart failure (HF), acute appendicitis (AcApp), venous thromboembolism (VTE), primary open-angle glaucoma (POAG), uterine cancer (UtC), abdominal aortic aneurysm (AAA), idiopathic pulmonary fibrosis (IPF), thyroid cancer (ThC).