Global Biobank Meta-analysis Initiative: powering genetic discovery across human diseases

Wei Zhou[1,2,3], Masahiro Kanai[1,2,3,4,5], Kuan-Han H Wu[6], Rasheed Humaira[7,8,9], Kristin Tsuo[1,2,3], Jibril B Hirbo[10,11], Ying Wang[1,2,3], Arjun Bhattacharya[12], Huiling Zhao[13], Shinichi Namba[5], Ida Surakka[14], Brooke N Wolford[6], Valeria Lo Faro[15,16,17], Esteban A Lopera-Maya[18], Kristi Läll[19], Marie-Julie Favé[20], Sinéad B Chapman[2,3], Juha Karjalainen[1,2,3,21], Mitja Kurki[1,2,3,21], Maasha Mutaamba[1,2,3,21], Ben M Brumpton[22], Sameer Chavan[23], Tzu-Ting Chen[24], Michelle Daya[23], Yi Ding[12,25], Yen-Chen A Feng[26], Christopher R Gignoux[23], Sarah E Graham[14], Whitney E Hornsby[14], Nathan Ingold[27], Ruth Johnson[12,28], Triin Laisk[19], Kuang Lin[29], Jun Lv[30], Iona Y Millwood[29,31], Priit Palta[19,21], Anita Pandit[32], Michael Preuss[33], Unnur Thorsteinsdottir[34], Jasmina Uzunovic[20], Matthew Zawistowski[32], Xue Zhong[10,35], Archie Campbell[36], Kristy Crooks[23], Geertruida h De Bock[37], Nicholas J Douville[38,39], Sarah Finer[40], Lars G Fritsche[32], Christopher J Griffiths[40], Yu Guo[41], Karen A Hunt[42], Takahiro Konuma[5,43], Riccardo E Marioni[36], Jansonius Nomdo[15], Snehal Patil[32], Nicholas Rafaels[23], Anne Richmond[44], Jonathan A Shortt[23], Peter Straub[10,35], Ran Tao[35,45], Brett Vanderwerff[32], Kathleen C Barnes[23], Marike Boezen[37], Zhengming Chen[29,31], Chia-Yen Chen[46], Judy Cho[33], George Davey Smith[13,47], Hilary K Finucane[1,2,3], Lude Franke[18], Eric Gamazon[35,48], Andrea Ganna[1,2,21], Tom R Gaunt[13], Tian Ge[49,50], Hailiang Huang[1,2], Jennifer Huffman[51], Clara Lajonchere[52,53], Matthew H Law[27], Liming Li[30], Cecilia M Lindgren[54], Ruth JF Loos[33], Stuart MacGregor[27], Koichi Matsuda[55], Catherine M Olsen[27], David J Porteous[36], Jordan A Shavit[56], Harold Snieder[37], Richard C Trembath[57], Judith M Vonk[37], David Whiteman[27], Stephen J Wicks[23], Cisca Wijmenga[18], John Wright[58], Jie Zheng[13], Xiang Zhou[32], Philip Awadalla[20,59], Michael Boehnke[32], Nancy J Cox[10,60], Daniel H Geschwind[52,61,62], Caroline Hayward[44], Kristian Hveem[22], Eimear E Kenny[63], Yen-Feng Lin[24,64,65], Reedik Mägi[19], Hilary C Martin[66], Sarah E Medland[27], Yukinori Okada[5,67,68,69,70], Aarno V Palotie[1,2,21], Bogdan Pasaniuc[12,25,52,61,71], Serena Sanna[18,72], Jordan W Smoller[73], Kari Stefansson[34], David A van Heel[42], Robin G Walters[29,31], Sebastian Zoellner[32], Biobank Japan, BioMe, BioVU, Canadian Partnership for Tomorrow, China Kadoorie Biobank Collaborative Group, Colorado Center for Personalized Medicine, deCODE Genetics, Estonian Biobank, FinnGen, Generation Scotland, Genes & Health, LifeLines, Mass General Brigham Biobank, Michigan Genomics Initiative, QIMR Berghofer Biobank, Taiwan Biobank, The HUNT Study, UCLA ATLAS Community Health Initiative, UK Biobank, Alicia R Martin[1,2,3], Cristen J Willer[6,14,74*], Mark J Daly[1,2,3,21*], Benjamin M Neale[1,2,3*]

*These authors jointly supervised this work

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA, [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, [5]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan, [6]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA, [7]K. G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health, NTNU, Norwegian University of Science and Technology, Trondheim, Norway, [8]Division of Medicine and Laboratory Sciences, University of Oslo, Norway, [9]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK, [10]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA, [11]Vanderbilt Genetic Institute, Vanderbilt University

Medical Center, Nashville, TN, USA, [12]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [13]MRC Integrative Epidemiology Unit (IEU), Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, UK, [14]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA, [15]University of Groningen, UMCG, Department of Ophthalmology, Groningen, the Netherlands, [16]Department of Clinical Genetics, Amsterdam University Medical Center (AMC), Amsterdam, the Netherlands, [17]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden, [18]University of Groningen, UMCG, Department of Genetics, Groningen, the Netherlands, [19]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia, [20]Ontario Institute for Cancer Research, Toronto, ON, Canada, [21]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland, [22]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim, 7030, Norway, [23]University of Colorado - Anschutz Medical Campus, Aurora, CO, USA, [24]Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan, [25]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA, [26]Division of Biostatistics, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taiwan, [27]QIMR Berghofer Medical Research Institute, Brisbane, Australia, [28]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA, [29]Nuffield Department of Population Health, University of Oxford, Oxford, UK, [30]Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China, [31]MRC Population Health Research Unit, University of Oxford, Oxford, UK, [32]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA, [33]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [34]deCODE Genetics/Amgen inc., 101, Reykjavik, Iceland, [35]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashviile, TN, USA, [36]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK, [37]University of Groningen, UMCG, Department of Epidemiology, Groningen, the Netherlands, [38]Department of Anesthesiology, Michigan Medicine, Ann Arbor, MI, USA, [39]Institute of Healthcare Policy & Innovation, University of Michigan, Ann Arbor, MI, USA, [40]Wolfson Institute of Population Health, Queen Mary University of London, London, UK, [41]Chinese Academy of Medical Sciences, Beijing, China, [42]Blizard Institute, Queen Mary University of London, London, UK, [43]Central Pharmaceutical Research Institute, JAPAN TOBACCO INC., Takatsuki 569-1125, Japan, [44]Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK, [45]Department of Biostatistics, Vanderbilt University Medical Center, Nashviile, TN, USA, [46]Biogen, Cambridge, MA, USA, [47]NIHR Bristol Biomedical Research Centre, Bristol, UK, [48]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK, [49]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, [50]Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA, [51]Centre for Population Genomics, VA Boston Healthcare System, Boston, MA, USA, [52]Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA, USA, [53]Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [54]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK, [55]Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan, [56]University of Michigan, Department of Pediatrics, Ann Arbor MI 48109, [57]School of Basic and Medical Biosciences, Faculty of Life Sciences and Medicine, King's College London, London, UK, [58]Bradford Institute for Health Research, Bradford Teaching Hospitals National Health Service (NHS) Foundation Trust, Bradford, UK, [59]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, [60]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA, [61]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [62]Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [63]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New

York, NY, USA, [64]Department of Public Health & Medical Humanities, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan, [65]Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan, [66]Medical and Population Genomics, Wellcome Sanger Institute, Hinxton, UK, [67]Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita 565-0871, Japan, [68]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan, [69]Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [70]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan, [71]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [72]Institute for Genetics and Biomedical Research, National Research Council, Cagliari 09100, Italy, [73]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, [74]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

Summary

Biobanks are being established across the world to understand the genetic, environmental, and epidemiological basis of human diseases with the goal of better prevention and treatments. Genome-wide association studies (GWAS) have been very successful at mapping genomic loci for a wide range of human diseases and traits, but in general, lack appropriate representation of diverse ancestries - with most biobanks and preceding GWAS studies composed of individuals of European ancestries. Here, we introduce the Global Biobank Meta-analysis Initiative (GBMI) -- a collaborative network of 19 biobanks from 4 continents representing more than 2.1 million consented individuals with genetic data linked to electronic health records. GBMI meta-analyzes summary statistics from GWAS generated using harmonized genotypes and phenotypes from member biobanks. GBMI brings together results from GWAS analysis across 6 main ancestry groups: approximately 33,000 of African ancestry either from Africa or from admixed-ancestry diaspora (AFR), 18,000 admixed American (AMR), 31,000 Central and South Asian (CSA), 341,000 East Asian (EAS), 1.4 million European (EUR), and 1,600 Middle Eastern (MID) individuals. In this flagship project, we generated GWASs from across 14 exemplar diseases and endpoints, including both common and less prevalent diseases that were previously understudied. Using the genetic association results, we validate that GWASs conducted in biobanks worldwide can be successfully integrated despite heterogeneity in case definitions, recruitment strategies, and baseline characteristics between biobanks. We demonstrate the value of this collaborative effort to improve GWAS power for diseases, increase representation, benefit understudied diseases, and improve risk prediction while also enabling the nomination of disease genes and drug candidates by incorporating gene and protein expression data and providing insight into the underlying biology of the studied traits.

Keywords
biobank meta-analysis, genetic association studies, phenotype harmonization, ancestry diversity

Introduction

Understanding the genetic basis of disease can elucidate the biology or underlying epidemiological risk factors, nominate genes as drug targets, and identify at-risk individuals for prevention strategies. Genetic association studies have been routinely performed genome-wide for over 15 years and have identified thousands of loci for hundreds of diseases and traits (see GWAS Catalog (MacArthur et al., 2017)). Meta-analysis across cohorts has been instrumental in making these discoveries. However, most genomics research has been performed primarily in cohorts of European ancestries and conducted in mostly high-resource countries. Although much remains to be done to address the lack of representation in genomics, here we present the Global Biobank Meta-analysis Initiative (GBMI), a small step in building a more comprehensive view of the impact of genetic variation on human health and disease.

Biobanks with health data linked with genomic information provide unprecedented resources for the genetic research community. The rapid drop in the cost of genotyping and sequencing has led to an increase in the number of genomically profiled biobanks worldwide. Compared to disease or trait-based cohorts centered around a particular phenotype or several relevant phenotypes, biobanks enable cost-effective genetic discovery for hundreds to thousands of phenotypes, curated from electronic health records (EHRs), registry-based data (e.g. pharmaceutical, death, or cancer registry data), and/or epidemiological questionnaires to understand the genetic etiology of human diseases (Bowton et al., 2014; Wolford et al., 2018).

In 2019, we formed the GBMI bringing together 19 biobanks to work together to understand the genetic basis of human health and disease (**Figure 1, Supplementary Table 1**). The goal was to jumpstart and align global efforts, particularly since meta-analysis of GWAS is simple in terms of data sharing yet enables a variety of scientific goals including: increasing the power of GWASs for common diseases, enabling the genetic investigation into less prevalent or understudied diseases, increasing the ancestral diversity of genetic association studies and in doing so analyzing a broader set of genetic variation, cross-validating new

At the heart of the GBMI is a community of investigators that have adopted seven principles to guide our collaboration:

1. Collaborate in an environment of honesty, fairness and trust;

2. Promote early-career researchers;

3. Respect other groups' data;

4. Operate transparently with a goal of no surprises;

5. Seek permission from each group to use results prior to public release;

6. Do not share another group's results with other parties without permission;

7. Do not inhibit any work being done within an individual biobank (or between pairs of biobanks).

findings across biobanks, and facilitating follow-up analyses such as polygenic risk scores or Mendelian Randomization.

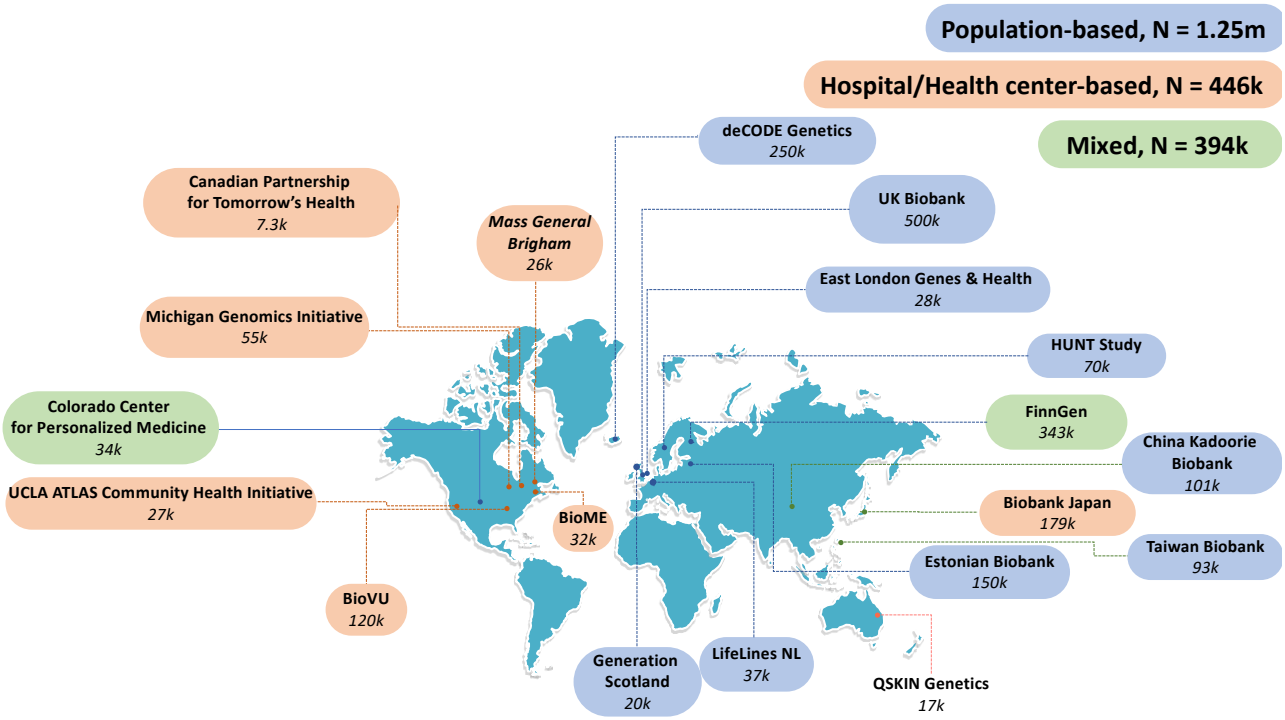

Figure 1. 19 biobanks across four continents have joined in GBMI as of November 2021, bringing the total number of samples with matched health data and genotypes to more than 2.1 million. Biobanks are colored based on the sample recruiting strategies.

Here we present the pilot effort of GBMI, in which we meta-analyzed GWAS results for 14 endpoints of common interest (**Supplementary Table 2**). These include diseases across a more than 30x prevalence range: asthma (153,763 cases (sample prevalence: 8.54%)), chronic obstructive pulmonary disease (COPD, 81,568 cases (5.86%)), heart failure (HF, 68,408 cases (5.05%)), and stroke (60,176 cases (4.39%)), gout (37,105 cases (2.50%)), venous thromboembolism (VTE, 27,987 cases (2.63%)), primary open-angle glaucoma (POAG, 26,848 cases (1.80%)), abdominal aortic aneurysm (AAA, 9,453 cases (0.65%)), idiopathic pulmonary fibrosis (IPF, 8,006 cases (0.64%)), thyroid cancer (ThC, 6,699 cases (0.41%)), and cardiomyopathy (HCM, 2,993 cases (0.25%)), a female-specific disease: uterine cancer (UtC, 8,295 cases (1.2%)), and to examine procedure-related phenotypes: acute appendicitis (AcApp, 32,706 cases (2.95%)) and the related appendectomy procedure (14,446 cases (1.86%)), which is an endpoint phenotype that can be extracted from EHR procedure codes but has not been widely studied in previous GWAS. As a proof of concept, using aligned phenotype definitions, analysis methods, sharing standards, and quality control, we demonstrate the advantages of aggregating biobanks together for genetic studies of human diseases.

Results

Overview of biobanks in GBMI

GBMI represents 2.1 million research participants with health and genetic data from nineteen biobanks across four continents: one from Australia, three from East Asian countries, eight from European countries, and seven ascertained in North America. Specifically, the biobanks included were: QSKIN Genetics (QSkin)(Olsen et al., 2012) in Australia, Biobank Japan (BBJ)(Nagai et al., 2017), China Kadoorie Biobank (CKB)(Chen et al., 2011), and Taiwan Biobank (TWB)(Feng et al., 2021) in East Asia, deCODE Genetics (deCode)(Gudbjartsson et al., 2015), Estonian Biobank (ESTBB)(Leitsalu et al., 2015), FinnGen, Generation Scotland (GS)(Smith et al., 2013), Genes & Health (GNH)(Finer et al., 2020), LifeLines(Scholtens et al., 2015), Trøndelag Health Study (HUNT)(Krokstad et al., 2013), and UK Biobank (UKBB)(Bycroft et al., 2018) in Europe, and BioME(Abul-Husn et al., 2021), BioVU(Bowton et al., 2015), Canadian Partnership for Tomorrow's Health (CanPath)(Dummer et al., 2018), Colorado Center for Personalized Medicine Biobank (CCPM)(Aquilante et al., 2020), Mass General Brigham Biobank (MGB)(Karlson et al., 2016), Michigan Genomics Initiative (MGI) (Zawistowski et al., 2021) , and UCLA ATLAS Community Health Initiative (UCLA)(Johnson et al., 2021) in North America. **Supplementary Table 1** presents a brief summary of the biobanks in GBMI, including basic information about each biobank (location, institute, cohort size, and sample recruiting approach), participants (ancestry and age), types of electronic health data (self-report data from epidemiological survey questionnaires, billing codes, doctors' narrative notes, and death registry, etc.) and genotypes (genotyping platforms and imputation reference), as well as data access and references (webpage if available).

Biobanks differ in many aspects, such as locations, sample sizes, genotyping and phenotyping approaches, follow-up time (longitudinal data and samples), and strategies to recruit participants: community/population-based, health center/hospital-based, or mixed (**Figure 1**). As a result, disease prevalence varies across biobanks (**Supplementary Figure 1**) and across sample recruiting strategy groups (**Supplementary Figure 2A**). Biobanks recruiting participants from health centers or hospitals, relative to those recruiting participants from the general population, had a significantly higher prevalence (Wilcoxon test p-value < 0.05) for six out of 13 examined diseases (Appendectomy was excluded from the test due to insufficient data shared from the hospital-based biobanks) (**Supplementary Figure 2B**), including asthma, heart failure (HF), stroke, venous thromboembolism (VTE), gout, and Idiopathic pulmonary fibrosis (IPF).

GBMI incorporates diverse ancestries in genetic studies by including biobank samples of 6 main ancestry groups: approximately 33,000 of African ancestries either from Africa or from admixed-ancestry diaspora (AFR), 18,000 admixed American (AMR), 31,000 Central and South Asian (CSA), 341,000 East Asian (EAS), 1.4 million European (EUR), and 1,600 Middle Eastern (MID) individuals. To compare the ancestries represented between the different biobanks, we projected biobanks' participants to the same principal components (PC) space (**Figure 2**) using the pre-computed loadings of genetic markers overlapping in all biobanks and the reference containing 1000 Genomes (1000 Genomes Project Consortium et al., 2015) and Human Genome Diversity Project (HGDP) (Cann et al., 2002). PCs projected in the same space enables the cross-comparison of the sample ancestry among all biobanks (**Methods).** Notably, the population labels used in GBMI were defined by global genetic

reference datasets, but GBMI is not globally representative; for example, given that the majority of individuals assigned to AMR and AFR ancestry groups are mostly from biobanks in the US, GBMI participants' ancestries are not currently representative of broader Central/South American or continental African ancestries, respectively.

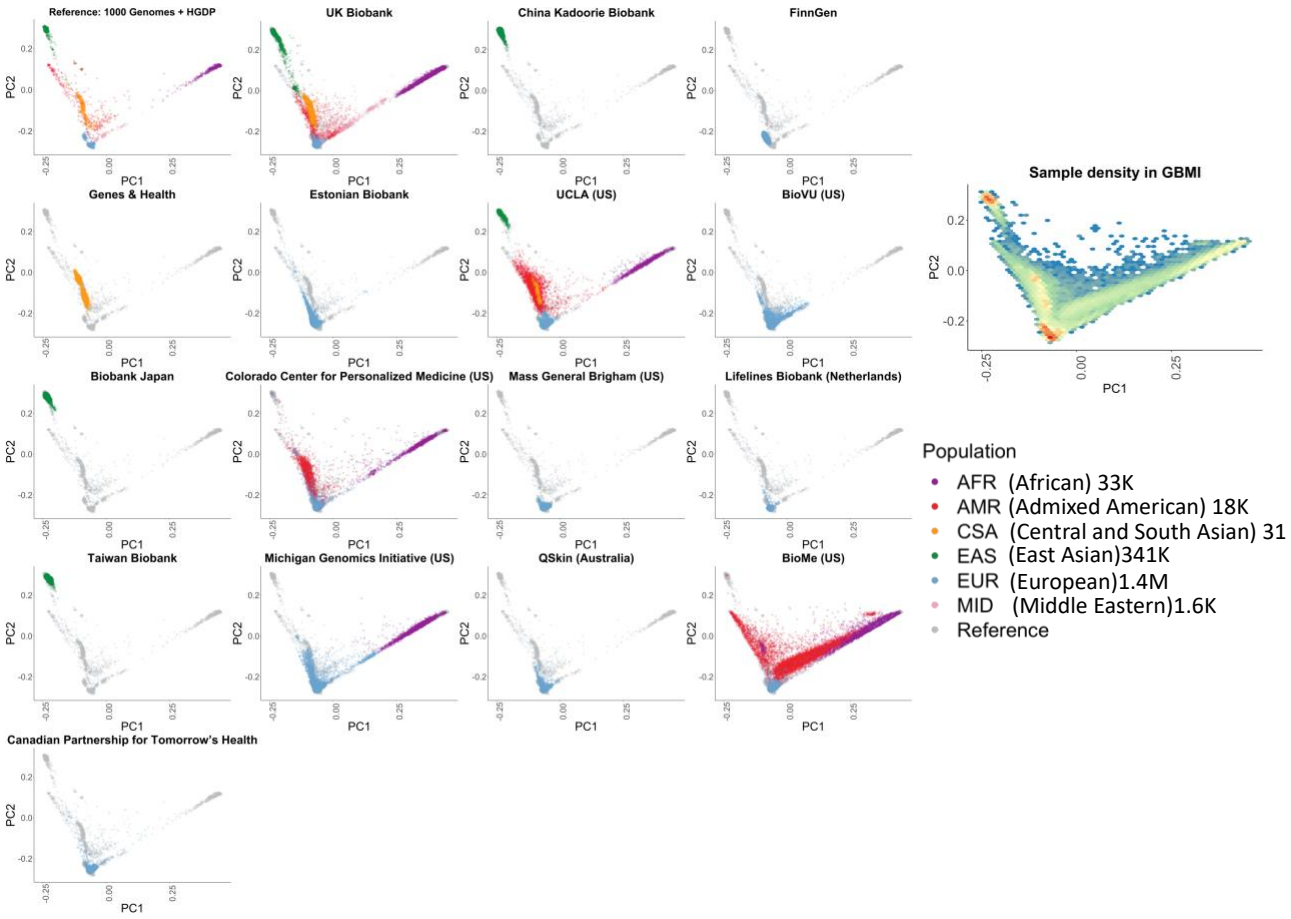

Figure 2. GBMI incorporates biobanks with diverse sample ancestry into genetic studies. Biobanks' participants were projected to the same principal components (PC) space using the pre-computed loadings of genetic markers.

All-Biobank Meta-analysis

The overview of the meta-analysis is presented in **Supplementary Figure 3**. We harmonized the phenotype definitions primarily by mapping the International Classification of Diseases (ICD) codes to phecodes (Denny et al., 2013) for diseases and Classification of Interventions and Procedures codes (OPCS) for procedures, which were used by biobanks to curate phenotypes based on the health data available (**Supplementary Table 3** and **4**). After standard sample-level and variant-level quality control (**Supplementary Table 1**), GWASs stratified by ancestry and sex were conducted in each biobank. The central analysis team performed post-GWAS variant-level quality control for each

biobank by flagging markers with different allele frequencies compared to gnomAD (Karczewski et al., 2020) and excluding markers with imputation quality score < 0.3 (**Methods**). Across all biobanks, 70.8 million genetic variants were tested for associations, of which 2.9 million variants are protein coding (**Supplementary Table 5**).

We first highlight the power of GBMI to uncover novel genetic associations through increased sample size. Inverse variance-based meta-analysis of all biobanks for the 14 endpoints successfully replicated 320 previously reported loci and identified 188 apparently novel loci, spanning the frequency spectrum of less frequent to common variants (**Supplementary Table 6, Figure 3**)**.** At 87 loci, a protein coding variant was the most significant variant (n=26) (**Table 1**) or in linkage disequilibrium with the most significant variant with $r^2 > 0.8$ (n=66 additional) at the locus. 18 of these 87 loci are novel (**Supplementary Table 7**). Furthermore, 13 out of 14 endpoints have SNP-based heritability significantly different from 0 on the liability scale, (under the assumption that the population prevalence matches the prevalence of all biobanks aggregated together), ranging from 1.79% (acute appendicitis) to 10.73% (Gout) (**Supplementary Table 8**).

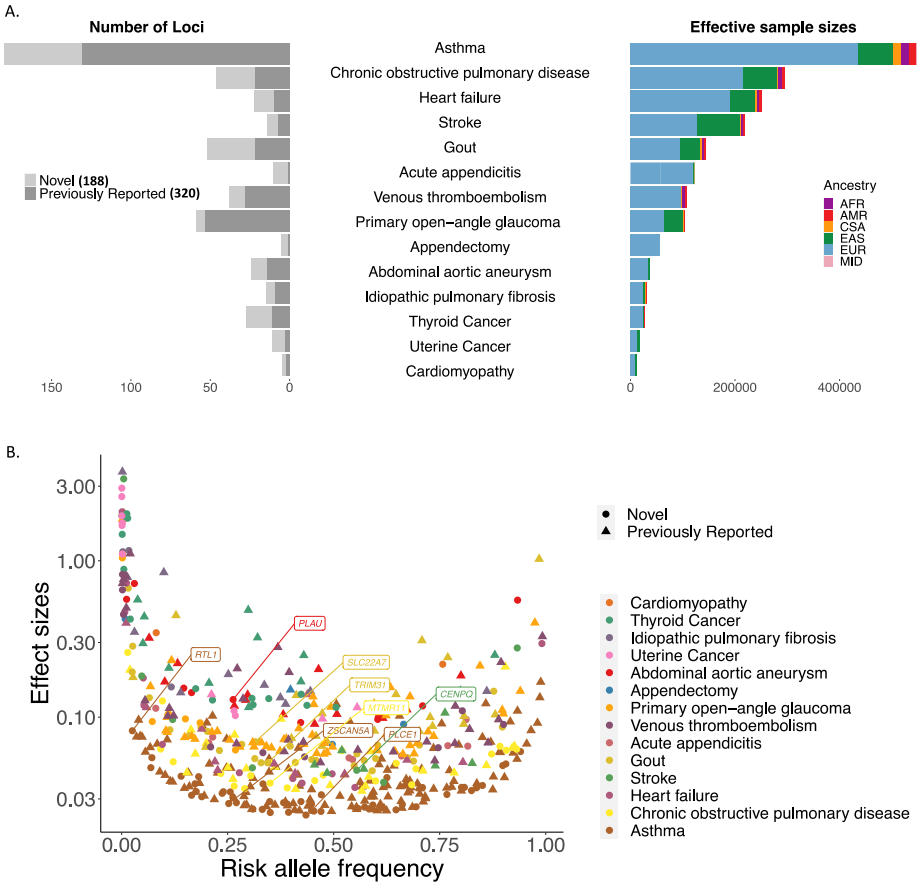

Figure 3: All-biobank meta-analysis for the 14 endpoints have successfully replicated 320 previously reported loci and identified 188 novel loci. A. number of loci were plotted for each endpoint (left panel) against the effective sample sizes 1/(4/cases + 4/controls) colored by the sample ancestry (right panel). B. Top hits spread over the entire allele frequency spectrum. Phenotypes are in ascending order by the effective sample sizes. 3 markers with beta > 5 are not shown. Gene names are labeled for the novel loci with protein-coding index variants.

Secondly, we show that identified associations were largely shared across biobanks. The lead variants at ~96% of the 508 genome-wide loci did not show evidence for heterogeneity in effect sizes across different data sets (per biobank and ancestry) (**Supplementary Table 6**) with p-value for Cochran's Q test ≥ 1/508, despite biobanks differing in many aspects, as discussed above. This suggests that harmonizing phenotyping and then integrating GWASs from different biobanks together with the analysis pipeline within GBMI enables reliable discoveries for genetic-disease associations. Out of the lead variants for 27 loci showing evidence for heterogeneity in effect sizes, 11 have heterogeneous effect sizes across ancestry groups (**Supplementary Table 6**). We additionally used the meta-regression approach implemented in MR-MEGA (Mägi et al., 2017) to conduct the all-biobank meta-analysis across all ancestries. In contrast with a fixed-effects, inverse variance-based meta-analysis, MR-MEGA accounts for the effect size heterogeneity across data sets. This led to the identification of 17 additional loci across 10 endpoints, including 12 that were novel (**Supplementary Figure 4, Supplementary Table 9**).

Power improved by incorporating samples with non-European ancestries

An additional 21.8 million genetic variants were analyzed in the all-biobank meta-analysis which were not present in the European HRC-imputed variant sets. The majority of these variants were rare, with 18 million having a MAF ≤ 1%, and the other 3.8 million were common in at least one ancestry group (**Supplementary Figure 5**). Incorporating samples with diverse ancestries to the biobank meta-analysis enabled the comparison of effect sizes of genomic loci across ancestry at associated loci. Nine out of the 499 loci that were tested in more than one ancestry showed evidence for heterogeneity in effect sizes across ancestry (p-value for Cochran's Q test across ancestry < 1/499) (**Supplementary Table 10**). 343 loci were identified in the European-only meta-analysis and the inclusion of the non-European samples yielded an additional 165 loci (**Supplementary Figure 6A, Supplementary Table 11)**, bringing the total number of loci to 508 in the all-biobank meta-analysis. While increase in sample size will drive some of this increase alone, the increased diversity allows to identify loci whose index variants are much more frequent in other ancestries than in European ancestries. In particular, in contrast to only 4 out of the 343 loci (1.16%), 21 out of the 165 additional loci (12.7%) have index variants that are at least 10 times more frequent in other ancestries than in European ancestries with the frequency lower than 5% in European ancestries (**Supplementary Table 11)**. Forest plots of several example loci are shown in **Supplementary Figure 6B,** highlighting loci whose index variants that are more frequent in East Asians than Europeans and other ancestries (*MIR2054/INTU* for POAG, *PNPT1/EFEMP1* for COPD, and *NAA38* for Asthma) as well as loci driven by variants that are more frequent in African ancestry than other ancestries, including *VPS13D/DHRS3* for VTE, *BCL2L12* for HF, and *MEIS2/TMCO5A* for stroke.

Sex-stratified meta-analyses

We performed sex-stratified meta-analysis to compare GWAS effect sizes between sexes. 479 loci were tested in more than one biobank for both male-only and female-only meta-analyses, of which eight loci showed evidence of heterogeneous effect sizes between males and females (p-value for

Cochran's Q test < 1/479) in the all-biobank meta-analysis (**Supplementary Table 12**). This included one novel locus for the studied diseases: a region containing *CTDP1/KCNG2* for asthma (**Supplementary Figure 7A**) and seven previously identified loci (**Supplementary Figure 7B**).

Environmental factors, such as smoking status and alcohol usage, that are different in males and females may play a role in GWAS effect size differences among sexes. For example, the most strongly associated variant in the *CTDP1/KCNG2* locus (rs11665567) is an intergenic variant with a female-specific association for asthma (in females allele frequency(AF) = 18.8%, effect size (se) = 0.05 (0.008), p-value = 5.62×10^{-10}; in males AF=18.7%, effect size (se) = 0.003 (0.01), p-value = 0.75 (p-value for difference=2.4×10^{-4}).This locus was previously shown to be associated with smoking status (Liu M et. al. 2019). In addition, we have replicated two previously reported loci that are located at the aldehyde dehydrogenase family genes for gout exhibiting stronger associations in males than in females (Mizuno et al., 2015; Sulem et al., 2011). One is the East Asian-specific intronic variant rs4646776 (r^2 = 0.99 with the missense variant rs671(Matoba et al., 2020)) located at the gene *ALDH2* with a stronger effect in males than in females (in females AF = 20.4%, effect size (se) = -0.10 (0.056), p-value = 0.07, in males AF=24.2%, effect size (se) = -0.29(0.023), p-value = 2.5×10^{-36}). This has been attributed to the higher frequency of habitual alcohol drinking in males than females among individuals of East Asian ancestries (Mizuno et al., 2015). The other one is the low-frequency European-specific intronic variant located in *ALDH16A1* that is associated with serum urate levels (Sulem et al., 2011) more strongly in men(rs752383928 intronic, in females AF = 0.74%, effect size (se) = 1.63 (0.29), p-value = 2.43×10^{-8}, in males AF=0.73%, effect size (se) = 2.70 (0.18), p-value = 1.33×10^{-50}). To clarify whether the sex-specific associations identified are due to pleiotropic effects of the genetic variants, environmental factors, or possible gene-environment interactions requires further study.

In addition, we uncovered significant sex differences for five previously reported loci: *RANBP6/IL33* for asthma(Demenais et al., 2017), *AFAP1* for COPD(Wyss et al., 2018), *PKD2* for gout (Matsuo et al., 2016), *MUC5AC/MUC5B* for IPF(Seibold et al., 2011), and *ARHGEF12* for POAG (Springelkamp et al., 2015).

Furthermore, there were 31 loci only identified in the sex-stratified meta-analyses but not in the sex-combined meta-analyses (p-value > 5×10^{-8}), of which 11 loci were detected in female-only meta-analyses and 20 loci only in male-only meta-analyses. 26 out of the 31 loci are potentially novel for the studied phenotypes (**Supplementary Table 13**). The female-only meta-analysis for stroke identified the previously reported locus *CETP* (Buraczynska et al., 2018) that did not reach the genome-wide significance threshold in the sex-combined meta-analysis. The top hit is an intronic rs7499892 variant with stronger association in females than males (in females: effect size (se) = 0.078 (0.014), p-value =1.08×10^{-8}, in males effect size (se) = 0.007(0.012), p-value = 0.56) .Previous studies have shown that the transgenic expression of *CETP* increases plasma triglyceride levels in females and males through distinct mechanisms (Palmisano et al., 2016, 2021).

Integration of association results across biobanks

Leveraging the heterogeneity among biobanks, we used the genetic association results to evaluate the integration of biobanks in the meta-analyses for genetic discovery. First, we compared the effect sizes of the index variants with p-value < $1x10^{-10}$ by the all-biobank meta-analysis in each biobank and leave-one-biobank-out meta-analysis (LOBO). For the biobank and LOBO pairs, we fit a Deming regression model (Deming, 1943), which accounts for the standard errors of effect sizes in both association datasets, with the intercept set to zero. In **Supplementary Figure 8,** the slope estimates for the biobank and LOBO pairs were plotted against the effective sample sizes with biobanks annotated by phenotype source (health records (e.g. ICD codes and physician's diagnosis) only or self-reported data included), sampling strategy, and sample ancestry. Most of the slope estimates were not significantly different from one across biobanks and phenotypes suggesting the genetic association results are robust despite differences between biobanks. However, we observed exceptions to this among biobanks with relatively smaller sample sizes and/or non-European or multiple ancestries. For example, the multi-ancestry biobanks BioMe, BioVU, and UCLA as well as GNH, which included SAS samples, have different effects compared to others for multiple phenotypes, including gout, HF, VTE, and POAG. Note that POAG tends to have more phenotypic heterogeneity due to glaucoma types not being well defined by self-reported data, leading to the inclusion of other types of glaucoma, such as the angle-closure glaucoma. As expected, the three biobanks using self-reported data for phenotype curation, Lifelines, TWB, and BBJ, also showed effect size differences for POAG compared to other biobanks (**Supplementary Figure 8**).

Next, we estimated the genetic correlation between individual biobanks and LOBO for the three endpoints with the highest heritability estimates: asthma, gout, and COPD. Genetic correlation estimates between biobanks and the LOBO were close to 1, although genetic correlation was only possible to estimate for biobanks with non-zero heritability estimates (p-value < 0.05) (**Methods**) (**Supplementary Figure 9**). We then compared the all-biobank meta-analyses with previously published GWAS studies. Among previously reported loci, we show consistent effect directions between GBMI and the previous largest studies. For example, all 18 loci for asthma that were previously identified by the Trans-National Asthma Genetic Consortium (TAGC) (Demenais et al., 2017) have consistent effect directions in GBMI. Similarly, 24 out of 25 previously identified loci by MVP for AAA (Klarin et al., 2020) and 40 out of 40 previously identified loci for gout (Tin et al., 2019) show effect size consistency between GBMI and the previous GWASs (**Supplementary Table 14**). Note that by cross-comparing the cohort lists in previous studies and GBMI, no sample overlap was noted for Asthma and AAA, while 3 biobanks in GBMI (BioVU, GS, and UKBB) were also included in the previous meta-analysis for gout (Tin et al., 2019), accounting for about 20% of the GBMI samples.

Biological implications of genetic associations

Pleiotropic effects of associated loci

We investigated the genetic relationship between the endpoints and other complex traits by examining the associations of the top variants identified by the all-biobank meta-analyses with 1,238 human diseases in UKBB (**Methods**). Of the 430 loci whose index variants were available in the UKBB

GWAS data, 78 variants identified from 13 GBMI endpoints (except for HCM) exhibited significant (p-value < 5×10^{-8}) pleiotropic associations with at least one other phenotype (**Supplementary Table 15**). Risk increasing alleles of the top variants at two asthma-associated loci, the known asthma locus *BACH2* and the novel locus *FGFR1OP*, are both associated with a reduced risk of hypothyroidism. The risk increasing allele of the top variant at the novel locus *GOT1/LINC01475* for acute appendicitis (AcApp) is associated with the decreased risk of ulcerative colitis. A previous study also observed a low risk of ulcerative colitis among people who had undergone an appendectomy for appendicitis and mesenteric lymphadenitis(Andersson et al., 2001), but the reason for this remains unclear.

Prioritization of cell types, tissues, and genes

To further understand the biology underlying the genetic associations, we first prioritized the tissues and cell types in which genes at the associated loci are likely to be highly expressed using the Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT)(Pers et al., 2015) (**Supplementary Table 16).** For example, at FDR < 0.05, the adrenal cortex, which releases the mineralocorticoid aldosterone, was prioritized for AAA consistent with previous functional studies which have shown that the mineralocorticoid aldosterone can induce aortic aneurysm and dissection in the presence of high salt(Liu et al., 2013). Prioritized tissue types for asthma included lymphoid tissue and immune systems (blood cells, antigen presenting cells, and myeloid cells) as well as nasal and respiratory mucosa. Besides muscle cells and connective tissue cells, heart and blood vessels were identified for POAG(Lo Faro et al., 2021).

Next, several methods were used to prioritize the potentially functional genes, including DEPICT (**Supplementary Table 17)**, the gene-level Polygenic Priority Score (PoPS) (Weeks et al., 2020) (**Supplementary Table 18)**, transcriptome-wide association studies (TWAS) (Bhattacharya et al., 2021a) (**Supplementary Table 19)**, and proteome-wide Mendelian randomization (PWMR) (Zhao et al., 2021) (**Methods**). Using asthma, POAG, and VTE as examples, the gene lists generated by these different methods showed quite little overlap (**Supplementary Figure 10**). For asthma, 618 genes were prioritized by at least one of the four approaches (FDR < 0.05 by DEPICT, top 1% scores in PoPS, P < 2.5×10^{-6} by TWAS, P < 0.001 by PWMR) (**Supplementary Figure 10A**). However no genes were prioritized by all four methods and 5 were prioritized by any three out of the four methods (*FCER1G, IL18R1, IL4R*, and *SMAD3* by DEPICT, TWAS, and PoPS and *IL2RB* by DEPICT, PoPS, and PWMR) (**Supplementary Table 20**). All these genes are located at the well-known asthma-associated loci. *FCER1G* encodes the Fc Fragment of IgE Receptor Ig, and the *IL18R1, IL4R,* and *IL2RB* encode Interleukin receptors, which are all involved in the immune system. Dupilumab, an anti-interleukin 4 receptor alpha monoclonal antibody, blocks IL-4 and IL-13 and decreases IgE over time, is an FDA approved add-on therapy for asthma (Castro et al., 2018; Rabe et al., 2018). *SMAD3* encodes a transcription factor whose methylation has been shown to be associated with neonatal production of IL-1β and childhood asthma risk (DeVries et al., 2017). Similarly, for POAG, 204 genes were prioritized, but no genes were prioritized by all four or any three methods (**Supplementary Figure 10B**) For VTE, 244 genes were prioritized, one well-known VTE-associated gene, *F2*, that encodes the coagulation factor II, was prioritized by all four methods, and 5 genes were prioritized by any three methods (*F5, PLCG2, PLEK, PROC*, and *PROS1* by DEPICT, PWMR, and PoPS) (**Supplementary Figure**

10C). In line with what has previously been discussed (Weeks et al., 2020), these results showed that the existing gene prioritization methods successfully prioritized relevant genes for diseases but have poor agreement. Note that besides adapting different statistical models and pipelines, these approaches prioritize genes based on different expression data types; DEPICT, PoPS, and TWAS are based on expression quantitative trait loci (eQTLs), while PWMR uses the protein quantitative trait loci (pQTLs) (**Methods**). Our results highlight the challenges in interpreting genome-wide significant loci and the clear need for robust in silico approaches and pipelines to nominate genes for experimental follow-up.

Biobank meta-analysis for genetic association studies

Improving power of genetic discovery for common diseases

Aggregating 18 biobanks in GBMI brings a substantial increase in samples sizes for genetic association studies for asthma (153,763 cases and 1,647,022 controls) compared to the previous largest meta-analysis by TAGC (Demenais et al., 2017) with 66 individual asthma cohorts (23,948 cases, 118,538 controls) (**Supplementary Table 2)**. The larger sample size leads to an increase in power for genetic discovery; 180 genome-wide significant loci for asthma were identified by GBMI, of which 49 are novel (**Supplementary Table 6**). Furthermore, all 18 loci that were first reported by TAGC have more significant association p-values in GBMI (**Supplementary Figure 11**). In addition, meta-analyzing GBMI biobanks and the existing disease consortia would further increase the discovery power to uncover genetic risks for human diseases. For example, we meta-analyzed 14 biobanks in GBMI with two previous meta-analysis studies for POAG (excluding three overlapped biobank data sets), which doubled the case numbers compared to the previous largest meta-analysis and successfully identified 103 significant loci, of which 19 are novel (Lo Faro et al., 2021).

Providing opportunities for genetic studies on less prevalent diseases

EHR-linked biobanks provide opportunities to assess less prevalent diseases that were understudied by previous GWAS studies. For example, although gout has caused an increased health burden in recent years, the largest meta-analysis for gout so far was conducted on 13,179 cases and 750,634 controls across 20 studies. By meta-analyzing 15 biobanks in GBMI comprising 37,105 cases and 1,448,128 controls with 5 ancestral populations (**Supplementary Table 2**), we have identified 52 significant loci, of which 30 are novel (**Supplementary Table 6**). As expected, a vast majority of these loci (n=40, either same SNP or multiple different SNPs) were associated with serum urate levels (Gill et al., 2021; Huffman et al., 2015; Sinnott-Armstrong et al., 2021; Tin et al., 2019) (**Supplementary Table 21),** including key urate transporter genes *SLC2A9* and *SLC22A12* (Xu et al., 2017). To further identify the potential biological explanation of the loci with gout risk in our study, we explored their associations with other relevant phenotypes. We find that 30 of these loci are associated with other relevant traits and diseases. For example, *RAB24, STC1*, and *MAF* are associated with BMI, *MPPED2*, *BCAS3* with kidney function, *HNF4A, PNPLA3, MC4R* with diabetes, *GCKR* with glycolysis and *ARID1A*, *MLXIPL*, *A1CF*, *INHBC* among others are associated with blood pressure and lipids (HDL-C, LDL-C, TG, and Apolipoprotein B). Previous studies have already speculated the possible mechanisms for the

involvement of these traits or processes in gout etiology. For example, coating of urate crystals with Apolipoprotein B can down-regulate the innate immune system by suppressing neutrophil activation(Terkeltaub et al., 1984) and neutrophil activation is needed for the endocytosis and lysis of urate crystals and thus the resolution of gout attack (So and Martinon, 2017). Similar biological links have also been proposed for other above-mentioned traits and uric acid metabolism and thus can explain the observed association of related genes with gout risk in our study.

Biobanks also enable genetic studies of different types of disease phenotypes. We conducted biobank meta-analysis for acute appendicitis and the relevant procedure endpoint appendectomy and observed high genetic correlation between the two endpoints (r^2 = 0.99). Out of the 10 loci identified for acute appendicitis by meta-analyzing 10 biobanks, 3 were also significant for appendectomy (**Supplementary Table 6**) even though the sample size was 3 times lower, suggesting that the procedure phenotypes may add meaningful information in biobank-based genetic studies and incorporating these phenotypes to traditional disease diagnosis phenotypes could improve discovery power.

Improving polygenic risk scores based on multi-biobank multi-ancestry meta-analyses

Based on the leave-one-biobank-out meta-analyses in GBMI, biobanks can estimate the predictive value of PRS for the endpoints. Using asthma as an example, we have demonstrated the improved PRS prediction based on the GBMI summary statistics compared to the previous meta-analysis by TAGC (Demenais et al., 2017) in 6 biobanks across 6 ancestral populations (**Supplementary Figure 12**). Wang et al has evaluated the performance of PRS using the multi-biobank multi-ancestry meta-analyses of GBMI for additional endpoints and demonstrated improved PRS prediction in individual biobanks based on GBMI results compared to previous GWASs. The Bayesian method PRS-CS(Ge et al., 2019) overall outperformed the classic pruning and thresholding method, P+T, especially for endpoints with higher SNP-based heritability, while the PRS prediction accuracy varies across biobanks and ancestries. The EUR-based LD reference panel provides comparable or better prediction accuracy relative to using other cosmopolitan LD panels for PRS construction methods based on GBMI association results as the EUR ancestry constitutes the largest proportion (about 69%) of GWAS participants (Wang et al., 2021).

Discussion

Genetic discovery benefits from the increasing numbers of EHR-linked biobanks, despite biobanks differing in many aspects. As of September 2021, 19 biobanks across four continents comprising six major ancestral groups have joined GBMI for the globally collaborative efforts to uncover genetic risk factors of human diseases. As the pilot effort in GBMI, with carefully harmonized phenotype definitions and analysis pipelines, we meta-analyzed GWASs in up to 18 biobanks for 14 endpoints, including both common diseases (asthma, COPD, VTE, etc.) and less prevalent diseases (gout, IPF, AAA, thyroid cancer, etc.). Over 500 genome-wide significant loci were detected of which 188 are novel. Sex-stratified meta-analysis allows for comparing effects between sexes and identified 8 loci with different effect sizes in men and women. Besides demonstrating the integration of genetic

association results from different biobanks, our results have illustrated the gains by meta-analyzing biobanks together. The increase in the sample sizes and sample diversity leads to higher discovery power. Incorporating non-European samples in the meta-analysis allows for the genetic association tests for 21.8 million additional markers, of which more than 85% are low-frequency variants (AF < 1%), which may facilitate functional follow-up studies to disentangle the causal variants at identified loci.

Gout has a prevalence of 1-4% worldwide and affects more than 8 million people in the United States (Zhu et al., 2011), but has been less well studied by previous GWASs. Meta-analyzing biobanks increases the case number of gout by three times compared to the previous GWAS, which uncovered 30 novel loci. In addition, biobank meta-analyses were done for different types of endpoints that can be accessed through EHRs in biobanks but not studied by previous GWASs, such as the procedure endpoint appendectomy and the relevant disease acute appendicitis. The high genetic correlation between the two endpoints suggests potential usage of the non-disease endpoints derived from EHRs for genetic discovery. As expected, based on the leave-one-biobank meta-analysis results, more accurate predictive polygenic risk scores, especially for biobank samples with European and East Asian ancestries, were constructed due to the increased genetic discovery power compared to previous studies. This gain can be further extended to non-European samples as the sample diversity continues to increase in GBMI. The collaborative efforts of biobanks in GBMI creates invaluable resources and opportunities to advance the understanding of the etiology of human diseases, leading to better treatment and prevention, and helps move toward the equitability of genetic studies in diverse ancestries.

We formed multiple working groups to 1. deepen the genetic investigation of the biological implications of results for several endpoints, including asthma (Tsuo et al., 2021), COPD (Tsuo et al., 2021), VTE (Wolford et al., 2021), POAG (Lo Faro et al., 2021), stroke (Surakka et al., 2021), and heart failure (Wu et al., 2021), 2. systematically characterize genome-wide significant loci via fine-mapping(Kanai et al., 2021), transcriptome-wide association (Bhattacharya et al., 2021a), protein QTL Mendelian randomization analysis (Zhao et al., 2021), prioritizing drug targets (Namba et al., 2021) (Namba et al., 2021), and improving the disease risk prediction by polygenic risk scores based on the multi-biobank multi-ancestry meta-analysis results (Wang et al., 2021).

Together, the pilot work conducted in GBMI has shown that despite the heterogeneities across biobanks in many aspects, such as locations, sample sizes, genotyping and phenotyping approaches, sample ancestries, and strategies to recruit participants, with standardized phenotype definitions and the analysis pipeline, biobanks can be meta-analyzed together to provide reliable genetic discoveries. Biobank meta-analysis can have substantial benefits to advance the genetic discoveries for human diseases with the larger sample sizes and the increased ancestry diversity. We have evaluated the challenges in multiple down-stream in silico studies that are used for prioritizing the functional genes and variants, provided the best practices and pipelines based on our lessons from GBMI, and highlighted the need for new method development to address the upcoming issues in the current analysis based on the biobank meta-analysis results.

Code and Data availability

The all-biobank meta-analysis results and plots for the 14 endpoints (including both ancestry-specific and cross-ancestry meta-analyses and sex stratified meta-analyses) are available for downloading at https://www.globalbiobankmeta.org/resources and browsed at the browser http://results.globalbiobankmeta.org. Custom scripts used for quality control, meta-analysis and summary of results are available at https://github.com/globalbiobankmeta. The optimized trans-ancestry and single-ancestry polygenic score weights will be deposited within the PGS Catalog (https://www.pgscatalog.org/).

Methods

Phenotype definition

A phenotype definition guideline was created and shared with all biobanks (**Supplementary Table 3**). The disease endpoints were defined following the phecode maps (Denny et al., 2013), which maps the ICD-9 or ICD-10 codes into hierarchical phecodes, each representing a specific disease group. Study participants were labeled a phecode if they had one or more of the phecode -specific ICD-9 or ICD-10 codes. Cases were all study participants with the phecode of interest and controls were all study participants without the phecode of interest or any related phecodes. For sex-specfic disease endpoint, which is uterine cancer (UtC) in the endpoint list, only females were included in the study samples. The procedure endpoint, appendectomy, was defined based on the OPCS. Any biobank participant with codes H01 (Emergency excision of appendix), H01.1(emergency excision of abnormal appendix and drainage), or H01.2 (emergency excision of a normal appendix) were cases, all other participants without these codes were controls. Biobanks which do not collect the ICD codes or OPCS codes define the phenotypes using the available EHRs according to the phenotype definitions in the guideline.

GWAS

Each biobank conducted genotyping, imputation, and quality controls independently, followed by running GWASs following the analysis plan shared in GBMI (information available at https://www.globalbiobankmeta.org/) with phenotypes curated according to the harmonized phenotype definitions (see the Phenotype definition section in Methods). We recommended to run GWAS analysis using Scalable and Accurate Implementation of GEneralized mixed model (SAIGE)(Zhou et al., 2018) or REGENIE (Mbatchou et al., 2021), which are scalable for biobank-scale data and account for sample relatedness and case-control imbalances. The suggested covariates were age, age2, sex, age*sex, 20 first principal components, and any biobank specific covariates, such as genotyping batches and recruiting centers.

Post-GWAS Quality control

Variant-level quality control was conducted for each data set containing GWAS summary statistics shared by biobanks (**Supplementary Table 22, Supplementary Figure 13 and 14**). Genetic variants with MAC < 20 and variants that are poorly imputed with an imputation score < 0.3 were firstly excluded. Genome coordinates of all genetic variants were lifted to GRCh38. For palindromic SNPs (with A/T or G/C alleles), we compared their allele frequencies of the aligned reference allele in the GWAS data set (AF-GWAS) to gnomAD (Karczewski et al., 2020) (AF-gnomAD) by ancestry. If a palindromic SNP meets any one of the following standards, we flip its alleles in the GWAS data set and indicate that this variant has the potential strand flip with a flag: 1. The fold difference is greater than two, 2. The allele frequency of the alternative allele in the GWAS data set is closer to AF-gnomAD than the reference allele, 3. AF-GWAS < 0.4 and AF-gnomAD > 0.6, 4. AF-GWAS > 0.6 and AF-gnomAD < 0.4. We then identified genetic variants with different allele frequencies compared to gnomAD. For each genetic variant, the Mahalanobis distance between AF-GWAS and AF-gnomAD was estimated and the variant was flagged to have different AF-GWAS and AF-gnomAD if the Mahalanobis distance is greater than 3 standard deviations away from the mean. We observed that across 18 biobanks that shared GWAS summary statistics to the meta-analysis for asthma, very small proportions (0.003% to 0.65%) of variants were flagged as either palindromic SNPs with flipped strands or variants having very different allele frequencies compared to gnomAD.

Meta-analysis

Fixed-effect meta-analyses based on inverse-variance weighting were performed for all endpoints with 1. all biobanks across all ancestries, 2. leave-one-biobank out across all ancestries 3. all biobanks by each ancestry, and 4. all biobanks by sex. Trans-ancestry meta-analysis was performed using MR-MEGA (Mägi et al., 2017) with 3 principal components of ancestry. We defined genome-wide significant loci by iteratively spanning the ± 500 kb region around the most significant variant and merging overlapping regions until no genome-wide significant variants were detected within ± 500 kb. The most significant variant in each locus is selected as the index variant. The nearest gene(s) to the index variant is used to name each locus. Cochran's *Q*-test for heterogeneity has been conducted to identify loci with index variants that have different effect sizes across GWAS data sets, ancestry, or in males and females.

PC projection

179,195 genetic variants have been genotyped/imputed in all biobanks, among which 168,899 are also in the 1000 Genomes (1000 Genomes Project Consortium et al., 2015) and HGDP (Cann et al., 2002). The weights corresponding to principal components for those markers were estimated based on the PCA analysis for the reference samples with known ancestry in 1000G and HGDP and shared among biobanks. Biobanks then generated PC loadings based on the pre-estimated weights of those markers.

Variant annotation

We annotated genetic variants using ANNOVAR (Wang et al., 2010) for the nearest genes. To obtain a more complete annotation for putative loss-of-function variants, VEP (McLaren et al., 2016) with the LOFTEE plug (Karczewski et al., 2020) as implemented in Hail was used

Heritability estimation and genetic correlation

We conducted LD score regression analyses using LDSC (Bulik-Sullivan et al., 2015) to estimate narrow-sense heritability based on the GWAS summary statistics for individual biobanks and to estimate the genetic correlation coefficients between each biobank and the all other biobanks together (LOBO). As most of the samples in the LOBO meta-analysis are of European ancestries, for biobanks with samples of non-EUR ancestries, such as BBJ, we estimated the trans-ancestry genetic correlation estimation using Popcorn(Brown et al., 2016) based on the LD scores pre-estimated using UK Biobank samples.

Prioritize functional genes

DEPICT

Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT) (Pers et al., 2015) was applied to investigate the results from genome-wide association studies of 14 endpoints. DEPICT uses three analyses to predict the gene functions: 1) prioritize the most likely causal genes, 2) identify enriched gene sets, and 3) discover tissues/cell types with highly expressed genes at associated loci. Two p-value thresholds were used to define genome-wide significance $1x10^{-5}$ and $5x10^{-8}$, for input summary statistics. A reference panel from individuals of European ancestry in 1000 Genomes was used to calculate LD and further identify the tag SNP from GWAS results. A minimum of 10 index variants from GWAS results was set to perform analysis using DEPICT. Enrichment results for significant findings from DEPICT were defined by FDR < 0.05. Sensitivity analysis was conducted with GWAS summary statistics derived from the meta-analysis of biobank data sets with samples of European ancestries (not including Finns) using LD information from the 1000 Genomes European panel to compare our findings with DEPICT results using multi-ancestry GWAS summary statistics.

PoPs

Polygenic Priority Score (PoPS) is a gene prioritization method used in our study to identify potential causal genes (Weeks et al., 2020) . PoPS integrates GWAS summary statistics with publicly available bulk and single-cell gene expression, biological pathway, and predicted protein-protein interaction data to comprehensively perform gene prioritization. PoPS first applied Multi-marker Analysis of GenoMic Annotation (MAGMA) (de Leeuw et al., 2015) to meta-analyze gene-level associations and create gene-gene correlation matrix. Gene-level associations were generated by meta-analyzing the variants across the same gene, using GWAS summary statistics and LD panel from 1000 Genomes European only dataset. Next, MAGMA integrated previously calculated gene-level associations and gene-gene correlation to perform enrichment analysis for gene features selection. Lastly, a PoPS

score was calculated by fitting a joint model with all the selected features simultaneously. In our study, genes with a PoPS score in the top one percentile were considered as the prioritized genes.

Transcriptome-wide association studies (TWAS)

Prediction of gene expression: Using genotypes and gene expression from 296 European donors from GTEx ver. 8 (Consortium, 2020), we trained predictive expression models using Joint-Tissue Imputation (JTI)(Zhou et al., 2020) and Multi-Omic Strategies for TWAS (MOSTWAS)(Bhattacharya et al., 2021b). Due to small eQTL sample sizes of non-European patients in GTEx, we restricted TWAS to European populations. We used gene expression from multiple relevant tissues for the analysis. For asthma, gene expression in Lung was used and for POAG, gene expression in Brain Cortex was used. For VTE, gene expressions in five most relevant tissues were used: Artery Aorta, Artery Coronary, Artery Tibial, Heart Atrial Appendage, and Heart Left Ventricle. JTI borrows information across transcriptomes of different tissues, leveraging shared genetic regulation, to improve prediction performance in a tissue-dependent manner (Zhou et al., 2020). MOSTWAS prioritizes distal-SNPs to a gene of interest that are mediated by biomarkers local to the distal-SNPs; these prioritized distal-SNPs are incorporated in the final model. We only considered genes with positive SNP heritability at p-value < 0.05 and adjusted cross-validation (CV) $R^2 > 0.01$ with p-value < 0.05; we considered the gene model from the method that showed larger CV R^2 for TWAS.

Association testing and probabilistic fine-mapping: Using meta-analyzed GBMI GWAS summary statistics from European-ancestry subjects, we detected gene-trait associations through the weighted burden test and 1000 Genomes Project CEU population as an LD reference (1000 Genomes Project Consortium et al., 2015; Gusev et al., 2016). We defined a transcriptome-wide significance using a Bonferroni correction across 20,000 tests (p-value < 2.5×10^{-6}) (Gusev et al., 2016, 2018; Mancuso et al., 2017). As complex correlations between predicted expression levels at a given region can yield multiple associated genes in TWAS, we used FOCUS, a probabilistic gene-level fine-mapping method, to define credible sets of genes that explain the expression-trait signal at a given locus(Mancuso et al., 2019). Here, we used the default non-informative priors implemented in FOCUS and estimated the posterior inclusion probability (PIP) and a 90% credible set of genes at a given locus.

Proteome-wide Mendelian randomization (PWMR)

We estimated the putative causal role of 1,310 proteins on eight diseases in the NFE samples using proteome-wide association study (PWMR) and sensitivity analyses. For the exposure of the analysis, 5,418 conditional independent pQTLs of 1,310 proteins in European samples from ARIC (Zhang et al., 2021) were selected as genetic predictors. For outcomes, eight of the 14 diseases from GBMI were selected since they had full GWAS summary statistics in both European and African ancestries and had relatively good sample size (> 100 cases). The eight disease outcomes included idiopathic pulmonary fibrosis (IPF), primary open-angle glaucoma (POAG), heart failure (HF), venous thromboembolism (VTE), stroke, gout, chronic obstructive pulmonary disease (COPD) and asthma in European and African ancestries. For the discovery PWMR analysis, we applied a generalised inverse variance weighted approach (Burgess et al., 2017) that takes into account the correlation between genetic predictors. To increase the possibility of identifying true causal links between proteins and diseases, we applied five sensitivity analyses. First, we applied generalised MR-Egger regression to

estimate the influence of horizontal pleiotropy (Burgess et al., 2017). For PWMR association with a p-value of the gEgger intercept term lower than 0.05, we considered these associations as influenced by horizontal pleiotropy and excluded them from the top finding list. Second, we applied Cochrane's Q test for gIVW results and Rocker's Q test for gEgger results to estimate the potential heterogeneity of PWMR estimates (Bowden et al., 2017; Greco M et al., 2015). Third, we applied three types of genetic colocalization analyses to distinguish causality from confounding by LD. The conventional colocalization, pairwise conditional and colocalization (PWCoCo) and LD check (Giambartolomei et al., 2014; Zheng et al., 2020). Fourth, to control for potential aptamer binding artificial effect of pQTLs, we listed all PWMR associations using pQTLs from the coding regions and flagged these associations with caution. Fifth, to estimate the influence of potential reverse causality, we applied MR-Steiger filtering (Hemani et al., 2017) and removed any PWMR associated with evidence of reverse causality from the top finding list. All the remaining PWMR associations with p-value < 0.001 were selected as candidate findings.

Phenome-wide association test

For all 508 identified index variants (in known or novel loci), we carried out look-ups for their association with 1,283 human diseases curated based on phecodes mapped to ICD codes in the UK Biobank(Zhou et al., 2018). We reported associations with p-value $< 5 \times 10^{-8}$. The index variant was lifted over to GRCh37 for comparison with the UKBB results.

Polygenic scores

The polygenic scores (PRS) were constructed using PRS-CS (Ge et al., 2019), which is based on the Bayesian framework. We used the auto model with default parameters implemented in the software to estimate the posterior mean SNP effects. The input for GWAS sample size was estimated as the total effective sample size. The LD matrix calculated using European individuals from 1000G Phase 3 (1KG) provided by PRS-CS was used here. Specifically, we used leave-one-biobank-out meta-analysis GBMI Asthma GWAS as the discovery GWAS and validated the PRS in 9 different biobanks, including: BBJ, BioVU, Lifelines, UKBB, CanPath, ESTBB, FinnGen, HUNT and MGI. To quantify the accuracy improvement attributable to GBMI, we also built PRS using public GWAS from (Demenais et al., 2017). The prediction performance of PRS was estimated using Nagelkerke's R^2 after regressing out the biobank-specific covariates with a logistic regression. It was further transformed to R^2 on the liability scale(Lee et al., 2012), with biobank-specific case proportion used as the disease population prevalence. The corresponding 95% confidence intervals (CIs) were calculated using bootstrap with 1000 replicates.

Acknowledgments

The work of the contributing biobanks was supported by numerous grants from governmental and charitable bodies, and biobank specific acknowledgements are included in the **Supplementary Notes**. We would like to thank the organizing committee of the International Common Disease

Alliance for intellectual contributions on the set up of the GBMI as a nascent activity to the larger effort. We would like to thank Daniel King from the Hail team and Sam Bryant from the Stanley Center Data Management team at the Broad Institute for helping with the Google bucket set up and data sharing, and Bethany Klunder from the University of Michigan Medical school for helping with the paper submission.

Competing Financial Interests Statement

M.J.D. is a founder of Maze Therapeutics. B.M.N. is a member of the scientific advisory board at Deep Genomics and consultant for Camp4 Therapeutics, Takeda Pharmaceutical, and Biogen. The spouse of C.J.W works at Regeneron Pharmaceuticals. C.Y.C. is employed by Biogen. C.R.G. owns stock in 23andMe, Inc. T.R.G. has received research funding from various pharmaceutical companies to support the application of Mendelian randomization to drug target prioritization. E.E.K. has received speaker fees from Regeneron, Illumina, and 23&Me, and is a member of the advisory board for Galateo Bio. R.E.M. has received speaker fees from Illumina and is a scientific advisor to the Epigenetic Clock Development Foundation. G.D.S has received research funding from various pharmaceutical companies to support the application of Mendelian randomization to drug target prioritization. K.S. and U.T. are employed by deCODE Genetics/Amgen inc. J.Z. has received research funding from various pharmaceutical companies to support the application of Mendelian randomization to drug target prioritization.

Table 1. Lead variants that are protein coding within 26 disease-associated loci identified in the multi-biobank multi-ancestry meta-analyses in GBMI

Endpoint	CHR/POS (hg38)	REF/ ALT	Freq[a]	Odds Ratio (95% CI)[b]	p-value	Heterogeneity p-value	Gene	Function	cases	controls	Number of biobanks
Novel											
AAA	10:73913343	T/C	0.737	0.88 (0.85-0.91)	5.93E-12	0.76	*PLAU*	missense	9,453	1,446,422	11
COPD	1:149934520	T/C	0.350	1.04 (1.03-1.05)	7.91E-10	0.54	*MTMR11*	missense	79,844	1,289,683	15
Stroke	6:49492265	A/G	0.446	0.96 (0.94-0.97)	1.8E-11	0.99	*CENPQ*	missense	60,176	1,310,725	16
Asthma	10:94279840	G/C	0.448	1.03 (1.02-1.03)	2.52E-09	0.98	*PLCE1*	missense	153,763	1,647,022	18
Asthma	14:100883117	G/T	0.025	1.09 (1.05-1.12)	2.61E-08	0.73	*RTL1*	missense	133,369	1,370,606	16
Asthma	19:56222056	C/A	0.253	1.03 (1.02-1.04)	2.35E-08	0.60	*ZSCAN5A*	missense	149,293	1,626,581	17
Known											
COPD	14:94378610	C/T	0.020	1.22 (1.16-1.29)	5.2E-15	9.27E-03	*SERPINA1*	missense	54,105	883,399	11
COPD	19:44908684	T/C	0.140	0.95 (0.94-0.97)	1.04E-08	0.36	*APOE*	missense	81,568	1,310,798	16
Gout	2:27508073	T/C	0.588	0.87 (0.86-0.88)	9.27E-64	0.11	*GCKR*	missense	37,105	1,448,128	15
Gout	11:64593747	G/A	0.016	0.36 (0.31-0.42)	1.19E-41	0.10	*SLC22A12*	stop gain	6,634	248,305	2
Gout	12:57449928	G/A	0.194	0.91 (0.89-0.93)	1.51E-17	0.45	*INHBC*	missense	37,105	1,448,128	15
IPF	5:1279370	T/C	0.001	862 (205-3618)	2.66E-20	0.09	*TERT*	missense	1,278	330,954	2
IPF	5:169588475	G/A	0.014	2.19 (1.81-2.66)	1.61E-15	0.02	*SPDL1*	missense	4,812	882,416	7
POAG	1:11193760	C/T	0.026	0.67 (0.6-0.74)	4.39E-14	0.90	*ANGPTL7*	missense	12,810	421,360	5
POAG	1:171636338	G/A	0.002	6.33 (4.71-8.51)	1.67E-34	4.33E-06	*MYOC*	stop gain	15,916	1,092,446	11
POAG	14:60509819	C/A	0.547	0.89 (0.87-0.91)	7.08E-30	0.31	*SIX6*	missense	26,848	1,460,599	15
Stroke	12:111803962	G/A	0.238	0.9 (0.88-0.92)	5.16E-18	0.37	*ALDH2*	missense	23,804	269,656	4
VTE	1:169549811	C/T	0.020	3.04 (2.85-3.24)	1.5E-245	5.52E-13	*F5*	missense	26,749	1,011,509	9
VTE	12:6034818	T/C	0.889	1.1 (1.07-1.13)	1.59E-10	0.21	*VWF*	missense	27,987	1,035,290	9
VTE	12:103742510	C/T	0.011	1.65 (1.45-1.88)	6.19E-14	0.25	*STAB2*	missense	10,353	341,418	2
Asthma	1:12115601	G/A	0.012	0.85 (0.8-0.89)	1.7E-11	0.46	*TNFRSF8*	missense	118,767	1,202,660	12
Asthma	1:31699894	G/T	0.573	1.03 (1.02-1.04)	1.61E-10	0.49	*COL16A1*	missense	148,045	1,579,632	17
Asthma	4:38797027	C/A	0.387	0.95 (0.94-0.96)	4.21E-21	0.54	*TLR1*	missense	138,764	1,458,022	15
Asthma	4:102267552	C/T	0.044	1.08 (1.06-1.1)	2.53E-12	0.81	*SLC39A8*	missense	129,434	1,256,670	14
Asthma	5:14610200	C/G	0.084	1.07 (1.05-1.09)	7.68E-15	0.16	*OTULINL*	missense	125,483	1,241,068	13
Asthma	9:128721272	T/A	0.068	0.95 (0.93-0.96)	5.61E-10	0.21	*ZDHHC12*	missense	152,469	1,638,824	18

[a]Frequencies are reported with respect to the alternate allele (ALT) in the combined meta-analysis data sets.
[b]Odds ratios are reported with respect to the alternate allele (ALT) in the meta-analyses.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Abul-Husn, N.S., Soper, E.R., Braganza, G.T., Rodriguez, J.E., Zeid, N., Cullina, S., Bobo, D., Moscati, A., Merkelson, A., Loos, R.J.F., et al. (2021). Implementing genomic screening in diverse populations. Genome Med. *13*, 17.

Andersson, R.E., Olaison, G., Tysk, C., and Ekbom, A. (2001). Appendectomy and protection against ulcerative colitis. N. Engl. J. Med. *344*, 808–814.

Aquilante, C.L., Kao, D.P., Trinkley, K.E., Lin, C.-T., Crooks, K.R., Hearst, E.C., Hess, S.J., Kudron, E.L., Lee, Y.M., Liko, I., et al. (2020). Clinical implementation of pharmacogenomics via a health system-wide research biobank: the University of Colorado experience. Pharmacogenomics *21*, 375–386.

Bhattacharya, A., Hirbo, J., Zhou, D., Zhou, W., Global Biobank Meta-analysis Initiative, Pasaniuc, B., Gamazon, E., and Cox, N.J. (2021a). Best practices of multi-ancestry, meta-analytic transcriptome-wide associations: lessons from the Global Biobank Meta-Initiative. In Preparation.

Bhattacharya, A., Li, Y., and Love, M.I. (2021b). MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. PLoS Genet. *17*, e1009398.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat. Med. *36*, 1783–1802.

Bowton, E., Field, J.R., Wang, S., Schildcrout, J.S., Van Driest, S.L., Delaney, J.T., Cowan, J., Weeke, P., Mosley, J.D., Wells, Q.S., et al. (2014). Biobanks and electronic medical records: enabling cost-effective research. Sci. Transl. Med. *6*, 234cm3.

Bowton, E.A., Collier, S.P., Wang, X., Sutcliffe, C.B., Van Driest, S.L., Couch, L.J., Herrera, M., Jerome, R.N., Slebos, R.J.C., Alborn, W.E., et al. (2015). Phenotype-Driven Plasma Biobanking Strategies and Methods. J Pers Med *5*, 140–152.

Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C.J., Price, A.L., and Zaitlen, N. (2016). Transethnic Genetic-Correlation Estimates from Summary Statistics. Am. J. Hum. Genet. *99*, 76–88.

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Buraczynska, K., Rejdak, K., and Buraczynska, M. (2018). Cholesteryl Ester Transfer Protein Gene Polymorphism (I405V) and Risk of Ischemic Stroke. J. Stroke Cerebrovasc. Dis. *27*, 2887–2891.

Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B.B., and Hopewell, J.C. (2017). Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. Genet.

Epidemiol. *41*, 714–725.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.

Castro, M., Corren, J., Pavord, I.D., Maspero, J., Wenzel, S., Rabe, K.F., Busse, W.W., Ford, L., Sher, L., FitzGerald, J.M., et al. (2018). Dupilumab Efficacy and Safety in Moderate-to-Severe Uncontrolled Asthma. N. Engl. J. Med. *378*, 2486–2496.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L., and China Kadoorie Biobank (CKB) collaborative group (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int. J. Epidemiol. *40*, 1652–1666.

Consortium, T.G. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et al. (2017). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat. Genet. *50*, 42–53.

Deming, W.E. (1943). Statistical adjustment of data. *261*.

Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. *31*, 1102–1110.

DeVries, A., Wlasiuk, G., Miller, S.J., Bosco, A., Stern, D.A., Lohman, I.C., Rothers, J., Jones, A.C., Nicodemus-Johnson, J., Vasquez, M.M., et al. (2017). Epigenome-wide analysis links SMAD3 methylation at birth to asthma in children of asthmatic mothers. J. Allergy Clin. Immunol. *140*, 534–542.

Dummer, T.J.B., Awadalla, P., Boileau, C., Craig, C., Fortier, I., Goel, V., Hicks, J.M.T., Jacquemont, S., Knoppers, B.M., Le, N., et al. (2018). The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. CMAJ *190*, E710–E717.

Feng, Y.-C.A., Chen, C.-Y., Chen, T.-T., Kuo, P.-H., Hsu, Y.-H., Yang, H.-I., Chen, W.J., Shen, C.-Y., Ge, T., Huang, H., et al. (2021). Taiwan Biobank: a rich biomedical research database of the Taiwanese population. In Preparation.

Finer, S., Martin, H.C., Khan, A., Hunt, K.A., MacLaughlin, B., Ahmed, Z., Ashcroft, R., Durham, C., MacArthur, D.G., McCarthy, M.I., et al. (2020). Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. Int. J. Epidemiol. *49*, 20–21i.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.

Gill, D., Cameron, A.C., Burgess, S., Li, X., Doherty, D.J., Karhunen, V., Abdul-Rahim, A.H., Taylor-Rowan, M., Zuber, V., Tsao, P.S., et al. (2021). Urate, Blood Pressure, and Cardiovascular Disease: Evidence From Mendelian Randomization and Meta-Analysis of Clinical Trials. Hypertension *77*, 383–392.

Greco M, F.D., Minelli, C., Sheehan, N.A., and Thompson, J.R. (2015). Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. Stat. Med. *34*, 2926–2940.

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. *47*, 435–444.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B.M., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. *50*, 538–548.

Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. *13*, e1007081.

Huffman, J.E., Albrecht, E., Teumer, A., Mangino, M., Kapur, K., Johnson, T., Kutalik, Z., Pirastu, N., Pistis, G., Lopez, L.M., et al. (2015). Modulation of genetic associations with serum urate levels by body-mass-index in humans. PLoS One *10*, e0119752.

Johnson, R., Ding, Y., Venkateswaran, V., Bhattacharya, A., Chiu, A., Schwarz, T., Freund, M., Zhan, L., Burch, K.S., Caggiano, C., et al. (2021). Leveraging genomic diversity for discovery in an EHR-linked biobank: the UCLA ATLAS Community Health Initiative.

Kanai, M., Elzur, R., Global Biobank Meta-analysis Initiative, Daly, M.J., and Finucane, H.K. (2021). Inter-cohort heterogeneity significantly undermines fine-mapping a meta-analysis. In Preparation.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Karlson, E.W., Boutin, N.T., Hoffnagle, A.G., and Allen, N.L. (2016). Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. J Pers Med *6*.

Klarin, D., Verma, S.S., Judy, R., Dikilitas, O., Wolford, B.N., Paranjpe, I., Levin, M.G., Pan, C., Tcheandjieu, C., Spin, J.M., et al. (2020). Genetic Architecture of Abdominal Aortic Aneurysm in the Million Veteran Program. Circulation *142*, 1633–1646.

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T.L., Midthjell, K., Stene, T.R., Bratberg, G., Heggland, J.,

and Holmen, J. (2013). Cohort Profile: the HUNT Study, Norway. Int. J. Epidemiol. *42*, 968–977.

Lee, S.H., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2012). A better coefficient of determination for genetic profile analysis. Genet. Epidemiol. *36*, 214–224.

de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. *11*, e1004219.

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147.

Liu, S., Xie, Z., Daugherty, A., Cassis, L.A., Pearson, K.J., Gong, M.C., and Guo, Z. (2013). Mineralocorticoid receptor agonists induce mouse aortic aneurysm formation and rupture in the presence of high salt. Arterioscler. Thromb. Vasc. Biol. *33*, 1568–1579.

Lo Faro, V., Bhattacharya, A., Zhou, W., Zhou, D., Wang, Y., Läll, K., Kanai, M., Lopera-Maya, E., Straub, P., Pawar, P., et al. (2021). 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.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M.I., COGENT-Kidney Consortium, T2D-GENES Consortium, and Morris, A.P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum. Mol. Genet. *26*, 3639–3650.

Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am. J. Hum. Genet. *100*, 473–487.

Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. Nat. Genet. *51*, 675–682.

Matoba, N., Akiyama, M., Ishigaki, K., Kanai, M., Takahashi, A., Momozawa, Y., Ikegawa, S., Ikeda, M., Iwata, N., Hirata, M., et al. (2020). GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. Nat Hum Behav *4*, 308–316.

Matsuo, H., Yamamoto, K., Nakaoka, H., Nakayama, A., Sakiyama, M., Chiba, T., Takahashi, A., Nakamura, T., Nakashima, H., Takada, Y., et al. (2016). Genome-wide association study of clinically defined gout identifies multiple risk loci and its association with clinical subtypes. Ann. Rheum. Dis. *75*, 652–659.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Mizuno, Y., Harada, E., Morita, S., Kinoshita, K., Hayashida, M., Shono, M., Morikawa, Y., Murohara, T., Nakayama, M., Yoshimura, M., et al. (2015). East asian variant of aldehyde dehydrogenase 2 is associated

with coronary spastic angina: possible roles of reactive aldehydes and implications of alcohol flushing syndrome. Circulation *131*, 1665–1673.

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al. (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27*, S2–S8.

Namba, S., Konuma, T., Wu, K.-H., Zhou, W., and Okada, Yukinori, on behalf of Global Biobank Meta-analysis Initiative (2021). A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis. In Preparation.

Olsen, C.M., Green, A.C., Neale, R.E., Webb, P.M., Cicero, R.A., Jackman, L.M., O'Brien, S.M., Perry, S.L., Ranieri, B.A., Whiteman, D.C., et al. (2012). Cohort profile: the QSkin Sun and Health Study. Int. J. Epidemiol. *41*, 929–929i.

Palmisano, B.T., Le, T.D., Zhu, L., Lee, Y.K., and Stafford, J.M. (2016). Cholesteryl ester transfer protein alters liver and plasma triglyceride metabolism through two liver networks in female mice. J. Lipid Res. *57*, 1541–1551.

Palmisano, B.T., Anozie, U., Yu, S., Neuman, J.C., Zhu, L., Edington, E.M., Luu, T., and Stafford, J.M. (2021). Cholesteryl Ester Transfer Protein Impairs Triglyceride Clearance via Androgen Receptor in Male Mice. Lipids *56*, 17–29.

Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.-J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. *6*, 5890.

Rabe, K.F., Brusselle, G., Castro, M., Sher, L., Zhu, H., Dong, Q., Hamilton, J.D., Brian, W., Jagerschmidt, A., Pirozzi, G., et al. (2018). Dupilumab shows rapid and sustained suppression of inflammatory biomarkers in corticosteroid (CS)-dependent severe asthma patients in LIBERTY ASTHMA VENTURE. In Allergy and Immunology, (European Respiratory Society),.

Scholtens, S., Smidt, N., Swertz, M.A., Bakker, S.J.L., Dotinga, A., Vonk, J.M., van Dijk, F., van Zon, S.K.R., Wijmenga, C., Wolffenbuttel, B.H.R., et al. (2015). Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int. J. Epidemiol. *44*, 1172–1180.

Seibold, M.A., Wise, A.L., Speer, M.C., Steele, M.P., Brown, K.K., Loyd, J.E., Fingerlin, T.E., Zhang, W., Gudmundsson, G., Groshong, S.D., et al. (2011). A common MUC5B promoter polymorphism and pulmonary fibrosis. N. Engl. J. Med. *364*, 1503–1512.

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet. *53*, 185–194.

Smith, B.H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S.M., Deary, I.J., Macintyre, D.J., Campbell, H., McGilchrist, M., et al. (2013). Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. Int. J. Epidemiol. *42*, 689–700.

So, A.K., and Martinon, F. (2017). Inflammation in gout: mechanisms and therapeutic targets. Nat. Rev.

Rheumatol. *13*, 639–647.

Springelkamp, H., Iglesias, A.I., Cuellar-Partida, G., Amin, N., Burdon, K.P., van Leeuwen, E.M., Gharahkhani, P., Mishra, A., van der Lee, S.J., Hewitt, A.W., et al. (2015). ARHGEF12 influences the risk of glaucoma by increasing intraocular pressure. Hum. Mol. Genet. *24*, 2689–2699.

Sulem, P., Gudbjartsson, D.F., Walters, G.B., Helgadottir, H.T., Helgason, A., Gudjonsson, S.A., Zanon, C., Besenbacher, S., Bjornsdottir, G., Magnusson, O.T., et al. (2011). Identification of low-frequency variants associated with gout and serum uric acid levels. Nat. Genet. *43*, 1127–1130.

Surakka, I., Wu, K.-H., Wolford, B.N., Shen, F., Zhou, W., Pandit, A., Hornsby, W., Brumpton, B., Skogholt, A.H., Gabrielssen, M., et al. (2021). Multi-ancestry meta-analysis identifies 5 novel loci associated with ischemic stroke and reveals association heterogeneity between sexes and ancestries. In Preparation.

Terkeltaub, R., Curtiss, L.K., Tenner, A.J., and Ginsberg, M.H. (1984). Lipoproteins containing apoprotein B are a major regulator of neutrophil responses to monosodium urate crystals. J. Clin. Invest. *73*, 1719–1730.

Tin, A., Marten, J., Halperin Kuhns, V.L., Li, Y., Wuttke, M., Kirsten, H., Sieber, K.B., Qiu, C., Gorski, M., Yu, Z., et al. (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. Nat. Genet. *51*, 1459–1474.

Tsuo, K., Zhou, W., Wang, Y., Kanai, M., Namba, S., Gupta, R., Majara, L., Nkambule, L.L., Okada, Y., Morisaki, T., et al. (2021). Multi-ancestry meta-analysis of asthma identifies novel associations and highlights shared genetic architecture across biobanks and traits. In Preparation.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wang, Y., Namba, S., Lopera, E., Kerminen, S., Tsuo, K., Läll, K., Kanai, M., Zhou, W., Wu, K.-H., Favé, M.-J., et al. (2021). Global biobank analyses provide lessons for computing polygenic risk scores across diverse cohorts. In Preparation.

Weeks, E.M., Ulirsch, J.C., Cheng, N.Y., Trippe, B.L., Fine, R.S., Miao, J., Patwardhan, T.A., Kanai, M., Nasser, J., Fulco, C.P., et al. (2020). Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases (medRxiv).

Wolford, B.N., Willer, C.J., and Surakka, I. (2018). Electronic health records: the next wave of complex disease genetics. Hum. Mol. Genet. *27*, R14–R21.

Wolford, B.N., Lab, S., Wu, K.-H.H., Surakka, I., Zhao, Y., Yu, X., Richter, C.E., Bhatta, L., Brumpton, B., Desch, K., et al. (2021). Multi-ancestry GWAS for venous thromboembolism identifies novel loci followed by experimental validation. In Preparation.

Wu, K.-H.H., Douville, N.J., Konerman, M.C., Mathis, M.R., Hummel, S.L., Wolford, B.N., Surakka, I., Graham, S.E., Joo, H., Hirbo, J., et al. (2021). Polygenic risk score from a multi-ancestry GWAS uncovers cases of heart failure. In Preparation.

Wyss, A.B., Sofer, T., Lee, M.K., Terzikhan, N., Nguyen, J.N., Lahousse, L., Latourelle, J.C., Smith, A.V., Bartz, T.M., Feitosa, M.F., et al. (2018). Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. Nat. Commun. *9*, 2976.

Xu, L., Shi, Y., Zhuang, S., and Liu, N. (2017). Recent advances on uric acid transporters. Oncotarget *8*, 100852–100862.

Zawistowski, M., Fritsche, L.G., Pandit, A., Vanderwerff, B., Patil, S., Schmidt, E.M., VandeHaar, P., Brummett, C.M., Keterpal, S., Zhou, X., et al. (2021). The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine patients. In Preparation.

Zhang, J., Dutta, D., Köttgen, A., Tin, A., Schlosser, P., Grams, M.E., Harvey, B., CKDGen Consortium, Yu, B., Boerwinkle, E., et al. (2021). Large Bi-Ethnic Study of Plasma Proteome Leads to Comprehensive Mapping of cis-pQTL and Models for Proteome-wide Association Studies.

Zhao, H., Rasheed, H., Nøst, T.H., Cho, Y., Liu, Y., Bhatta, L., Bhattacharya, A., Global Biobank Meta-analysis Initiative, Hemani, G., Smith, G.D., et al. (2021). Proteome-wide Mendelian randomization in global biobank meta-analysis reveals trans-ancestry drug targets for common diseases. In Preparation.

Zheng, J., Haberland, V., Baird, D., Walker, V., Haycock, P.C., Hurle, M.R., Gutteridge, A., Erola, P., Liu, Y., Luo, S., et al. (2020). Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nat. Genet. *52*, 1122–1131.

Zhou, D., Jiang, Y., Zhong, X., Cox, N.J., Liu, C., and Gamazon, E.R. (2020). A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat. Genet. *52*, 1239–1246.

Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. *53*, 1097–1103.

Zhu, Y., Pandya, B.J., and Choi, H.K. (2011). Prevalence of gout and hyperuricemia in the US general population: the National Health and Nutrition Examination Survey 2007-2008. Arthritis Rheum. *63*, 3136–3141.