

University of Groningen

Global Biobank Meta-analysis Initiative

deCODE Genetics; Estonian Biobank; FinnGen; Generation Scotland; Genes & Health Research Team; LifeLines; Mass General Brigham Biobank; Michigan Genomics Initiative; National Biobank of Korea; Penn Medicine BioBank

Published in:
 Cell Genomics

DOI:
[10.1016/j.xgen.2022.100192](https://doi.org/10.1016/j.xgen.2022.100192)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
 Publisher's PDF, also known as Version of record

Publication date:
 2022

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

deCODE Genetics, Estonian Biobank, FinnGen, Generation Scotland, Genes & Health Research Team, LifeLines, Mass General Brigham Biobank, Michigan Genomics Initiative, National Biobank of Korea, Penn Medicine BioBank, Qatar Biobank, The QSkin Sun and Health Study, Taiwan Biobank, The HUNT Study, UCLA ATLAS Community Health Initiative, Uganda Genome Resource, UK Biobank, Biobank of the Americas, BioBank Japan Project, ... Sanna, S. (2022). Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics*, 2(10), Article 100192. <https://doi.org/10.1016/j.xgen.2022.100192>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

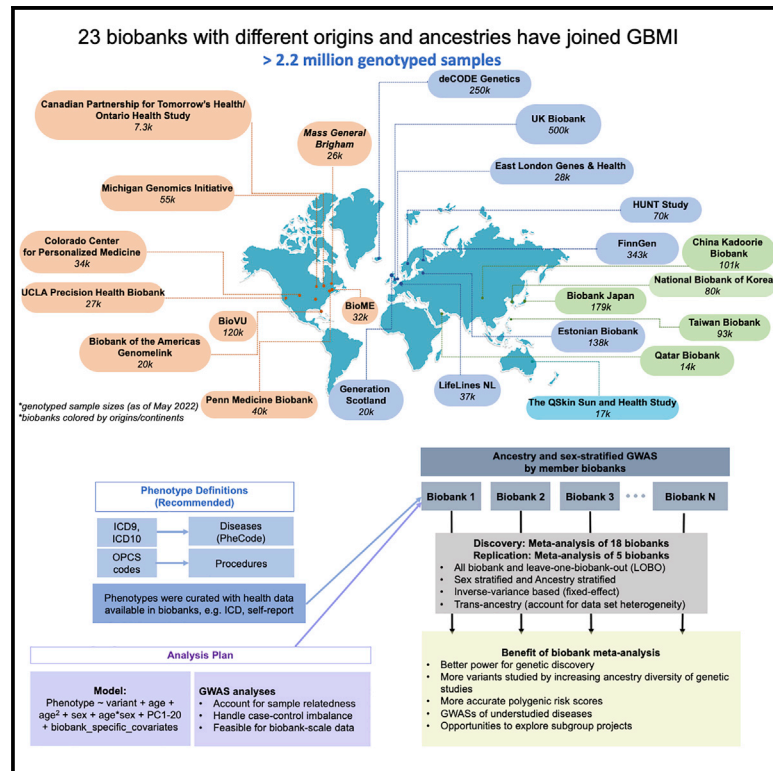
Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease

Graphical abstract



Authors

Wei Zhou, Masahiro Kanai, Kuan-Han H. Wu, ..., Cristen J. Willer, Mark J. Daly, Benjamin M. Neale

Correspondence

wzhou@broadinstitute.org (W.Z.), cristen@umich.edu (C.J.W.), mjdaly@broadinstitute.org (M.J.D.), bneale@broadinstitute.org (B.M.N.)

In brief

Zhou et al. present the flagship project of the Global Biobank Meta-analysis Initiative (GBMI). They demonstrate the substantial benefits of the collaborative efforts of 23 biobanks worldwide to advance genetic discoveries for human diseases with larger sample sizes and increased ancestry diversity and highlight issues and challenges in biobank meta-analyses.

Highlights

- GBMI is a collaborative network of 24 biobanks with >2.2 M individuals
- GWASs in different biobanks worldwide can be successfully integrated
- Biobank meta-analyses identified 317 known and 183 novel loci for 14 endpoints
- GBMI publicly releases summary statistics of biobank meta-analyses



Article

Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease

Wei Zhou,^{1,2,3,92,*} Masahiro Kanai,^{1,2,3,4,5} Kuan-Han H. Wu,⁶ Humaira Rasheed,^{7,8,9} Kristin Tsuo,^{1,2,3} Jibril B. Hirbo,^{10,11} Ying Wang,^{1,2,3} Arjun Bhattacharya,¹² Huiling Zhao,⁹ Shinichi Namba,⁵ Ida Surakka,¹³ Brooke N. Wolford,^{6,7} Valeria Lo Faro,^{14,15,16} Esteban A. Lopera-Maya,¹⁷ Kristi Läll,¹⁸ Marie-Julie Favé,¹⁹ Juulia J. Partanen,²⁰ Sinéad B. Chapman,²³ Juha Karjalainen,^{1,2,3,20} Mitja Kurki,^{1,2,3,20} Mutaamba Maasha,^{1,2,3,20} Ben M. Brumpton,^{7,21,22} Sameer Chavan,²³ Tzu-Ting Chen,²⁴ Michelle Daya,²³ Yi Ding,^{12,25} Yen-Chen A. Feng,²⁶ Lindsay A. Guare,²⁷ Christopher R. Gignoux,²³ Sarah E. Graham,¹³ Whitney E. Hornsby,¹³ Nathan Ingold,^{28,29} Said I. Ismail,³⁰ Ruth Johnson,^{31,12} Triin Laisk,¹⁸ Kuang Lin,³² Jun Lv,³³ Iona Y. Millwood,^{32,34} Sonia Moreno-Grau,³⁵ Kisung Nam,³⁶ Priit Palta,^{18,20} Anita Pandit,³⁷ Michael H. Preuss,³⁸ Chadi Saad,³⁰ Shefali Setia-Verma,³⁹ Unnur Thorsteinsdottir,⁴⁰ Jasmina Uzunovic,¹⁹ Anurag Verma,^{41,42} Matthew Zawistowski,³⁷ Xue Zhong,^{10,11} Nahla Affi,⁴³ Kawthar M. Al-Dabhani,⁴³ Asma Al Thani,⁴³ Yuki Bradford,²⁷ Archie Campbell,⁴⁴ Kristy Crooks,²³ Geertruida H. de Bock,⁴⁵ Scott M. Damrauer,^{27,42,46} Nicholas J. Douville,^{47,48} Sarah Finer,⁴⁹ Lars G. Fritsche,³⁷ Eleni Fthenou,⁴³ Gilberto Gonzalez-Arroyo,^{35,50} Christopher J. Griffiths,⁴⁹ Yu Guo,⁵¹ Karen A. Hunt,⁵² Alexander Ioannidis,^{35,53} Nomdo M. Jansonius,¹⁴ Takahiro Konuma,^{5,54} Ming Ta Michael Lee,³⁵ Arturo Lopez-Pineda,^{35,50} Yuta Matsuda,⁵⁵ Riccardo E. Marioni,⁴⁴ Babak Moatamed,³⁵ Marco A. Nava-Aguilar,^{35,50} Kensuke Numakura,⁵⁵ Snehal Patil,³⁷ Nicholas Rafaels,²³

(Author list continued on next page)

¹Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

²Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

³Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

⁴Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

⁵Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan

⁶Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

⁷K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

⁸Division of Medicine and Laboratory Sciences, University of Oslo, Oslo, Norway

⁹MRC Integrative Epidemiology Unit (IEU), Bristol Medical School, University of Bristol, Bristol, UK

¹⁰Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

¹¹Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

(Affiliations continued on next page)

SUMMARY

Biobanks facilitate genome-wide association studies (GWASs), which have mapped genomic loci across a range of human diseases and traits. However, most biobanks are primarily composed of individuals of European ancestry. We introduce the Global Biobank Meta-analysis Initiative (GBMI)—a collaborative network of 23 biobanks from 4 continents representing more than 2.2 million consented individuals with genetic data linked to electronic health records. GBMI meta-analyzes summary statistics from GWASs generated using harmonized genotypes and phenotypes from member biobanks for 14 exemplar diseases and endpoints. This strategy validates that GWASs conducted in diverse biobanks can be integrated despite heterogeneity in case definitions, recruitment strategies, and baseline characteristics. This collaborative effort improves GWAS power for diseases, benefits understudied diseases, and improves 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 human diseases and traits.

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. Genome-wide association studies (GWASs) have identified thousands of genetic loci for hundreds of human diseases and traits (see GWAS Catalog¹). Meta-analysis across



Anne Richmond,⁵⁶ Agustin Rojas-Muñoz,³⁵ Jonathan A. Shortt,²³ Peter Straub,^{10,11} Ran Tao,^{11,57} Brett Vanderwerff,³⁷ Manvi Vernekar,⁵⁵ Yogasudha Veturi,²⁷ Kathleen C. Barnes,²³ Marike Boezen,^{45,90} Zhengming Chen,^{32,34} Chia-Yen Chen,⁵⁸ Judy Cho,³⁸ George Davey Smith,^{9,59} Hilary K. Finucane,^{1,2,3} Lude Franke,¹⁷ Eric R. Gamazon,^{10,11,60} Andrea Ganna,^{1,2,20} Tom R. Gaunt,^{9,59} Tian Ge,^{61,62} Hailiang Huang,^{1,2} Jennifer Huffman,⁶³ Nicholas Katsanis,³⁵ Jukka T. Koskela,²⁰ Clara Lajonchere,^{64,65} Matthew H. Law,^{28,29} Liming Li,³³ Cecilia M. Lindgren,⁶⁶ Ruth J.F. Loos,^{38,67} Stuart MacGregor,²⁸ Koichi Matsuda,⁶⁸ Catherine M. Olsen,²⁸ David J. Porteous,⁴⁴ Jordan A. Shavit,^{69,89} Harold Snieder,⁴⁵ Tomohiro Takano,⁵⁵ Richard C. Trembath,⁷⁰ Judith M. Vonk,⁴⁵ David C. Whiteman,²⁸ Stephen J. Wicks,²³ Cisca Wijmenga,¹⁷ John Wright,⁷¹ Jie Zheng,⁹ Xiang Zhou,³⁷ Philip Awadalla,^{19,72} Michael Boehnke,³⁷ Carlos D. Bustamante,^{35,53,73} Nancy J. Cox,^{10,11} Segun Fatumo,^{74,75,76} Daniel H. Geschwind,^{64,77,78} Caroline Hayward,⁵⁶ Kristian Hveem,^{7,21} Eimear E. Kenny,⁷⁹ Seunggeun Lee,³⁶ Yen-Feng Lin,^{24,80,81} Hamdi Mbarek,³⁰ Reedik Mägi,¹⁸ Hilary C. Martin,⁸² Sarah E. Medland,²⁸ Yukinori Okada,^{5,83,84,85,86} Aarno V. Palotie,^{1,2,20} Bogdan Pasaniuc,^{12,25,64,77,87} Daniel J. Rader,^{27,41} Marylyn D. Ritchie,²⁷ Serena Sanna,^{17,88} Jordan W. Smoller,^{61,62} Kari Stefansson,⁴⁰ David A. van Heel,⁵² Robin G. Walters,^{32,34} Sebastian Zöllner,³⁷ Biobank of the Americas, Biobank Japan Project, BioMe, BioVU, CanPath - Ontario Health Study, China Kadoorie Biobank Collaborative Group, Colorado Center for Personalized Medicine, deCODE Genetics, Estonian Biobank, FinnGen, Generation Scotland, Genes & Health Research Team,

(Author list continued on next page)

¹²Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

¹³Department of Internal Medicine, Division of Cardiology, University of Michigan, Ann Arbor, MI, USA

¹⁴University of Groningen, UMCG, Department of Ophthalmology, Groningen, the Netherlands

¹⁵Department of Clinical Genetics, Amsterdam University Medical Center (AMC), Amsterdam, the Netherlands

¹⁶Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

¹⁷University of Groningen, UMCG, Department of Genetics, Groningen, the Netherlands

¹⁸Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia

¹⁹Ontario Institute for Cancer Research, Toronto, ON, Canada

²⁰Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

²¹HUNT Research Centre, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Levanger, Norway

²²Clinic of Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

²³University of Colorado - Anschutz Medical Campus, Aurora, CO, USA

²⁴Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan

²⁵Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA

²⁶Division of Biostatistics, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

²⁷Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

²⁸QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

²⁹Faculty of Health, School of Biomedical Sciences, Queensland University of Technology, Brisbane, QLD, Australia

³⁰Qatar Genome Program, Qatar Foundation Research, Development and Innovation, Qatar Foundation, Doha, Qatar

³¹Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA

³²Nuffield Department of Population Health, University of Oxford, Oxford, UK

³³Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China

³⁴MRC Population Health Research Unit, University of Oxford, Oxford, UK

³⁵Galatea Bio, Inc., Hialeah, FL, USA

³⁶Graduate School of Data Science, Seoul National University, Seoul, South Korea

³⁷Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA

³⁸The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

³⁹Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

⁴⁰deCODE Genetics/Amgen, Inc., 101 Reykjavik, Iceland

⁴¹Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

⁴²Corporal Michael Crescenz VA Medical Center, Philadelphia, PA, USA

⁴³Qatar Biobank for Medical Research, Qatar Foundation for Education, Science, and Community, Doha, Qatar

⁴⁴Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

⁴⁵Department of Epidemiology, University Medical Center Groningen, Groningen, the Netherlands

⁴⁶Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

⁴⁷Department of Anesthesiology, Michigan Medicine, Ann Arbor, MI, USA

⁴⁸Institute of Healthcare Policy & Innovation, University of Michigan, Ann Arbor, MI, USA

⁴⁹Wolfson Institute of Population Health, Queen Mary University of London, London, UK

⁵⁰Amphora Health, Morelia, Michoacan, Mexico

⁵¹Chinese Academy of Medical Sciences, Beijing, China

⁵²Blizard Institute, Queen Mary University of London, London, UK

(Affiliations continued on next page)

LifeLines, Mass General Brigham Biobank, Michigan Genomics Initiative, National Biobank of Korea, Penn Medicine BioBank, Qatar Biobank, The QSkin Sun and Health Study, Taiwan Biobank, The HUNT Study, UCLA ATLAS Community Health Initiative, Uganda Genome Resource, UK Biobank, Alicia R. Martin,^{1,2,3} Cristen J. Willer,^{6,13,89,91,*} Mark J. Daly,^{1,2,3,20,91,*} and Benjamin M. Neale^{1,2,3,91,*}

⁵³Stanford University School of Medicine, Stanford, CA, USA

⁵⁴Central Pharmaceutical Research Institute, Japan Tobacco, Inc., Takatsuki 569-1125, Japan

⁵⁵Genomelink, Inc., Berkeley, CA, USA

⁵⁶Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

⁵⁷Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

⁵⁸Biogen, Cambridge, MA, USA

⁵⁹NIHR Bristol Biomedical Research Centre, Bristol, UK

⁶⁰MRC Epidemiology Unit, University of Cambridge, Cambridge, UK

⁶¹Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

⁶²Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA

⁶³Centre for Population Genomics, VA Boston Healthcare System, Boston, MA, USA

⁶⁴Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA, USA

⁶⁵Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

⁶⁶Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK

⁶⁷Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Medicine and Health Sciences, University of Copenhagen, Copenhagen, Denmark

⁶⁸Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

⁶⁹University of Michigan, Department of Pediatrics, Ann Arbor, MI 48109, USA

⁷⁰School of Basic and Medical Biosciences, Faculty of Life Sciences and Medicine, King's College London, London, UK

⁷¹Bradford Institute for Health Research, Bradford Teaching Hospitals National Health Service (NHS) Foundation Trust, Bradford, UK

⁷²Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

⁷³Chan Zuckerberg Biohub, San Francisco, CA, USA

⁷⁴The African Computational Genomics (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda

⁷⁵London School of Hygiene & Tropical Medicine, London, UK

⁷⁶Medical Research Council/Uganda Virus Research Institute/London School of Hygiene and Tropical Medicine (MRC/UVRI/LSHTM) Uganda Research Unit, Entebbe, Uganda

⁷⁷Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

⁷⁸Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

⁷⁹Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁸⁰Department of Public Health & Medical Humanities, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

⁸¹Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan

⁸²Human Genetics Programme, Wellcome Sanger Institute, Hinxton, UK

⁸³Department of Genome Informatics, Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan

⁸⁴Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan

⁸⁵Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

⁸⁶Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan

⁸⁷Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

⁸⁸Institute for Genetics and Biomedical Research (IRGB), National Research Council (CNR), Cagliari, Italy

⁸⁹Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

⁹⁰Deceased

⁹¹These authors contributed equally

⁹²Lead contact

*Correspondence: wzhou@broadinstitute.org (W.Z.), cristen@umich.edu (C.J.W.), mjdaly@broadinstitute.org (M.J.D.), bneale@broadinstitute.org (B.M.N.)

<https://doi.org/10.1016/j.xgen.2022.100192>

cohorts has been instrumental in making these discoveries. However, most genomics research has been performed primarily in cohorts of European ancestry in 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 step toward 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 resources for the genetic research community. The

drop in the cost of genotyping and sequencing has led to an increase in the number of genomically profiled biobanks worldwide. Compared with 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.^{2,3}

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)

reference), as well as data access and references (webpage if available).

Disease prevalence varies across biobanks (Figure S1) and across sample recruiting strategy groups (Figure S2A). Biobanks recruiting participants from health centers or hospitals, relative to those recruiting participants from the general population, had a significantly higher prevalence (Wilcoxon test $p < 0.05$) for 6 out of 13 examined diseases (appendectomy was excluded from the test because of insufficient data shared from the hospital-based biobanks) (Figure S2B), including asthma, HF, stroke, VTE, gout, and IPF.

GBMI incorporates diverse genetic ancestries in genetic studies by including biobank samples of 6 main ancestry groups: approximately 42,000 of African ancestries from admixed-ancestry diaspora (AFR), 18,000 admixed American (AMR), 31,000 Central and South Asian (CSA), 415,000 East Asian (EAS), 1.4 million European (EUR), and 12,000 Middle Eastern (MID) individuals (Table S3). To compare the genetic ancestries represented between different biobanks, we projected biobanks' participants to the same principal component (PC) space (Figure 3) using pre-computed loadings of genetic markers overlapping in all biobanks and the reference containing 1000 Genomes⁵ and Human Genome Diversity Project (HGDP).⁶ PCs projected in the same space enable a cross-comparison of the sample genetic ancestry among all biobanks (STAR Methods). Notably, the population labels used in GBMI were defined by global genetic reference datasets, although GBMI is not globally representative; for example, the majority of individuals assigned to AMR and AFR ancestry groups are mostly from biobanks in the US, and GBMI participants' ancestries are not currently representative of broader Central/South American or continental African ancestries, respectively.

Biobank meta-analyses

Biobank meta-analyses (Figure 4) were performed. We harmonized phenotype definitions primarily by mapping the International Classification of Diseases (ICD) codes to phecodes⁷ for diseases and using Classification of Interventions and Procedures (OPCS) codes for procedures. We shared the definitions with member biobanks to curate phenotypes (Table S4). Biobanks that have not collected ICD or OPCS codes for their participants used the shared phenotype definitions as a guideline to create phenotypes with any available health data, such as self-report data (Table S5). After standard quality control and the estimation of ancestry groups (Table S1), GWASs

Figure 2. Seven collaboration principles in GBMI

stratified by ancestry and sex were conducted in each biobank (Table S2) with the first 20 genetic PCs adjusted as covariates, which are continuous measures of sample ancestries (STAR Methods). The central analysis team performed post-GWAS variant-level quality control for

each biobank by flagging markers with different allele frequencies compared with gnomAD⁸ and excluding markers with an imputation quality score < 0.3 (STAR Methods). Across all biobanks, 70.7 million genetic variants were tested for associations, of which 39.4 million variants were tested in at least two biobanks (Table S6). The discovery meta-analyses contain up to 18 biobanks, and for each endpoint, all-biobank meta-analysis as well as ancestry- and sex-stratified meta-analyses were conducted. In addition, we performed the leave-one-biobank-out (LOBO) meta-analyses for each biobank, estimated genetic correlation, and compared effect size estimates between GWASs in individual biobanks and the corresponding LOBO (see [integration of association results across biobanks](#)). LOBO results have been used by analyses that are sensitive to sample overlap, such as developing and testing polygenic risk scores (PRSs) for disease prediction.⁹ Post-meta-analysis filters were applied to genome-wide significant loci (STAR Methods). Five biobanks (BBofA, PMBB, CanPath, NBK, and QBB) were meta-analyzed to replicate loci identified by the discovery meta-analysis.

Inverse variance-based meta-analyses of all biobanks for 14 endpoints successfully replicated 317 previously reported loci¹ and identified 183 apparently novel loci, spanning the variant frequency spectrum (STAR Methods; Table S7; Figure 5). 431 loci were tested for 12 endpoints (except for VTE and appendectomy) in the replication meta-analysis containing up to 73,596 samples (9,991 cases and 63,605 controls) (Tables S2 and S7). Despite that for 360 out of 431 loci, the case numbers in the replication data were less than 10% of the case numbers in the discovery data, 127 loci (30%) had a p value < 0.05 in the replication meta-analyses. Out of the 127 loci, 124 loci had consistent effect direction in discovery and replication meta-analyses (Table S7).

At 87 loci, a protein-coding variant was either the most significant one ($n = 26$) (Table 1) or in linkage disequilibrium with the most significant variant with $r^2 > 0.8$ ($n = 61$ additional). 18 of these 87 loci were novel (Table S8). 13 endpoints had 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% (AcApp) to 10.73% (gout) (Table S9). The heritability of cardiomyopathy was estimated to be 0. This could be because the heritability estimation was underpowered based on the low prevalence (0.25%) with a low number of cases (2,993 cases) and because the disease has heterogeneous subtypes, including dilated and hypertrophic cardiomyopathy, with different genetic causes.¹⁰

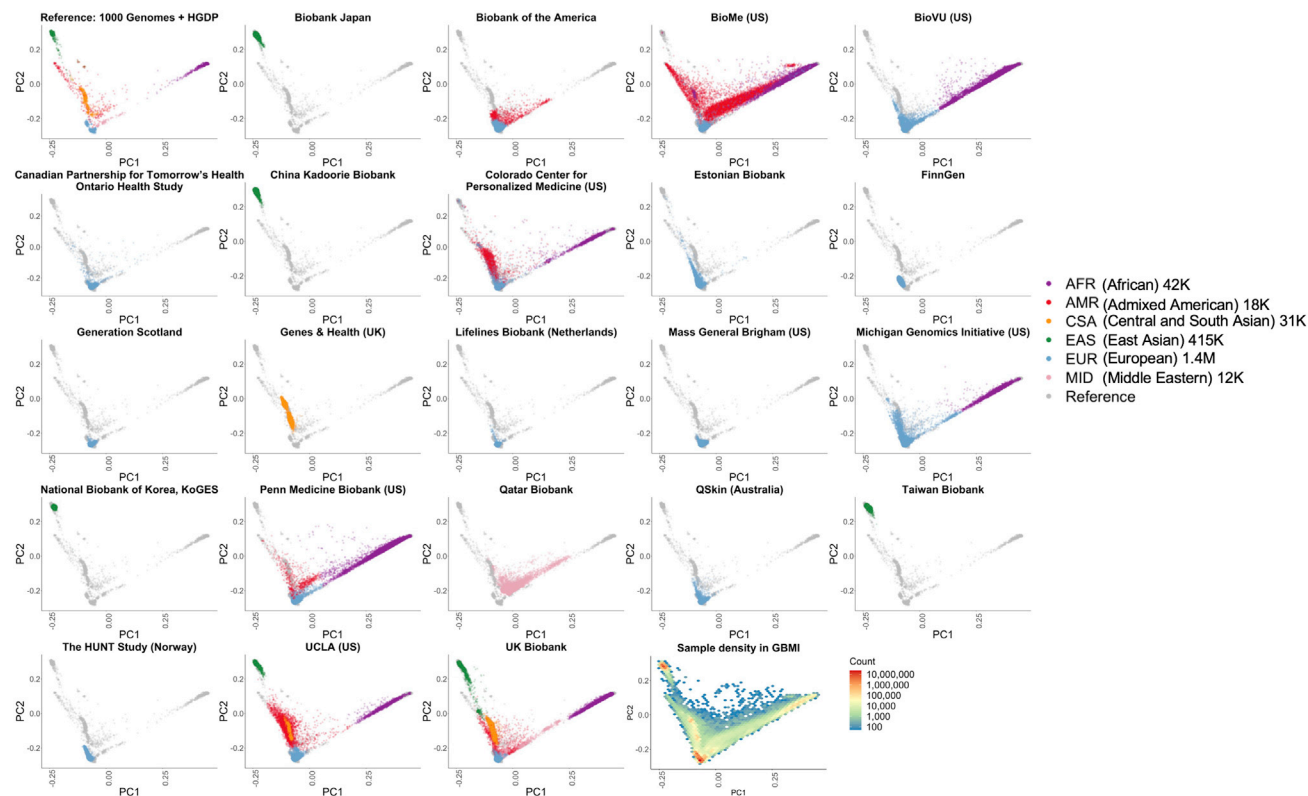


Figure 3. GBMI incorporates biobanks with diverse sample ancestry into genetic studies

Biobanks' participants were projected to the same principal component (PC) space using the pre-computed loadings of genetic markers.

Identified associations were largely shared across biobanks. The lead variants at 95% ($n = 476$) of the 500 genome-wide loci did not show evidence for heterogeneity in effect sizes across different datasets (per biobank and ancestry) (Table S7) with a p value for Cochran's Q test $\geq 1/500$, despite biobanks differing in many aspects, as discussed above. This suggests that harmonizing phenotyping and then integrating GWASs from different biobanks together using the analysis pipeline within GBMI enables reliable discoveries for genetic-disease associations. We also used the meta-regression approach implemented in MR-MEGA¹¹ for all-biobank meta-analyses. In contrast with fixed-effects, inverse variance-based meta-analyses, MR-MEGA accounts for the effect size heterogeneity across datasets, which identified 17 additional loci across 10 endpoints, including 12 that were novel (Figure S3; Table S10).

Power improved by incorporating samples with non-EUR ancestries

An additional 21.8 million genetic variants were analyzed in the all-biobank meta-analyses that were not present in the EUR-only meta-analyses with variant sets imputed from Haplotype Reference Consortium (HRC) and/or population-specific reference panels (Tables S1 and S6). The majority of these variants were rare, with 18.3 million having a minor allele frequency (MAF) $\leq 1\%$, and the other 3.4 million were common in at least one ancestry group (Figure S4). Incorporating samples with diverse ancestries to the meta-analyses allowed us to

compare effect sizes of genomic loci across ancestry. 486 out of the 500 loci were tested in more than one ancestry (Table S7). 16 out of the 486 loci showed evidence for heterogeneity in effect sizes across ancestry (p value for Cochran's Q test across ancestry $< 1/486$) (Table S11). 337 loci were identified in the EUR-only meta-analyses, and including non-EUR samples yielded 163 more loci (Figure S5A; Table S12), bringing the total number of loci to 500. While an increase in sample size drives some of our ability to detect variants, the increased diversity allows the identification of loci whose index variants are much more frequent in non-EUR ancestries. In contrast to only 4 out of the 337 loci identified in the EUR-only meta-analyses (1.19%), 21 out of the 163 additional loci (12.9%) had index variants that are at least 10 times more frequent in other ancestries than in EUR ancestries and have a MAF $< 5\%$ in EUR ancestries (Table S12). Forest plots (Figure S5B) highlight analyses with index variants more frequent in EAS than other ancestries (*MIR2054/INTU* for POAG, *PNPT1/EFEMP1* for COPD, and *NAA38* for asthma) as well as loci 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-analyses 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. 8 loci showed evidence of heterogeneous effect sizes between

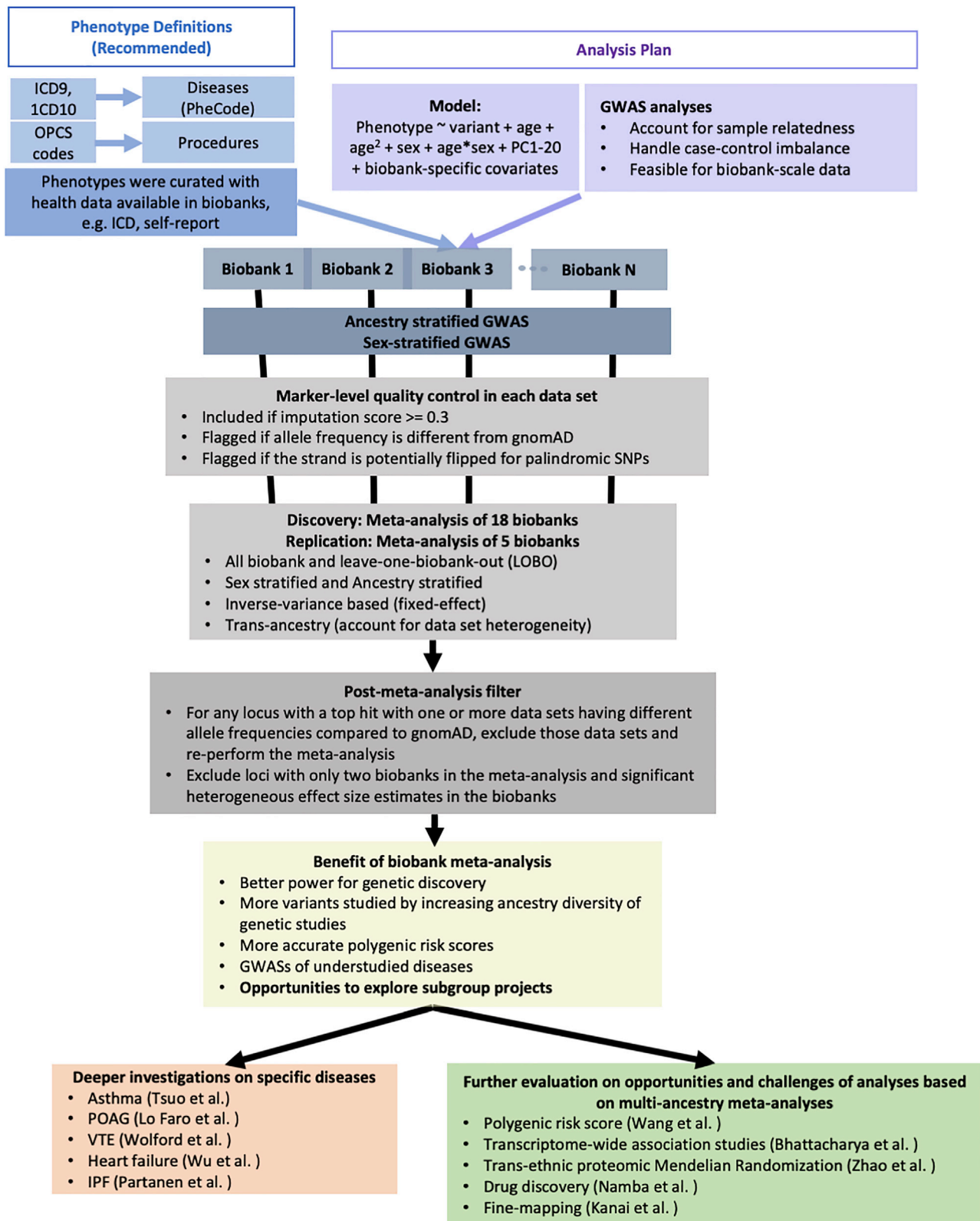


Figure 4. Workflow of the flagship project in GBMI

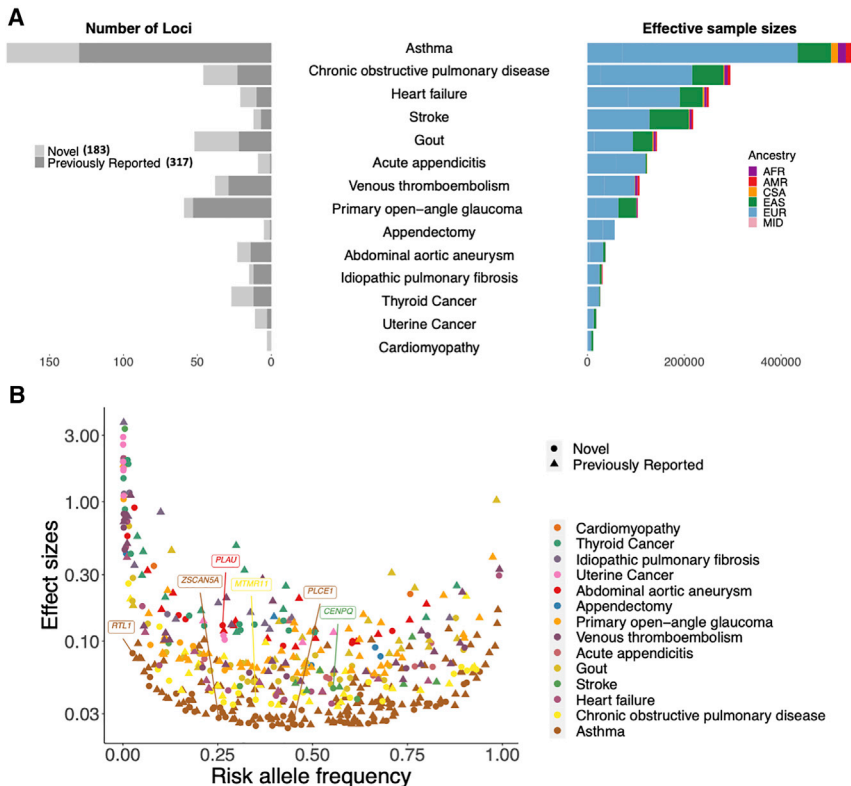


Figure 5. All-biobank meta-analyses for the 14 endpoints have successfully replicated 317 previously reported loci and identified 183 novel loci

(A) Number of loci was plotted for each endpoint (left panel) against the effective sample sizes $1/(4/\text{cases} + 4/\text{controls})$ 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. One marker with $\beta > 5$ is not shown. Gene names are labeled for the novel loci with protein-coding index variants.

males and females (p value for Cochran's Q test $< 1/479$) (Table S13; Figure S6).

Environmental factors, such as alcohol usage, that have differences in males and females may play a role in GWAS effect size differences between sexes. We have replicated two previously reported loci that are located at the aldehyde dehydrogenase family genes for gout and exhibit stronger associations in males than in females.^{12,13} The top hit of one locus was an EAS-specific intronic variant rs4646776 ($r^2 = 0.99$ with the missense variant rs6711¹⁴) 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], $p = 0.07$; in males AF = 24.2%, effect size [SE] = -0.29 [0.023], $p = 2.5 \times 10^{-36}$). The top hit of the other locus was a low-frequency EUR-specific intronic variant located in *ALDH16A1*, which has been previously identified to be associated with serum urate levels.¹³ The variant was more strongly associated with gout in males and in females (rs752383928 intronic, in females AF = 0.74%, effect size [SE] = 1.63 [0.29], $p = 2.43 \times 10^{-8}$; in males AF = 0.73%, effect size [SE] = 2.70 [0.18], $p = 1.33 \times 10^{-50}$). We have also uncovered significant sex differences for loci that were previously reported for disease association: *RANBP6/IL33* for asthma,¹⁵ *AFAP1* for COPD,¹⁶ *PKD2* for gout,¹⁷ *MUC5AC/MUC5B* for IPF,¹⁸ and *ARHGEF12* for POAG.¹⁹

The top variant in the *CTDP1/KCNG2* locus (rs11665567) was an intergenic variant with a female-specific association for asthma (in females AF = 18.8%, effect size [SE] = 0.05 [0.008],

$p = 5.62 \times 10^{-10}$; in males AF = 18.7%, effect size [SE] = 0.003 [0.01], $p = 0.75$ [p value for difference = 2.4×10^{-4}]). Interestingly, the allele associated with an increased risk in asthma has been reported to be associated with an increased risk in smoking,²⁰ but the relationship between smoking and asthma risk remains unclear. Clarifying whether the sex-specific associations are due to pleiotropic effects of the genetic variants, environmental factors, and/or possible gene-environment interactions requires further study.

There were 31 loci only identified in the sex-stratified meta-analyses but not in the sex-combined meta-analyses

(p value $> 5 \times 10^{-8}$), 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 were potentially novel for the studied phenotypes (Table S14). Female-only meta-analysis for stroke identified the previously reported locus *CETP*²¹ that did not reach the genome-wide significance threshold in the sex-combined meta-analysis. The top hit is an intronic variant rs7499892 with stronger association in females than in males (in females: effect size [SE] = 0.078 [0.014], $p = 1.08 \times 10^{-8}$; in males: effect size [SE] = 0.007 [0.012], $p = 0.56$). Transgenic expression of *CETP* increases plasma triglyceride levels in females and males through distinct mechanisms,^{22,23} while the role of triglyceride levels in the risk of stroke remains elusive. None of the 31 loci had significant heterogeneity in effect size estimates across ancestries within females or males with $p < 0.05/31$ (Table S14). In between-sex heterogeneity tests conducted within each ancestry, the significant heterogeneity of effect sizes in males and females was observed in some ancestries, suggesting that the between-sex effect heterogeneity of these loci was unlikely confounded by differences across ancestries in the proportion of each sex studied (Figure S7). Some loci with sex-specific effects were also ancestry specific, such as the EAS-specific locus *ALDH2* and the EUR-specific locus *ALDH16A1* for gout.

Integration of association results across biobanks

We evaluated the integration of different biobanks in the meta-analyses (STAR Methods; Data S1). We compared the effect sizes of top variants in individual biobanks and the

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	Heterogeneity p	Gene	Function	Cases	Controls	Number of biobanks
Novel											
AAA	10:73913343	T/C	0.737	0.88 (0.85–0.91)	5.93×10^{-12}	0.76	<i>PLAU</i>	missense	9,453	1,446,422	11
COPD	1:149934520	T/C	0.350	1.04 (1.03–1.05)	7.91×10^{-10}	0.54	<i>MTMR11</i>	missense	79,844	1,289,683	15
Stroke	6:49492265	A/G	0.446	0.96 (0.94–0.97)	1.8×10^{-11}	0.99	<i>CENPQ</i>	missense	60,176	1,310,725	16
Asthma	10:94279840	G/C	0.448	1.03 (1.02–1.03)	2.52×10^{-9}	0.98	<i>PLCE1</i>	missense	153,763	1,647,022	18
Asthma	14:100883117	G/T	0.025	1.09 (1.05–1.12)	2.61×10^{-8}	0.73	<i>RTL1</i>	missense	133,369	1,370,606	16
Asthma	19:56222056	C/A	0.253	1.03 (1.02–1.04)	2.35×10^{-8}	0.60	<i>ZSCAN5A</i>	missense	149,293	1,626,581	17
Known											
COPD	14:94378610	C/T	0.020	1.22 (1.16–1.29)	5.2×10^{-15}	9.27×10^{-3}	<i>SERPINA1</i>	missense	54,105	883,399	11
COPD	19:44908684	T/C	0.140	0.95 (0.94–0.97)	1.04×10^{-8}	0.36	<i>APOE</i>	missense	81,568	1,310,798	16
Gout	2:27508073	T/C	0.588	0.87 (0.86–0.88)	9.27×10^{-64}	0.11	<i>GCKR</i>	missense	37,105	1,448,128	15
Gout	11:64593747	G/A	0.016	0.36 (0.31–0.42)	1.19×10^{-41}	0.10	<i>SLC22A12</i>	stop gain	6,634	248,305	2
Gout	12:57449928	G/A	0.194	0.91 (0.89–0.93)	1.51×10^{-17}	0.45	<i>INHBC</i>	missense	37,105	1,448,128	15
IPF	5:1279370	T/C	0.001	862 (205–3618)	2.66×10^{-20}	0.09	<i>TERT</i>	missense	1,278	330,954	2
IPF	5:169588475	G/A	0.014	2.19 (1.81–2.66)	1.61×10^{-15}	0.02	<i>SPDL1</i>	missense	4,812	882,416	7
POAG	1:11193760	C/T	0.026	0.67 (0.6–0.74)	4.39×10^{-14}	0.90	<i>ANGPTL7</i>	missense	12,810	421,360	5
POAG	1:171636338	G/A	0.002	6.33 (4.71–8.51)	1.67×10^{-34}	4.33×10^{-6}	<i>MYOC</i>	stop gain	15,916	1,092,446	11
POAG	14:60509819	C/A	0.547	0.89 (0.87–0.91)	7.08×10^{-30}	0.31	<i>SIX6</i>	missense	26,848	1,460,599	15
Stroke	12:111803962	G/A	0.238	0.9 (0.88–0.92)	5.16×10^{-18}	0.37	<i>ALDH2</i>	missense	23,804	269,656	4
VTE	1:169549811	C/T	0.020	3.04 (2.85–3.24)	1.5×10^{-245}	5.52×10^{-13}	<i>F5</i>	missense	26,749	1,011,509	9
VTE	12:6034818	T/C	0.889	1.1 (1.07–1.13)	1.59×10^{-10}	0.21	<i>VWF</i>	missense	27,987	1,035,290	9
VTE	12:103742510	C/T	0.011	1.65 (1.45–1.88)	6.19×10^{-14}	0.25	<i>STAB2</i>	missense	10,353	341,418	2
Asthma	1:12115601	G/A	0.012	0.85 (0.8–0.89)	1.7×10^{-11}	0.46	<i>TNFRSF8</i>	missense	118,767	1,202,660	12
Asthma	1:31699894	G/T	0.573	1.03 (1.02–1.04)	1.61×10^{-10}	0.49	<i>COL16A1</i>	missense	148,045	1,579,632	17
Asthma	4:38797027	C/A	0.387	0.95 (0.94–0.96)	4.21×10^{-21}	0.54	<i>TLR1</i>	missense	138,764	1,458,022	15
Asthma	4:102267552	C/T	0.044	1.08 (1.06–1.1)	2.53×10^{-12}	0.81	<i>SLC39A8</i>	missense	129,434	1,256,670	14
Asthma	5:14610200	C/G	0.084	1.07 (1.05–1.09)	7.68×10^{-15}	0.16	<i>OTULINL</i>	missense	125,483	1,241,068	13
Asthma	9:128721272	T/A	0.068	0.95 (0.93–0.96)	5.61×10^{-10}	0.21	<i>ZDHHC12</i>	missense	152,469	1,638,824	18

^aFrequencies are reported with respect to the alternate allele (ALT) in the combined meta-analysis datasets.

^bOdds ratios are reported with respect to the alternate allele (ALT) in the meta-analyses.

corresponding LOBO meta-analyses by fitting a Deming regression.²⁴ Most of the slope estimates were not significantly different from 1 across biobanks and phenotypes with exceptions among biobanks with smaller sample sizes and non-EUR or multiple ancestries (Figure S8; Data S1). Genetic correlation estimates between individual biobanks and LOBO for diseases with the highest heritability estimates, asthma, gout, and COPD, were close to 1 (STAR Methods; Figure S9; Table S9). We also compared population-based biobanks and hospital-/healthcare-based biobanks (STAR Methods) with meta-analyses for biobank groups separately for gout, ThC, asthma, and POAG, and consistent effect sizes between two biobank groups were observed (Figure S10; Data S1). Robust genetic association results despite differences among biobanks suggested the integration of genetic association results across biobanks.

For previously reported loci, consistent effect size direction and magnitude were observed between GBMI and the previous largest meta-analyses for AAA²⁵ and gout²⁶ (Table S15; Data S1) as well as for VTE (see Wolford et al.²⁷). All 18 loci that were previously identified by the Trans-National Asthma Genetic Consortium (TAGC)¹⁵ for asthma had more significant association p values in GBMI (Figure S11), and attenuation of genetic effects in GBMI was observed (Table S15; see Tsuo et al.²⁸). Attenuation of effect estimates in GBMI compared with disease-specific cohorts that generally study highly ascertained patients was also observed for IPF (see Koskela et al.,²⁹ Data S1). These results suggested that the impact of using EHR-curated phenotypes in biobanks on effect size estimates varies across phenotypes and that biobank studies can show attenuated genetic effects where phenotyping is often, by necessity, more pragmatic.

Biological implications of genetic associations Pleiotropic effects of associated loci

We investigated the genetic relationship between endpoints studied in this project and other complex traits by examining associations of the top variants identified by all-biobank meta-analyses with 1,283 human diseases in UKBB (STAR Methods). 78 variants identified from 12 GBMI endpoints (except for HCM and UtC) exhibited significant ($p < 5 \times 10^{-8}$) pleiotropic associations with at least one other phenotype (Table S16; Data S1). We further investigated pleiotropic effects of the 52 loci (30 novel) identified for gout by all-biobank meta-analysis. 40 of these loci were associated with serum urate levels,^{26,30–32} and most of these loci were also associated with other relevant traits and diseases (Table S17; Data S1).

Prioritization of cell types, tissues, and genes

To further understand the biology underlying the genetic associations, we prioritized tissues and cell types in which genes at the associated loci are likely to be highly expressed using Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT)³³ (Table S18).

Prioritizing potentially functional genes with genetic variant associations is a large challenge for genomic research. We applied several methods to prioritize potentially functional genes, including DEPICT (Table S19), the gene-level polygenic priority score (PoPS)³⁴ (Table S20), transcriptome-wide association studies (TWASs)³⁵ (Table S21), and proteome-wide Mendelian randomization (PWMR)³⁶ (Table S22; STAR Methods). Using

asthma, POAG, and VTE as examples, the gene lists generated by these different methods showed little overlap (Figure S12). For asthma, 618 genes were prioritized by at least 1 of the 4 approaches (Figure S12A). However, no genes were prioritized by all 4 methods, and only 5 were prioritized by any 3 methods (Table S22). All these genes are located at well-known asthma-associated loci. We then extracted the nearest genes of the most significant variants (for intergenic variants, the nearest genes on both sides were included if both are located within 50 kb from the top hits), which brings the total number of prioritized genes to 729. *FCER1G*, *IL4R*, and *SMAD3*, which were prioritized by DEPICT, TWAS, and PoPS, were also the nearest genes of top hits at those loci. 17 more genes were prioritized by any of the two methods and the naive nearest genes approach (Figure S12A; Table S22): *BCL2*, *CD247*, *CD28*, *GSDMB*, *HDAC7*, *IL13*, *IL2RA*, *IL6R*, *IL7R*, *ITPKB*, *JAZF1*, *NEK6*, *PTPRC*, *RUNX3*, *STAT6*, *TLR1*, and *TNFSF8*. Low overlap of the prioritized gene lists by different methods was also seen for POAG and VTE (Figures S12B and S12C; Table S22; Data S1).

In line with previous discussions,³⁴ these results suggest that existing gene prioritization methods successfully prioritized relevant genes for diseases but had poor agreement. Note that besides adapting different statistical models and pipelines, these approaches prioritize genes with different expression data types (STAR Methods).

In addition, a gold standard set of 41 VTE genes was curated blindly from the meta-analysis results²⁷ (Figure S13A; Table S23). Based on this gene set, the nearest gene approach had comparable precision and recall to other methods (Figure S13B; Data S1). When using 13 genes in the gold standard set that are located within 1 Mb around the VTE top hit, as expected, we observed an increase in the recall of DEPICT and the nearest gene approach (Figures S13B and S13C). This is because both approaches tend to prioritize genes that are located at GWAS loci.

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.

Prioritization of functional variants through fine mapping

While previous meta-analysis studies have applied existing fine-mapping methods to further prioritize functional variants at disease association loci,^{37–39} it is unclear whether heterogeneous characteristics of biobanks in the meta-analysis affect fine-mapping calibration and recall. We investigated the impacts on fine mapping of the heterogeneity across biobanks via simulation studies and demonstrated that different sample sizes, ancestries, phenotyping, genotyping, and imputation can lead to mis-calibrated fine-mapping results.⁴⁰ Thus, we developed a summary statistics-based quality control (QC) method, SLALOM, to identify suspicious loci for meta-analysis fine mapping. Applying SLALOM to the all-biobank meta-analysis results for 14 endpoints in GBMI found that 68% of loci showed suspicious patterns that call into question fine-mapping accuracy.⁴⁰ These results suggest the need for development of methods that take the cohort heterogeneity into account for reliable meta-analysis fine mapping. We thus urge caution when

interpreting fine-mapping results from meta-analysis until improved methods are available. Because of the lack of robust fine-mapping methods to obtain credible sets for most loci, we report the protein-coding variants at association loci (Tables 1 and S8) to shed light on potentially functional variants.

Biobank meta-analysis for genetic association studies Improving power of genetic discovery for common diseases

Aggregating 18 biobanks in GBMI substantially increases sample sizes for genetic association studies for asthma (Table S2), leading to an increase in power for genetic discovery; 179 genome-wide significant loci for asthma were identified by GBMI, of which 49 are novel (Table S7). Notably, all 18 loci that were reported by TAGC¹⁵ have more significant association *p* values in GBMI (Figure S11). 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 (three overlapped biobank datasets were excluded from GBMI), which doubled the case numbers compared with the previous largest meta-analysis and successfully identified 103 significant loci, of which 19 are novel.⁴¹

Providing opportunities for genetic studies on less prevalent diseases

EHR-linked biobanks provide opportunities to assess less prevalent diseases that were understudied by previous GWASs. For example, the largest meta-analysis for gout so far was conducted on 13,179 cases and 750,634 controls across 20 studies. Here, meta-analysis of 15 biobanks in GBMI achieved a sample size of 37,105 cases and 1,448,128 controls across 5 genetic ancestries (Table S2) and identified 52 significant loci, of which 30 are novel (Table S7).

Biobanks also enable genetic studies of different types of disease phenotypes. Meta-analyses of biobanks for AcApp and the relevant procedure endpoint appendectomy, respectively, demonstrated high genetic correlation between the two endpoints ($r^2 = 0.99$). Out of the 9 loci identified for AcApp, 3 were also significant for appendectomy (Table S7) even with a 3 times lower sample size, suggesting that the procedure phenotypes may add meaningful information in biobank-based genetic studies. Incorporating these phenotypes to traditional disease diagnosis phenotypes could improve discovery power.

Improving PRSs with multi-biobank multi-ancestry meta-analyses

The LOBO meta-analyses in GBMI allow us to construct biobank-specific PRSs and evaluate PRSs for disease risk prediction. We have investigated PRS construction, evaluation, and interpretation using the multi-biobank multi-ancestry summary statistics resource and shared lessons and methodological considerations (see details in Wang et al.⁹).

Using asthma as an example, we observed improved PRS prediction accuracy with GBMI summary statistics, compared with a previous meta-analysis,¹⁵ in 6 biobanks across 6 ancestral populations (Figure S14). The accuracy improvements were greatest for EUR samples (0.017–0.047), followed by SAS and EAS samples (0.01–0.037), and were least for AFR samples

(0.005–0.014). Improvements in prediction accuracy using GBMI compared with previous GWASs were also observed for more endpoints.⁹ Note that the PRS prediction accuracy varies across biobanks and ancestries, which might be attributable to factors such as non-genetic factors and differences among biobanks. How to better account for these factors in disease prediction models remains an open question for future research and methods development.

DISCUSSION

Genetic discovery benefits from the increasing numbers of EHR-linked biobanks, despite differences among biobanks. As of January 2022, 23 biobanks across four continents comprising six major ancestral groups have joined GBMI to uncover genetic risk factors of human diseases. Researchers worked to address challenges in large-scale genetic studies related to biobanks, such as harmonizing phenotypes to account for different sources of heterogeneity and developing analytic pipelines to account for sample relatedness, case-control imbalance, and large data sizes. GBMI is an important initiative aiming to integrate large-scale biobanks for genetic studies. With carefully harmonized phenotype definitions and analysis pipelines, we meta-analyzed GWASs in up to 18 biobanks for 14 endpoints, including common diseases (asthma, COPD, VTE, etc.), less prevalent diseases (gout, IPF, AAA, ThC, etc.), and the procedure endpoint appendectomy for primary discovery, and then conducted replication in up to 5 additional biobanks. 500 genome-wide significant loci were detected, of which 183 are novel. Sex-stratified meta-analysis allows for comparing effects between sexes and identified 8 loci with different effect sizes in men and women. Not only have we demonstrated the integration of genetic association results from different biobanks, but we have also illustrated the gains by meta-analyzing biobanks together. The increase in the sample size and sample diversity leads to higher discovery power. Incorporating non-EUR samples in the meta-analyses allows for genetic association tests of 21.8 million additional markers. 85% of those markers are low-frequency ones ($AF < 1\%$), which may further facilitate functional follow-up studies to disentangle the causal variants at identified loci.

Several loci identified in our biobank meta-analyses have associations with other human diseases, which could be due to pleiotropy, disease comorbidities, or linkage disequilibrium. Follow-up statistical analyses, such as colocalization analysis, are needed to obtain clearer biological implications. As expected, based on biobank meta-analysis results, more accurate predictive PRSs can be constructed because of the increased genetic discovery power compared with previous studies. This gain can be further extended to non-EUR 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 (1) to deepen the genetic investigation of the biological implications of results for several

endpoints,^{27–29,42,43} (2) to systematically characterize genome-wide significant loci via fine mapping,⁴⁰ transcriptome-wide association,³⁵ protein QTL Mendelian randomization analysis,³⁶ and drug target prioritization,⁴⁴ and (3) to improve the disease risk prediction with PRSs based on the multi-biobank multi-ancestry meta-analysis results.⁹

Together, the pilot work conducted in GBMI shows that biobanks can be meta-analyzed to provide reliable genetic discoveries despite the heterogeneous characteristics 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 analysis pipelines. We have evaluated the challenges in downstream *in silico* studies to prioritize 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 current analyses based on the biobank meta-analysis results.

Limitations of the study

Although we have demonstrated how biobank meta-analysis can help advance the understanding of genetic risk factors for human diseases, the current study has several limitations. (1) Most samples (73%) in the current meta-analysis are of EUR ancestry, and, because of the small proportion of samples of non-EUR ancestries, the power to identify and investigate ancestry-specific genetic and environmental factors for human traits and diseases remains limited. (2) Phenotype curation in biobanks was harmonized according to the phecode map, which led to attenuation of effect size estimates for several phenotypes compared with disease-specific cohorts that generally study highly ascertained patients. (3) As has been demonstrated in detail,⁴⁰ due to heterogeneous characteristics across biobanks, prioritization of functional variants through fine mapping based on the summary statistics generated by biobank meta-analyses can have mis-calibrated results. (4) Different gene prioritization methods do not always agree, highlighting the challenges in interpreting genome-wide significant loci and nominating genes for functional follow-up. (5) Only three biobanks used self-report data to curate phenotypes (two are EA biobanks). This challenged our ability to evaluate the impact of phenotyping on genetic associations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **METHOD DETAILS**
 - Phenotype definition
 - GWAS
 - Post-GWAS quality control
 - Meta-analysis

- Post-meta-analysis quality control
- PC projection
- Variant annotation
- Heritability estimation
- Evaluate the integration genetic associations from diverse biobanks
- Prioritize functional genes
- Phenome-wide association test
- Polygenic scores

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.xgen.2022.100192>.

ACKNOWLEDGMENTS

The work of the contributing biobanks was supported by numerous grants from governmental and charitable bodies. Biobank-specific acknowledgments and more detailed acknowledgments are included in [Data S2](#).

AUTHOR CONTRIBUTIONS

Initiative management, S.B.C., J.C., N.J.C., M.J.D., E.E.K., A.R.M., B.M.N., Y.O., A.V.P., D.A.v.H., R.G.W., C.J.W., W.Z., and S.Z.; individual biobank analysis, A.B., Y.B., B.M.B., C.D.B., S.C., T.-T.C., K.C., S.M.D., M.D., G.H.d.B., Y.D., N.J.D., M.-J.F., Y.-C.A.F., S.F., V.L.F., L.G.F., E.R.G., T.R.G., D.H.G., C.R.G., G.G.-A., S.E.G., L.A.G., C.H., J.B.H., W.E.H., H.H., K.H., N.I., A.I., R.J., M. Kurki, J.K., N.K., E.E.K., J.T.K., M. Kanai, T.L., K.L., M.H.L., S.L., K.L., Y.-F.L., V.L.F., R.J.F.L., E.A.L.-M., A.R.-M., S.M.-G., R.M., R.E.M., H.C.M., A.R.M., Y.M., H.M., S.E.M., I.Y.M., B.M., S.M., K.N., S.N., M.A.N.-A., K.N., Y.O., P.P., A.L.-P., A.P., B.P., S.P., M.H.P., D.J.R., N.R., M.D.R., A.R., C.S., S.S., S.S.S., J.A.S., P.S., I.S., T.T., R.T., K.T., J.U., D.A.v.H., B.V., M.V., Y.V., J.M.V., R.G.W., Y.W., S.J.W., B.N.W., K.-H.H.W., M.Z., X.Z., and S.Z.; individual biobank management, N.A., A.A.T., K.M.A.-D., P.A., K.C.B., M. Boehnke, M. Boezen, C.D.B., A.C., Z.C., C.-Y.C., J.C., N.J.C., S.M.D., S.F., Y.-C.A.F., S.F., E.F., T.G., C.R.G., C.J.G., Y.G., H.H., K.A.H., K.H., S.I.I., N.M.J., N.K., E.E.K., J.T.K., C.L., M.H.L., M.T.M.L., L.L., K.L., Y.-F.L., R.J.F.L., J.L., S.M., Y.M., K.M., I.Y.M., Y.O., C.M.O., A.V.P., B.P., D.J.P., D.J.R., M.D.R., S.S., J.W.S., H.S., K.S., T.T., U.T., R.C.T., D.A.v.H., M.V., R.G.W., D.C.W., C.W., J.W., M.Z., X.Z., and S.Z.; study design and interpretation of results, A.B., M. Boehnke, M. Boezen, B.M.B., T.-T.C., C.-Y.C., M.J.D., G.D.S., N.J.D., S.F., M.-J.F., H.K.F., E.R.G., A.G., T.G., J.B.H., J.H., K.H., R.J., M.K., E.E.K., T.K., C.M.L., V.L.F., E.A.L.-M., A.R.M., S.N., B.M.N., C.M.O., J.J.P., B.P., N.R., H.R., J.A.S., I.S., K.T., D.A.v.H., R.G.W., Y.W., D.C.W., S.J.W., C.J.W., B.N.W., J.W., K.-H.H.W., M.Z., H.Z., J.Z., W.Z., X.Z., and S.Z.; drafted and edited the paper, A.B., M. Boehnke, M. Boezen, M.J.D., G.H.d.B., N.J.D., T.R.G., J.B.H., N.I., N.M.J., M.K., V.L.F., S.M., A.R.M., H.M., S.N., B.M.N., C.M.O., B.P., H.R., C.S., J.A.S., J.W.S., K.T., Y.W., D.C.W., C.J.W., K.-H.H.W., H.Z., J.Z., W.Z., and S.Z.; primary meta-analysis and quality control, M.J.D., H.K.F., M. Kanai, J.K., J.T.K., M. Kurki, M.M., B.M.N., C.J.W., K.-H.H.W., and W.Z.; drug discovery: S.N., T.K., K.-H.H.W., W.Z., and Y.O.; fine mapping, M. Kanai, W.Z., M.J.D., and H.K.F.; polygenic risk score, Y.W., S.N., E.A.L.-M., S.K., K.T., K.L., M. Kanai, W.Z., K.W., M.-J.F., L.B., P.A., P.D., V.L.F., R.M., Y.M., B.B., S.S., J.U., E.R.G., N.J.C., I.S., Y.O., A.R.M., and J.B.H.; proteome-wide Mendelian randomization, H.Z., H.R., A.B., G.H., G.D.S., B.M.B., W.Z., B.M.N., T.R.G., and J.Z.; transcriptome-wide association study, A.B., J.B.H., W.Z., J.Z., M. Kanai, B.P., E.R.G., and N.J.C.; asthma, K.T., W.Z., Y.W., M. Kanai, S.N., Y.O., B.M.N., M.J.D., and A.R.M.; heart failure, K.-H.H.W., N.J.D., B.N.W., I.S., S.E.G., J.B.H., N.J.C., M.P., R.J.F.L., M.J.D., B.M.N., W.Z., W.E.H., and C.J.W.; idiopathic pulmonary fibrosis, J.J.P., W.Z., M.J.D., J.T.K., N.J.C., and J.B.H.; primary open-angle glaucoma, V.L.F., A.B., W.Z., Y.W., K.L., M. Kanai, E.A.L.-M., P.S., R.T., X.Z., S.N., S.S., Y.O., N.I., S.M., H.S., I.S., C.W., A.R.M., E.R.G., N.M.J., N.J.C., and J.B.H.; stroke, I.S.,

K.-H.H.W., W.H., B.N.W., W.Z., J.E.H., A.P., B.B., A.H.S., M.E.G., R.G.W., K.H., C.K., S.Z., M.J.D., B.M.N., and C.J.W.; venous thromboembolism, B.N.W., I.S., K.-H.H.W., B.B., V.L.F., K.T., M.D., B.N., W.Z., J.A.S., and C.J.W. All authors reviewed the manuscript.

DECLARATION OF INTERESTS

M.J.D. is a founder of Maze Therapeutics. B.M.N. is a member of the scientific advisory board at Deep Genomics and a 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 23andMe 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. S.M. is a co-founder of and holds stock in Seonix Bio.

Received: November 29, 2021

Revised: June 19, 2022

Accepted: September 9, 2022

Published: October 12, 2022

REFERENCES

- 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. <https://doi.org/10.1093/nar/gkw1133>.
- 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. <https://doi.org/10.1126/scitranslmed.3008604>.
- 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. <https://doi.org/10.1093/hmg/ddy081>.
- Fatumo, S., Mugisha, J., Soremekun, O.S., Kalungi, A., Mayanja, R., Kintu, C., Makanga, R., Kakande, A., Abaasa, A., Asiki, G., et al. (2022). Uganda Genome Resource: a rich research database for genomic studies of communicable and non-communicable diseases in Africa. Preprint at bioRxiv. Published online May 7. <https://doi.org/10.1101/2022.05.05.22274740>.
- 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. <https://doi.org/10.1038/nature15393>.
- 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. <https://doi.org/10.1126/science.296.5566.261b>.
- 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 phenotype-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* 31, 1102–1110. <https://doi.org/10.1038/nbt.2749>.
- 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. <https://doi.org/10.1038/s41586-020-2308-7>.
- Wang, Y., Namba, S., Lopera-Maya, E.A., et al. (2021). Global biobank analyses provide lessons for computing polygenic risk scores across diverse cohorts. Preprint at medRxiv. 2021.11.18.21266545. <https://doi.org/10.1101/2021.11.18.21266545>.
- Puckelwartz, M.J., Pesce, L.L., Dellefave-Castillo, L.M., Wheeler, M.T., Pottinger, T.D., Robinson, A.C., Kearns, S.D., Gacita, A.M., Schoppen, Z.J., Pan, W., et al. (2021). Genomic context differs between human dilated cardiomyopathy and hypertrophic cardiomyopathy. *J. Am. Heart Assoc.* 10, e019944. <https://doi.org/10.1161/JAHA.120.019944>.
- Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M.I., and 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. <https://doi.org/10.1093/hmg/ddx280>.
- Mizuno, Y., Harada, E., Morita, S., Kinoshita, K., Hayashida, M., Shono, M., Morikawa, Y., Murohara, T., Nakayama, M., Yoshimura, M., and Yasue, H. (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. <https://doi.org/10.1161/CIRCULATIONAHA.114.013120>.
- Sulem, P., Gudbjartsson, D.F., Walters, G.B., Helgadóttir, H.T., Helgason, A., Gudjonsson, S.A., Zanon, C., Besenbacher, S., Björnsdóttir, 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. <https://doi.org/10.1038/ng.972>.
- 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. <https://doi.org/10.1038/s41562-019-0805-1>.
- Deménais, F., Margeritte-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. (2018). Multiethnicity association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* 50, 42–53. <https://doi.org/10.1038/s41588-017-0014-7>.
- 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. <https://doi.org/10.1038/s41467-018-05369-0>.
- 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. <https://doi.org/10.1136/annrheumdis-2014-206191>.
- 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. <https://doi.org/10.1056/NEJMoa1013660>.
- 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. <https://doi.org/10.1093/hmg/ddv027>.
- Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* 51, 237–244. <https://doi.org/10.1038/s41588-018-0307-5>.

21. 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. <https://doi.org/10.1016/j.jstrokecerebrovasdis.2018.06.020>.
22. 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. <https://doi.org/10.1194/jlr.M069013>.
23. 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. <https://doi.org/10.1002/lipd.12271>.
24. Deming, W.E. (1943). Statistical Adjustment of Data²⁶¹. <https://psycnet.apa.org/fulltext/1944-00642-000.pdf>.
25. 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. <https://doi.org/10.1161/CIRCULATIONAHA.120.047544>.
26. 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. <https://doi.org/10.1038/s41588-019-0504-x>.
27. Wolford, B.N., Zhao, Y., Surakka, I., Wu, K.H.H., Yu, X., Richter, C.E., Bhatta, L., Brumpton, B., Desch, K., Thibord, F., et al. (2022). Multi-ancestry GWAS for venous thromboembolism identifies novel loci followed by experimental validation in zebrafish. Published online June 27. <https://doi.org/10.1101/2022.06.21.22276721>.
28. Tsuo, K., Zhou, W., Wang, Y., et al. (2021). Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. Preprint at medRxiv. Published online December 7. <https://doi.org/10.1101/2021.11.30.21267108>.
29. Partanen, J.J., Häppölä, P., Zhou, W., et al. (2021). Leveraging global multi-ancestry meta-analysis in the study of Idiopathic Pulmonary Fibrosis genetics. Preprint at medRxiv. Published online December 31. <https://doi.org/10.1101/2021.12.29.21268310>.
30. 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. <https://doi.org/10.1161/HYPERTENSIONAHA.120.16547>.
31. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Lillila, H.M., Kiiskinen, T., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* 53, 185–194. <https://doi.org/10.1038/s41588-020-00757-z>.
32. 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. <https://doi.org/10.1371/journal.pone.0119752>.
33. 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. <https://doi.org/10.1038/ncomms6890>.
34. 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. Preprint at bioRxiv. Published online September 10. <https://doi.org/10.1101/2020.09.08.20190561>.
35. Bhattacharya, A., Hirbo, J.B., Zhou, D., et al. (2021). Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: lessons from the Global Biobank Meta-analysis Initiative. Preprint at medRxiv. Published online November 29. <https://doi.org/10.1101/2021.11.24.21266825>.
36. Zhao, H., Rasheed, H., Nøst, T.H., Cho, Y., Liu, Y., Bhatta, L., Bhattacharya, A., Hemani, G., Smith, G.D., Brumpton, B.M., et al. (2022). Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. Preprint at medRxiv. 2022.01.09.21268473. <https://doi.org/10.1101/2022.01.09.21268473>.
37. Gharahkhani, P., Jorgenson, E., Hysi, P., Khawaja, A.P., Pendergrass, S., Han, X., Ong, J.S., Hewitt, A.W., Segrè, A.V., Rouhana, J.M., et al. (2021). Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. *Nat. Commun.* 12, 1258. <https://doi.org/10.1038/s41467-020-20851-4>.
38. Wightman, D.P., Jansen, I.E., Savage, J.E., Shadrin, A.A., Bahrami, S., Holland, D., Rongve, A., Børte, S., Winsvold, B.S., Drange, O.K., et al. (2021). A genome-wide association study with 1, 126, 563 individuals identifies new risk loci for Alzheimer’s disease. *Nat. Genet.* 53, 1276–1282. <https://doi.org/10.1038/s41588-021-00921-z>.
39. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* 50, 1505–1513. <https://doi.org/10.1038/s41588-018-0241-6>.
40. Kanai, M., Elzur, R., Zhou, W., Daly, M.J., Finucane, H.K., and Finucane, H.K. (2022). Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. Preprint at medRxiv. 2022.03.16.22272457. <https://doi.org/10.1101/2022.03.16.22272457>.
41. Lo Faro, V., Bhattacharya, A., Zhou, W., et al. (2022). Genome-wide association meta-analysis identifies novel ancestry-specific primary open-angle glaucoma loci and shared biology with vascular mechanisms and cell proliferation. Preprint at medRxiv. 2021.12.16.21267891. <https://doi.org/10.1101/2021.12.16.21267891>.
42. Surakka, I., Wu, K.H., Hornsby, W., Wolford, B.N., Shen, F., Zhou, W., Huffman, J.E., Pandit, A., Hu, Y., Brumpton, B., et al. (2022). Multi-ancestry meta-analysis identifies 2 novel loci associated with ischemic stroke and reveals heterogeneity of effects between sexes and ancestries. Preprint at medRxiv. 2022.02.28.22271647. <https://doi.org/10.1101/2022.02.28.22271647>.
43. Wu, K.H.H., Douville, N.J., Konerman, M.C., et al. (2021). Polygenic risk score from a multi-ancestry GWAS uncovers susceptibility of heart failure. Preprint at medRxiv. 2021.12.06.21267389. <https://doi.org/10.1101/2021.12.06.21267389>.
44. Namba, S., Konuma, T., Wu, K.H., Zhou, W., and Global Biobank Meta-analysis Initiative; and Okada, Y. (2021). A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis. Preprint at medRxiv. 12.03.21267280. <https://doi.org/10.1101/2021.12.03.21267280>.
45. 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. <https://doi.org/10.1038/s41588-018-0184-y>.
46. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyadinov, 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. <https://doi.org/10.1038/s41588-021-00870-7>.
47. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. <https://doi.org/10.1093/nar/gkq603>.
48. 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. <https://doi.org/10.1186/s13059-016-0974-4>.

49. 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. <https://doi.org/10.1038/ng.3211>.
50. Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium; Ye, C.J., Price, A.L., and Zaitlen, N. (2016). Asian genetic epidemiology network type 2 diabetes Consortium, ye CJ, price AL, zaitlen N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* *99*, 76–88. <https://doi.org/10.1016/j.ajhg.2016.05.001>.
51. 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. <https://doi.org/10.1371/journal.pcbi.1004219>.
52. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* *369*, 1318–1330. <https://doi.org/10.1126/science.aaz1776>.
53. 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. <https://doi.org/10.1038/s41588-020-0706-2>.
54. Bhattacharya, A., Li, Y., and Love, M.I. (2021). MOSTWAS: multi-omic strategies for transcriptome-wide association studies. *Zhu X. PLoS Genet.* *17*, e1009398. <https://doi.org/10.1371/journal.pgen.1009398>.
55. 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. <https://doi.org/10.1038/ng.3506>.
56. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Schizophrenia Working Group of the Psychiatric Genomics Consortium; and McCarroll, S., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* *50*, 538–548. <https://doi.org/10.1038/s41588-018-0092-1>.
57. 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. <https://doi.org/10.1016/j.ajhg.2017.01.031>.
58. 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. <https://doi.org/10.1038/s41588-019-0367-1>.
59. Zhang, J., Dutta, D., Köttgen, A., et al. (2021). Large Bi-ethnic study of plasma proteome leads to comprehensive mapping of cis-pQTL and models for proteome-wide association studies. Preprint at bioRxiv. 2021.03.15.435533. <https://doi.org/10.1101/2021.03.15.435533>.
60. 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. <https://doi.org/10.1002/gepi.22077>.
61. 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. <https://doi.org/10.1002/sim.7221>.
62. 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. <https://doi.org/10.1002/sim.6522>.
63. 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. <https://doi.org/10.1371/journal.pgen.1004383>.
64. 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. <https://doi.org/10.1038/s41588-020-0682-6>.
65. 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. <https://doi.org/10.1371/journal.pgen.1007081>.
66. 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. <https://doi.org/10.1038/s41467-019-09718-5>.
67. 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. <https://doi.org/10.1002/gepi.21614>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Association results	This paper	https://www.globalbiobankmeta.org/resources
PRS weights	This paper; 9	https://www.pgscatalog.org/publication/PGP000262/
gnomAD	8	https://www.nature.com/articles/s41586-020-2308-7
GTE _x ver. 8	52	https://www.science.org/doi/10.1126/science.aaz1776
1000 Genomes	5	https://www.nature.com/articles/nature15393
Human Genome Diversity Project (HGDP)	6	https://www.science.org/doi/10.1126/science.296.5566.261b
Software and algorithms		
Scripts used for quality control, meta-analysis and summary of results	This paper	https://zenodo.org/badge/latestdoi/295461030
PC projection	This paper	https://zenodo.org/badge/latestdoi/353203447
SAIGE	45	https://www.nature.com/articles/s41588-018-0184-y
REGENIE	46	https://www.nature.com/articles/s41588-021-00870-7
MR-MEGA	11	https://academic.oup.com/hmg/article/26/18/3639/3976569
ANNOVAR	47	https://academic.oup.com/nar/article/38/16/e164/1749458
LDSC	49	https://www.nature.com/articles/ng.3211
DEPICT	33	https://www.nature.com/articles/ncomms6890
PoPS	34	https://doi.org/10.1101/2020.09.08.20190561
PRS-CS	66	https://www.nature.com/articles/s41467-019-09718-5

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Wei Zhou wzhou@broadinstitute.org.

Materials availability

This study did not generate new unique reagents.

Data and code 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 browsing at the browser <http://results.globalbiobankmeta.org>. The PRS weights estimated using all-biobank multi-ancestry meta-analyses and leave-UKBB-out multi-ancestry meta-analyses have been deposited within the PGS Catalog with study ID PGP000262 (<https://www.pgscatalog.org/>).
- All original code has been deposited to Zenodo with DOIs as below and is publicly available as of the date of publication. Links are listed in the [key resources table](#).
- Scripts used for quality control, meta-analysis, and summary of results are available at <https://github.com/globalbiobankmeta> and deposited at <https://zenodo.org/badge/latestdoi/295461030>.
- Scripts for PC projection are deposited at <https://zenodo.org/badge/latestdoi/353203447>.
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

METHOD DETAILS

Phenotype definition

A phenotype definition guideline was created and shared with all biobanks (Table S4). The disease endpoints were defined following the phecode maps⁷ to map 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-specific 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, while all other participants without these codes were controls. Biobanks which do not collect the ICD codes or OPCS codes define phenotypes using the available EHRs according to the phenotype definitions in the guideline.

GWAS

Each biobank conducted genotyping, imputation and quality controls and estimated sample ancestry independently. Then biobank run GWAS 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 [STAR Methods](#)). We recommended to run GWAS analysis using Scalable and Accurate Implementation of GEneralized mixed model (SAIGE)⁴⁵ or REGENIE,⁴⁶ 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 ([Table S24](#) and [Figures S15](#) and [S16](#)). Genetic variants with MAC < 20 and variants that were 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 gnomAD⁸ (AF-gnomAD) by ancestry. If a palindromic SNP met any one of the following standards, we flipped its alleles in the GWAS data set and indicated that this variant had the potential strand flip with a flag: 1. The fold difference was greater than two, 2. The allele frequency of the alternative allele in the GWAS data set was 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 was greater than three 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 (LOBO) ([Figure S8](#)) 3. all biobanks stratified by each ancestry, and 4. all biobanks stratified by sex. Trans-ancestry meta-analysis was performed using MR-MEGA¹¹ with three 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 ± 1 Mb. A locus was categorized as “previously reported” if the region after merging is within ± 500 kb of variants for the corresponding phenotype in GWAS Catalog,¹ otherwise, it was categorized as “novel” ([Table S7](#)). The most significant variant in each locus was selected as the index variant. The nearest gene(s) to the index variant was 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.

Post-meta-analysis quality control

For genome-wide significant loci, post meta-analysis quality control has been conducted. 1. For any locus with the top hit tested in one or more data sets that have different allele frequencies compared to gnomAD,⁸ we excluded those data sets and re-performed the meta-analysis. 2. We excluded loci with the top hits that were tested in only two biobanks and had significant heterogeneous effect size estimates in the biobanks.

PC projection

179,195 genetic variants have been genotyped/imputed in all biobanks, among which 168,899 are also in the 1000 Genomes⁵ and HGDP.⁶ The weights corresponding to principal components for those markers were estimated based on the PCA analysis for the reference samples with known ancestry in 1000 Genomes and HGDP and shared among biobanks. Biobanks then generated PC loadings based on the pre-estimated weights of those markers.

Variant annotation

Genetic variants were annotated using ANNOVAR⁴⁷ for the nearest genes. To obtain a more complete annotation for putative loss-of-function variants, VEP⁴⁸ with the LOFTEE plug⁸ as implemented in Hail was used.

Heritability estimation

LD score regression analyses were conducted using LDSC⁴⁹ to estimate narrow-sense heritability based on the summary statistics of all-biobank meta-analyses based on the LD scores pre-estimated using UK Biobank samples (Table S9).

Evaluate the integration genetic associations from diverse biobanks

Compare individual biobank with LOBO meta-analysis

The effect sizes of top variants with association p-value $< 1 \times 10^{-10}$ by all-biobank meta-analyses in individual biobanks were compared to the effect sizes estimated in the corresponding LOBO meta-analyses. For each biobank and LOBO pair, we fit a Deming regression model,²⁴ which accounts for standard errors of effect size estimates in both association datasets, with the intercept set to zero (Figure S8). We estimated genetic correlation between individual biobanks and LOBO (Figure S9) for the three endpoints with highest heritability estimates: asthma, gout, and COPD, using LDSC.⁴⁹ For biobanks with samples of non-EUR ancestries, such as BBJ, we estimated the trans-ancestry genetic correlation estimation using Popcorn.⁵⁰

Compare GBMI with published GWAS studies

The effect size estimates of previously known loci in all-biobank meta-analyses were compared to the effect size estimated in previous GWAS (Table S15). This analysis was done for 18 loci that were previously identified by TAGC¹⁵ for asthma, 24 previously identified loci for AAA by MVP, and 40 previously identified loci for gout, respectively. For the 18 asthma loci, p-values by all-biobank meta-analyses in GBMI were compared those reported by TAGC¹⁵ (Figure S11).

Prioritize functional genes

DEPICT

Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT)³³ was applied to investigate the results from GWAS of 14 endpoints. DEPICT uses three analyses to predict the gene functions: 1) prioritizing the most likely causal genes, 2) identifying enriched gene sets, and 3) discovering tissues/cell types with highly expressed genes at associated loci (Tables S18, S19, S22, and S23 and Figures S12 and S13). Two p-value thresholds were used to define genome-wide significance 1×10^{-5} and 5×10^{-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 were 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 genes³⁴ (Tables S20 and S22, and S23 and Figures S12 and S13). 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 applies Multi-marker Analysis of GenoMic Annotation (MAGMA)⁵¹ to meta-analyze gene-level associations and create a 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 the 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. A PoPS score cutoff, the top 0.1 percentile, was also used in the gene prioritization method evaluation.

Transcriptome-wide association studies (TWASs)

Prediction of gene expression: Using genotypes and gene expression from 296 European donors from GTEx ver. 8,⁵² we trained predictive expression models using Joint-Tissue Imputation (JTI)⁵³ and Multi-Omic Strategies for TWAS (MOSTWAS).⁵⁴ 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.⁵³ 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 biobank meta-analysis summary statistics in GBMI from European-ancestry subjects, we detected gene-trait associations through the weighted burden test and 1000 Genomes Project CEU population as an LD reference^{5,55}. We defined a transcriptome-wide significance using a Bonferroni correction across 20,000 tests (p-value $< 2.5 \times 10^{-6}$)⁵⁵⁻⁵⁷ (Tables S21, S22, S23 and Figures S12 and S13). 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.⁵⁸ 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)

The putative causal role of 1,310 proteins on eight diseases in the NFE samples were estimated using proteome-wide association study (PWMR) and sensitivity analyses. For the exposure of the analysis, 5,418 conditional independent protein quantitative trait loci (pQTLs) of 1,310 proteins in European samples from ARIC⁵⁹ were selected as genetic predictors. For outcomes, eight of the 14 endpoints 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 approach⁶⁰ 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.⁶⁰ 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 Cochran's Q test for gIVW results and Rucker's Q test for gEgger results to estimate the potential heterogeneity of PWMR estimates^{61,62}. 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^{63,64}. Fourth, to control for potential aptamer binding artificial effects 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 filtering⁶⁵ 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 (Tables S22 and S23 and Figures S12 and S13).

Phenome-wide association test

For top variants at the 500 identified loci (known or novel), look-ups were carried out for their association with 1,283 human diseases curated based on phecodes mapped to ICD codes in the UK Biobank.⁴⁵ We reported associations with p-value < 5×10^{-8} .

Polygenic scores

The polygenic scores (PRS) were constructed using PRS-CS,⁶⁶ 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 matrices calculated using European individuals from 1000 Genomes Phase 3 (1KG) provided by PRS-CS were used. Specifically, we used leave-one-biobank-out (LOBO) meta-analysis in GBMI for asthma 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 built PRS using a published GWAS by TAGC.¹⁵ The prediction performance of PRS was estimated using Nagelkerke's R^2 after regressing out all biobank-specific covariates with a logistic regression. It was further transformed to R^2 on the liability scale,⁶⁷ with biobank-specific case proportion used as the disease population prevalence. The corresponding 95% confidence intervals (CIs) were calculated using bootstrap with 1000 replicates (Figure S14).