HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division of Medical Sciences

Biomedical Informatics

have examined a dissertation entitled

Fine-mapping complex traits in large-scale biobanks across diverse populations

presented by Masahiro Kanai

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: Soumya Raychaudhuri
Soumya Raychaudhuri (May 4, 2022 11:16 EDT)

Typed Name: Dr. Soumya Raychaudhuri

Signature: Joel Hirschhorn
Joel Hirschhorn (May 4, 2022 10:05 EDT)

Typed Name: Dr. Joel Hirschhorn

Signature: Po-Ru Loh
Po-Ru Loh (May 4, 2022 09:20 EDT)

Typed Name: Dr. Po-Ru Loh

Signature: Nancy J. Cox
Nancy J. Cox (May 4, 2022 09:14 CDT)

Typed Name: Dr. Nancy Cox

Date: May 02, 2022

Fine-mapping complex traits in large-scale biobanks across diverse populations

A DISSERTATION PRESENTED
BY
MASAHIRO KANAI
TO
THE DIVISION OF MEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOMEDICAL INFORMATICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2022

Thesis advisor: Professor Mark J. Daly and Hilary K. Finucane Masahiro Kanai

Fine-mapping complex traits in large-scale biobanks across diverse populations

Abstract

Identifying causal variants for complex traits is a major goal of human genetics research. Despite the great success of genome-wide association studies (GWAS) in locus discovery, individual causal variants in associated loci remain largely unresolved, limiting the biological inference possible from follow-up experimentation. In this dissertation, I present our fine-mapping analyses of complex traits in large-scale biobanks across diverse populations to create an atlas of causal variants.

We first fine-mapped complex traits using 361,194 European individuals from UK Biobank (UKBB) and gene expression using 49 tissues from GTEx (**Chapter 1**). We then extended our fine-mapping of complex traits to multiple populations, using 178,726 Japanese individuals from BioBank Japan and 271,341 Finnish individuals from FinnGen (**Chapter 2**). In total, we identified 4,518 variant-trait pairs with high posterior probability (> 0.9) of causality across the three populations. Aggregating data across populations enabled replication of 285 high-confidence variant-trait pairs as well as identification of 1,492 unique fine-mapped coding variants and 176 genes in which multiple independent coding variants influence the same trait. These results demonstrate that fine-mapping in diverse populations enables novel insights into the biology of complex traits by pinpointing high-confidence causal variants for further characterization.

Next, we investigated fine-mapping accuracy in GWAS meta-analysis (**Chapter 3**). We demonstrated that meta-analysis fine-mapping is substantially miscalibrated in simulations and proposed a novel quality-control method, SLALOM, that identifies suspicious loci for meta-analysis fine-mapping. Having validated SLALOM performance in simulations, we found widespread suspicious

patterns in existing GWAS significant loci that call into question fine-mapping accuracy. We thus urge extreme caution when interpreting fine-mapping results from meta-analysis.

Finally, we introduce a new polygenic risk score (PRS) method, PolyPred, that improves cross-population polygenic prediction by combining a new fine-mapping-based predictor and a published BOLT-LMM predictor (**Chapter 4**). Leveraging estimated causal effects from fine-mapping enabled higher PRS transferability in non-European populations, achieving up to +32% improvement in prediction accuracy vs. BOLT-LMM using UKBB Africans.

Altogether, this work demonstrates key advances in fine-mapping complex traits across diverse populations and provides insights into further variant characterization as well as improved polygenic prediction based on fine-mapping.

Contents

Listing of figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisors, Mark Daly and Hilary Finucane, for their dedicated mentorship throughout my PhD. Their guidance, encouragement, and extreme generosity helped me to develop a strong sense of scientific rigor and shaped me into the scientist I am today. I also feel incredibly lucky to be co-mentored by Mark and Hilary. In fact, I initially planned to rotate with Mark in my first year—however, by an incredible twist of fate, I happened to stop by Hilary's poster at ASHG 2017, and as we got to know each other as the first-year student and PI, we realized that the co-mentorship might be magically possible. The synergy of having two advisors was indeed magnificent for my dissertation and I am deeply indebted to them for the years of guidance.

Beyond incredible amount of guidance and resources that Mark and Hilary provided me, I was very privileged to have many additional mentors. I first met Yuki Okada in 2014 and joined the lab as his first student. Since then, Yuki has been a great mentor and collaborator in Japan, who originally taught me statistical genetics, encouraged me to study abroad, and continues to provide me resources and opportunities for collaboration. Ben Neale, Alkes Price, and Soumya Raychaudhuri, who served on my DAC, always provided me guidance and knowledge of statistical genetics. I am also particularly grateful to them for giving me an opportunity to work with their lab members,

in addition to my dissertation work, that helped me to expand my local network. Alicia Martin has inspired me in many ways, for her passion and effort to increase diversity in human genetics. I was incredibly fortunate to have worked with her in my early PhD for the PRS and health disparity project, and later for the Pan-UKBB project.

A PhD is a long journey. I am extremely proud of conquering it together with Jacob Ulirsch, by fine-mapping everything in our hands. Thanks so much for being a close buddy for this endeavor and inspiring me through endless scientific (and casual) conversations in Slack. Current and former PhD students in the ATGU have also been great buddies for this journey—Jack Kosmicki, Beryl Cummings, Qingbo Wang, Ryan Collins, Sherif Gerges, Rahul Gupta, Kristin Tsuo, Hannah Jacobs, and Layla Siraj. Thank you all for sharing the joy and fun with me in science and socials. I am also particularly owe Qingbo for his kind guidance during the PhD application and Jack for his initial assistance connecting me with Mark. Without their help, I could not even join the lab!

One of the greatest joy that I had during my PhD is interacting closely with many people within and beyond the ATGU. These include but not limited to the current and former members of the Finucane lab—Ran Cui, Roy Elzur, Nathan Cheng, Carlos Albors, Elle Weeks, and Katherine Tashman; the Daly lab—Wei Zhou, Amy Elliott, TJ Singh, Nikita Artomov, Kyle Satterstrom, and Henrike Heyne; and the Pan-UKBB team—Konrad Karczewski, Elizabeth Atkinson, Raymond Walters, Patrick Turley, Duncan Palmer, Ying Wang, and Nikolas Baya. I cannot recall how many times we regularly met at our weekly meetings, and am very grateful to their support and helpful feedback. I also had a privilege to closely follow the formation of two large-scale international consortia from scratch, the Global Biobank Meta-analysis Initiative (GBMI) and the COVID-19 Host Genetics Initiative (HGI). Having contributed to them as one of the central analysts and the writing group leads, respectively, I am especially grateful to Wei and Andrea Ganna for their leadership, in addition to the senior leadership provided by Mark and Ben.

Furthermore, I feel extremely privileged to closely work with outstanding scientists in two large-

scale biobanks, the BioBank Japan (BBJ) and FinnGen. I am grateful to the BBJ team—Koichi Matsuda, Michiaki Kubo, Yoichiro Kamatani, Saori Sakaue, Kazuyoshi Ishigaki, Masato Akiyama, and many other colleagues, as well as the FinnGen team—Aarno Palotie, Samuli Ripatti, Mitja Kurki, Juha Karjalainen, Arto Lehisto, Priit Palta, Juha Mehtonen, Mutaamba Maasha, and many other academia/pharma contributors, for their excellent science and incredible support to my projects. I am especially owe a beer to Mitja and Juha, and a Japanese sake to Saori, who analyze the data and drink beer/sake together with me. My warmest gratitude also extends to a beautiful city of Helsinki (as well as my home town Tokyo) which welcomes me with sauna and beer every time. I really hope to visit there soon to continue our close ties after the pandemic.

I would like to give a special thanks to the ATGU admin team—Jill Doucette, Carla Hammond, Elizabeth Raynard, and Autumn Taylor-Kelley, who manages our requests in the lab so smoothly; the Hail team—Cotton Seed, Daniel King, Tim Poterba, Jackie Goldstein, John Compitello, and many other contributors to the Hail package, who provides an essential tool for our everyday science; our consortium admins—Sinéad Chapman, Christine Stevens, and Amy Trankiem for their administrative support in the GBMI and the COVID-19 HGI; our data manager, Sam Bryant for his continuous monitoring of our Google Cloud usage; and the Nakajima Foundation and the Masason Foundation for their fellowships. Their everyday support has made my PhD life much easier.

Finally, I cannot express how thankful I am to Naoko Sawada and to have her in my life. While pursuing our PhDs separately in Japan and the U.S. had numerous challenges, I am extremely proud that we both managed it together, accompanied by many joys and excitements that we shared along the way. I am also deeply grateful to my parents, Yoshisada and Tamiko Kanai, for their continuous encouragement and unconditional support to pursue my dream in the U.S.

As already demonstrated, the work presented in this dissertation is only made possible by incredible team science. I am extremely grateful to numerous colleagues and collaborators who contributed

to each chapter, as shown below:

Chapter 1 Jacob C. Ulirsch*, **Masashiro Kanai***, Qingbo S. Wang, Roy Elzur, Ran Cui, Christian Benner, Ryan L. Collins, Elle M. Weeks, Steven Gazal, François Aguet, Zack R. McCaw, John Compitello, Daniel King, Juha Karjalainen, Caleb A. Lareau, Layla Siraj, Cotton Seed, Mitja Kurki, Steven K. Reilly, Kristin G. Ardlie, Benjamin M. Neale, Ryan Tewhey, Pardis C. Sabeti, Mark J. Daly, and Hilary K. Finucane.

Chapter 2 **Masahiro Kanai**, Jacob C. Ulirsch, Juha Karjalainen, Mitja Kurki, Konrad J. Karczewski, Eric B. Fauman, Qingbo S. Wang, Hannah Jacobs, François Aguet, Kristin G. Ardlie, Nurlan Kerimov, Kaur Alasoo, Christian Benner, Kazuyoshi Ishigaki, Saori Sakaue, Steven Reilly, The BioBank Japan Project, FinnGen, Yoichiro Kamatani, Koichi Matsuda, Aarno Palotie, Benjamin M. Neale, Ryan Tewhey, Pardis C. Sabeti, Yukinori Okada, Mark J. Daly, and Hilary K. Finucane.

Chapter 3 **Masahiro Kanai**, Roy Elzur, Wei Zhou, Global Biobank Meta-analysis Initiative, Mark J. Daly, and Hilary K. Finucane.

Chapter 4 Omer Weissbrod*, **Masahiro Kanai***, Huwenbo Shi*, Steven Gazal, Wouter J. Peyrot, Amit V. Khera, Yukinori Okada, The Biobank Japan Project, Alicia R. Martin, Hilary K. Finucane and Alkes L. Price.

(* denotes equal contribution)

0

Introduction

Overview

A primary goal of human genetics is to determine the genetic causes of rare and common diseases. Identifying genetic causal variants provides the key mechanistic insights into disease biology, facilitates the diagnosis of individual patients, and may lead to the development of new therapeutic treatments. While many human diseases are heritable[1], each disease has a different degree of heritability (phenotypic variance that is due to genotypic variance in a population) and causal genetic architecture (from monogenic to polygenic, as explained by the number of causal loci)[1–3]. Rare severe diseases are in particular caused by a single or few causal loci with large effects[4], allowing the successful identification of risk loci via linkage mapping in a small number of families (*e.g.*, Huntington's disease and the CAG repeats in *HTT*[5]; cystic fibrosis and ΔF508 in *CFTR*[6]). On the other hand, common complex diseases (*e.g.*, type 2 diabetes[7,8], rheumatoid arthritis[9,10], and schizophrenia[11]) are often caused by numerous variants spread across the genome, each of which has a small (but nonzero) effect on a phenotype (*i.e.*, polygenicity)[3]. The polygenicity of common complex diseases make it particularly challenging to identify causal loci via linkage mapping, requiring a large-scale genetic association study to identify associated loci and a follow-up fine-mapping analysis to determine individual causal variants[12–14]. In this dissertation, I present the work to fine-map complex traits in large-scale biobanks across diverse populations.

Genome-wide association study (GWAS)

Genome-wide association studies (GWAS) aim to identify associations between genetypes and phenotypes of interest using a large number of individuals[15]. GWAS typically test hundreds of thousands of genetic variants (*e.g.*, single nucleotide variants, insertion/deletion, copy number variants, etc.) to find those statistically associated with a phenotype. Since each causal variant only has a small effect, GWAS requires thousands of samples to achieve a statistical significance after multiple testing

correction (typically $P < 5.0 \times 10^{-8}$)[16,17], and thus early GWAS had limited power to detect associated loci. However, as genomic profiling technologies have progressed (*e.g.*, genotyping microarray and next-generation sequencing), GWAS have successfully identified thousands of loci associated with complex traits—according to the GWAS Catalog[18], more than 372,000 locus-trait associations have been reported to date, including many well-established risk loci, such as *FTO* for obesity[19] and *TCF7L2* for type 2 diabetes[20], as well as risk loci for emerging infectious diseases, such as the 3p21.31 locus for COVID-19 severity[21].

By design, GWAS highly rely on linkage disequilibrium (LD, the correlation among genetic variants)[22] to identify an associated locus that contains causal variant(s) tagged by genotyped variants. While LD made early GWAS possible by allowing a study to use only "marker" variants, it also inevitably prevent us to pinpoint individual causal variants from the correlated variants. Even today, when dense genotype imputation and whole genome sequencing allow direct genotyping of the majority of causal variants, high LD among the associated variants limits the identification of causal variants and the subsequent biological inference possible from follow-up experimentation.

Statistical fine-mapping

To disentangle LD from the correlated GWAS associations, a follow-up statistical analysis, known as statistical fine-mapping, is employed to prioritize individual causal variants[13]. While multiple methods have been proposed (*e.g.*, approximate Bayes factor [ABF][23,24], CAVIAR[25], PAINTOR[26,27], FINEMAP[28,29], and SuSiE[30]), the current state-of-the-art are primarily Bayesian methods that essentially compare the evidence of association vs. prior expectations and provide posterior inclusion probability (PIP) for each variant as well as a set of variants that accounts for a certain probability of causality (typically 95% credible set). As individual methods differ by models (*e.g.*, single vs. multiple causal variants; prior effect size distribution) and algorithms (*e.g.*, exhaustive vs. shotgun stochas-

tic search), their fine-mapping calibration, recall, and computational efficiency vary substantially. Recent development of scalable fine-mapping methods (FINEMAP[28,29] and SuSiE[30]) finally enables biobank-scale fine-mapping of hundreds of phenotypes with high confidence, which provides a foundation of this dissertation.

Polygenic risk score (PRS)

It is worth noting that fine-mapping has direct relevance to polygenic risk scores (PRS). PRS is a genetic predictor that estimates individual's risk for complex diseases and traits[31–33]. As its name suggests, PRS typically combines hundreds to thousands of (small) genetic effects to derive a single predictor, which is often trained on large-scale European GWAS. Similar to PRS weights produced by existing PRS methods (*e.g.*, LD pruning + *P*-value thresholding [P+T][34], LDpred[35], SBayesR[36], and PRS-CS[37]), fine-mapping results provide posterior (causal) effect sizes that can be used as PRS weights and may improve prediction accuracy.

Despite the limited predictive power in early studies[34], PRS of a certain diseases (*e.g.*, breast cancer[38], prostate cancer[39], type 1 diabetes[40]) already demonstrated higher prediction accuracy than the current clinical risk factors in Europeans. However, many studies[41–60] (including our work[47] that I contributed during my PhD) have shown that PRS trained on European GWAS has limited accuracy in non-European populations (*i.e.*, low transferability). This loss of accuracy is driven by many factors (*e.g.*, differences in LD[46–49], allele frequencies[47,48,61], causal effect sizes[46–48,62–65], and heritabilities[47,48,66])—however, it is partially addressable by using the fine-mapping posterior effect sizes which have already disentangled LD differences. This motivated us to the development of a new fine-mapping based PRS predictor in this dissertation.

This dissertation consists of four chapters. In **Chapter 1**, I describe our fine-mapping analysis of complex traits using 361,194 Europeans from UK Biobank (UKBB) and gene expression using 49 tissues from GTEx[67]. By systematically evaluating fine-mapping methods in simulations, we propose the best practice of fine-mapping complex traits and gene expression (*e.g.*, quality-control, the use of [covariate-adjusted] in-sample LD, and the choice of fine-mapping methods), and apply it to UKBB and GTEx. We then conduct a series of analyses to nominate disease relevance and/or molecular mechanisms underlying fine-mapped variants using a phenome-wide association study in GWAS Catalog, cis-eQTL colocalization, and both experimental and machine learning predictions of accessible chromatin and transcription factor occupancy.

In **Chapter 2**, I then describe our extended analysis of complex traits fine-mapping in multiple populations[68], using additional 178,726 Japanese individuals from BioBank Japan (BBJ) and 271,341 Finnish individuals from FinnGen. Taken together, we identify thousands of variant-trait pairs with high posterior probability (> 0.9) of causality across the three populations, which allows us, for the first time, the comparison and replication of fine-mapping results across three large-scale independent cohorts. We then characterize multiple contributors to both successful and failed replication of fine-mapped variants across biobanks. Focusing on coding variants, we further demonstrate identification of tens of putative causal coding variants with extreme allele frequency enrichment (> 10-fold) in the Japanese and Finnish populations. Aggregating both common and population-enriched coding variants across populations enables us to identify hundreds of genes with an allelic series, which demonstrates the significant value of diverse populations in fine-mapping studies.

In **Chapter 3**, I describe our analyses of fine-mapping accuracy in GWAS meta-analysis[69]. Unlike GWAS in biobanks, GWAS meta-analysis consists of cohorts that are heterogeneous in many

ways (*e.g.*, ancestry, sample size, phenotyping, genotyping, or imputation) and it is unclear how these characteristics affect fine-mapping calibration and recall. Using systematic simulations, we first demonstrate that meta-analysis fine-mapping is substantially miscalibrated when different genotyping arrays and imputaion panels are included. We then propose a novel quality-control method, SLALOM, that identifies suspicious loci for meta-analysis fine-mapping. We validate SLALOM performance in simulations, and show widespread suspicious patterns in the GWAS Catalog as well as the GBMI summary statistics that call into question fine-mapping accuracy.

In **Chapter 4**, I describe a new polygenic risk score (PRS) method, PolyPred, that improves cross-population polygenic prediction based on fine-mapping results in the European population[70]. PolyPred leverages estimated posterior (causal) effect sizes from fine-mapping in addition to a published polygenic predictor (*e.g.*, BOLT-LMM, SBayesR, and PRS-CS), which addresses low cross-population transferability of PRS due to different LD structure. When large-scale training samples are available in non-European populations, we introduces a method PolyPred+ which further combines a published polygenic predictor from the non-European training data. We apply PolyPred and PolyPred+ to complex traits from UKBB, BBJ, and Uganda-APCDR cohorts and demonstrate significant improvement in prediction accuracy.

In summary, the work presented in this dissertation demonstrates key advances in fine-mapping complex traits across diverse populations. The insights, resources, and methods generated here facilitate future application and interpretation of biobank-scale fine-mapping, and provide a guide for further functional characterization efforts as well as improved polygenic prediction.

The work presented in this chapter will be published as

Ulirsch, JC.* & Kanai, M.* *et al.*[67]

1

An annotated atlas of causal variants underlying complex traits and gene expression

Abstract

Genome-wide association studies have successfully identified thousands of genomic loci associated with human traits and diseases[18,71], but the delineation of causal variants and their mechanisms has lagged sorely behind[72]. To advance our understanding of these loci, we systematically evaluated and applied genetic fine-mapping algorithms[25,28–30,73–75] to 96 complex traits from the UK Biobank[76], narrowing in on 2,519 likely causal variants (posterior inclusion probability > 0.9). Likely causal variants were 2.4-fold more likely to alter protein-coding DNA or lie within regulatory genomic regions and often affected multiple traits. Colocalization[77–79] with an improved atlas of fine-mapped eQTLs refined our understanding of the genomic mechanisms of complex trait variants but provided modest identification of causal genes. Combining bulk and single cell maps[80–86] of accessible chromatin and active histone modifications across diverse cell types and states, we determined that 47% of likely causal non-coding variants lie in biochemically supported *cis* regulatory elements (CREs), many of which have trait-relevant tissue-specific regulatory capabilities[87]. Integrating predictive models[88,89] with comprehensive maps of transcription factor (TF) occupancy[84,90,91] and chromatin accessibility, we nominate molecular mechanisms for 61% of colocalized CRE single nucleotide variants with a background rate of 11%. Of note, only 14% of colocalized CRE variants disrupt the canonical motif of an occupying TF. In total, we provide a modern annotated atlas of putative causal regulatory variants underlying complex human traits and use this to nominate likely causal variants underlying complex diseases[92].

1.1 Introduction

Hundreds of thousands of genetic loci associated with complex human traits and diseases have been identified at a rapid pace since the first genome-wide association studies (GWAS) in the early 2000s [18,71,93]. Moving from associated locus to biological understanding has advanced far more incrementally: the full path from causal variant to physiological function has been resolved for a tiny fraction of significant loci [12,72], the majority of which contain protein-coding variants.

Filling in these causal pathways requires overcoming a number of challenges [72]. First, associated variants in a locus are often inherited together, making it difficult to identify the causal variant(s) due to correlation known as linkage disequilibrium (LD) [94]. Second, most associated LD blocks do not contain a variant that would alter protein structure or function; instead, approximately 80% of variants [95,96] appear to act in non-genic regions, likely to regulate gene expression and function [97,98]. Although much progress has been made in understanding this functional architecture across the genome [80,85,95,99,100], a lack of unambiguous functional consequences for non-coding variants often precludes confidently identifying the molecular mechanisms and cellular contexts by which individual variants act.

An increasingly common first step when probing variant-to-function pathways is to quantitatively disentangle the effects of LD and association strength at a locus using genetic fine-mapping [13,73,74,87]. Bayesian fine-mapping methods [25,28–30,73–75] estimate the probability that a variant is causal for a trait (**Fig. 1.1a**), typically referred to as a posterior inclusion probability (PIP), and finds the smallest set of variants in LD that contains the causal variant with 95% probability, known as the 95% credible set (CS). Similarly, colocalization methods estimate the joint probability that a variant is causal for a trait and a molecular quantitative trait loci (QTL). With current sample sizes, these methods typically do not fully resolve loci but often narrow down to a handful of genetic variants which can then be combined systematically with genomic annotations or used in functional characterization

studies to discover the regulatory mechanisms, genes, and cell types underlying individual trait associations[87,101].

In this work, we apply state-of-the-art fine-mapping methods, optimized in large-scale simulations, to 96 complex traits and diseases in the UK Biobank (UKB)[76]. Across traits, we observe widespread pleiotropy[102] and show how fine-mapping quantitative traits can inform putative causal variant identification in complex diseases. In order to increase our confidence in specific non-coding causal variants, we perform colocalization[77–79] for gene expression changes across 49 tissues from the Genotype-Tissue Expression (GTEx) project[103], carefully considering the effects of population stratification[104–106] and priors[77,78]. While we observe increased confidence in colocalized variants, we determine that the complex regulatory capabilities of genetic variants in part confounds the use of colocalization as a gene prioritization tool. Finally, we annotate non-coding fine-mapped and colocalized variants for molecular function using canonical motifs[88], (allele-specific) transcription factor (TF) occupancy[84,90], (allele-specific) DNase footprinting[91], and neural network predictions of TF and chromatin changes[89]. In total, our study provides a contemporary annotated atlas[99,102] of the genomic functions of likely causal complex trait and disease variants.

1.2 Results

1.2.1 Identification of thousands of likely causal variants for complex human traits and diseases

We selected 96 well-powered complex traits and diseases across 10 phenotypic domains[102] from the UK Biobank[76] (UKB) for inclusion in our GWAS and fine-mapping study (**Fig. 1.1b,c**, **Supplementary Fig. A.2,A.3**, **Supplementary Tables A.1–A.3**). We performed association studies for well-imputed (INFO[107] > 0.8) common and low frequency variants (MAF > 0.001; MAF > 10^{-6} for coding variants) in up to 361,194 unrelated individuals in the previously-defined "white British"

cohort using generalized linear mixed models[108,109]. In total, we identified 11,745 genome-wide significant loci ranging from 3 to 22 Mb in which to perform fine-mapping.

To evaluate different method and parameter choices for fine-mapping, we simulated 50 realistic biobank-scale traits, drawing true genotypes from the observed genotype probability distribution of the "white British" subset of the UKB dataset (n = 361,194) and using causal variant density[110], minor allele frequency (MAF)-dependent causal effect sizes[111], and total SNP-heritability[112] consistent with the architecture of a typical complex trait (see **1.4 Methods**). Our simulations support the use of in-sample imputed dosage genotypes for both association and LD rather than hard-called genotypes or reference panel LD[87,113], generally less restrictive MAF and imputation quality thresholds to capture more causal variants, relatively large 1–3 Mb windows to better model LD between causal and tag variants, and methods that jointly model multiple causal variants (MCVs) like FINEMAP and SuSiE rather than conditional[1147] or single causal variant[24] methods (**Supplementary Fig. A.1**). Our simulations also enabled us to explore a few additional scenarios in detail, including the calibration of fine-mapping for poorly imputed causal variants, the particular importance of in-sample LD for rare variants, and the basis for improved power in MCV fine-mapping (**Supplementary Fig. A.1**).

We performed fine-mapping using this optimized approach, identifying 3,785 variant-trait pairs (2,519 unique variants) with a posterior inclusion probability (PIP) > 0.90 and 25,174 independent 95% CS-trait pairs (min r^2 > 0.25 in CS, SuSiE only; 15,103 merged CSs [MCSs]; see **1.4 Methods**) with a median size of 12 variants (**Fig. 1.1d,e**, **Supplementary Table A.4**). Consistent with our simulations, FINEMAP and SuSiE PIPs of real complex traits were very well correlated (**Fig. 1.1f**) but restricting to likely causal variants identified by both improves confidence in their identification (**Supplementary Fig. A.3**). These results provide improvements over previous large-scale fine-mapping studies[102,115] and a comprehensive baseline for future comparisons.

Our fine-mapped variants showed large and significant enrichments in several functional cat-

egories, with 63% of likely causal variants (PIP > 0.9) annotated as coding (LoF, missense, or synonymous), 3' or 5' UTR, or regulatory (promoter or *cis*-regulatory element [CRE]; **Fig. 1.1g**). The non-genic set of variants (*i.e.*, not included in any of the above categories) is enriched for variants that are evolutionarily conserved[95,116] (**Supplementary Fig. A.3**), suggesting that many of these variants may act through regulatory elements in other cell-types or states, or through regulatory mechanisms not captured by these annotations. Because our fine-mapping pipeline is agnostic to functional annotations, these functional enrichments serve as orthogonal validation of our results. Approaches that estimate the contribution of variants in functional genomic categories have shown that certain categories are enriched for SNP heritability (h^2_g)[95,97]. Across a set of 40 independent traits and 39 genomic annotations (see **1.4 Methods**), we observed that genome-wide polygenic enrichments are generally concordant with fine-mapped variant enrichments (**Supplementary Fig. A.3**, **Supplementary Table A.5**).

Finally, we explore several notable observations from fine-mapping. First, we find a small number of variants with PIP > 0.5 that do not meet genome-wide significance thresholds; functional genomic enrichment suggests that a subset of these variants are truly causal (**Supplementary Fig. A.3**, **Supplementary Table A.6**). Second, we find examples where the posterior variant effect sizes are in the opposite direction than expected based upon marginal effect sizes; these variants likely represent a mix of real LD masked effects as well as limitations of current fine-mapping approaches (**Supplementary Fig. A.3**, **Supplementary Table A.7**). Third, consistent with the simulation results showing the importance of jointly modeling multiple causal variants, we observed that 49% of regions contained multiple CSs and 12% of all CSs physically overlapped with another CS for the same trait (**Fig. 1.1h**). Moreover, we found non-trivial (r^2 > 0.1) LD between 5% of CSs for the same trait (see **1.4 Methods**) and predict that jointly modeling multiple causal variants when fine-mapping will become even more necessary for pinpointing likely causal variants as sample size increases (**Supplementary Fig. A.3**).

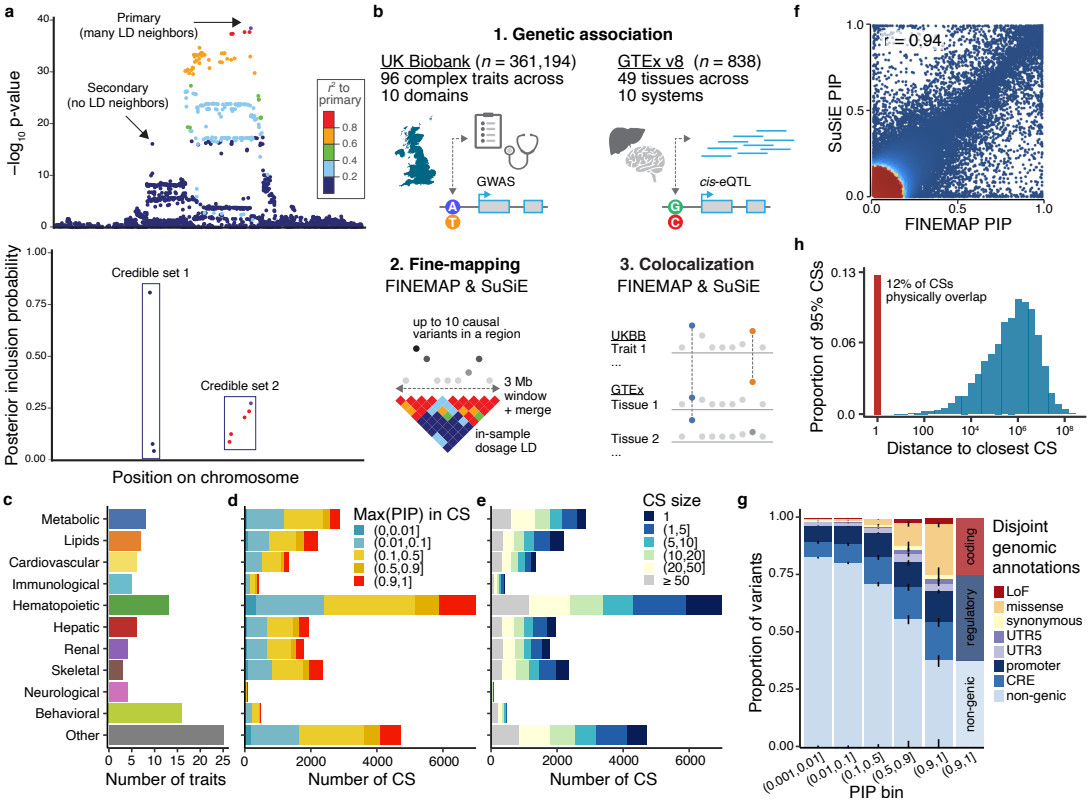


Figure 1.1: Genetic fine-mapping of complex traits in UK Biobank. **a**. Hypothetical example of fine-mapping starting from a locus zoom marginal association plot (top) to PIPs and independent 95% CSs (bottom). **b**. Overview of the fine-mapping and colocalization studies across UK biobank phenotypes and GTEx tissues. **c**. Number of traits in each phenotypic domain. **d**,**e**. Number of 95% CSs across traits in each domain colored by the highest PIP of a variant in the CS (**d**) and the size of the CS (**e**). **f**. Scatter plot and Pearson correlation of PIPs from FINEMAP and SuSiE methods. Red indicates higher density of variants, and blue indicates lower. **g**. Overlap of fine-mapped variants in each PIP bin and 8 disjoint genomic annotations. The max PIP for each variant across traits is used. Error bars represent 95% CIs. A column summarizing these annotations into coding, regulatory and non-genic is also shown. **h**. Physical distance between 95% CS defined as the distance between their closest variants.

1.2.2 Extensive pleiotropy across complex traits and diseases

We hypothesized that our large-scale fine-mapping study would allow for better identification of the shared causal variant effects[102,117,118] across complex human traits. To this end, we identified 6,394 unique variants that were fine-mapped (PIP > 0.1) for at least two traits, comprising 5,039 unique MCSs (**Supplementary Table A.8**). These pleiotropic variants were most commonly observed within a single phenotypic domain, but 1,323 had effects across more than 3 domains (**Fig. 1.2a**). We estimated genetic correlation (r_g)[119,120], a measure of the genome-wide sharing of polygenic effects, for all trait pairs in our study. With these estimates, we found that 45% of pleiotropic variants and 55% of pleiotropic MCSs were observed across predominately independent traits (all pairwise $|r_g| < 0.2$), highlighting the widespread re-use of small-effect common variants across the phenotypic landscape[12].

To gain insight into the mechanisms of pleiotropic variants, we next investigated the thirty-four fine-mapped variants that demonstrated extensive pleiotropy across over 10 phenotypic domains (**Fig. 1.2c**). These include well-described missense variants in the gene encoding for GCKR, which regulates glucose metabolism in hepatocytes[121]; SH2B3, which regulates cytokine signaling broadly[122]; APOE, which transports cholesterol in the liver and brain[123]; and ADH1B, which metabolizes alcohol and other substrates[124]. Other extremely pleiotropic variants appeared to act via gene regulation, such as rs998584, which is associated with multiple skeletal traits such as body fat percentage, in addition to lipid and blood cell traits. Consistent with these associations, rs998584 lies downstream of *VEGFA*, which encodes for the primary regulator of angiogenesis[125], in a regulatory element that is predominately accessible in blood and connective tissue cell-types (**Supplementary Fig. A.4**). Another example is rs76895963, which is a likely causal variant for 18 traits across 6 domains (PIP > 0.9 for all). This variant lies within a regulatory element in the first intron of *CCND2*, a key regulator of the cell cycle[126], which is partially accessible across a broad range

of tissues during development (**Supplementary Fig. A.4**).

Although pairs of traits with higher genome-wide polygenic correlation tended to have a higher proportion of pleiotropic variants with shared effect directions (**Fig. 1.2b**), we observed 44 variants where the variant effect directions disagreed with the expected direction based upon r_g for genetically similar ($|r_g| > 0.5$) traits (**Supplementary Table A.9**). For example, we found unexpected effect direction disagreement at rs1047891 (a missense variant in *CPS1*) for serum creatinine-based and cystatin C-based eGFR [127], rs76895963 (see above) for T2D and BW [128], and rs2740488 (an intronic variant to *ABCA1*) for HDLC and TG [129,130]. Together, these results suggest that functional variants are typically repurposed to have similar effects in related traits but can on occasion have opposite effects in highly correlated traits.

Finally, we investigated whether the widespread pleiotropy that we observed could allow us to make inferences about causal variants in diseases not well represented in the UKB, a relatively healthy cohort, since fine-mapping is particularly difficult in well-powered meta-analytic studies. To these ends, we conducted a phenome-wide association study (PheWAS) of all fine-mapped (PIP > 0.1) variants within our UK Biobank study (see **1.4 Methods**) across 1,183 unique traits from 3,079 previous studies (collected by OpenTargets [92], see **1.4 Methods**).

For 169 distinct diseases and disease-relevant traits, we identified at least one genome-wide significant hit (lead and tagging variants with $r^2 > 0.7$, $P < 5.0 \times 10^{-8}$) that was also fine-mapped in our UKB analysis (**Supplementary Table A.10**). Specifically, this approach allowed us to identify fine-mapped variants for 220 independent loci for coronary artery disease, 190 independent loci for autoimmune disorders such as Crohn's diseases and Psoriasis, and 135 independent loci for Schizophrenia, in addition to many other complex diseases (**Fig. 1.2e**). For example, we fine-map rs2076295, a regulatory variant of *DSP* in lung epithelial cells [131], for FEV1/FVC ratio and find that this reaches genome-wide significance for COPD. We also fine-map several variants, including rs71368508, for eosinophil count that are genome-wide significant for Eczema [132]. We conclude

that combining systematic fine-mapping with large meta-analytic efforts compellingly nominates potential causal variants underlying complex diseases.

1.2.3 Colocalization of complex trait variants and expression-associated loci

To gain insight into the molecular effects and gene targets of fine-mapped complex trait variants, we integrated our fine-mapping results with expression quantitative trait loci (eQTLs) using colocalization[79] analysis. Rather than simply overlapping these associations, colocalization models the effects of LD and provides estimates of the joint probability that a variant or set of variants causally influences both a complex and molecular trait[115]. Based upon recent simulations[78,133], we reasoned that our improved multiple causal variant (MCV) complex trait fine-mapping, coupled with MCV eQTL fine-mapping, would improve colocalization[78], which is most commonly performed assuming at most one causal variant[79] or with incompatible reference LD[115]. Although MCV fine-mapping is available for gene-tissue pairs across 49 tissues in GTEx v8[103] using other methods[25,134,135], they showed extensive disagreement[103]. Thus, we reasoned that applying FINEMAP and SuSiE, which we validated in complex trait simulations, may better resolve expression trait variants in GTEx v8.

To account for potential confounding due to uncorrected population stratification[105] in fine-mapping this ancestrally heterogeneous cohort[103], we used covariate-adjusted (cov-adj) LD, analogous to an approach we developed for heritability estimation in heterogeneous cohorts[136]. We show that using cov-adj LD is theoretically justified[104,136] (see **1.4 Methods**), that it leads to the identification of variants with greater functional enrichments[103,137] in the very heterogeneous GTEx cohort (**Fig. 1.3c**, **Supplementary Fig. A.5**), and that it has little impact in the much more homogeneous "white British" subset of the UKB (**Supplementary Fig. A.5**). Thus, we suggest that future fine-mapping studies of heterogeneous cohorts use cov-adj genotypes or cov-adj LD.

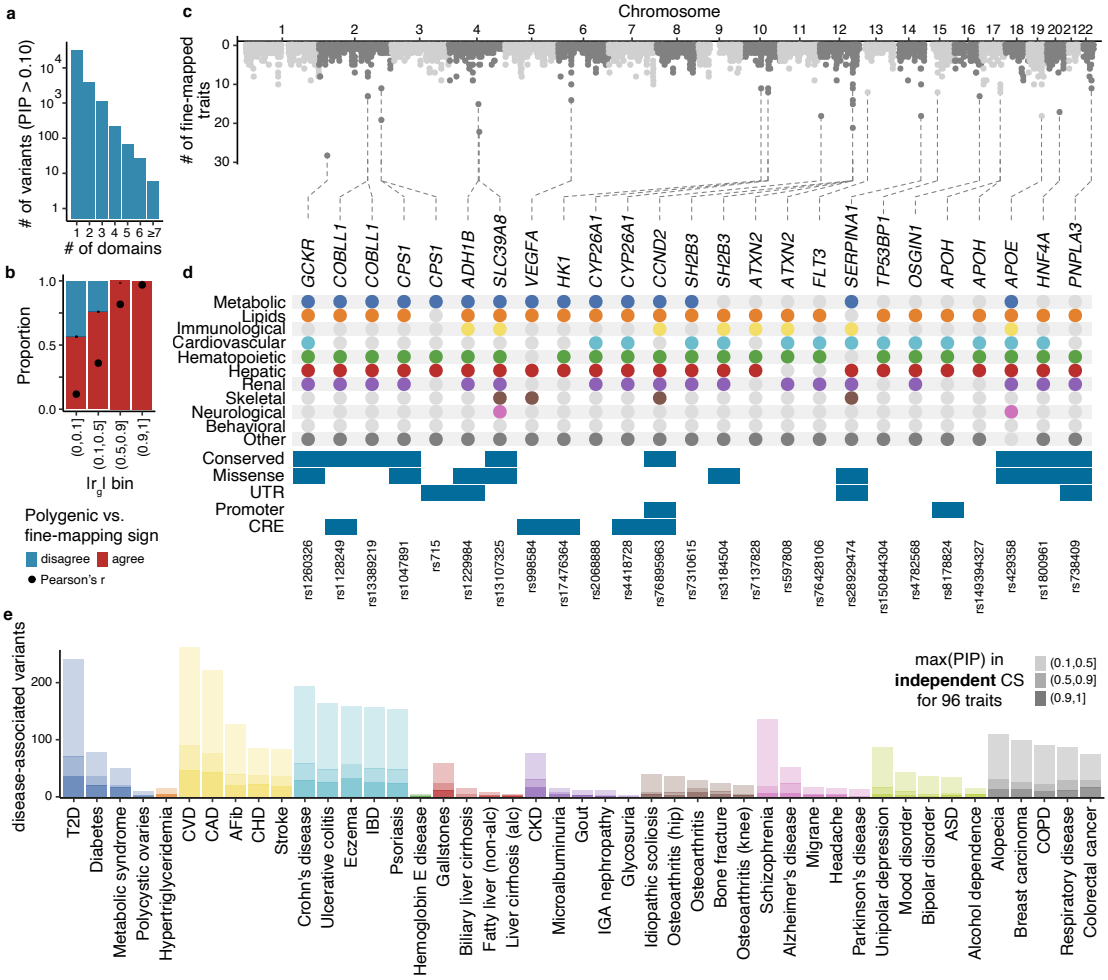


Figure 1.2: Widespread pleiotropy of candidate causal variants. **a**. Number of fine-mapped variants shared across phenotypic domains. **b**. Comparison of effect size agreement for fine-mapped (PIP > 0.1) variants and polygenic expectation for trait pairs. Stacked bar plots and 95% CIs of the proportion of variants where effect size direction of the trait pairs agree or disagree across sign of the polygenic correlation (r_g), binned by $|r_g|$. Points and 95% CIs indicate Pearson correlation for fine-mapped variant posterior effect sizes in each bin. **c**. A Manhattan-like plot of fine-mapped (PIP > 0.1) variants for the number of pleiotropic traits. **d**. The top 25 most pleiotropic variants are highlighted and annotated using phenotypic domains and the same genomic annotations as in **Fig. 1.1h**. **e**. The number of variants and max PIP in UKB for those variants that are also genome-wide significant ($P < 5.0 \times 10^{-8}$) sentinel variants or in LD with sentinel variants in a selected non-UKB disease. To prevent double counting, we include only one overlap per merged 95% CS across the 96 traits in our study.

Having validated our use of cov-adj LD, we applied SuSiE and FINEMAP and discovered 303,906 95% CSs for 25,005 unique protein coding genes and lncRNAs with a median 95% CS size of 8 (interquartile range: 2–22). Nearly a quarter (21%) of 95% SuSiE CSs contained a fine-mapped (PIP > 0.9) variant, resulting in 19,410 distinct, putative causal variants underlying differential gene expression (**Fig. 1.3b**, **Supplementary Table A.11**). Up to one-third (34% for Tibial Nerve) of tissue-gene pairs harbored more than one 95% CS. When compared to other methods[24,25,135], MCV methods that control for covariates properly[28,30,134] identify more fine-mapped variants, and these variants are almost always better enriched for relevant genomic annotations (**Supplementary Fig. A.5**)[138].

Having successfully performed and validated our MCV fine-mapping in both UKBB and GTEx v8, we turned to colocalization. We performed colocalization using both a conservative, non-informative eCAVIAR prior[77], assuming that there is no excess overlap between causal variants for complex and molecular traits, and a more powerful, non-independence prior that we estimated for each trait-tissue pair using fastENLOC[78]. We observe that fastENLOC often finds that the most relevant tissues are amongst the most enriched for each trait (**Supplementary Fig. A.6**, **Supplementary Table A.12**). Overall, accounting for non-independence leads to a 3.2-fold increase in colocalized trait-gene pairs at the 95% CS level, accounting for MCVs in complex traits leads to a 1.6-fold increase, and accounting for MCVs in eQTL leads to a 1.1-fold increase (**Fig. 1.3e**). In summary, we observe that allowing for MCVs and non-independence for both molecular and complex trait fine-mapping together identifies the largest number of colocalized trait-gene pairs (**Fig. 1.3e**).

Next, we performed functional enrichment analyses on likely causal complex trait variants, likely causal eQTLs, and high confidence colocalized variants (variants with a colocalized posterior probability [CLPP] > 0.9 of being causal for both a complex and molecular trait; see **1.4 Methods**). Previous studies have highlighted key differences between the genomic localization of complex trait and eQTL variants[99]; consistent with this previous work, we find that fine-mapped eQTLs are more enriched in 5' and 3' UTRs, promoters, and synonymous variants, while fine-mapped complex

trait variants are more enriched for LoF, missense, and CRE variants (**Fig. 1.3f**, **Supplementary Fig. A.6**, **Supplementary Table A.13**). We also find that high confidence colocalized variants are more likely to fall within a functional genomic annotation than likely causal complex trait or eQTL variants alone (79% vs. 63% and 42% for colocalization vs. UKBB and GTEx, respectively), including being more enriched for LoF, missense, 5' and 3' UTRs, and distal CREs, (**Supplementary Fig. A.6**). This suggests that the set of colocalized variants is of particularly high confidence, and we turn our focus to characterizing these high confidence variants and assessing their use for gene prioritization.

First, we investigate several factors affecting the accuracy of colocalization for causal gene prioritization. We observe that genetic regulation of expression in *cis* is often complex: 19% (181 / 914) of precisely colocalized (CLPP > 0.9) variant-trait pairs and 31% (2118 / 6792) of colocalized (CLPP > 0.1 for at least one variant) 95% CSs nominate two or more distinct gene products (**Fig. 1.3g**, **Supplementary Fig. A.6**, **Supplementary Table A.14**,A.15). When a variant colocalizes across multiple tissues, we find that 6% of the time (14 / 218) the marginal direction of effect is inconsistent for at least one tissue. As an example of complex genetic regulation, we find that rs12740374 colocalizes with *SORT1*, *PSRC1*, and *CELSR2* for multiple lipid traits (all CLPP > 0.9), although only *SORT1* has been shown experimentally to modulate cholesterol levels[139]. In contrast, we observe that the hepatic control region variant[140], rs35136575, colocalizes (all CLPP > 0.9) with 3 *cis* apoliproteins for LDL-C and ApoB levels: *APOE*, *APOC1*, and *APOC2*. In this case, it is unclear whether one or more genes lie on the causal pathway, although evidence is emerging of loci where regulation of multiple genes in *cis* is important[141]. In another example, we find that the variant rs8012, fine-mapped for mean corpuscular volume, lies in the 3' UTR of *GCDH* on the forward strand but also in the last intron of *SYCE2* on the reverse strand, colocalizing with changes in the expression of both. The exact mechanism by which this variant affects mRNA expression here is unclear: it could be direct transcriptional regulation or it could affect how RNA polymerase processes

these overlapping transcripts. Finally, we observe multiple missense mutations that colocalize with other genes. Of particular note, we find that rs760077, a missense variant in *MTX1* fine-mapped for RBC indices, Urea, and eGFR levels, colocalizes with nearby gene *THBS3*. This variant may regulate *THBS3* transcription directly, as it also lies in accessible chromatin, or indirectly, through a more complex mechanism.

To quantify the extent to which colocalization could be used for causal gene prioritization, we restricted our analyses to non-coding 95% eQTL CSs in the physical neighborhood (< 500 kb) of validation genes with fine-mapped (PIP > 0.50) coding variants and asked what proportion of our colocalizations identified the validation gene (precision) and for how many distinct complex trait signals did colocalization identify the validation gene (recall)[142]. Across methods, we found that precision was moderate (range: 0.37–0.59) while recall was relatively poor (range: 0.04–0.10). Allowing for MCVs improved both, and while the fastENLOC prior improved recall with a loss in precision, we were able to achieve similar improvements in recall with smaller losses in precision simply by using a more lenient threshold on the eCAVIAR results. This likely reflects the identification of multiple genes in validation loci (**Fig. 1.3h**, **Supplementary Fig. A.6**, **Supplementary Table A.16**). Variants of *F* statistics allow researchers to quantify precision/recall tradeoff and optimise decision-making; for example, when precision is 3-fold more valued than recall, such as for scalable GWAS experimental follow-up, our analyses suggest that using CLPP > 0.01 with an eCAVIAR prior or CLPP > 0.1 with a fastENLOC prior are preferred ($F_{0.33} = 0.34$ and 0.32, respectively). Taken together, our results suggest that current colocalization methods can nominate genes with moderate confidence at a small fraction of loci. We anticipate that increased power due to better priors, larger sample sizes, and new tissues and cell types will increase the number of colocalizations detected, increasing both the number of disease-relevant genes identified and the number of non-disease-relevant genes identified.

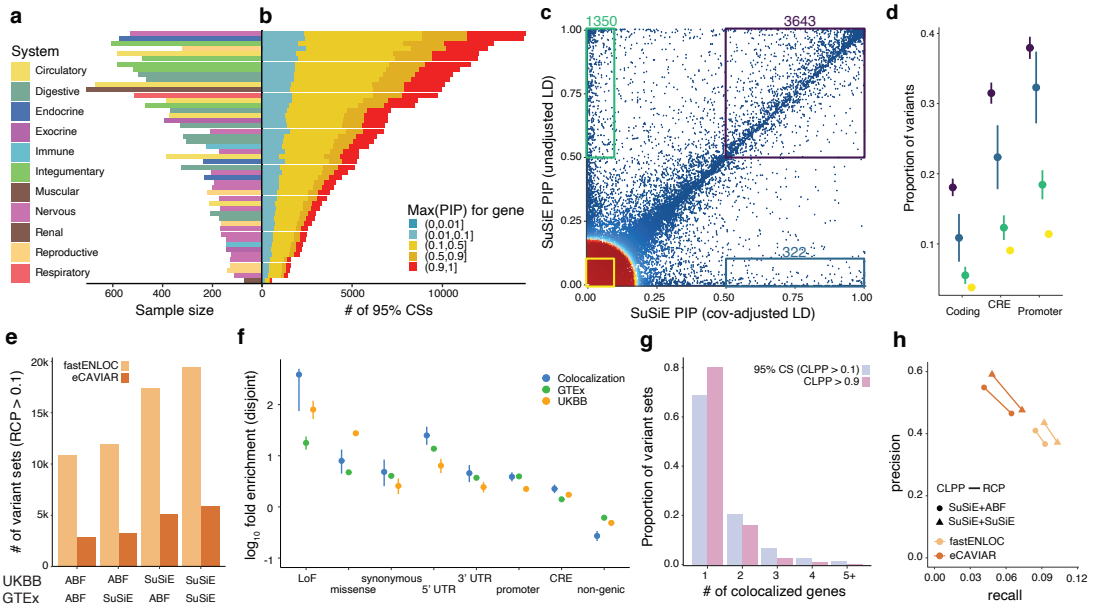


Figure 1.3: Improved colocalization of complex and molecular traits. a. Sample size and physiological system of the 49 tissues from GTEx v8 fine-mapped in our study. **b**. Number of 95% CSs across tissues in each system colored by the highest PIP in the CS **c**. Comparison of PIPs from SuSiE fine-mapping using unadjusted in-sample LD or covariate-adjusted in-sample LD. Red indicates higher density of variants, and blue indicates lower. **d**. Proportions and 95% CIs of the corresponding subset of fine-mapped variants from **c** for a subset of genomic annotations from **Fig. 1.2h**. **e**. Number of variant sets with regional colocalization probability (RCP) > 0.1 for each colocalization method. Fine-mapping using a single causal variant (ABF) and allowing for multiple causal variants (SuSiE) for both complex traits and eQTLs is varied as well as priors assuming independence between complex and molecular traits (eCAVIAR) or informative empirical priors (fastENLOC). **f**. Enrichment comparing variants with probability > 0.9 (for fine-mapping or colocalization) to variants with probability < 0.1 for disjoint annotations from **Fig. 1.2h**. Error bars represent 95% CIs. **g**. Number of genes that each variant or 95% CS (approximated by using colocalization posterior probability [CLPP] > 0.1 for at least one variant) colocalizes with. **h**. Using a validation set of fine-mapped coding variants with PIP > 0.5, precision and recall estimated for non-coding 95% CSs within 500 kb across different colocalization methods varying single or multiple causal variant assumptions, independent or informative priors, and estimand (RCP vs. CLPP). Estimates are restricted to estimated probability values (CLPP or RCP) equal to 0.1.

1.2.4 Cell types and proximal mechanisms of fine-mapped regulatory variants

Complex trait heritability is enriched in accessible chromatin (AC)[95,97], so we annotated the possible regulatory regions of our fine-mapped variants with maps of AC in an extensive set of tissues and cell types. To do this, we combined and re-QC'd bulk and single cell AC maps, including four datasets covering a broad diversity of tissues[80,84–86], in addition to fine-scale investigations of AC during blood cell production[81], immune system activation[83], and in the adult human brain[82] (see **1.4 Methods**, **Supplementary Fig. A.7**). We further restricted each dataset to elements that are likely *cis*-regulatory enhancer elements (CREs) by requiring H3K27ac in at least one of 327 distinct cell types[84,85] We observed enrichments across these datasets ranging from 2.2-4.2 fold, although no single dataset annotated more than 34% of non-coding variants (**Fig. 1.4a**). Combined, we are able to annotate 47% of likely causal (PIP > 0.9) variants and 68% of high confidence colocalized (CLPP > 0.9) variants. Importantly, these annotations still miss a large fraction of non-coding variants, possibly due to missing important tissue-specific CREs, alternate regulatory mechanisms, or model misspecification when fine-mapping.

To inform functional follow-up experiments on non-coding variants, we explored whether our fine-mapped variants were more likely to lie in trait-relevant cell type-specific accessible chromatin regions[99,143] using our previously developed method g-chromVAR[87]. Across all accessible chromatin atlases, we identified 2,557 trait-cell type enrichments that survived atlas-specific Bonferroni correction (**Supplementary Table A.17**). We recapitulated many previously reported trait-tissue enrichments[143] such as lymphocyte subset counts and myeloid subset indices (*e.g.*, RBC and platelet counts) in white blood cell types and myeloid cell types[87] (**Fig. 1.4b**), respectively, musculoskeletal traits (*e.g.*, eBMD) in bone and muscle tissues, cholesterol (*e.g.*, LDL-C) in liver cell types, lung function (*e.g.*, FEV1/FVC ratio) in lung tissues, glucose metabolism phenotypes (*e.g.*, T2D) in kidney and islet[144] tissues, atrial fibrillation in cardiomyocytes[145], BMI in several brain regions[146], and

Alzheimer's in myeloid cells[147]. Our fine-mapping was also able to distinguish between testosterone levels in XX individuals, which were enriched in adrenal tissues[148], from testosterone levels in XY individuals, which were most highly enriched in renal tissues, in addition to identifying several novel trait-tissue enrichments, such as cholelethiasis (gallstones), in the pancreas, which further informs the shared genetic etiology of gallstones and acute pancreatitis[149]. In total, we identified 485 likely causal variants lying in an enriched trait-cell type accessible chromatin region (**Supplementary Table A.18**). Functional dissection of the variants in our annotated atlas is likely a high yield step into understanding gene regulatory mechanisms underlying complex traits.

Finally, we investigated the proximal mechanisms by which single nucleotide fine-mapped regulatory variants act. We first investigated whether likely causal complex trait variants (PIP > 0.9) or high confidence colocalized (CLPP > 0.9) variants within CREs were more likely to disrupt one of 426 TF binding motifs[88], observing a small enrichment (1.3-fold) with a very high background (**Fig. 1.4c**). When we restricted to 37 motifs that were enriched across combined traits (**Supplementary Table A.19**), we observed substantially reduced background and better enrichment (2.1-fold), noting that this estimate suffers from "double-dipping". As an unbiased approach, we increased the confidence that our fine-mapped variants truly disrupt TF binding or activity by additionally requiring that the TF itself occupies the CRE in at least one of 768 measured cell types[84]. This further increased our enrichment (2.7-fold), while retaining 14% of all non-coding fine-mapped (PIP > 0.9) variants (**Fig. 1.4c**); restricting again to enriched motifs increases enrichment (5.9-fold), but annotates only 6% of fine-mapped variants. We observe generally similar results using TF footprints[91] inferred from accessible chromatin but observe even stronger enrichments for variants that exhibit allele-specific binding or allele-specific chromatin accessibility (3.4- and 4.5-fold, respectively).

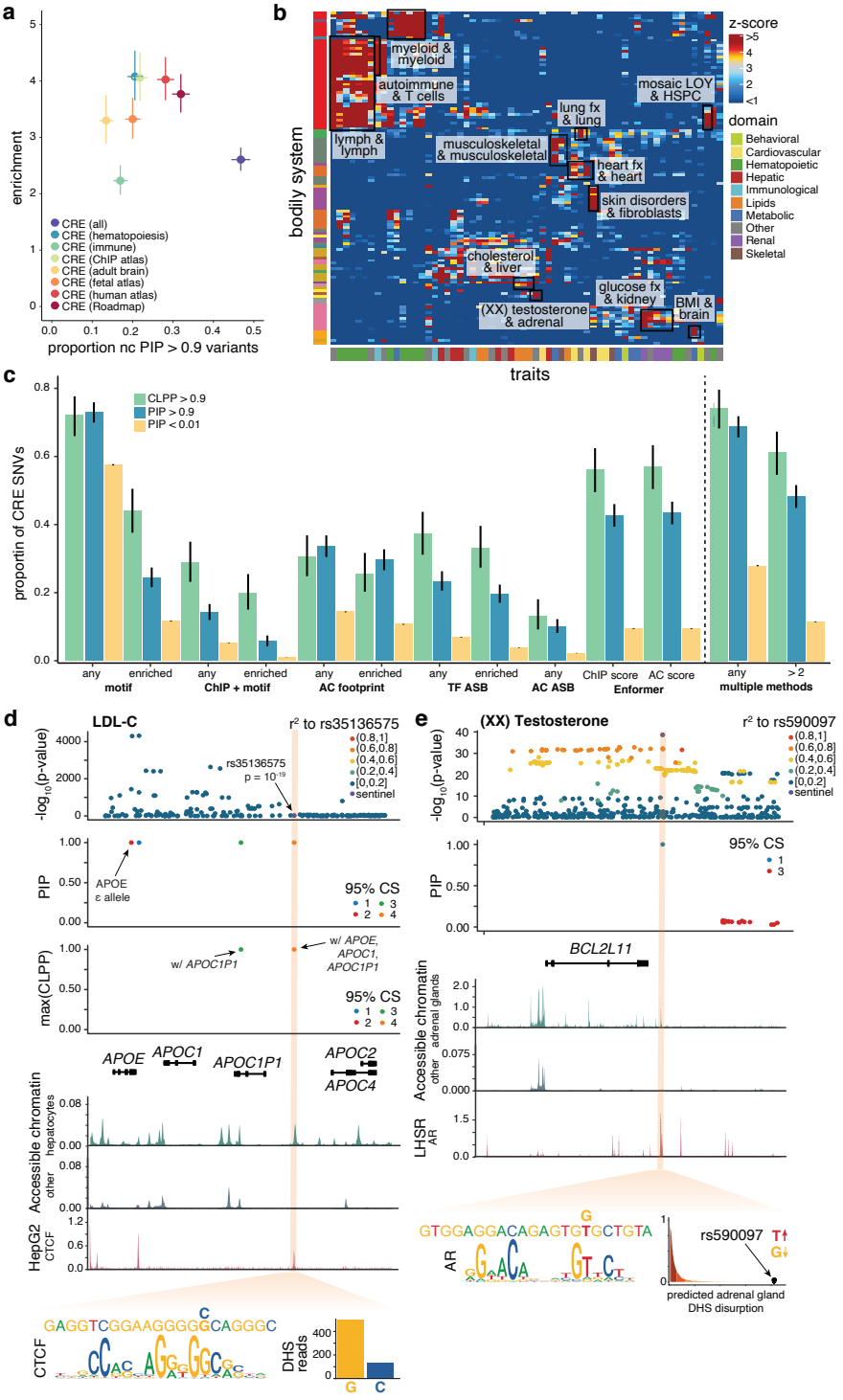
To identify regulatory variants with effects both in[99] and outside[150] of canonical motifs, we turn to a recent deep learning approach, Enformer, that can predict the effects of variants on both

TF occupancy and AC levels[89]. We determine a threshold for total TF occupancy disruption (see **1.4 Methods**) such that exactly 90% of low PIP variants fall below this, observing that 43-57% of high PIP or CLPP variants exceed it (4.5- to 6.0-fold enrichment). Importantly, even after excluding variants that disrupt canonical motifs, fall into TF footprints, or exhibit allele specific activity, we still observe a 2.5- to 3.0-fold enrichment in TF occupancy score (**Supplementary Fig. A.7**). Collectively, these data provide evidence that common trait-associated regulatory variants can act by disrupting known TF binding motifs but also by tuning their occupancy outside of standard position weight matrices (**Supplementary Table A.20**).

Altogether, we are able to nominate a proximal mechanism for 49% and 61% of single nucleotide CRE fine-mapped variants and colocalized variants (**Supplementary Table A.21**), respectively, when combining 3 or more methods (vs. 11% for CRE variants with PIP < 0.01). As validation, our approach identifies that the likely molecular mechanism of rs2814778's association with blood cell traits at the well-known Duffy locus is disrupting the canonical motif of bound GATA1/2[151], rs11257655's association with diabetes-related biomarkers at the *CDC123/CAMK1D* locus is by disrupting the canonical motif of bound FOXA1/2[152], and rs1414660's association with bone mineral density at the *GREM2* locus is by disrupting the canonical motif of bound CEBPB[153]. Our approach also highlights a number of novel mechanisms, including the disruption of CTCF and RAD21 occupancy by the colocalized hepatic control region variant rs35136575 and the disruption of an androgen receptor TF complex by the (XX) testosterone-level fine-mapped variant rs590097 at the *BCL2L11* locus (**Fig. 1.4d,e**). At the *IRF8* locus, SuSiE and FINEMAP identify 5 independent signals with PIP > 0.9 for monocyte count. All 5 of these variants are within CREs, and 3 of these variants have 3 or more lines of evidence for their molecular function. We suggest that rs11640143 alters SP1 / PAX5 binding, rs56177354 affects STAT5A/B binding, and rs11642657 acts through a complex mechanism to impact activity of the EZH2 / REST complex.

Figure 1.4 *(following page)*: Comprehensive annotation of fine-mapped non-coding variants. a. Proportion and enrichment (PIP > 0.9 vs. < 0.1) of fine-mapped variants across 7 different accessible chromatin datasets and a combination of all 7. CREs are defined as the intersection of accessible chromatin with H3K27ac from any cell-type. Error bars represent 95% CIs. **b**. Cell-type specific enrichment of fine-mapped variants in accessible chromatin for the atlas from Meuleman *et al.*[80] using g-chromVAR. Only cell-types and traits with at least one enrichment (Bonferroni-adjusted P-value < 0.05) are shown. Bodily systems are colored the same as in **Fig. 1.4a**. **c**. Proportion of CRE variants with one or more annotated molecular mechanisms. Molecular mechanisms include motif breaking (motif), motif breaking and occupied by the corresponding TF (ChIP + motif), residing within an accessible chromatin footprint (AC footprint), exhibiting allele specific TF binding (TF ASB), allele specific accessible chromatin (AC ASB), or TF occupancy or accessible chromatin predicted changes from deep neural network models (Enformer). When multiple methods are used, motif breaking only variants are not considered nor are enriched mechanism categories. **d,e**. Example fine-mapped variants with accessible chromatin, TF occupancy, and motif alteration, and other functional information tracks. Control accessible chromatin tracks were obtained from liver tissue[86] but not annotated as hepatocytes. CCCTC-binding factor (CTCF) occupancy is shown for the hepatoblastoma cell line, HepG2 and androgen receptor (AR) occupancy is shown for the prostate cell line, LHSR. The alternative allele is shown above the reference allele. The empirical density for the absolute predicted change in accessible chromatin from the Enformer model across all PIP < 0.01 CRE SNVs is shown in **e**.

Figure 1.4: (continued)



1.3 Discussion

To identify likely causal variants underlying complex human traits and diseases, we evaluated and applied state-of-the-art multiple-causal-variant statistical fine-mapping to 96 phenotypes in the UKBB, identifying 3,785 variant-trait pairs with PIP > 0.9. We observe that variants are often fine-mapped for multiple traits and show how this widespread pleiotropy can be leveraged to pinpoint causal variants in complex disease where fine-mapping is complicated by fundamental complexities present in large meta-analyses. At the genomic level, these variants are enriched in the exons of protein coding genes, which we explore in our companion manuscript; their untranslated mRNA regions, whose functional consequences we have experimentally probed in other work [154], and in trait-relevant tissue accessible chromatin, which we annotate and investigate the potential mechanisms of here. In total, our study provides both a set of recommendations for fine-mapping large genetic studies as well as a deeply annotated resource of likely causal protein coding and regulatory variants.

We then systematically evaluated methods for colocalization analysis, which estimate the convergence of causal complex and molecular trait variants to nominate causal genes. Notably, our state-of-the-art approach, which allows for multiple causal variants for both types of trait, properly accounts for population stratification, and uses powerful empirical priors, identified the correct gene only 50% of the time for at most 10% of loci, adding further evidence to the limitations of colocalization [78,155,156]. We show that this is due in part to our ever-improving ability to identify likely causal variants underlying changes in expression of multiple genes, confounding downstream gene prioritization. On the other hand, we found that colocalized variants present a more high-confidence and interpretable set than fine-mapped complex trait variants alone and may provide a more fruitful foothold for variant to function studies.

Finally, we improved our ability to annotate regulatory variants with extensive maps of chro-

matin state[80–85] but found that only 50% of fine-mapped variants were within a CRE, defined by accessible chromatin and H3K27ac. Within CREs, we integrated predictive models of variant function[88,89] with large-scale maps of TF occupancy[84,90,91] to propose molecular mechanisms for 49% and 61% of likely causal complex trait CRE variants and colocalized CRE variants, respectively, with a background rate of 11%. Notably, the majority of variants did not disrupt known binding motifs for occupied TFs. Although these analyses allowed us to identify many likely mechanisms of complex trait regulatory variants, it is clear that improved measurement and prediction of gene regulatory alteration across additional cell types, states, and factors is needed to further elucidate the underlying actions of complex trait variants.

Our work is subject to several limitations. First, while our approach performs well in simulations which we endeavored to make as representative of real complex trait GWAS as possible, additional complexities likely remain unaccounted, leading to model misspecification. Second, our colocalization analysis is conducted on a limited set of tissues and cell types; for many loci in many traits, these data omit the relevant tissue or cell type at the locus. Finally, we have focused on the well-studied "white British" subset of the UK Biobank, whose ancestral lack of heterogeneity makes it well-suited for application of currently-available fine-mapping approaches. Our companion paper[68] extends this study to include FinnGen and Biobank Japan, although more representative studies of the global majority are desperately needed. Overall, our annotated atlas of likely causal variants underlying complex human traits, diseases, and gene expression provides insights into the molecular mechanisms of regulatory variants and is a contemporary resource for studies of variant to function.

1.4 Methods

1.4.1 Cohorts

UK Biobank (UKBB)

The UK Biobank (UKBB)[76] is a prospective population-based cohort in the United Kingdom that recruited approximately 500,000 individuals aged between 40–69 years old from 2006 to 2010. In this study, we analyzed 366,194 unrelated "white British" individuals that were previously defined in the Neale Lab GWAS (https://github.com/Nealelab/UK_Biobank_GWAS). Briefly, this cohort includes individuals of British ancestry, based on the PCA-based sample selection criteria (https://github.com/Nealelab/UK_Biobank_GWAS/blob/master/ukb31063_eur_selection.R), and is further filtered to those who self-reported as "white British", "Irish", or "white".

The genotypes were obtained using either i) the Applied Biosystems UK BiLEVE Axiom Array or ii) UKB Axiom Array, and were further imputed using IMPUTE4 with a combination of reference panels: i) the Haplotype Reference Consortium and ii) UK10K and the 1000 Genomes Phase 3. We applied variant-level quality-control (QC) criteria as previously defined in the Neale Lab GWAS (https://github.com/Nealelab/UK_Biobank_GWAS), which retained 13,791,467 variants with INFO > 0.8, MAF > 0.001, and Hardy-Weinberg equilibrium P value > 1.0×10^{-10}, with exception for the VEP-annotated coding variants where we allowed MAF > 1.0×10^{-6}. All the variants were processed on the human genome assembly GRCh37.

We defined phenotypes using various data types available in UKBB, including biomarkers, body measures, and disease case-control status mapped on phecodes[157] (https://phewascatalog.org/phecodes). We summarized detailed phenotype definitions in **Supplementary Table A.1**. The UK Biobank analysis was conducted via application number 31063.

Genotype Tissue Expression Project (GTEx) v8

The Genotype Tissue Expression Project (GTEx) v8 is a cohort of 838 individuals on whom genotype and gene expression have been measured. In this study, we obtained genotypes for 838 individuals (dbGAP accession ID: phs000424.v8) who had been included in the v8 release for *cis*-eQTLs across 49 tissues. Of note, genotypes from the GTEx v8 project were obtained using whole genome sequencing and gene expression profiles were obtained using RNA-seq on post-mortem samples. Additional details on individuals, tissue samples, and sequencing have been previously described [103].

As input to fine-mapping, we obtained *cis*-eQTL summary statistics from the GTEx portal (`https://gtexportal.org/`). We also obtained the covariate matrix containing sex, whether PCR was used to amplify DNA to create libraries, sequencing platform (HiSeq 2000 or HiSeq X), the first five genotype PCs, and up to 60 Probabilistic Estimation of Expression Residuals (PEER) factors [158]. After fine-mapping, all variants were lifted over from GRCh38 to hg19; variants with no allele matching the reference sequence were excluded.

1.4.2 Genome-wide association

We conducted GWAS using a generalized linear mixed model as implemented in SAIGE [108] (for binary traits) or BOLT-LMM [109,159] (for quantitative traits). We used age, sex, age^2, age $\times$ sex, age^2 $\times$ sex, and top 20 principal components as covariates, while excluding sex-adjusting covariates from sex-specific or stratified traits (*i.e.*, age at menarche/menopause, breast cancer, testosterone levels). For mosaic loss of chromosome Y, we used summary statistics publicly available from UKBB [160].

1.4.3 LD score regression

We used LD score regression to estimate common and low-frequency variant heritability, genetic correlation, and confounding in UKBB complex traits [95,119]. Following Gazal *et al.* [161], we com-

puted LD scores from 3,567 unrelated individuals in the UK10K cohort [162] for all variants down to MAF of 0.05 across the baselineLD v2.2 annotations. For binary traits, liability threshold heritability was calculated from observed heritability as previously described [163] using in sample prevalence estimates. The attenuation ratio, defined as (LDSC intercept $-$ 1) / (mean χ^2 $-$ 1), and the LD score intercept were used to assess residual confounding in traits, similar to Loh *et al.* [109]. for different traits colored by phenotypic domain. Of note, we found that height had the largest LD score intercept, but a low attenuation ratio, suggesting that much of the presumed residual confounding is driven through polygenicity. On the other hand, educational attainment (college Y/N) had both high LD score intercepts and attenuation ratio, suggesting that the phenotype as defined suffers from serious residual confounding. Other potentially problematic traits include deep vein thrombosis (DVT) and total bilirubin levels. We encourage caution when interpreting results about these phenotypes.

To define a set of genetically independent traits, we investigated all groups of traits such that all pairwise traits have low genetic correlation ($|r_g| < 0.2$). To find the largest group of such traits, we found the largest independent vertex set [164] induced by the binary (0 if $|r_g| < 0.2$, 1 otherwise) adjacency matrix. Since there was no largest unique set, we chose the set that maximized the median number of 95% CSs across the set of independent traits. Meta-analysis across independent traits was performed using the rmeta package with random effects.

1.4.4 Fine-mapping simulations

To benchmark fine-mapping performance, we simulated a biobank-scale GWAS and performed fine-mapping under various QC and fine-mapping parameters. First, we randomly draw "true" genotypes for chromosome 1 based on the genotype probabilities (GP) in the imputed bgen provided by UKBB. To do this efficiently, probabilistic "true" genotypes (pGTs) for a given variant i were computed via $\text{pGT}_i = \text{ceiling}(u_i - \text{GP}(X_i = 0)) + \text{ceiling}(u_i - \text{GP}(X_i = 0) - \text{GP}(X_i = 1))$

where $\mathrm{GP}(X_i = k)$ $[k = 0, 1, 2]$ represents the GP for having k copies of alternative alleles and $u_i \sim \mathrm{Unif}(0, 1)$ represents a uniform random variable.

We then simulated 50 true phenotypes that resemble the observed complex trait genetic architecture, including causal variant density, MAF-dependent causal effect sizes, and total SNP-heritability. Based on the previous literature, we set parameters as the followings: 1) 50% of 1 Mb loci have 1.5 causal variants on average[110]; 2) per-allele causal effect sizes have variance proportional to $[2p(1-p)]^{\alpha}$ where p represents MAF and α is set to be -0.38 (ref.[111]); and 3) total SNP-heritability h^2_{g} for chromosome 1 equals to 0.025 (ref.[112]). The true causal effect size β_j for a randomly drawn true causal variant j was drawn from $\beta_j \sim N(0, \sigma^2_{\mathrm{g}} \cdot [2p(1-p)]^{\alpha})$ where σ^2_{g} was determined by $\sigma^2_{\mathrm{g}} = h^2_{\mathrm{g}} / \sum_j [2p(1-p)]^{(1+\alpha)}$. The true phenotype y was computed via $y = X\beta + \varepsilon$ where X is the above true genotype (pGT) matrix and $\varepsilon_i \sim N(0, 1 - \sigma^2_{\mathrm{g}})$ i.i.d. We note that, to prevent inflation in true causal effect sizes due to extremely rare variants, we used a lower-bounded MAF p' instead of p where $p' = \max(p, 0.001)$ when simulating effect sizes.

We ran 50 simulated GWAS via a standard linear regression in Hail v0.2 using the above true phenotypes. For genotypes, we used 1) the above true genotypes, 2) the imputed dosage, and 3) hard-called genotypes based on the imputed dosage with a hard-call threshold set to be 0.1. Before fine-mapping, we applied sumstats QC with various thresholds: 1) Hardy-Weinberg equilibrium P value $> 1.0 \times 10^{-10}$; 2) MAF $> 1.0 \times 10^{-5}$, 1.0×10^{-4}, 0.001, 0.01, and 0.05; 3) INFO > 0.2, 0.4, 0.6, 0.8, 0.9, and 0.99. We computed LD matrices using 1) full-sample dosage LD (n = 366,194); 2) down-sampled dosage LD (n = 10,000 and 100,000); and 3) external reference LD using the 1000 Genomes Phase 3 Europeans (n = 503). We defined fine-mapping regions based on four different window sizes (100 kb, 1 Mb, 3 Mb, and 5 Mb) around each lead variant. We used four different fine-mapping methods (FINEMAP[28], SuSiE[138], ABF[24], and COJO-ABF[7]) and one ensemble method ("Average") which takes an average of PIP from FINEMAP and SuSiE while excluding variants with a substantial PIP difference (> 5%) between the methods.

To assess the performance of fine-mapping we used two metrics: calibration, the difference between the observed and expected proportion of causal variants in a range of PIP values, and recall, which we quantified as the number of true causal variants detected in the top ranked variants by PIP.

1.4.5 Complex trait statistical fine-mapping

We performed statistical fine-mapping using FINEMAP[28,29] and SuSiE[30] with GWAS summary statistics from SAIGE/BOLT-LMM and in-sample dosage LD computed by LDstore 2 (ref.[113]). Fine-mapping regions were defined by adding a 1.5 Mb window upstream/downstream of each lead variant and were merged if they overlapped. We excluded the major histocompatibility complex (MHC) region (chr 6: 25–36 Mb) from analysis due to extensive LD structure in the region. We allowed up to 10 causal variants per region. Using the default uniform prior probability of causality, we estimated posterior inclusion probabilities (PIP) of each variant and derived up to 10 independent 95% credible sets (CS).

In most analyses, we combined fine-mapping results from FINEMAP and SuSiE by taking the average of PIP for each variant across methods and excluding variants with a substantial PIP difference (> 5%). This improved fine-mapping accuracy, since we observed lower functional enrichment for the variants with inconsistent PIPs across the methods (**Supplementary Fig. A.3**). On rare occasions when the fine-mapping method(s) failed (*e.g.*, due to conversion failure or available memory restrictions), we used successful results from either of the methods or excluded the regions from analysis if both methods failed.

We also applied ABF and COJO+ABF to simulations and for other comparative analyses as previously described[7,24].

1.4.6 Credible set merging

We defined independent merged CSs across traits by merging SuSiE 95% CS from each trait using hierarchical clustering with the weighted Jaccard similarity index. Briefly, we computed the PIP-weighted Jaccard similarity index between all the pairs of CS across the studied traits. For a pair of CS, the similarity index is defined as $\sum_i \min(x_i, y_i) / \sum_i \max(x_i, y_i)$ where x_i and y_i are PIP values (or zero if missing) in each CS for the same variant i. We then used $1 -$ the similarity index as a distance to conduct hierarchical clustering of the CS using the complete linkage method. We cut a dendrogram tree at a height of 0.9 so that any CSs with PIP-weighted Jaccard similarity above 0.1 are merged into a single CS.

1.4.7 Fine-mapped variant enrichment

To compute the enrichment of fine-mapped variants, we compared the proportion of variants with PIP > 0.9 to the proportion of variants with PIP < 0.01. Confidence intervals estimated from the Fisher exact test. For estimates of proportions, confidence intervals were estimated using Wilson's method. To compute the enrichment of fine-mapped variants in accessible chromatin regions across different cell-types, we used g-chromVAR. Briefly, g-chromVAR creates a test statistic (PIP-weighted chromatin accessibility fragments) for each trait cell-type combination, computes the deviation of each test statistic from expected across all input cell-types, and then creates a null distribution for these deviations across a set of 50 background sets of regions (peaks) matched on GC content and fragment count. *Z*-scores and *P*-values from a 1-sample test for each trait cell-type combination are derived from comparing the observed deviation to this empirical null distribution. Only traits with a total probability sum of > 5 in accessible chromatin peaks are included in the analysis. For each accessible chromatin dataset, a Bonferroni correction (0.05 / # of traits / # of cell-types) is used.

1.4.8 PheWAS

To explore trait associations that are not covered in our study, we used the fine-mapped variants with PIP > 0.1 from our study to conduct a PheWAS in the Open Targets Genetics [165] (version 20.02.01). The Open Targets Genetics provides a collection of disease-associated loci (the V2D resource), compiling associations from the GWAS Catalog [18], the Neale Lab UK Biobank summary statistics (http://www.nealelab.is/uk-biobank/), and the SAIGE UK Biobank summary statistics [108]. To avoid potential double counting, we excluded associations from the UKBB summary statistics (the studies with ID starting either from "NEALE2" or "SAIGE"). All the traits were mapped to the Experimental Factor Ontology (EFO) [166] terms, resulting in a collection of 1,183 unique traits from 3,079 previous studies. We defined disease traits based on the EFO hierarchy by taking all the terms that belong to the term "disease" (EFO_0000408). We took all the lead variants and tagging variants ($r^2 > 0.7$) as defined by the Open Targets Genetics to see whether they overlap with our fine-mapped variants with PIP > 0.1.

1.4.9 eQTL statistical fine-mapping

We performed fine-mapping of eQTLs in GTEx v8 similar to how we performed fine-mapping in complex traits from the UKBB with several key differences. First, regions were defined as genes with at least one genome-wide significant ($P < 5.0 \times 10^{-8}$) *cis*-eQTL. Second, in-sample LD was computed using BlockMatrices in Hail 0.2 (ref. [167]). Third, covariates were projected out of the genotypes prior to LD calculation. Fourth, we focus primarily on results from SuSiE rather than an average of FINEMAP and SuSiE.

Fine-mapping methods are typically derived assuming the following normalized, simple linear relationship between causal genotypes and a phenotype: $Y \mid X, \beta, \sigma^2_{\text{res}} \sim N(X\beta, \sigma^2_{\text{res}})$. However, in GWAS, we usually use a linear model that includes covariates: $Y = X\beta + C\alpha + \varepsilon$. This controls

for the effects of covariates C on Y in our summary statistics $(\hat{\beta}, \text{se}[\hat{\beta}])$, which are often used as input to fine-mapping. Applying the Frisch-Waugh-Lovell theorem [104,136], we show that current fine-mapping approaches can be straightforwardly used with covariate-adjusted summary statistics when covariate-adjusted LD is also used, analogous to previous work considering covariate-adjusted LD for heritability estimation [136].

Specifically, we define the orthogonal projection matrix as $P_\perp = I - C(C^TC)^{-1}C^T$. We can then project covariates out of both sides of the linear regression model accordingly:

$$P_\perp Y = P_\perp X\beta + P_\perp C\alpha + P_\perp \varepsilon \tag{1.1}$$

Since C is invariant under its own projection matrix, $P_\perp C = 0$. Defining $Y_\perp = P_\perp Y, X_\perp = P_\perp X$, and $\varepsilon_\perp = P_\perp \varepsilon$, we can rewrite the above equation as:

$$Y = X_\perp \beta + \varepsilon_\perp \tag{1.2}$$

Assuming large sample size GWAS ($\gg \dim[C]$), $\varepsilon \approx \varepsilon_\perp$ and the summary statistics obtained from regressing Y on X, including C as covariates will be approximately equal to those obtained from regressing $Y_\perp$ on $X_\perp$. Thus, we have re-written the linear regression model with covariates as a transformed model without covariates. We can then apply standard fine-mapping methods to summary statistics derived from univariate regression with covariates when we additionally project the covariates out of the genotypes. In practice, this means that the covariate adjusted LD matrix is $R_\perp = X_\perp^T X_\perp$.

We finally note that $R \approx R_\perp$ in the case of "white British" UK Biobank individuals and this adjustment is not necessary in practice. In the more heterogeneous GTEx v8 cohort, genotype PCs are associated with many genotypes and thus covariate-adjusted and unadjusted LD are substantially different. In this case, $R_\perp$ is no longer on the same scale as R (*e.g.*, values are no longer Pearson cor-

relations and diagonals may be < 1). This can introduce complexities to interpretation of prior and posterior effect sizes.

1.4.10 Colocalization

Colocalization models the effects of LD and provides estimates of the joint probability that a variant or set of variants causally influences both a complex and molecular trait[115]. Letting $\gamma_{\text{GWAS}} = 1$ be the event that a specific variant is causal for a complex trait and $\gamma_{\text{eQTL}} = 1$ the event that the variant is causal for a complex trait. The joint probability can then be written as:

$$P(\gamma_{\text{GWAS}} = 1, \gamma_{\text{eQTL}} = 1) = P(\gamma_{\text{GWAS}} = 1 \mid \gamma_{\text{eQTL}} = 1)P(\gamma_{\text{eQTL}} = 1) \tag{1.3}$$

In its simplest form, assuming an independent prior on the causal effects of complex traits and molecular traits (known as the eCAVIAR prior), colocalization reduces to a multiplication of the fine-mapped posteriors of traits:

$$P(\gamma_{\text{GWAS}} = 1, \gamma_{\text{eQTL}} = 1) = P(\gamma_{\text{GWAS}} = 1)P(\gamma_{\text{eQTL}} = 1) \tag{1.4}$$

Alternatively, we can estimate a more informative prior using fastENLOC[78]. Specifically, fastENLOC fits a logistic regression model to the complex trait PIPs to estimate the enrichment of eQTL PIPs: $\text{logit}\left[P(\gamma_{\text{GWAS}} = 1 \mid \gamma_{\text{eQTL}})\right] = \beta_0 + \beta_1 \gamma_{\text{eQTL}}$. In practice, we don't know $P(\gamma_{\text{GWAS}} = 1 \mid \gamma_{\text{eQTL}})$, so fastENLOC (1) fits the logistic model to γ_{GWAS} from our study assuming an eCAVIAR prior, (2) approximately fine-maps with the current iteration fastENLOC prior, (3) iterates until convergence of colocalization values. Other details, such as the fine-mapping approximation, calculation of the fastENLOC prior, use of multiple imputation, and prior on enrichment effect size ($\beta_1 \sim N[0, 1]$) are described in the fastENLOC study[78].

Since the fastENLOC method gains an improvement in speed by only using variants in complex

trait 95% CSs as input, we restricted our analyses to SuSiE 95% CSs and ABF 95% CSs for complex traits and similarly for molecular traits. To compare our results to released fine-mapping data, we downloaded CAVIAR, CaVEMaN, and dap-g results from the GTEx portal (https://gtexportal.org/home/datasets). We restricted our comparisons to gene-tissue pairs that had fine-mapping results for all methods.

We also evaluated the usefulness of colocalization methods for gene prioritization. Instead of using curated gold standard gene sets which can be biased towards more well-studied and more easily identifiable genes and even include genes prioritized based upon the non-coding GWAS loci themselves, we used the validation set of likely causal genes for each trait we previously developed [142]. First, we identified fine-mapped (PIP > 0.5) protein-coding variants for 589 genes. Leveraging our intuition that multiple signals in a region likely act through the same gene(s), we identified 1,348 non-coding CSs within 500kb of a validation gene for the same trait. We then evaluated how often each colocalization method correctly identified the validation gene at a pre-specified colocalization threshold. When colocalization identified multiple genes in a region, we took the gene with the highest colocalization value. The precision, defined as the number of 95% CSs that colocalized with a gene in the validation set divided by the total number of colocalized gene-CS pairs, and recall, defined as the number of colocalized gene-CS pairs that were also in the validation set divided by the total number of gene-CS pairs in the validation set, of each method for each threshold was then computed. Finally, we calculated the weighted harmonic mean of precision and recall, typically referred to as the F_x value, for each evaluation as $F_x = (1+x)^2 \frac{(\text{precision})(\text{recall})}{x^2(\text{precision})+\text{recall}}$. We chose $x = 0.33$ to formalize our preference for higher (3-fold) precision than recall when comparing methods.

1.4.11 Genomic annotations

Genic annotations (LoF, missense, synonymous, 5' UTR, and 3' UTR) were obtained using the Ensembl Variant Effect Predictor (VEP) [168] v85. When a variant had multiple annotations, the most

severe consequence on the canonical transcript (GENCODE v19) was used. Further, LoF variants were refined by LOFTEE, which included only high confidence stop gained, splice acceptor, splice donor, and frameshift variants. Promoter annotations were obtained from the S-LDSC baseline model. Cis-regulatory element (CRE) annotations are defined as the intersection of accessible chromatin with H3K27ac (across any cell-type, see details below). Finally, non-genic variants are defined as variants not in any of the above annotations.

To define CREs, we retrieved and performed additional quality control on the following accessible chromatin and histone modification atlases:

- ROADMAP Epigenomics[85] — DNase I hypersensitivity for 39 broad cell-types and H3K27ac for 98 broad cell-types
- Meuleman *et al.*[80] — DNase I hypersensitivity for 438 broad cell-types
- Domcke *et al.*[86] — single cell ATAC-seq for ~720,000 cells representing 54 broad cell-types
- Corces *et al.*[81] — ATAC-seq for 18 cell-types in the hematopoietic lineage
- Corces *et al.*[82] — single cell ATAC-seq of ~70,000 cells representing 24 distinct brain cell-types
- Calderon *et al.*[83] — ATAC-seq for 25 immune cell subsets
- ChiP-Atlas[84] — DNase I hypersensitivity for 284 broad cell-types and and H3K27ac for 720 broad cell-types

We performed multiple additional QC steps to harmonize the accessible chromatin data. First, we lifted over all peak calls in GRCh38 to hg19, keeping only 1-to-1 matches, and normalized peak sizes to 300 bps. With the exception of the ChIP-Atlas dataset, all data were available in the matrix format of peak x cell-types with a count or normalized count value. To remove low quality peaks

from each dataset, we required that each peak was within the top 100,000 for at least one annotated cell-type in the dataset, first whitening low count values (0, 1, or 2). With the exception of the Corces *et al.* brain dataset, we found limited enrichment for fine-mapped variants in the removed peaks (**Supplementary Fig. A.7**), so we proceeded with the QC'd data for all except the Corces *et al.* brain dataset for which we kept all provided peaks. For ChIP-Atlas, we took a stringent MACS2 $-\log_{10}(q\text{-value})$ score > 500 given the lack of a count matrix.

1.4.12 Regulatory variant functional analysis

To determine if fine-mapped variants disrupted canonical TF binding PWMs, we identified variants effects on 426 PWMs from HOCOMOCOv11 [88] with motifbreakR [169] using the information content ("ic") method, a p-value threshold of 0.0001 to identify PWM matches at either allele, and a difference of > 0.4 for the scaled motif matrix between alleles. To increase our confidence in motif disruption calls, we also investigated disrupted motifs where the predicted binding site was occupied by the TF. We downloaded uniformly reprocessed ChIP-seq peaks for 1,009 TFs across 768 cell-types from ChIP-Atlas [84]. An overlap was called when a TF had an occupancy across its exact corresponding motif or a similar motif (Pearson's $r > 0.7$ for similarity of the PWM). Additionally, we investigated if fine-mapped variants have evidence of allele specific TF occupancy. Using the ADASTRA database covering 1,025 human TFs and 566 cell types, we consider a variant as having allele specific TF occupancy if the FDR adjusted *P*-value is < 0.05 for an increased read count compared to expected read count for either allele.

We also investigated whether fine-mapped variants fell within accessible chromatin footprints. We downloaded consensus footprints from 243 cell-types [91], lifted coordinates over from GRCh38 to hg19, keeping only 1-to-1 matches, and overlapped these regions with fine-mapped variants. Similarly, we investigated allele-specific changes in accessible chromatin [91], keeping only variants with a detectable imbalance at a false positive rate < 0.05.

To investigate the effects of TF binding outside of canonical PWMs, we lifted over all variants in our analysis (PIP > 0.001) to GRCh38 and scored them using Enformer[89], a state-of-the-art convolutional neural net that employs transformer layers to predict TF occupancy and accessible chromatin levels. After lifting back to hg19, we filtered to 1,447 tissue-TF pairs and 320 accessible chromatin datasets. We then combined the scores into a simple single metric by first standardizing (z-scoring) each dataset prediction and then taking the sum of squares (SS) of predictions for each variant. To "call" a disruption of TF occupancy or accessible chromatin, we identified the 90th or 95th percentile of the Enformer SSs in the control set of variants (PIP < 0.01 and in a CRE) and called a "disruption" as any value above this threshold.

Enrichment analyses across all traits and TFs may result in deflated estimates. Ideally, cell-type relevant TFs would be identified and investigated. In lieu of this difficult task, we perform an enrichment of each TF (comparing PIP > 0.9 to PIP < 0.01 variants for multiple data types) and take all TFs with a marginal p-value < 0.05 for the Fisher's exact test. This method suffers from "double-dipping" into the data, but can be compared to previous enriched motif estimates[99] and represents a possible upper bound on fine-mapped variant enrichment for each data type. Finally, we investigated whether combining multiple methods would improve our ability to annotate CRE variants. When combining, we excluded the motif disruption only method and all enrichment-based methods.

1.5 Data availability

Fine-mapping results produced by this study will be publicly available at https://www.finucanelab.org/data, the ENCODE data portal (https://www.encodeproject.org/), and the GWAS catalog (https://www.ebi.ac.uk/gwas/home). Individual-level data for UKBB participants is accessible on request through the UK Biobank Access Management System (https://www.ukbiobank.ac.

uk/). The UKBB analysis in this study was conducted via application number 31063. GTEx v8 summary statistics are available at the GTEx Portal (https://gtexportal.org/home/datasets). GTEx individual-level data is accessible on request through the dbGAP application (accession code: phs000424.v8.p2; https://gtexportal.org/home/protectedDataAccess).

1.6 Code availability

The fine-mapping pipeline is available at https://github.com/mkanai/finemapping-pipeline. Scripts to perform most analyses and generate main data figures will be provided at https://github.com/julirsch/annotatedatlas.

1.7 Acknowledgements

We acknowledge the contribution of all the participants of the UK Biobank and GTEx studies. We thank O. Weissbrod, A. Price, J. Engreitz, J. Nasser, T. Jones, C. Vockley, E. Bao, D. Kelley, and all members of the Finucane lab for their helpful feedback. L.S. was supported by a NIH training grant (T32GM007753). This study was supported as a project under the ENCODE Functional Characterization Center grant UM1HG009435 (to P.C.S, R.T., and H.K.F.). M.K. was supported by a Nakajima Foundation Fellowship and the Masason Foundation. H.K.F. was funded by NIH grant DP5 OD024582 and by Eric and Wendy Schmidt.

1.8 Author contributions

J.C.U. and H.K.F. designed the study. J.C.U., M.Kanai, Q.S.W., R.E., S.G., F.A., and L.S. performed analyses. C.B., R.L.C., E.M.W., Z.R.M., J.K., C.A.L., M.Kurki, S.K.R., B.M.N., R.T., P.C.S., and M.J.D. contributed ideas and insights. H.K.F supervised this work. H.K.F. and P.C.S. obtained funding. J.C.U., M.Kanai, and H.K.F. wrote the manuscript with input from all authors.

The work presented in this chapter will be published as Kanai, M. *et al.* and was posted online as a preprint (*medRxiv*, 2021)[68].

2

Insights from complex trait fine-mapping across diverse populations

Abstract

Despite the great success of genome-wide association studies (GWAS) in identifying genetic loci significantly associated with diseases, the vast majority of causal variants underlying disease-associated loci have not been identified [12–14]. To create an atlas of causal variants, we performed and integrated fine-mapping across 148 complex traits in three large-scale biobanks (BioBank Japan [170,171], FinnGen [172], and UK Biobank [67,76]; total n = 811,261), resulting in 4,518 variant-trait pairs with high posterior probability (> 0.9) of causality. Of these, we found 285 high-confidence variant-trait pairs replicated across multiple populations, and we characterized multiple contributors to the surprising lack of overlap among fine-mapping results from different biobanks. By studying the bottlenecked Finnish and Japanese populations, we identified 21 and 26 putative causal coding variants with extreme allele frequency enrichment (> 10-fold) in these two populations, respectively. Aggregating data across populations enabled identification of 1,492 unique fine-mapped coding variants and 176 genes in which multiple independent coding variants influence the same trait (*i.e.*, with an allelic series of coding variants). Our results demonstrate that fine-mapping in diverse populations enables novel insights into the biology of complex traits by pinpointing high-confidence causal variants for further characterization.

2.1 Introduction

Identifying causal variants for complex traits is a major goal of human genetics research, but most genome-wide association studies (GWAS) do not pinpoint specific variants, limiting the biological inference possible from follow-up experimentation [12–14]. Identifying causal variants from GWAS associations (*i.e.*, fine-mapping) is challenging due to extensive linkage disequilibrium (LD)

among associated variants, effect sizes that are often small, and the presence of multiple independent causal variants at a locus. Fine-mapping methods assign to each variant a posterior probability of being a causal variant (posterior inclusion probability, PIP)[23,25,26,28–30,134,173], and recently-developed methods for fine-mapping use scalable, sophisticated algorithms[28–30] that allow for multiple causal variants in a locus and can be applied to the very large data sets necessary to overcome the challenges listed above. Previous studies, performed almost exclusively in cohorts of European ancestry[7,73,87,99,118,174] or meta-analyses of majority European ancestry[52,63,175–180], have used fine-mapping methods to identify putative causal variants, enabling novel biological insights into diseases such as inflammatory bowel disease[73] and type 2 diabetes[7] and traits such as blood cell counts[87] and kidney function[180].

The recent development of large-scale biobanks worldwide[63,76,170,172] provides an exciting opportunity for well-powered fine-mapping of multiple phenotypes in diverse populations of both European and non-European ancestries. Unlike results from most meta-analyses, biobanks allow access to individual-level genotypes at large scale, enabling more accurate fine-mapping results[87,118], and often include hundreds of complex diseases and quantitative traits. For example, BioBank Japan (BBJ)[170,171], the largest non-European biobank, has recruited 200,000 individuals with >200 phenotypes, which is sufficient to achieve powerful fine-mapping in a cohort of East Asian ancestry. Within Europe, there is also substantial genetic diversity[181]; for example, FinnGen[172], a biobank in Finland, currently combines genotype data with electronic health records for 270,000 individuals in a population that has undergone strong population bottleneck followed by subsequent isolation and rapid expansion, making it genetically distinct from mainland Europe[182]. Moreover, because both Japan and Finland have recently undergone population bottlenecks, these populations harbor deleterious alleles with high frequency that are rare or absent in other populations[183–185].

Here, for the first time, we compare and combine fine-mapping results across large-scale biobanks in three distinct populations. To this end, we apply state-of-the-art multiple-causal-variant fine-

mapping methods at scale in BBJ [170,171] and FinnGen [172], and we analyze these results in conjunction with results from our parallel effort performing fine-mapping in UK Biobank (UKBB) [67,76]. Our multiple-biobank fine-mapping enables us to identify high-confidence putative causal variants that replicate in multiple populations, to compare fine-mapping results across biobanks, and to identify population-specific putative causal variants and the genes these variants converge on.

2.2 Results

2.2.1 Expanded atlas of putative causal variants across three populations

In a companion paper [67], we describe our fine-mapping in UKBB [76] (n = 361,194; 119 traits); here, to create an atlas of causal variants of complex traits, we extended our analysis to additionally include 148 complex diseases and traits available in BBJ [170,171] (n = 178,726; 79 traits) and FinnGen [172] (n = 271,341; 67 traits from release 6) (**Fig. 2.1a**; **Supplementary Tables B.1,B.2**). These traits were manually curated in each biobank to cover a wide spectrum of human phenotypes ranging from common complex diseases to biomarkers. Of these, 26 traits (*e.g.*, height and type 2 diabetes) are available in all the three cohorts, 65 traits (*e.g.*, lab tests and biomarkers) are available in any two of the three, and the rest are specific to a single cohort (**Fig. 2.1b**). We performed GWAS in BBJ and FinnGen using a generalized linear mixed model as implemented in SAIGE [108] or BOLT-LMM [109,159] (**2.4 Methods**). We identified 2,611 and 1,698 genome-wide significant locus-trait pairs ($P < 5.0 \times 10^{-8}$; 3 Mb regions excluding the major histocompatibility complex [MHC]; **2.4 Methods**) in BBJ and FinnGen, respectively. We then conducted multiple-causal-variant fine-mapping using FINEMAP [28,29] and SuSiE [30] (**2.4 Methods**).

In total, our expanded atlas included 476, 342, and 3,847 fine-mapped variant-trait pairs (posterior inclusion probability [PIP] > 0.9), and 3,558, 2,348, and 27,276 95% credible set (CS)-trait pairs (median CS size = 11, 9, and 12) in BBJ, FinnGen, and UKBB, respectively (**Fig. 2.1c–e**).

These consisted of 4,518 unique variant-trait pairs (PIP > 0.9 in any population) and 31,598 unique 95% CS-trait pairs (median CS size = 12; independent SuSiE CS merged across populations; **2.4 Methods**) in aggregate, of which 23,563 CS-trait pairs (75%) contained at least one variant with PIP > 0.1 (**Supplementary Tables B.3,B.4**). Notably, our expanded atlas included 66 unique variant-trait pairs (PIP > 0.9 in any population) and 601 CS-trait pairs on the understudied X chromosome. The three biobanks displayed similar and strong enrichment of high-PIP (> 0.9) variants in seven main distinct functional categories (defined as non-overlapping regions; **2.4 Methods**): predicted loss-of-function (pLoF), missense, synonymous, 5'/3' UTR, promoter, and *cis*-regulatory element (CRE) regions (DNase I hypersensitive sites [DHS] and H3K27ac [186]; **Supplementary Fig. B.1a–h**; **Supplementary Table B.5**). In addition, our combined results recapitulated the functional enrichments of 35 additional annotations as previously reported [95,161,187,188], including conserved regions in mammals [116,189] and ancient putative promoter/enhancer [190]; these enrichments remained significant even when analysis is restricted to the "non-genic" variants that do not belong to any of the seven main functional categories listed above (**Supplementary Fig. B.1i**; **Supplementary Table B.6**).

We additionally performed eQTL colocalization in BBJ and FinnGen, using fine-mapped *cis*-eQTLs from GTEx [67,103] v8 and eQTL catalog [191] release 4, identifying 719 variant-trait-gene triples; in our companion paper [67], we identified 4,420 triples in UKBB. We aggregated these results into a combined 4,957 unique variant-trait-gene triples in which the variant was fine-mapped for both the trait and expression of the gene (colocalized posterior probability [CLPP] $=$ $\text{PIP}_{\text{GWAS}} \times \text{PIP}_{\text{cis-eQTL}} > 0.1$), spanning 117 traits and 3,937 genes (**Fig. 2.1f**; **Supplementary Table B.7**). We defined the rate of colocalization as the proportion of variants with PIP > 0.1 in each biobank that showed at least one *cis*-eQTL colocalization (CLPP > 0.1 across any trait, gene, or tissue) in our study; this rate was 5.3%, 5.6%, and 7.3% for BBJ, FinnGen, and UKBB, respectively. We investigated the MAF distribution of colocalized variants in each biobank and observed that 85%,

74%, and 89% of colocalized variants showed MAF > 5% in BBJ, FinnGen, and UKBB, respectively (**Fig. 2.1g**). This is in contrast to the coding variants with PIP > 0.1, of which 56%, 42%, and 55% had MAF > 5% in BBJ, FinnGen, and UKBB, respectively (**Fig. 2.1h**).

2.2.2 High-PIP variants are largely non-overlapping across populations

We set out to investigate what proportion of variants with PIP > 0.9 in one population are associated or fine-mapped in other populations. Fine-mapping methods employ a model in which there are a small number of causal variants driving the association signal at the locus, all of which are measured without error, and there are no uncorrected confounding or non-linear effects. When the model is perfectly specified and inference is perfectly accurate, we would expect, for example, 90% of variants with PIP = 0.9 to be truly causal; however, this will not always be the case. We systematically classified variants based on several hierarchical criteria (**Fig. 2.2a**; **2.4 Methods**). First, what proportion of high-PIP (PIP > 0.9) variants in one population (the "discovery population") reach genome-wide significance ($P_{\text{GWAS}} < 5.0 \times 10^{-8}$) in either of the other two ("secondary") populations, permitting a well-powered comparison of fine-mapping results at the same locus. Second, of these variants where association is strongly replicated, what proportion have replicated fine-mapping, defined by the same variant having PIP > 0.1 in the secondary population (that is, the variant is also fine-mapped in the second population, though at a lower threshold of confidence). For this analysis, we utilized only the 26 traits analyzed in all three cohorts.

Out of 646 unique variant-trait pairs with PIP > 0.9 in at least one of the three populations, we found that 45% (291 / 646) achieved genome-wide significance ($P_{\text{GWAS}} < 5.0 \times 10^{-8}$) in at least one of the other two populations (**Fig. 2.2b**). Of these, we found that 55% (160 / 291) had replicating fine-mapping (PIP > 0.1) in at least one of the other two populations, while the other 45% (131 / 291) did not (PIP ≤ 0.1). We took the proportion of fine-mapping replication (= # replication / [# replication + # non-replication] among the variants reaching $P_{\text{GWAS}} < 5.0 \times 10^{-8}$) and defined it

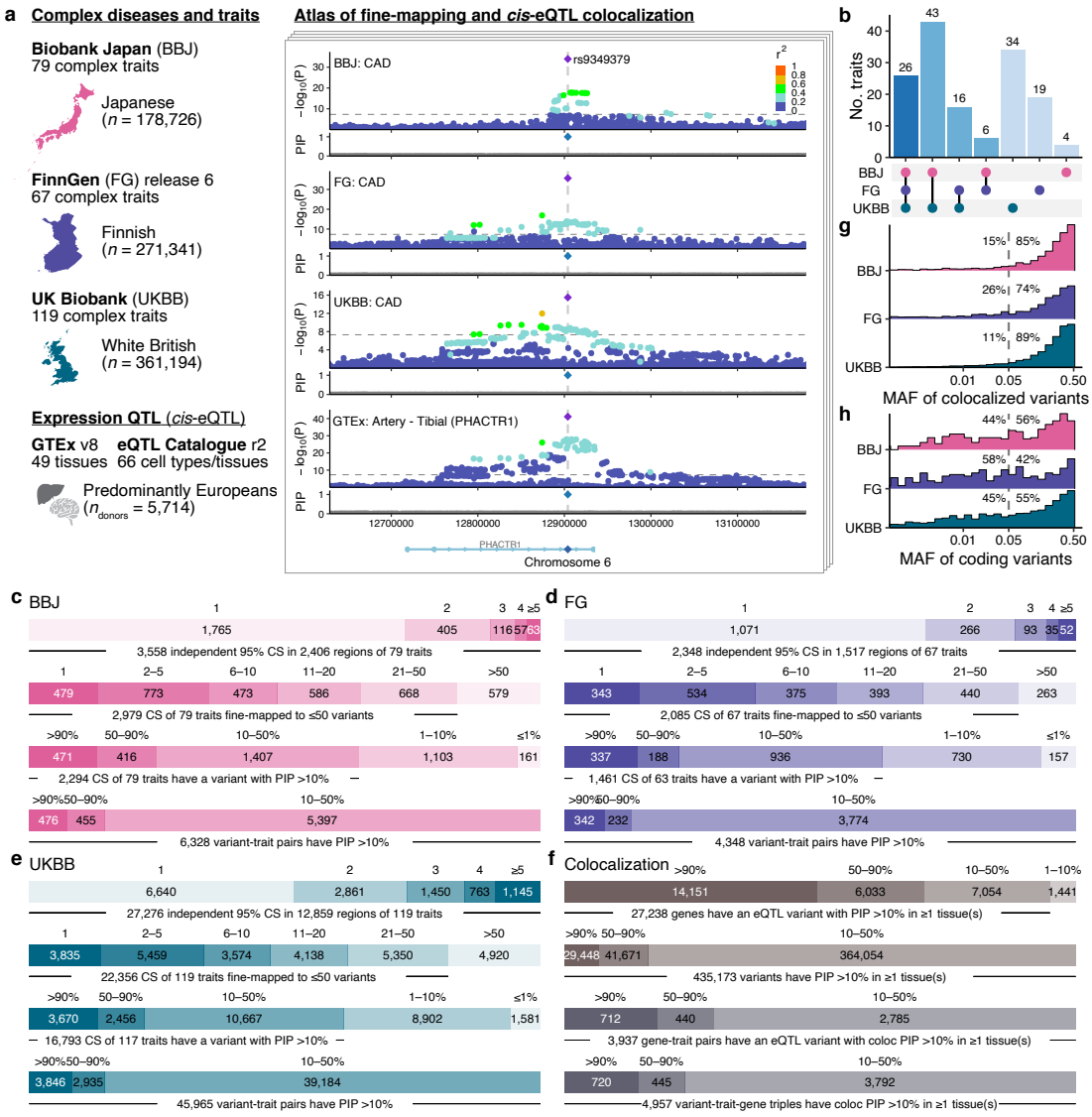


Figure 2.1: Expanded atlas of putative causal variants across three populations. **a**. Overview of the studied cohorts and *cis*-eQTL datasets. As an illustrative example, the 6p24.1 locus was shown for coronary artery disease (CAD) association in BBJ, FinnGen, and UKBB with *cis*-eQTL association of *PHACTR1* in tibial artery from GTEx. **b**. Number of traits shared across the cohorts. **c–e**. For each cohort, number of independent 95% CS per region, number of fine-mapped variants per 95% CS, number of 95% CS binned by the best PIP variant in each CS, and number of fine-mapped variants binned by PIP. All numbers are counted against unique trait pairs. **f**. (Top two rows) number of genes or variants binned by the best $PIP_{cis\text{-}eQTL}$ across tissues. (Bottom two rows) number of gene-trait pairs or variant-trait-gene triples binned by the best CLPP across tissues. **g**. MAF distribution of colocalized variants (the best CLPP > 0.1) in each biobank. **h**. MAF distribution of coding variants (the best PIP > 0.1) in each biobank. Labels represent proportions of variants with MAF > 5% and ≤ 5% in each biobank.

as the cross-biobank fine-mapping replication rate. This proportion was relatively consistent across all the pairs of populations, ranging from 38% to 57% (**Fig. 2.2b**). The cross-biobank fine-mapping replication rate was relatively insensitive to the specific threshold, increasing only slightly when considering a fine-mapping result to be replicated if it had PIP > 0.05 or was in a 95% credible set, as opposed to PIP > 0.1 (**Supplementary Fig. B.2a,b**). While mean PIP in a secondary population was positively correlated with PIP in the discovery population, the underlying distribution of PIP in the secondary populations were bimodal, particularly for variants with PIP > 0.9 in the discovery population (**Supplementary Fig. B.2c–e**).

To further interpret these observations, we simulated GWAS and fine-mapping in a secondary population as if fine-mapped variants in a discovery population were all and only true causal variants with the same causal effect sizes (estimated posterior effect sizes in a discovery population; **2.4 Methods**). In our simulations, we observed a substantially higher cross-biobank replication rate compared to real data (**Supplementary Fig. B.2f,g**; **2.4 Methods**). The observed inconsistency of cross-biobank fine-mapping replication rates in real data and simulations could be explained by lack of calibration in real data fine-mapping, overestimated causal effect sizes in a discovery population (*i.e.*, winner's curse[192]) used in simulations, and/or another complex discrepancy not well-simulated. In examining specific examples in real data, we found that lack of replication was sometimes due to differences in LD structure and effect sizes across populations that lower power in the secondary population, or likely non-causal variants that nonetheless achieve high PIP in the discovery population, as expected given the PIP threshold of 0.9. We illustrated a few examples in **Supplementary Fig. B.3**.

Of the remaining 55% (355 / 646) of variant-trait pairs that did not reach genome-wide significance ($P_{\text{GWAS}} < 5.0 \times 10^{-8}$) in either of the secondary populations, 42% (150 / 355) had an association that replicated at the more permissive threshold of $P_{\text{GWAS}} < 0.01$ (**Fig. 2.2c**), suggesting the association is present but at a level insufficient to perform fine-mapping reliably. An additional

14% (51 / 352) had high power to detect association (power > 0.9 for achieving $P_{GWAS} < 0.01$; Methods) in at least one of the secondary populations, assuming the same causal effect size from the discovery population and a standard linear regression, but were not associated at $P_{GWAS} < 0.01$ in either population. These variants may include causal variants with heterogeneous effect sizes across populations (likely due to differences in phenotyping and ascertainment) or false positive variants that nonetheless achieved high PIP in the discovery population (which is not unexpected given the number of traits studied and the PIP threshold of 0.9). A few causal variants would also be expected not to reach this threshold due only to random sampling, even with equal effect sizes and estimated power of 0.9. We note that three variant-trait pairs had replicated fine-mapping (PIP > 0.1) but not genome-wide significance in either of the secondary populations (Note that these are in genome-wide significant loci). Lastly, 42% (151 / 355) had low power or were missing from the GWAS summary statistics due mostly to differences in allele frequencies across populations (**Fig. 2.2c**; **B.1 Supplementary Note**). This proportion was different for different pairs of populations, ranging from 19% (UKBB and FinnGen) to 62% (BBJ and UKBB). Importantly, our results indicate that these missing causal variants are undiscoverable through standard GWAS fine-mapping in other populations, re-emphasizing the desperate need for data generation in diverse populations.

We further confirmed strong functional enrichment of our fine-mapped variants with replication compared to those non-replicated (**Supplementary Fig. B.2h**). Therefore, for the remainder of this manuscript, we mainly focus on several subsets of PIP > 0.9 variants with highest confidence: fine-mapped variants replicated in multiple populations, coding variants with PIP > 0.9, and genes supported by multiple fine-mapped variants.

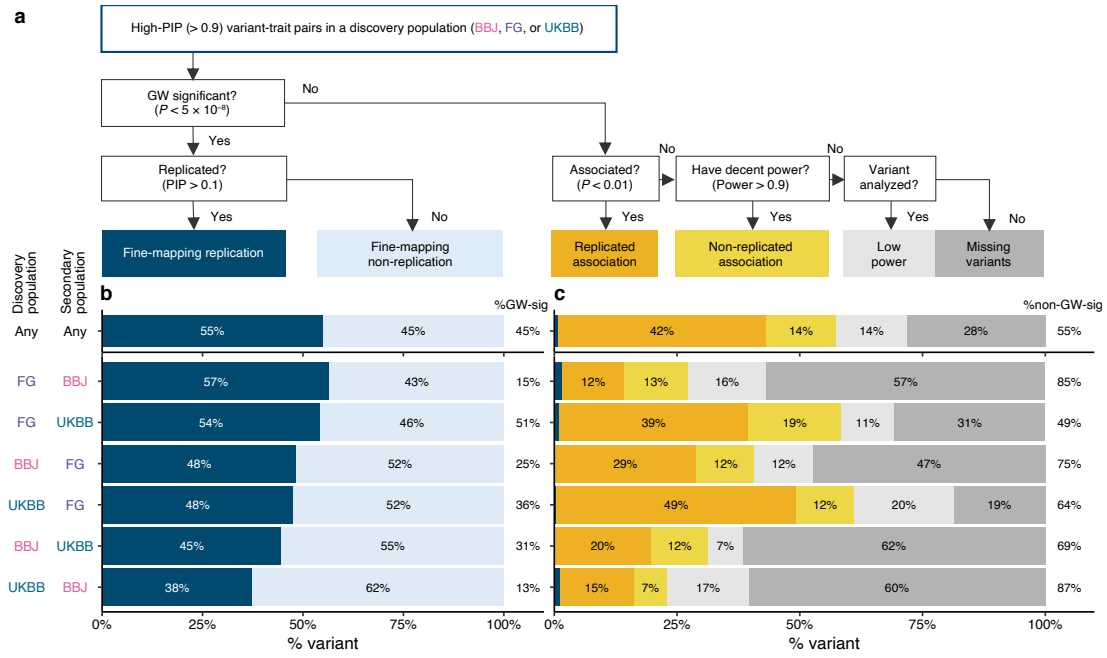


Figure 2.2: Overview of replication status for high-PIP fine-mapped variants across populations. a. Schematic flowchart of our classification criteria. Starting from the high-PIP (> 0.9) variant-trait pairs in a discovery population, we categorized each pair into the six categories: fine-mapping replication, fine-mapping non-replication, replicated association, non-replicated association, low power, and missing variants (**2.4 Methods**). **b,c**. Barplots showing a fraction of the high-PIP (> 0.9) variant-trait pairs identified in each discovery population, stratified by the above replication categories tested in the other two secondary populations. Labels in the bar represent a proportion for each category, while labels on the right represent a proportion of the genome-wide significant and non-genome-wide significant variant-trait pairs. **b**. Breakdowns for the genome-wide significant variant-trait pairs ($P_{\text{GWAS}} < 5.0 \times 10^{-8}$) in a secondary population. **c**. Breakdowns for the non genome-wide significant variant-trait pairs ($P_{\text{GWAS}} \geq 5.0 \times 10^{-8}$) in a secondary population. Note that there were three variant-trait pairs in total that had replicated fine-mapping (PIP > 0.1) but not genome-wide significance in either of the secondary populations (dark blue).

2.2.3 Common putative causal variants implicate shared biological mechanisms across populations

Restricting to 91 traits available in two or more populations, we identified 285 high-confidence variant-trait pairs (204 unique variants including 56 variants that are only polymorphic in Europeans) that achieve replicated fine-mapping across multiple populations analyzed (PIP > 0.9 in at least one population and PIP > 0.1 in at least one of the others; **Supplementary Tables B.8,B.9**). We observed 100% directional consistency for posterior effect sizes between populations (P for sign test = 2.7×10^{-107}). These replicated fine-mapped variants represent a set of common putative causal variants (**Supplementary Fig. B.4a,b**) with the highest confidence in our dataset, providing excellent candidates for functional characterization and therapeutic targets.

We observed a significant enrichment of coding variants in high-confidence variant-trait pairs: of the 285 high-confidence variant-trait pairs, 94 pairs (60 unique variants) are coding variants (**Supplementary Table B.8**), whereas 4 pairs would be expected by chance (Fisher's exact test $P < 0.05$). These variants include well-known pLoF and missense variants such as rs429358 (*APOE* ε4-tagging missense variant) for Alzheimer's disease [193]; rs2066847 (*NOD2*: p.Leu980ProfsTer2) for Crohn's disease [194,195]; rs855791 (*TMPRSS6*: p.Val736Ala) for blood hemoglobin levels and erythrocyte volume [196]; rs2642438 (*MARC1*: p.Ala165Thr) for alkaline phosphatase [197,198]; and rs4149056 (*SLCO1B1*: p.Val174Ala) for total bilirubin [199]. Notably, we found that rs9379084 (*RREB1*: p.Asp1171Asn) showed PIP > 0.9 for height in every population; this variant was previously implicated for type 2 diabetes [7] but not for height. We also found that a common synonymous variant rs55714927 on *ASGR1* (canonical transcript ENST00000269299.3) was fine-mapped for alkaline phosphatase in both BBJ and UKBB (PIP = 1.0 for both; **Supplementary Fig. B.5a**). The same variant was significantly associated with other traits in our dataset, such as albumin, cholesterol levels, and sex hormone binding globulin (**Supplementary Fig. B.5b**). *ASGR1* was

previously reported for having a rare non-coding 12-base-pair deletion within intron 4 (del12; c.284-36_283+33delCTGGGGCTGGGG, NM_001671.4; MAF = 0.41% in 398,000 Icelanders), which was associated with a reduced risk of coronary artery disease (CAD), lowering LDL cholesterol, and increasing alkaline phosphatase and vitamin B12 levels[200]. However, the reported del12-tagging variant rs186021206 is independent from the synonymous variant rs55714927 ($r^2 = 0.001$ in Europeans) and is monomorphic in East Asians, implying that the del12 variant does not contribute to the identified rs55714927 association here. Instead, we observed rs55714927 has a significant splicing QTL effect in GTEx liver[103] ($P = 2.4 \times 10^{-46}$) for the same isoform as del12 (**Supplementary Fig. B.5c,d**).

We also characterized 191 non-coding variant-trait pairs (144 unique variants) with replicated fine-mapping as described above (**Supplementary Table B.9**). These variants are primarily located within CRE (48%) followed by promoter (16%) and 3' UTR (8%) regions, and are enriched for predicted *cis*-regulatory expression modifier score[138], suggesting that most of these variants act through transcriptional or by post-transcriptional regulation (**Supplementary Fig. B.4b–d**). In total, we identified 48 out of 144 putative causal non-coding variants that co-localized with *cis*-eQTL associations (CLPP > 0.1 in at least one tissue; **Supplementary Table B.9**), including well-known variants, *e.g.*, rs2070895 (intronic variant of *LIPC*) for HDL cholesterol; and rs78378222 (3' UTR variant of *TP53*) for skin cancer; as well as under-characterized variants, *e.g.*, rs1497406 (intergenic variant, 22 kb upstream of *EPHA2*) for γ-glutamyl transferase; and rs34778241 (intronic variant of *EIF4E3*) for loss of Y chromosome (**Supplementary Fig. B.6**). Notably, we identified a well-known intronic variant rs9349379 in *PHACTR1* that was fine-mapped for CAD in every population (**Fig. 2.1b**; PIP = 1.0; MAF = 0.35, 0.45, and 0.41 for BBJ, FinnGen, and UKBB, respectively). This intronic variant also co-localized with a fine-mapped *cis*-eQTL association of *PHACTR1* in GTEx tibial artery[103] (CLPP = 1.0), consistent with previous work[201]. We note that it was previously demonstrated that rs9349379 also regulates expression of *EDN1* (located 600 kb upstream of *PHACTR1*)

in CRISPR-edited endothelial cells[202–204].

The 144 putative causal non-coding variants also included seven intergenic variants located in gene deserts; *i.e.*, that are more than 250 kb away from the closest gene[205] (**Supplementary Table B.10**). For example, rs77541621 and rs183373024 (349 kb and 322 kb upstream of *POU5F1B*, respectively) were fine-mapped for prostate cancer (PIP = 1.0 in FinnGen and UKBB), and are located within the 8q24 locus, a well-known gene desert associated with many complex diseases[206,207] (**Supplementary Fig. B.7a**). These variants are one of the 12 independent variants for prostate cancer that were previously identified at the 8q24 locus, but the exact functional mechanism of each variant is still under active investigation[208]. Other examples include rs1434282 (284 kb downstream of *PTPRC*) for mean corpuscular volume, rs116376456 (269 kb downstream of *IRS1*) for height, and rs35009121 (1.2 Mb downstream of *GATA3*) for serum calcium levels (**Supplementary Fig. B.7b–d**). Although these loci are also known as gene deserts, none of the fine-mapped variants are well-characterized in the current literature, nor do they overlap with enhancer-gene mappings predicted by the activity-by-contact (ABC) model[209].

We also found nine examples where a variant was fine-mapped in every population even though it was not significantly associated in every population. Five of these were significant at a more permissive threshold of $P < 1.0 \times 10^{-5}$, but in other cases the marginal effect sizes were substantially lower, due to LD with another causal variant(s). For example, rs244711 (4.7 kb upstream of *FGFR4*) is consistently fine-mapped for height but not significantly associated in BBJ (marginal $\beta = 9.0 \times 10^{-3}$; $P = 4.1 \times 10^{-4}$; **Supplementary Fig. B.8a–d**). We found that rs244711 is partially correlated with a nearby fine-mapped missense variant rs1966265 (*FGFR4*: p.Val10Ile) in every population (r^2 = 0.14, 0.08, and 0.13 in BBJ, FinnGen, and UKBB, respectively) but the correlation is only negative in BBJ ($r = -0.37$). The causal effect of rs244711 is thus partially cancelled out by the tagged effect of rs1966265 in BBJ, where the correlation between the two variants is negative, but not in UKBB and FinnGen, where the correlation is positive, leading to a non-significant asso-

ciation in BBJ but significant associations in UKBB and FinnGen. Another example is rs1801706 (3' UTR variant of *CETP*), which is consistently fine-mapped for HDL cholesterol but not significantly associated in BBJ (marginal $\beta = 6.0 \times 10^{-3}$; $P = 0.43$; **Supplementary Fig. B.8e–h**). This is owing to partial correlation with Japanese-enriched splice donor and missense variants rs5742907 (c.1321+1G>A) and rs2303790 (p.Asp459Gly). These two variants showed large effect sizes (marginal $\beta = 0.76$ and 0.39; $P = 4.9 \times 10^{-122}$ and 5.5×10^{-206}; respectively) and are negatively correlated with rs1801706 in BBJ ($r = -0.03$ and -0.06, respectively; this corresponds to -16.6 and -56.3 decrease in marginal χ^2 statistics of rs1801706 by partial tagging). These examples illustrate that, when a region contains multiple independent associations, differences in LD between two sites can create differences in the marginal effect size and observed association in univariate analyses between populations.

2.2.4 Identification of population-enriched putative causal variants

Given that a substantial number of the variants with high PIP (> 0.9) in one population are rare/absent (and therefore undiscoverable) in the other populations (**Fig. 2.2c**), we investigated allele frequency (AF)-enriched variants from the two bottlenecked populations included in our study, Finland[210,211] and Japan[212,213]. To quantify AF enrichment (AFE) in the Finnish and Japanese populations, we used the gnomAD[214] v2 and the GEM-J WGS[215] to compute a ratio of AF in Japanese vs. non-Japanese-Korean East Asians (NJKEA) for BBJ and in Finnish vs. non-Finnish-Swedish-Estonian Europeans (NFSEE) for FinnGen (**2.4 Methods**).

Past studies have noted that variants stochastically boosted through a bottleneck are enriched for functional categories[183–185,216–218]. Consistent with these previous studies, we found that there were significantly more variants with AFE > 10 than with AFE < 1/10 in both FinnGen and BBJ, and that variants with AFE > 10 were enriched for coding variants (2.2- and 4.8-fold enrichment over variants with AFE ≤ 10; **2.4 Methods**). Of 140,416 and 91,564 coding variants tested

in FinnGen and BBJ GWAS, 29,656 (21%) and 14,802 (16%) showed AFE > 10 in the Finnish or Japanese population, respectively (**Fig. 2.3a,b**). Furthermore, high-PIP (> 0.9) coding variants were significantly more likely to have high AFE than low-PIP (≤ 0.01) coding variants (**Fig. 2.3c,d**; Fisher's exact test $P < 0.05$; **2.4 Methods**); and showed substantially younger estimated allele age based on GEVA[219] (**Fig. 2.3e,f**). These observations are consistent with recent bottleneck events and negative selection on the putative causal variants studied here, because deleterious variants boosted in frequency through these bottlenecks have had insufficient time to be brought back down in frequency by selection[220,221].

Notably, we identified seven pLoF variants and 40 missense high-PIP (> 0.9) variants with extreme AF enrichment (> 10-fold) in BBJ or FinnGen (**Table 2.1**). These variants are more likely to be deleterious and impactful given their extreme enrichment. Indeed, the list includes several known pathogenic variants or genes in related autosomal recessive disorders. For example, rs75326924, a Japanese-enriched missense variant (p.Pro90Ser) on *CD36* is a known pathogenic variant for platelet glycoprotein IV (CD36) deficiency (PIP = 1.0 for platelet count; MAF = 0.047 in GEM-J WGS), contributing to high prevalence of CD36 deficiency in Japanese (2–3%)[222]; and rs386833873, a Finnish-enriched frameshift variant (p.Leu41AspfsTer50) on *NPHS1* is a well-known causal variant for the congenital nephrotic syndrome of the Finnish type (PIP = 1.0 for nephrotic syndrome; MAF = 0.011 in gnomAD Finnish)[223]. Interestingly, we found two novel population-enriched deleterious variants on *PLOD2*, fine-mapped for height: i) a Japanese-enriched missense variant rs148051196 (p.Gln553Arg; PIP = 1.0; MAF = 7.3×10^{-3} in GEM-J WGS) and ii) a Finnish-specific stop-gained variant rs201501322 (p.Ser166Ter; PIP = 0.58; MAF = 1.9×10^{-3} in gnomAD FIN). *PLOD2* is a known recessive gene for Bruck syndrome 2 (osteogenesis imperfecta with congenital joint contractures; OMIM: 609220)[224]. We identified additional population-enriched variants for height in 27 genes, including known recessive genes such as *ADAMTS17* (causal gene for Weill-Marchesani syndrome 4; OMIM: 613195) and IHH (brachydactyly type A1; OMIM:

112500). Furthermore, we identified fine-mapped variants on genes that were not previously implicated, such as rs199935580 (*THBS3*: p.Arg520Trp; MAF = 1.0 $\times$ 10^{-3} in gnomAD FIN) fine-mapped for carpal tunnel syndrome (PIP = 1.0); rs191692991 (*LUM*: p.Arg310Cys; MAF = 5.5 $\times$ 10^{-3} in gnomAD FIN) fine-mapped for fibroblastic disorders (PIP = 1.0); and rs200939713 (*POF1B*: p.Arg339Trp; MAF = 1.7 $\times$ 10^{-3} in gnomAD FIN) fine-mapped for varicose veins (PIP = 0.99). Detailed biological annotations of each gene are summarized in the **B.1 Supplementary Note**.

On the other hand, the high-PIP non-coding variants were not significantly more likely to have high AFE than low-PIP non-coding variants (**Supplementary Fig. B.9**; Fisher's exact test $P > 0.05$), partly because non-coding variants tend to be less deleterious and thus less likely to undergo strong negative selection. However, we identified 23 population-enriched (> 10-fold) high-PIP (> 0.9) non-coding variants that are independent of population-enriched coding variants ($r^2 < 0.1$) in each population (**Supplementary Table B.11**). While we are not able to replicate these population-enriched variants in other populations due to low AF, we identified several variants that might have biological significance. For example, a Finnish-enriched rs748670681 in an intron of *TNRC18* (MAF = 0.042 in gnomAD FIN) is fine-mapped for inflammatory bowel disease (IBD) and psoriasis (PIP = 1.0). Despite very significant association in FinnGen (P = 6.2 $\times$ 10^{-69} for IBD), this locus was not previously reported, and its biological function is not well-characterized.

2.2.5 Allelic series of putative causal variants across populations

Given that many fine-mapped variants are population-specific, we aggregated results across populations to identify genes harboring fine-mapped coding variants for one or more traits. Overall, we identified 1,492 unique putative causal pLoF/missense variants (best PIP > 0.1) that mapped onto 1,113 genes (**Supplementary Table B.12**). Of these genes, 240 have two or more putative causal pLoF/missense variants located on the same gene, and 113 have variants identified from mul-

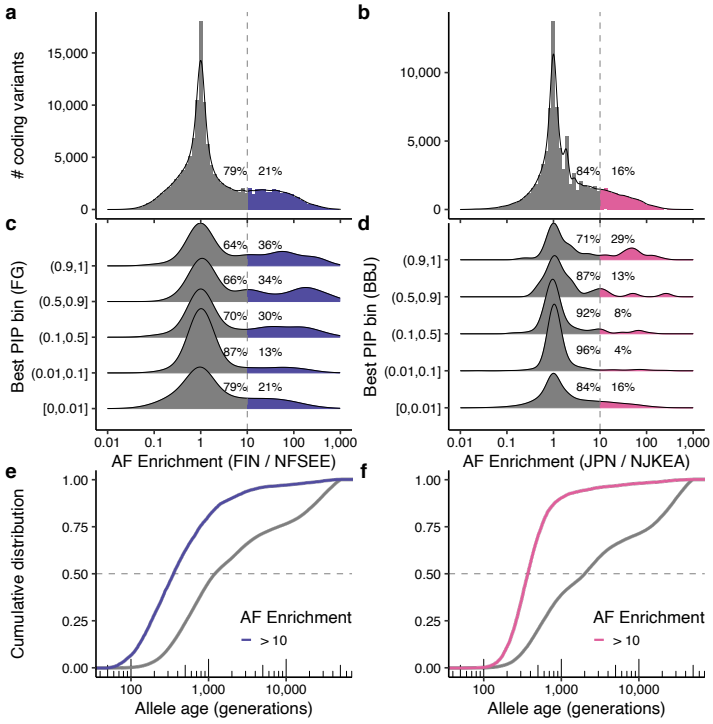


Figure 2.3: Population-enriched putative causal coding variants. a–d. Histograms showing a distribution of allele frequency (AF) enrichment metric in (**a**) Finnish (n = 10,824) and (**b**) Japanese (n = 7,609) populations. A ratio of AF was computed against NFSEE (n = 43,697) and NJKEA (n = 7,212) for coding variants analyzed in BBJ or FinnGen GWAS that exist in gnomAD WES or GEM-J WGS. For a subset of variants that are fine-mapped in our analysis (see **2.4 Methods**), we show AF enrichment distribution across maximum PIP bins computed in (**c**) FinnGen or (**d**) BBJ. **e–f.** Cumulative distribution of estimated allele age for coding variants, stratified by AF enrichment in (**e**) Finnish or (**f**) Japanese. FIN: Finnish, JPN: Japanese, NFSEE: Non-Finnish-Swedish-Estonian European, NJKEA: Non-Japanese-Korean East Asian.

Table 2.1: Population-enriched putative causal coding variants. Nonsynonymous coding variants (PIP > 0.9) with AFE > 10 in the Japanese or Finnish populations are shown.

Variant	rsid	Gene	Consequence	AF (pop)	AF (ref)	AF enrichment	Best PIP	Fine-mapped traits (PIP >0.9)
BBJ								
1:21890590:G:A	rs199669988	*ALPL*	Missense	0.015	0.00035	43.1	1	ALP
1:55505604:G:A	rs564427867	*PCSK9*	Missense	0.012	NA	Inf	1	LDLC, TC
2:21242731:G:A	rs13306206	*APOB*	Missense	0.039	0.00049	79.5	1	LDLC, MI, TC
2:44051573:T:TA	rs142037828	*ABCG5*	Splice region	0.051	0.00042	123.2	1	Cholelithiasis
2:120231070:C:G	rs3731600	*SCTR*	Missense	0.048	0.00069	69.2	0.99	T2D
2:219919943:C:T	rs200216644	*IHH*	Missense	0.0036	0.00027	13.4	0.99	Height
3:145794588:T:C	rs148051196	*PLOD2*	Missense	0.0073	0.00014	52.4	1	Height
4:6303731:G:A	rs147834269	*WFS1*	Missense	0.053	0.0012	43.7	1	T2D
6:158484904:G:C	rs141160611	*SYNJ2*	Missense	0.032	0.0025	12.9	0.96	GGT
7:45954540:C:T	rs17847676	*IGFBP3*	Missense	0.0061	6.90E-05	88.5	1	Height
7:80286003:C:T	rs75326924	*CD36*	Missense	0.047	0.00083	56.9	1	Plt
8:118184855:A:T	rs770224130	*SLC30A8*	Missense	0.0062	0.00014	44.7	0.97	T2D
9:107593923:C:G	rs754040394	*ABCA1*	Missense	0.0019	NA	Inf	1	HDLC, TC
11:64361219:G:A	rs121907892	*SLC22A12*	Stop gained	0.021	0.00042	51.6	1	UA
11:116661394:G:C	rs201229911	*APOA5*	Missense	0.01	NA	Inf	1	HDLC, TG
12:16510581:GAA:G	rs779999476	*MGST1*	Frameshift	0.015	NA	Inf	1	HDLC
12:109690842:C:T	rs17848833	*ACACB*	Missense	0.0032	NA	Inf	1	HDLC, TG
16:57016150:G:A	rs5742907	*CETP*	Splice donor	0.0037	NA	Inf	1	HDLC, TC
16:84872195:G:C	rs965984074	*CRISPLD2*	Missense	0.00086	NA	Inf	0.99	Height
17:7462468:G:T	rs201860460	*TNFSF13*	Missense	0.0022	7.00E-05	31.1	1	AG, NAP
17:48545926:C:A	rs201158957	*CHAD*	Missense	0.0081	6.90E-05	116.1	1	Height
17:78358945:G:A	rs112735431	*RNF213*	Missense	0.01	0.00049	21.2	1	CAD, MAP, PP, SBP
19:11241988:C:T	rs13306505	*LDLR*	Missense	0.0085	0.00021	41.1	1	LDLC, TC
19:42855705:G:A	rs200485103	*MEGF8*	Missense	0.0039	NA	Inf	0.95	Glucose
19:46178043:G:T	rs13306398	*GIPR*	Missense	0.02	6.90E-05	286.9	1	BMI, BW
20:44507112:G:A	rs139396693	*ZSWIM3*	Missense	0.015	0.00014	106.6	1	MCV
FG								
1:21890632:G:A	rs121918007	*ALPL*	Missense	0.017	0.0011	15.7	0.94	Urolithiasis
1:155170392:G:A	rs199935580	*THBS3*	Missense	0.001	1.10E-05	88.9	1	Carpal_Tunnel_Syndrome
1:192779303:G:T	rs201233692	*RGS2*	Missense	0.0088	1.10E-05	761.5	0.96	Statin
4:120528397:C:T	rs202226125	*PDE5A*	Missense	0.007	1.20E-05	612	1	Height
5:1272362:G:A	rs770066110	*TERT*	Stop gained	0.00052	NA	Inf	1	IPF
5:1279485:T:C	rs776981958	*TERT*	Missense	0.0016	NA	Inf	0.96	IPF
6:155450779:A:G	rs148543891	*TLAM2*	Missense	0.031	6.90E-05	455.3	1	Height
9:35609378:C:T	rs777777413	*TESK1*	Missense	0.0025	2.40E-05	101.4	1	Height
9:136501728:C:T	rs77273740	*DBH*	Missense	0.051	0.0023	21.7	1	Hypertension
10:13040400:A:G	rs199848893	*CCDC3*	Missense	0.0021	NA	Inf	1	Height
11:36248678:T:TG	rs767680853	*LDLRAD3*	Frameshift	0.0019	2.30E-05	82.3	1	Height
12:6882498:C:A	rs149722682	*LAG3*	Missense	0.00061	NA	Inf	1	AID, Hypothyroidism
12:91498031:G:A	rs191692991	*LUM*	Missense	0.0053	1.20E-05	450.7	1	Fibroblastic_Disorders, Height
14:100134609:G:A	rs201483470	*HHIPL1*	Missense	0.0093	0.00013	74.2	0.97	Height
15:28228553:C:T	rs74653330	*OCA2*	Missense	0.048	0.0014	33.4	1	Malignant_Neoplasms, SkC
15:101569374:C:T	rs41531245	*LRRK1*	Missense	0.0076	0.00073	10.4	1	Fibroblastic_Disorders, Inguinal_Hernia
17:56436130:C:T	rs199598395	*RNF43*	Missense	0.012	5.00E-05	239.7	1	Iron_Deficiency_Anaemia
17:60493445:C:T	rs552441218	*EFCAB3*	Stop gained	0.001	6.90E-05	14.7	0.98	Depression_medications
19:36342510:CAG:C	rs386833873	*NPHS1*	Frameshift	0.011	2.40E-05	473.3	1	Nephrotic_Syndrome
19:58421417:ACT:A	rs774674736	*ZNF417*	Frameshift	0.0018	4.60E-05	39.8	0.93	Chronic_Tonsillitis
X:84563165:G:A	rs200939713	*POF1B*	Missense	0.0016	NA	Inf	0.99	Varicose_Veins

tiple populations (**Fig. 2.4a**). The genes with the most putative causal pLoF/missense variants include *APOB* (13 missense variants; the loss-of-function observed/expected upper bound fraction [LOEUF][214] = 0.46), *TFR2* (1 pLoF and 6 missense variants; LOEUF = 0.77), and *PIEZO1* (7 missense variants; LOEUF = 0.58); despite containing many variants that impact human phenotypes, these genes are modestly constrained (**Fig. 2.4b**, **Supplementary Fig. B.10a,b**).

Out of 1,113 genes with at least one fine-mapped pLoF/missense variant (PIP > 0.1), 176 genes had multiple independent pLoF/missense variants with associations to the same trait(s), forming an allelic series (**Fig. 2.4c**, **Supplementary Table B.13**). We found that 69 of these genes contained variants fine-mapped in multiple populations, of which 34 genes had at most one variant per population, and so were only conclusively implicated when multiple populations were analyzed. The cross-population allelic series include *e.g.*, *ABCG2*, a known pathogenic gene for gout, where we identified two pLoF/missense variants (p.Gln126Ter and p.Phe489Leu) in BBJ, two missense variants (p.Asp620Asn and p.Ala528Thr) in UKBB, and one missense variant (p.Gln141Lys) in BBJ, FinnGen, and UKBB (**Supplementary Fig. B.10c**).

We further investigated allelic series including both coding and non-coding variants with associations to the same trait(s), assuming that non-coding causal variants proximal to deleterious coding variants (< 100 kb) might act through regulation of the same gene[142]. This facilitates understanding of unknown non-coding functions and enables us to identify allelic series for an additional 195 genes through coding/non-coding allelic series, of which 87 genes contained variants fine-mapped across multiple populations (**Supplementary Table B.14**). For example, we identified coding/non-coding allelic series around *EPX* (eosinophil peroxidase) for eosinophil count (**Supplementary Fig. B.10d**), where we found European-specific missense variant rs149610649 (*EPX*: p.Phe308Leu; MAF = 0.083 in gnomAD NFE) and Japanese-specific intergenic variant rs536070968 (MAF = 0.011 in GEM-J WGS). The intergenic variant rs536070968 is located 33 kb downstream of *EPX* and 11 kb upstream of LPO (lactoperoxidase), an ortholog of EPX, illustrating the value of allelic

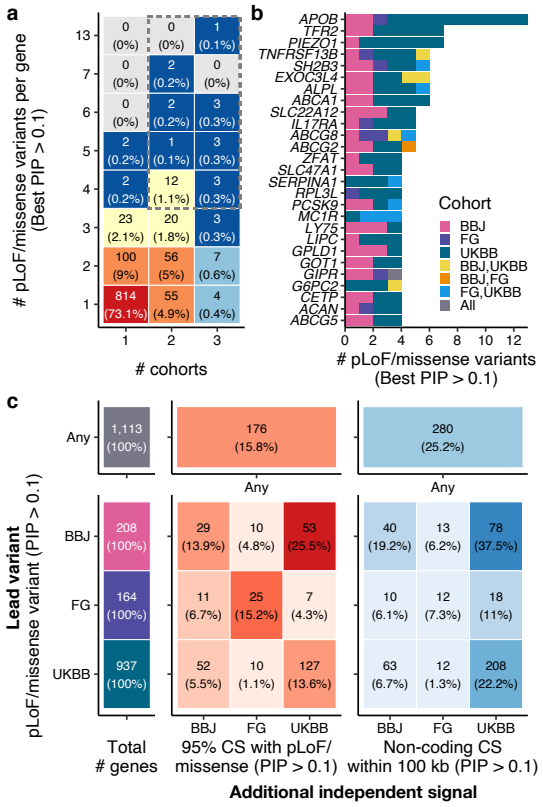


Figure 2.4: Allelic series of putative causal variants across multiple populations. a. Number of fine-mapped pLoF/missense variants (best PIP > 0.1) per gene, stratified by the number of cohorts identified. **b.** Top list of genes that have four or more fine-mapped pLoF/missense variants (best PIP > 0.1). **c.** Number of genes with fine-mapped pLoF/missense variants (PIP > 0.1), stratified by a discovery cohort. For genes with at least one fine-mapped pLoF/missense variant, we counted how many of them contained additional independent 95% CS with pLoF/missense and non-coding variants (PIP > 0.1) for the same gene and trait in each cohort.

series across multiple populations to assign a potential causal gene from nearby genes.

2.3 Discussion

In this study, we performed statistical fine-mapping in BBJ and FinnGen, and aggregated these results with our parallel fine-mapping of UKBB[67], providing an extensive list of candidate causal variants for 148 complex diseases and traits across diverse populations. By integrating fine-mapped

variants from deeply-phenotyped biobanks and eQTL studies, we expanded both the depth and breadth of the resource to explore biological mechanisms of complex traits at single-variant resolution, with replication across multiple populations and colocalization with different tissues. We make these resources publicly available for the community to further accelerate variant prioritization and characterization.

Examination of fine-mapping from three biobanks enabled the identification of 285 high-confidence variant-trait pairs that are replicated across multiple populations. However, the majority of high-PIP (> 0.9) variants are non-overlapping across populations. Many of the variants with high PIP in one population but not in the other two populations were trivially explained by the fact they are rare or monomorphic in the other two populations. However, our simulations suggest that more fine-mapping results would have been replicated if all high-PIP variants had been truly causal; we see further exploration of this phenomenon as an important direction for future work. The abundance of population-enriched variants exemplifies the significant value of diverse populations in fine-mapping studies, contributing to identification of population-specific discoveries and deeper allelic series of multiple variants at the same locus across populations. Our comparison of fine-mapping results across biobanks guides interpretation of these results.

Our study has several limitations that suggest directions for future work. First, the current fine-mapping methods rely on modeling assumptions that are not all met in real-world fine-mapping (*e.g.*, no genotyping or imputation errors). While we have focused here on a high confidence subset of results—high-PIP variants that replicate across biobanks, and fine-mapped coding variants—we see further exploration of potential misspecification of fine-mapping models as an important area for future work. Second, our sample sizes are still limited, especially for non-European populations, emphasizing the desperate need for more diversity in human genetics. Here, we were powered to fine-map variants with large or moderate effect sizes; more samples will be required to fine-map causal variants with small effect sizes. Moreover, molecular data from non-European samples are

vastly limited, which fundamentally inhibits variant interpretation of population-enriched variants. Third, systematic differences in study design, genotyping and imputation across cohorts limited our ability to integrate data from the three biobanks. We see method development for cross-population fine-mapping that takes into account this heterogeneity as an important direction for future work.

To our knowledge, this study provides the largest and the most comprehensive comparison of fine-mapping results from multiple large-scale biobanks of diverse ancestries. Although these data still remain limited to identify common but small-effect causal variants shared across populations, we have demonstrated that the use of diverse populations facilitates the identification of high-confidence causal variants shared across populations, population-enriched fine-mapped variants, and allelic series of high-impact variants across populations. With fast-evolving biobanks and high-throughput assays under development, our atlas of candidate causal variants provides a valuable resource for future functional characterization efforts.

2.4 Methods

2.4.1 Study cohorts

BioBank Japan (BBJ)

The BioBank Japan (BBJ) is a hospital-based cohort that collected DNA, serum, and clinical information of approximately 200,000 individuals from 66 hospitals in Japan between 2003 and 2007. All the study participants had been diagnosed with one or more of 47 target diseases by physicians at the cooperating hospitals. Written informed consent was obtained from all the participants, as approved by the ethics committees of the RIKEN Center for Integrative Medical Sciences, and the Institute of Medical Sciences, the University of Tokyo. Details of study design, sample collection, and baseline clinical information were described elsewhere [170,225].

We genotyped samples using i) the Illumina HumanOmniExpressExome BeadChip or ii) a com-

bination of the Illumina HumanOmniExpress and the HumanExome BeadChip. We applied standard quality-control criteria for samples and variants as described elsewhere[226] (summarized in **Supplementary Table B.1**). We analyzed 178,726 individuals of Japanese ancestry, chosen based on sample selection criteria using principal component analysis (PCA)[171]. The genotypes were prephased using Eagle[227] and imputed using Minimac3[228] with a reference panel that consists of the 1000 Genomes Project Phase 3 (version 5) samples (n = 2,504)[229] and whole-genome sequencing (WGS) data of Japanese individuals (n = 1,037)[230]. We excluded variants with low imputation quality (Rsq ≤ 0.7) and used 13,531,752 variants in this study. All the variants were processed on the human genome assembly GRCh37.

We defined phenotypes based on clinical information retrieved from medical records and interviews using a standardized questionnaire. Detailed phenotype definitions are described elsewhere[171] and summarized in **Supplementary Table B.2**.

FinnGen

FinnGen is a public-private partnership project combining genotype data from Finnish biobanks and digital health records from Finnish health registries[172]. This study used the Data Freeze 6 which contains 271,341 individuals of Finnish ancestry. Patients and control subjects in FinnGen provided informed consent as described in **B.1 Supplementary Note**). Detailed characteristics of the cohort are described in our companion paper[172].

Samples were primarily genotyped using the FinnGen ThermoFisher Axiom custom array. The samples from legacy cohorts have previously been genotyped using various generations of Illumina GWAS arrays. The genotypes were prephased using Eagle 2.3.5 and imputed using Beagle 4.1 with a reference panel of Finnish WGS data, the SISu v3 reference panel (n = 3,775). We applied post-imputation quality control as described in our companion paper[172], excluding variants with INFO < 0.6 and MAF < 0.001, and used 16,311,902 variants in our study. All the variants were originally

processed on the human genome assembly GRCh38, and lifted over to GRCh37 for comparison with other cohorts used in this study.

Clinical endpoints were defined based on medical records from multiple national health registries. Detailed phenotype definitions are described in our companion paper[172] and summarized in **Supplementary Table B.2**.

UK Biobank (UKBB)

The UK Biobank (UKBB) is a population-based cohort that recruited approximately 500,000 individuals in the United Kingdom between 2006 and 2010. This study analyzed a set of 366,194 unrelated "white British" individuals defined previously in the Neale Lab GWAS (https://github.com/Nealelab/UK_Biobank_GWAS). The individuals of British ancestry were determined by the PCA-based sample selection criteria (https://github.com/Nealelab/UK_Biobank_GWAS/blob/master/ukb31063_eur_selection.R), and were further filtered to self-reported "white British", "Irish", or "white". The UK Biobank analysis was conducted via application number 31063. The cohort characteristics were extensively described elsewhere[76].

Genotyping was performed using either i) the Applied Biosystems UK BiLEVE Axiom Array or ii) UKB Axiom Array. The genotypes were imputed using IMPUTE4 with a combination of reference panels: i) the Haplotype Reference Consortium and ii) UK10K and the 1000 Genomes Phase 3. We retained 13,791,467 variants with INFO > 0.8, MAF > 0.001, and Hardy-Weinberg equilibrium P value $> 1.0 \times 10^{-10}$, with exception for the VEP-annotated coding variants where we allowed MAF $> 1.0 \times 10^{-6}$. The detailed quality-control criteria were described in the Neale Lab GWAS (https://github.com/Nealelab/UK_Biobank_GWAS). All the variants were processed on the human genome assembly GRCh37.

We derived phenotypes based on multiple data sources available in UKBB, *e.g.*, biomarkers, body measures, and disease case-control status mapped on phecodes[157] (https://phewascatalog.org/

phecodes). Detailed phenotype definitions are described in our companion paper[67] and summarized in **Supplementary Table B.2**.

2.4.2 Genome-wide association analysis

We performed GWAS using a generalized linear mixed model as implemented in SAIGE[108] (for binary traits; v.0.37 or later) or BOLT-LMM[109,159] (for quantitative traits; v.2.3.4) with age, sex, top principal components, and other study-specific covariates as detailed in **Supplementary Table B.1**. We excluded sex-adjusting covariates from sex-specific or stratified traits (*i.e.*, age at menarche/menopause, breast cancer, testosterone levels, and uterine fibroid; **Supplementary Table B.2**). For mosaic loss of chromosome Y, we used summary statistics publicly available from BBJ[231] and UKBB[160].

2.4.3 Statistical fine-mapping

We conducted statistical fine-mapping using FINEMAP[28,29] v.1.3.1 and SuSiE[30] v.0.9.1 with GWAS summary statistics and in-sample dosage LD. We defined fine-mapping regions based on a 3 Mb window around each lead variant and merged regions if they overlapped. We excluded the major histocompatibility complex (MHC) region (chr 6: 25–36 Mb) from analysis due to extensive LD structure in the region. Allowing up to 10 causal variants per region, we derived up to 10 independent 95% credible sets (CS) and posterior inclusion probabilities (PIP) of each variant using the default uniform prior probability of causality. The 95% CS reported by FINEMAP and SuSiE each have 95% posterior probability of containing a causal variant; in a locus with multiple causal variants identified, there will be multiple CS. This definition of CS differs from the definition given in Hormozdiari *et al.*[77], in which each CS has 95% posterior probability of containing all causal variants in a locus. We computed in-sample dosage LD using LDstore v.2.0 (ref. [113]).

We combined fine-mapping results from the two methods by taking an average of PIP, excluding variants with a substantial PIP difference (> 5%) to further improve fine-mapping accuracy. We justify our approach based on functional enrichment analysis that demonstrates that the variants with inconsistent PIP across the methods show little functional enrichment (as described in our companion paper[67]). If either fine-mapping method failed (*e.g.*, due to conversion failure or available memory restrictions), we used successful results from the other method. If both of the methods failed, we excluded these regions from analysis.

To define independent CS merged across populations, we merged SuSiE 95% CS from each population using hierarchical clustering based on the weighted Jaccard similarity index. Briefly, we computed the PIP-weighted Jaccard similarity index between all the pairs of CS for the same trait identified from each cohort. For a pair of CS, we computed the similarity index as $\sum_i \min(x_i, y_i) / \sum_i \max(x_i, y_i)$ where x_i and y_i are PIP values (or zero if missing) in each CS for the same variant i. We then used $1 -$ the similarity index as a distance to conduct hierarchical clustering of the CS using the complete linkage method. We cut a dendrogram tree at a height of 0.9 so that any two CS with PIP-weighted Jaccard similarity above 0.1 are merged into a single CS.

2.4.4 Colocalization

We conducted colocalization of fine-mapped variants from complex trait and *cis*-eQTL associations. Based on fine-mapping results from complex trait and *cis*-eQTL, we computed a posterior inclusion probability of colocalization for a variant as a product of PIP for GWAS and for *cis*-eQTL ($\text{CLPP} = \text{PIP}_{\text{GWAS}} \times \text{PIP}_{\text{cis-eQTL}}$)[77]. We assembled fine-mapping results of *cis*-eQTL associations from GTEx[103] v8 (detailed in our companion paper[67]) and eQTL catalogue[191] release 4, both of which used the same or the functionally-equivalent fine-mapping pipelines to our GWAS fine-mapping (see **2.6 Code availability**). All the variants were originally processed on the human genome assembly GRCh38 and lifted over to GRCh37 to colocalize with GWAS results in this

study.

2.4.5 Functional enrichment

We performed functional enrichment analysis for fine-mapped variants from each population. We first defined seven main distinct functional categories: pLoF (predicted loss-of-function), missense, synonymous, 5' UTR, 3' UTR, promoter, *cis*-regulatory element (CRE), and non-genic. We assign fine-mapped variants to these categories in the sequential order so that each category is non-overlapping from each other. Variant-based categories (pLoF, missense, synonymous, and 5'/3' UTR variants) are defined based on the most severe consequence for a variant on a canonical transcript, predicted by the Ensembl Variant Effect Predictor (VEP)[168] v85 (using GRCh37 and GENCODE v19). The pLoF category represents stop-gained, splice site disrupting, and frameshift variants predicted as high-confidence by LOFTEE[214]. The missense category includes missense-like variants such as low-confidence LoF. Region-based categories (promoter and CRE) are defined using region-based annotations. The promoter annotation is retrieved from the baseline annotations in Finucane & Bulik-Sullivan *et al.*[95], originally from the UCSC Genome Browser[232] and post-processed by Gusev et al[97]. The CRE annotation is defined as intersection of DNase I hypersensitive sites (DHS) and H3K27ac regions from the Roadmap Epigenomics Project[85], ChIP-Atlas[84], Meuleman *et al.*[80], Domcke, *et al.*[86], Corces *et al.*[81], Buenrostro, *et al.*[233], and Calderon, *et al.*[83], reprocessed in our companion paper[67]. Lastly, the non-genic category represents any variants that do not belong to the other six categories. In addition, we annotated each variant using 35 binary annotations from the baselineLD v2.2 model[188].

For each variant, we computed the maximum PIP across traits in BBJ, FinnGen, UKBB, and all cohorts combined. We estimated functional enrichment for each category as a relative risk (*i.e.*, a ratio of proportion of variants) between being in an annotation and fine-mapped (PIP ≤ 0.01 or PIP > 0.9). That is, a relative risk = (proportion of variants with PIP > 0.9 that are in the annotation) /

(proportion of variants with PIP ≤ 0.01 that are in the annotation). The 95% confidence intervals are calculated using bootstrapping with 5,000 replicates.

2.4.6 Fine-mapping replication analysis

To investigate fine-mapping replication, we systematically evaluated the consistency of fine-mapping results across populations for the 26 traits analyzed in all three populations (**Supplementary Table B.2**), using all six pairs of discovery population and distinct secondary population. Starting from high-PIP (> 0.9) variant-trait pairs in the discovery population, we first split them by whether the association is genome-wide significant ($P < 5.0 \times 10^{-8}$) in the secondary population, and then categorized each pair into the following categories, based on criteria evaluated in the secondary population:

For genome-wide significant ($P < 5.0 \times 10^{-8}$) variant-trait pairs,

1. Pairs for which the fine-mapping result is replicated (PIP > 0.1).
2. Pairs for which the fine-mapping is not replicated (PIP ≤ 0.1)

For non-genome-wide significant ($P \geq 5.0 \times 10^{-8}$) variant-trait pairs,

3. Pairs for which the association is replicated ($P < 0.01$).
4. Pairs for which the association is not replicated ($P \geq 0.01$) but the variant is included in the study and has decent statistical power (estimated power > 0.9 for achieving $P = 0.01$). We estimated statistical power via the non-centrality parameter (NCP) of the chi-square distribution[2,3,4]. We defined NCP $= 2f(1-f)n\beta^2$ where f is MAF, n is the effective sample size, and β is a posterior effect size estimated by SuSiE in the discovery population. Here, we assumed the variant has the same causal effect size in a second population. For quantitative traits, effective sample size equals the number of samples. For binary traits, effective sample

size is calculated via $n\varphi(1-\varphi)$ where φ is the number of cases divided by the number of total samples. We note that this power estimation does not account for linear mixed models adopted by BOLT-LMM or SAIGE.

5. Whether the variant is analyzed in the study (i.e., exists in summary statistics). The missingness is mainly due to low frequency or monomorphism (non-existence) in the secondary population, which is described in the Supplementary Note.

The schematic flowchart of this process is illustrated in **Fig. 2.2a**. We note that there could be a case where non-genome-wide significant variant-trait pairs ($P \geq 5.0 \times 10^{-8}$) in a secondary population still had fine-mapping replication (PIP > 0.1) due to LD with another causal variant(s).

Simulation

To further investigate cross-biobank fine-mapping replication rates, we simulated GWAS and fine-mapping based on our fine-mapping results for the 26 traits analyzed in all three populations. Here, we assumed our fine-mapped variants in a discovery population were all and only true causal variants with the same allelic-scale causal effect sizes in a secondary population. For each pair of discovery and secondary populations, we obtained all the non-zero posterior mean effect size estimates b from SuSiE in a discovery population and computed true simulated phenotypes y in a secondary population via $y = Xb + \varepsilon$ where X is a dosage genotype matrix and ε is a random noise variable which follows $N(0, 1 - \text{var}(Xb))$. We performed GWAS using BOLT-LMM [109,159] v.2.3.4 with the same covariates used in the real data analysis. Note that we used BOLT-LMM even when the original phenotypes are binary because our true simulated phenotypes are always continuous as defined above.

We then conducted statistical fine-mapping using the exact same pipeline used in the real data analysis and investigated whether high-PIP (> 0.9) variant-trait pairs in a discovery population

showed simulated PIP > 0.1 in a secondary population. We note that we obtained simulation replication status for each variant-trait-discovery cohort trio, instead of each variant-trait pair, because the true causal effect sizes that we simulated are dependent on the discovery cohorts. To make an apple-to-apple comparison of the cross-biobank fine-mapping replication rates between real data and simulations, we made a slight modification to how we count the number of fine-mapping replications. Here, only for this analysis, we instead counted the numbers of variant-trait-discovery cohort trios for fine-mapping replication and non-replication. This modification slightly increased the cross-biobank replication rate in real data from 55% to 60% (**Supplementary Fig. B.2g**) but did not affect our conclusion.

2.4.7 High-confidence and low-confidence fine-mapping results

We annotated high-confidence and low-confidence high-PIP (> 0.9) variant-trait pairs for the 91 traits analyzed in two or more populations (**Supplementary Table B.3**). High-confidence pairs are defined as having PIP > 0.9 in at least one population and PIP > 0.1 in all the other populations analyzed in this study. Low confidence pairs are defined as having PIP > 0.9 in one population and $P < 5.0 \times 10^{-8}$ but PIP ≤ 0.1 and not in 95% CS in one of the other populations. Those categorized otherwise (*e.g.*, population-specific variants) were not assigned either annotation.

2.4.8 Allele frequency enrichment

To identify population-enriched variants, we defined allele frequency (AF) enrichment metrics as a ratio of pseudo AF between ancestral and founder populations. To do this, we retrieved allele counts from gnomAD[214] v2 and GEM-J WGS[215]. To account for finite sample sizes, we computed pseudo AF by constantly adding one to allele count (AC), *i.e.*, pseudo AF = (AC + 1) / allele number. Due to the disparity in available sample sizes between gnomAD v2 exomes and genomes, we

computed enrichment metrics separately for coding and non-coding variants using exomes and genomes, respectively. Coding and non-coding variants are defined as having VEP-predicted coding consequences or not (see the previous section).

For coding variants, we used gnomAD v2 exomes for the Finnish (n = 10,824), non-Finnish-Swedish-Estonian Europeans (NFSEE; n = 43,697), and non-Japanese-Korean East Asians (NJKEA; n = 7,212). For non-coding variants, we used gnomAD v2 genomes for the Finnish (n = 1,738), NFSEE (n = 5,421), and NJKEA (n = 780). We used the GEM-J WGS for both coding and non-coding variants, which contains WGS data from the Japanese population (n = 7,609). To account for coverage differences across data sources, we excluded regions from GEM-J WGS with a median coverage < 10 in gnomAD exomes or genomes. To eliminate non-coding enrichment due to tagging coding variants, we excluded non-coding variants in LD ($r^2 > 0.1$) with coding variants using gnomAD v2 LD matrices for the Finnish and East Asian populations. We restricted our analysis to 140,416 and 91,564 coding variants and 11,732,074 and 9,539,454 non-coding variants tested in FinnGen and BBJ GWAS, respectively. To annotate estimated allele age, we retrieved point estimates of allele age (mode of the composite posterior distribution) from the Genealogical Estimation of Variant Age (GEVA)[219].

2.4.9 Allelic series analysis

We investigated an allelic series of fine-mapped variants within and across populations. We first took nonsynonymous coding variants (pLoF and missense predicted by VEP as described in the previous section) that had PIP > 0.1 for at least one of the studied traits. We then counted the number of these variants falling in each gene, identified allelic series of two or more such variants in a single gene for the same trait, and categorized allelic series according to whether they were discoverable in a single population or only by combining data across populations. Furthermore, we investigated non-coding variants that are proximal to these fine-mapped nonsynonymous coding variants (< 100 kb),

assuming they might act through the same gene.

2.5 Data availability

The fine-mapping results produced by this study will be publicly available at https://www.finucanelab.org/data. The BBJ summary statistics are available at the National Bioscience Database Center (NBDC) Human Database (accession code: hum0197) and at the GWAS catalog[18] (https://www.ebi.ac.uk/gwas/home). They are also browsable at our PheWeb[235] website (https://pheweb.jp/). The BBJ genotype data is accessible on request at the Japanese Genotype-phenotype Archive (http://trace.ddbj.nig.ac.jp/jga/index_e.html) with accession code JGAD00000000123 and JGAS00000000114. The UKBB summary statistics will be available at the ENCODE data portal (https://www.encodeproject.org/) and at the GWAS catalog[18] (https://www.ebi.ac.uk/gwas/home). The UKBB individual-level data is accessible on request through the UK Biobank Access Management System (https://www.ukbiobank.ac.uk/). The UKBB analysis in this study was conducted via application number 31063. The FinnGen release 6 was used in this study and is still subject to embargo according to the FinnGen consortium agreement; thus the FinnGen summary statistics are available on request (https://www.finngen.fi/en/access_results) and are being prepared for public release by Q4 2021. The GTEx v8 summary statistics is available at the GTEx Portal (https://gtexportal.org/home/datasets). The GTEx individual-level data is accessible on request through the dbGAP application (accession code: phs000424.v8.p2; https://gtexportal.org/home/protectedDataAccess). The eQTL catalogue results are available at https://www.ebi.ac.uk/eqtl/Data_access/.

2.6 Code availability

Our fine-mapping pipeline is available at https://github.com/mkanai/finemapping-pipeline, and the code to perform all analyses and generate the figures is provided at https://github.com/mkanai/finemapping-insights. Custom fine-mapping pipelines for FinnGen is available at https://github.com/FINNGEN/finemapping-pipeline and for eQTL catalogue is available at https://github.com/eQTL-Catalogue/susie-workflow; both of which has implemented functionally-equivalent pipelines with a dataset-specific custom code.

2.7 Acknowledgements

We acknowledge all the participants of BioBank Japan, FinnGen, and UK Biobank. We thank all the members of Finucane and Daly labs for their helpful feedback. This study was supported as an ENCODE Functional Characterization Center (UM1HG009435). The BioBank Japan Project was supported by the Tailor-Made Medical Treatment program of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), the Japan Agency for Medical Research and Development (AMED). The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Celgene Corporation, Celgene International II Sàrl, Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, and Novartis AG. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_

Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi). Finnish Biobank Cooperative -FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. M.Kanai was supported by a Nakajima Foundation Fellowship and the Masason Foundation. Y.O. was supported by JSPS KAKENHI (19H01021, 20K21834), and AMED (JP21km0405211, JP21ek0109413, JP21ek0410075, JP21gm4010006, and JP21km0405217), JST Moonshot R&D (JPMJMS2021, JPMJMS2024), Takeda Science Foundation, and Bioinformatics Initiative of Osaka University Graduate School of Medicine, Osaka University. H.K.F. was funded by NIH grant DP5 OD024582 and by Eric and Wendy Schmidt.

2.8 Author contributions

M.Kanai, M.J.D, and H.K.F. designed the study. M.Kanai, J.C.U., J.K., M.Kurki, K.J.K., Q.S.W., and H.J. performed analyses. E.B.F. contributed biological interpretation. F.A., K.G.A., N.K., K.A., K.I., S.S., Y.K., K.M., A.P., and Y.O. contributed data acquisition. C.B., S.R., A.P., B.M.N., R.T., P.C.S., and Y.O. contributed ideas and insights. M.J.D. and H.K.F jointly supervised this work. H.K.F., P.C.S., and M.Kanai obtained funding. M.Kanai, M.J.D., and H.K.F. wrote the manuscript with input from all authors.

The work presented in this chapter will be published as Kanai, M. *et al.* and was posted online as a preprint (*medRxiv*, 2022)[69].

3

Meta-analysis fine-mapping is often miscalibrated at single-variant resolution

Abstract

Meta-analysis is pervasively used to combine multiple genome-wide association studies (GWAS) into a more powerful whole. To resolve causal variants, meta-analysis studies typically apply summary statistics-based fine-mapping methods as they are applied to single-cohort studies. However, it is unclear whether heterogeneous characteristics of each cohort (*e.g.*, ancestry, sample size, phenotyping, genotyping, or imputation) affect fine-mapping calibration and recall. Here, we first demonstrate that meta-analysis fine-mapping is substantially miscalibrated in simulations when different genotyping arrays or imputation panels are included. To mitigate these issues, we propose a summary statistics-based QC method, SLALOM, that identifies suspicious loci for meta-analysis fine-mapping by detecting outliers in association statistics based on ancestry-matched local LD structure. Having validated SLALOM performance in simulations and the GWAS Catalog, we applied it to 14 disease endpoints from the Global Biobank Meta-analysis Initiative and found that 68% of loci showed suspicious patterns that call into question fine-mapping accuracy. These predicted suspicious loci were significantly depleted for having likely causal variants, such as nonsynonymous variants, as a lead variant (2.8x; Fisher's exact $P = 6.2 \times 10^{-4}$). Compared to fine-mapping results in individual biobanks, we found limited evidence of fine-mapping improvement in the GBMI meta-analyses. Although a full solution requires complete synchronization across cohorts, our approach identifies likely spurious results in meta-analysis fine-mapping. We urge extreme caution when interpreting fine-mapping results from meta-analysis.

3.1 Introduction

Meta-analysis is pervasively used to combine multiple genome-wide association studies (GWAS) from different cohorts[236]. Previous GWAS meta-analyses have identified thousands of loci associated with complex diseases and traits, such as type 2 diabetes[7,8], schizophrenia[11,237], rheumatoid arthritis[9,10], body mass index[146], and lipid levels[238]. These meta-analyses are typically conducted in large-scale consortia (*e.g.*, the Psychiatric Genomics Consortium [PGC], the Global Lipids Genetics Consortium [GLGC], and the Genetic Investigation of Anthropometric Traits [GIANT] consortium) to increase sample size while harmonizing analysis plans across participating cohorts in every possible aspect (*e.g.*, phenotype definition, quality-control [QC] criteria, statistical model, and analytical software) by sharing summary statistics as opposed to individual-level data, thereby avoiding data protection issues and variable legal frameworks governing individual genome and medical data around the world. The Global Biobank Meta-analysis Initiative (GBMI)[239] is one such large-scale, international effort, which aims to establish a collaborative network spanning 19 biobanks from four continents (total n = 2.1 million) for coordinated GWAS meta-analyses, while addressing the many benefits and challenges in meta-analysis and subsequent downstream analyses.

One such challenging downstream analysis is statistical fine-mapping[12–14]. Despite the great success of past GWAS meta-analyses in locus discovery, individual causal variants in associated loci are largely unresolved. Identifying causal variants from GWAS associations (*i.e.*, fine-mapping) is challenging due to extensive linkage disequilibrium (LD, the correlation among genetic variants), the presence of multiple causal variants, and limited sample sizes, but is rapidly becoming achievable with high confidence in individual cohorts[67,68,87,118] owing to the recent development of large-scale biobanks[76,170,172] and scalable fine-mapping methods[28–30] that enable well-powered, accurate fine-mapping using in-sample LD from large-scale individual-level data.

After conducting GWAS meta-analysis, previous studies[7,10,52,176–178,238,240–242] have applied ex-

isting summary statistics-based fine-mapping methods (*e.g.*, approximate Bayes factor [ABF][23,24], CAVIAR[25], PAINTOR[26,27], FINEMAP[28,29], and SuSiE[30]) just as they are applied to single-cohort studies, without considering or accounting for the unavoidable heterogeneity among cohorts (*e.g.* differences in sample size, phenotyping, genotyping, or imputation). Such heterogeneity could lead to false positives and miscalibration in meta-analysis fine-mapping (**Fig. 3.1**). For example, case-control studies enriched with more severe cases or ascertained with different phenotyping criteria may disproportionately contribute to genetic discovery, even when true causal effects for genetic liability are exactly the same between these studies and less severe or unascertained ones. Quantitative traits like biomarkers could have phenotypic heterogeneity arising from different measurement protocols and errors across studies. There might be genuine biological mechanisms too, such as gene–gene (GxG) and gene–environment (GxE) interactions and (population-specific) dominance variation (*e.g.*, rs671 and alcohol dependence[243]), that introduce additional heterogeneity across studies[64,244]. In addition to phenotyping, differences in genotyping and imputation could dramatically undermine fine-mapping calibration and recall at single-variant resolution, because differential patterns of missingness and imputation quality across constituent cohorts of different sample sizes can disproportionately diminish association statistics of potentially causal variants. Finally, although more easily harmonized than phenotyping and genotyping data, subtle differences in QC criteria and analytical software may further exacerbate the effect of heterogeneity on fine-mapping.

An illustrative example of such issues can be observed in the *TYK2* locus (19p13.2) in the recent meta-analysis from the COVID-19 Host Genetics Initiative (COVID-19 HGI; **Supplementary Fig. C.1**)[21]. This locus is known for protective associations against autoimmune diseases[9,240], while a complete *TYK2* loss of function results in a primary immunodeficiency[245]. Despite strong LD ($r^2 = 0.82$) with a lead variant in a locus (rs74956615; $P = 9.7 \times 10^{-12}$), a known functional missense variant rs34536443 (p.Pro1104Ala) that reduces *TYK2* function[246,247] did not achieve genome-wide significance ($P = 7.5 \times 10^{-7}$), primarily due to its missingness in two more cohorts

than rs74956615. This serves as just one example of the major difficulties with meta-analysis fine-mapping at single-variant resolution. Indeed, the COVID-19 HGI cautiously avoided an in-silico fine-mapping in the flagship to prevent spurious results[21].

Only a few studies have carefully addressed these concerns in their downstream analyses. The Schizophrenia Working Group of PGC, for example, recently updated their largest meta-analysis of schizophrenia[11] (69,369 cases and 236,642 controls), followed by a downstream fine-mapping analysis using FINEMAP[28]. Unlike many other GWAS consortia, since PGC has access to individual-level genotypes for a majority of samples, they were able to apply standardized sample and variant QC criteria and impute variants using the same reference panel, all uniformly processed using the RICOPILI pipeline[248]. This harmonized procedure was crucial for properly controlling inter-cohort heterogeneity and thus allowing more robust meta-analysis fine-mapping at single-variant resolution. Furthermore, PGC's direct access to individual-level data enabled them to compute in-sample LD matrices for multiple causal variant fine-mapping, which prevents the significant miscalibration that results from using an external LD reference[67,87,118]. A 2017 fine-mapping study of inflammatory bowel disease also benefited from access to individual-level genotypes and careful pre- and post-fine-mapping QC[73]. For a typical meta-analysis consortium, however, many of these steps are infeasible as full genotype data from all cohorts is not available. For such studies, a new approach to meta-analysis fine-mapping in the presence of the many types of heterogeneity is needed. Until such a method is developed, QC of meta-analysis fine-mapping results deserves increased attention.

While existing variant-level QC procedures are effective for limiting spurious associations in GWAS (**C.1 Supplementary Note**)[249], they do not suffice for ensuring high-quality fine-mapping results. In some cases, they even hurt fine-mapping quality, because they can i) cause or exacerbate differential patterns of missing variants across cohorts, and ii) remove true causal variants as well as suspicious variants. Thus, additional QC procedures that retain consistent variants across cohorts for consideration but limit poor-quality fine-mapping results are needed. A recently proposed

method called DENTIST[250], for example, performs summary statistics QC to improve GWAS downstream analyses, such as conditional and joint analysis (GCTA-COJO[114]), by removing variants based on estimated heterogeneity between summary statistics and reference LD. Although DENTIST was also applied prior to fine-mapping (FINEMAP[28]), simulations only demonstrated that it could improve power for detecting the correct number of causal variants in a locus, not true causal variants. This motivated us to develop a new fine-mapping QC method for better calibration and recall at single-variant resolution and to demonstrate its performance in large-scale meta-analysis.

Here, we first demonstrate the effect of inter-cohort heterogeneity in meta-analysis fine-mapping via realistic simulations with multiple heterogeneous cohorts, each with different combinations of genotyping platforms, imputation reference panels, and genetic ancestries. We propose a summary statistics-based QC method, SLALOM (suspicious loci analysis of meta-analysis summary statistics), that identifies suspicious loci for meta-analysis fine-mapping by detecting association statistics outliers based on local LD structure, building on the DENTIST method. Applying SLALOM to 14 disease endpoints from the Global Biobank Meta-analysis Initiative[239] as well as 467 meta-analysis summary statistics from the GWAS Catalog[18], we demonstrate that suspicious loci for fine-mapping are widespread in meta-analysis and urge extreme caution when interpreting fine-mapping results from meta-analysis.

3.2 Results

3.2.1 Large-scale simulations demonstrate miscalibration in meta-analysis fine-mapping

Existing fine-mapping methods[23,28,30] assume that all association statistics are derived from a single-cohort study, and thus do not model the per-variant heterogeneity in effect sizes and sample sizes

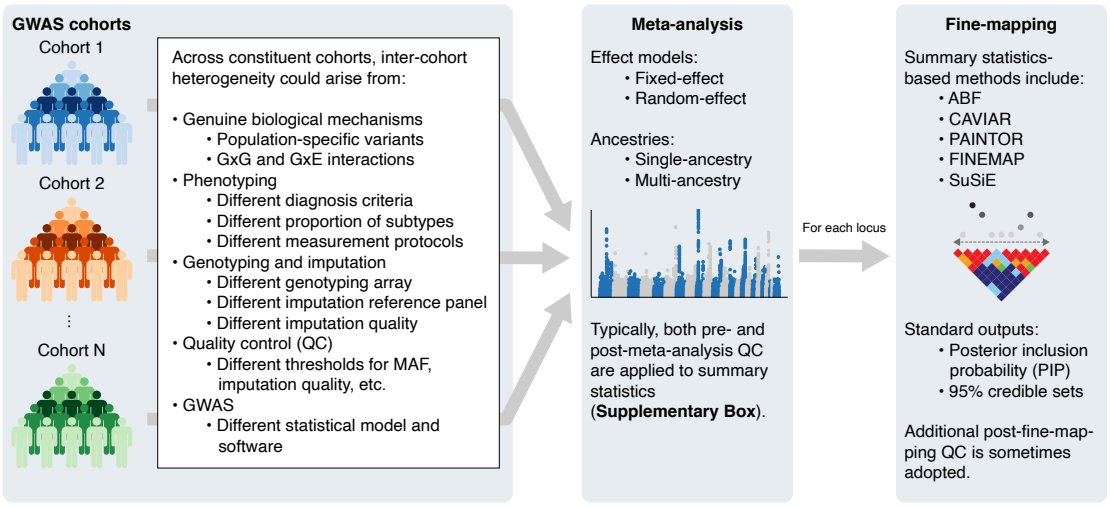


Figure 3.1: Schematic overview of meta-analysis fine-mapping.

that arise when meta-analyzing multiple cohorts (**Fig. 3.1a**). To evaluate how different characteristics of constituent cohorts in a meta-analysis affect fine-mapping calibration and recall, we conducted a series of large-scale GWAS meta-analysis and fine-mapping simulations (**Supplementary Tables C.1–C.4**; **3.4 Methods**). Briefly, we simulated multiple GWAS cohorts of different ancestries (10 European ancestry, one African ancestry and one East Asian ancestry cohorts; n = 10,000 each) that were genotyped and imputed using different genotyping arrays (Illumina Omni2.5, Multi-Ethnic Global Array [MEGA], and Global Screening Array [GSA]) and imputation reference panels (the 1000 Genomes Project Phase 3 [1000GP3][229], the Haplotype Reference Consortium [HRC][251], and the TOPMed[252]). For each combination of cohort, genotyping array, and imputation panel, we conducted 300 GWAS with randomly simulated causal variants that resemble the genetic architecture of a typical complex trait, including minor allele frequency (MAF) dependent causal effect sizes[111], total SNP heritability[112], functional consequences of causal variants[68], and levels of genetic correlation across cohorts (*i.e.*, true effect size heterogeneity; r_g = 1, 0.9, and 0.5; see **3.4 Methods**). We then meta-analyzed the single-cohort GWAS results across 10 independent

cohorts based on multiple configurations (different combinations of genotyping arrays and imputation panels for each cohort) to resemble realistic meta-analysis of multiple heterogeneous cohorts (**Supplementary Table C.4**). We applied ABF fine-mapping to compute a posterior inclusion probability (PIP) for each variant and to derive 95% and 99% credible sets (CS) that contain the smallest set of variants covering 95% and 99% of probability of causality. We evaluated the false discovery rate (FDR, defined as the proportion of variants with PIP > 0.9 that are non-causal) and compared against the expected proportion of non-causal variants if the meta-analysis fine-mapping method were calibrated, based on PIP. More details of our simulation pipeline are described in **3.4 Methods** and visually summarized in **Supplementary Fig. C.2**.

We found that FDR varied widely over the different configurations, reaching as high as 37% for the most heterogeneous configurations (**Fig. 3.2**). We characterized the contributing factors to the miscalibration. We first found that lower true effect size correlation r_g (*i.e.*, larger phenotypic heterogeneity) always caused higher miscalibration and lower recall. Second, when using the same imputation panel (1000GP3), use of less dense arrays (MEGA or GSA) led to moderately inflated FDR (up to FDR = 11% vs. expected 1%), while use of multiple genotyping array did not cause further FDR inflation (**Fig. 3.2a**). Third, when using the same genotyping array (Omni2.5), use of imputation panels (HRC or TOPMed) that does not match our simulation reference significantly affects miscalibration (up to FDR = 17% vs. expected 1%), and using multiple imputation panels further increased miscalibration (up to FDR = 35% vs. expected 2%, **Fig. 3.2c**); this setup is as bad as the most heterogeneous configuration using multiple genotyping arrays and imputation panels (FDR = 37%). When TOPMed-imputed variants were lifted over from GRCh38 to GRCh37, we observed FDR increases of up to 10%, likely due to genomic build conversion failures (**C.1 Supplementary Note**)[253]. Fourth, recall was not significantly affected by heterogeneous genotyping arrays or imputation panels (**Fig. 3.2b,d**). Fifth, including multiple genetic ancestries did not affect calibration when using the same genotyping array and imputation panel (Omni 2.5 and

1000GP3; **Fig. 3.2e**) but significantly improved recall if African ancestry was included (**Fig. 3.2f**). This is expected, given the shorter LD length in the African population compared to other populations, which improves fine-mapping resolution[254]. Finally, in the most heterogeneous configurations where multiple genotyping arrays and imputation panels existed, we observed a FDR of up to 37% and 28% for European and multi-ancestry meta-analyses, respectively (vs. expected 2% for both), demonstrating that inter-cohort heterogeneity can substantially undermine meta-analysis fine-mapping (**Fig. 3.2g,h**).

To further characterize observed miscalibration in meta-analysis fine-mapping, we investigated the availability of GWAS variants in each combination of ancestry, genotyping array, and imputation panel. Out of 3,285,617 variants on chromosome 3 that passed variant QC in at least one combination (per-combination MAF > 0.001 and Rsq > 0.6; **3.4 Methods**), 574,261 variants (17%) showed population-level gnomAD MAF > 0.001 in every ancestry that we simulated (African, East Asian, and European). Because we used a variety of imputation panels, we retrieved population-level MAF from gnomAD. Of these 574,261 variants, 389,219 variants (68%) were available in every combination (**Supplementary Fig. C.3a**). This fraction increased from 68% to 73%, 74%, and 76% as we increased gnomAD MAF thresholds to > 0.005, 0.01, and 0.05, respectively, but never reached 100% (**Supplementary Fig. C.4**). Notably, we observed a substantial number of variants that are unique to a certain genotyping array and an imputation panel, even when we restricted to 344,497 common variants (gnomAD MAF > 0.05) in every ancestry (**Supplementary Fig. C.3b**). For example, there are 34,317 variants (10%) that were imputed in the 1000GP3 and TOPMed reference but not in the HRC. Likewise, we observed 33,106 variants (10%) that were specific to the 1000GP3 reference and even 3,066 variants (1%) that were imputed in every combination except for East Asian ancestry with the GSA array and the TOPMed reference. When using different combinations of gnomAD MAF thresholds (> 0.001, 0.005, 0.01, or 0.05 in every ancestry) and Rsq thresholds (> 0.2, 0.4, 0.6, or 0.8), we observed the largest fraction of shared variants (78%) was achieved with gno-

mAD MAF > 0.01 and Rsq > 0.2 while the largest number of the shared variants (427,494 variants) was achieved with gnomAD MAF > 0.001 and Rsq > 0.2, leaving it unclear which thresholds would be preferable in the context of fine-mapping (**Supplementary Fig. C.4**).

The remaining 2,711,356 QC-passing variants in our simulations (gnomAD MAF ≤ 0.001 in at least one ancestry) further exacerbate variable coverage of the available variants (**Supplementary Fig. C.3c**). Of these, the largest proportion of variants (39%) were only available in African ancestry, followed by African and European (but not in East Asian) available variants (7%), European-specific variants (6%), and East Asian-specific variants (5%). Furthermore, similar to the aforementioned common variants, we found a substantial number of variants that are unique to a certain combination. Altogether, we observed that only 393,471 variants (12%) out of all the QC-passing 3,285,617 variants were available in every combination (**Supplementary Fig. C.3d**). These observations recapitulate that different combinations of genetic ancestry, genotyping array, imputation panels, and QC thresholds substantially affect the availability of common, well-imputed variants for association testing[255].

Thus, the different combinations of genotyping and imputation cause each cohort in a meta-analysis to have a different set of variants, and consequently variants can have very different overall sample sizes. In our simulations with the most heterogeneous configurations, we found that 66% of the false positive loci (where a non-causal [false positive] variant was assigned PIP > 0.9) had different sample sizes for true causal and false positive variants (median maximum/minimum sample size ratio = 1.4; **Supplementary Fig. C.5**). Analytically, we found that at common meta-analysis sample sizes and genome-wide significant effect size regimes, when two variants have similar marginal effects, the one with the larger sample size will usually achieve a higher ABF PIP (**C.1 Supplementary Note**). This elucidates the mechanism by which sample size imbalance can lead to miscalibration.

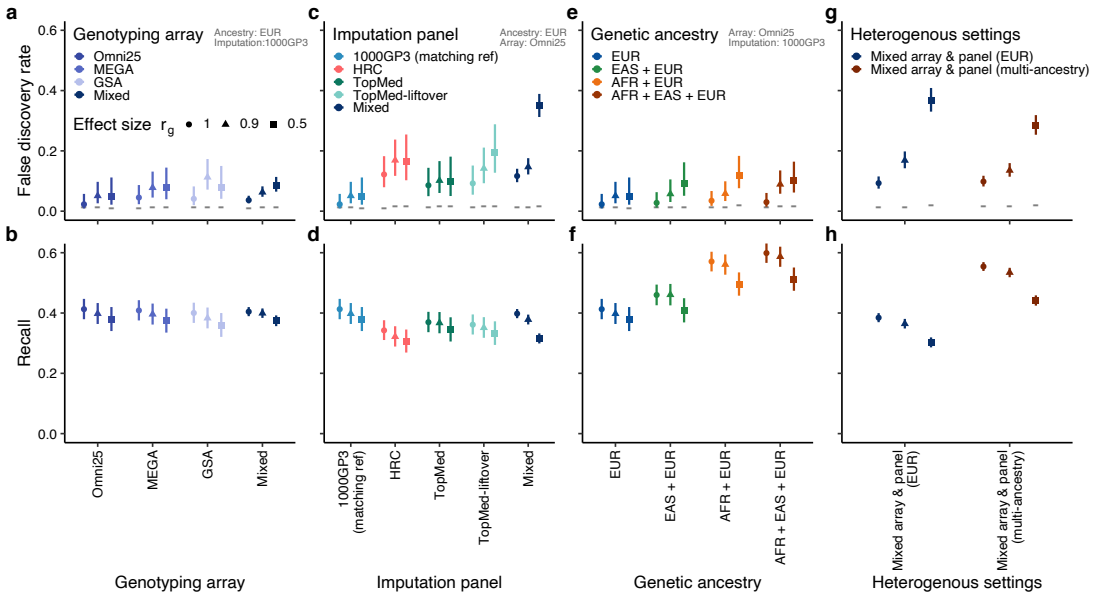


Figure 3.2: Evaluation of false discovery rate (FDR) and recall in meta-analysis fine-mapping simulations. We evaluated FDR and recall in meta-analysis fine-mapping using different genotyping arrays (**a**,**b**), imputation reference panels (**c**,**d**), genetic ancestries (**e**,**f**), and more heterogeneous settings by combining these (**g**,**h**). As shown in top-right gray labels, the EUR ancestry, the Omni2.5 genotyping array and/or the 1000GP3 reference panel were used unless otherwise stated. FDR is defined as the proportion of non-causal variants with PIP > 0.9. Horizontal gray lines represent 1 − mean PIP, *i.e.* expected FDR were the method calibrated. Recall is defined as the proportion of true causal variants in the top 1% PIP bin. Shapes correspond to the true effect size correlation r_g across cohorts which represent a phenotypic heterogeneity parameter (the lower r_g, the higher phenotypic heterogeneity).

3.2.2 Overview of the SLALOM method

To address the challenges in meta-analysis fine-mapping discussed above, we developed SLALOM (suspicious loci analysis of meta-analysis summary statistics), a method that flags suspicious loci for meta-analysis fine-mapping by detecting outliers in association statistics based on deviations from expectation, estimated with local LD structure (**3.4 Methods**). SLALOM consists of three steps, 1) defining loci and lead variants based on a 1 Mb window, 2) detecting outlier variants in each locus using meta-analysis summary statistics and an external LD reference panel, and 3) identifying suspicious loci for meta-analysis fine-mapping (**Fig. 3.3a,b**).

To detect outlier variants, we first assume a single causal variant per associated locus. Then the marginal z-score z_i for a variant i should be approximately equal to $r_{i,c} \cdot z_c$ where z_c is the z-score of the causal variant c, and $r_{i,c}$ is a correlation between variants i and c. For each variant in meta-analysis summary statistics, we first test this relationship using a simplified version of the DENTIST statistics[250], DENTIST-S, based on the assumption of a single causal variant. The DENTIST-S statistics for a given variant i is written as

$$T_i = \frac{(z_i - r_{i,c} \cdot z_c)^2}{1 - r_{i,c}^2} \tag{3.1}$$

which approximately follows a χ^2 distribution with 1 degree of freedom[250]. Since the true causal variant and LD structure are unknown in real data, we approximate the causal variant as the lead PIP variant in the locus (the variant with the highest PIP) and use a large-scale external LD reference from gnomAD[214], either an ancestry-matched LD for a single-ancestry meta-analysis or a sample-size-weighted LD by ancestries for a multi-ancestry meta-analysis (**3.4 Methods**).

SLALOM then evaluates whether each locus is "suspicious"—that is, has a pattern of meta-analysis statistics and LD that appear inconsistent and therefore call into question the fine-mapping accuracy. By training on loci with maximum PIP > 0.9 in the simulations, we determined that the

best-performing criterion for classifying loci as true or false positives is whether a locus has a variant with $r^2 > 0.6$ to the lead and DENTIST-S *P*-value $< 1.0 \times 10^{-4}$ (**3.4 Methods**). Using this criterion we achieved an area under the receiver operating characteristic curve (AUROC) of 0.74, 0.76, and 0.80 for identifying whether a true causal variant is a lead PIP variant, in 95% credible set (CS), and in 99% CS, respectively (**Fig. 3.3c**). We further validated the performance of SLALOM using all the loci in the simulations and observed significantly higher miscalibration in predicted suspicious loci than in non-suspicious loci (up to 16% difference in FDR at PIP > 0.9; **Fig. 3.3d**). Given the relatively lower miscalibration and specificity at low PIP thresholds (**Fig. 3.3d,e**), in subsequent real data analysis we restricted the application of SLALOM to loci with maximum PIP > 0.1 (**3.4 Methods**).

3.2.3 Widespread suspicious loci for fine-mapping in existing meta-analysis summary statistics

Having assessed the performance of SLALOM in simulations, we applied SLALOM to 467 meta-analysis summary statistics in the GWAS Catalog[18] that are publicly available with a sufficient discovery sample size ($N >$ 10,000; **Supplementary Table C.5**; **3.4 Methods**) to quantify the prevalence of suspicious loci in existing studies. These summary statistics were mostly European ancestry-only meta-analysis (63%), followed by multi-ancestry (31%), East Asian ancestry-only (3%), and African ancestry-only (2%) meta-analyses. Across 467 summary statistics from 96 publications, we identified 28,925 loci with maximum PIP > 0.1 (out of 35,864 genome-wide significant loci defined based on 1 Mb window around lead variants; **3.4 Methods**) for SLALOM analysis, of which 8,137 loci (28%) were predicted suspicious (**Supplementary Table C.6**).

To validate SLALOM performance in real data, we restricted our analysis to 6,065 loci that have maximum PIP > 0.1 and that contain nonsynonymous coding variants (predicted loss-of-function [pLoF] and missense) in LD with the lead variant ($r^2 > 0.6$). Given prior evidence[67,68,73] that such

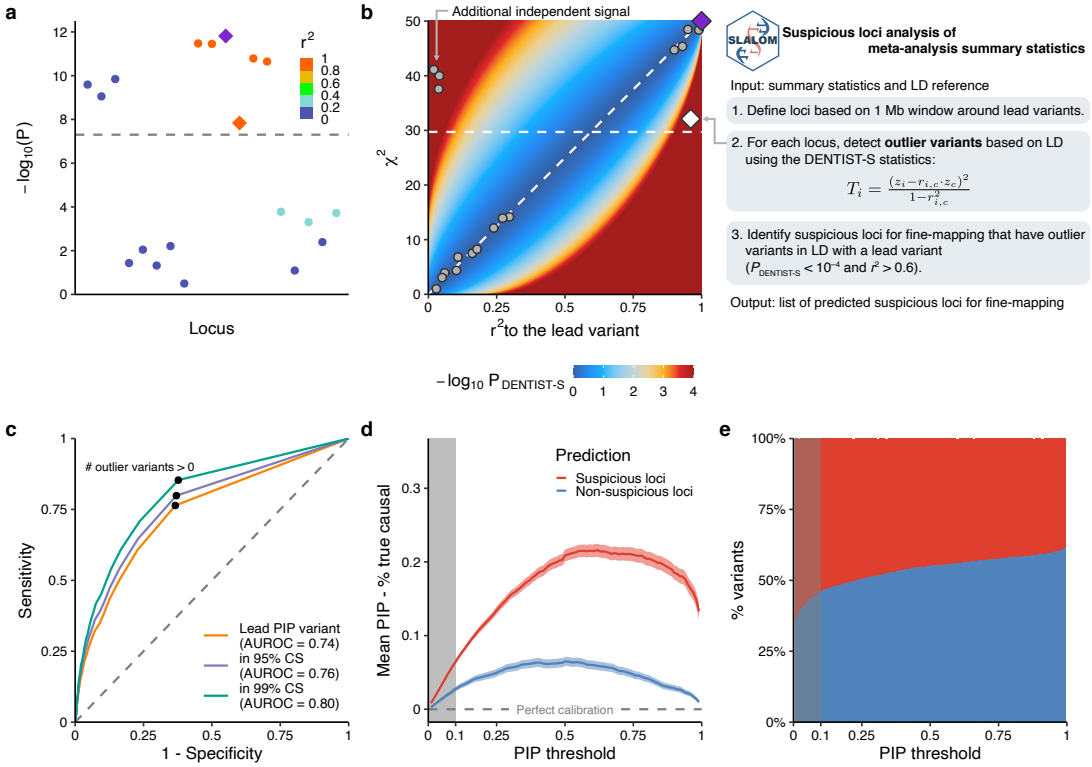


Figure 3.3: Overview of the SLALOM method. a,b. An illustrative example of the SLALOM application. **a.** In an example locus, two independent association signals are depicted: i) the most significant signal that contains a lead variant (purple diamond) and five additional variants that are in strong LD (r^2 > 0.9) with the lead variant, and ii) an additional independent signal (r^2 < 0.05). There is one outlier variant (orange diamond) in the first signal that deviates from the expected association based on LD. **b.** Step-by-step procedure of the SLALOM method. For outlier variant detection in a locus, a diagnosis plot of r^2 values to the lead variant vs. marginal χ^2 is shown to aid interpretation. Background color represents a theoretical distribution of $-\log_{10}P_{\text{DENTIST-S}}$ values when a lead variant has a marginal χ^2 of 50, assuming no allele flipping. Points represent the variants depicted in the example locus (**a**), where the lead variant (purple diamond) and the outlier variant (white diamond) were highlighted. Diagonal line represents an expected marginal association. Horizontal dotted lines represent the genome-wide significance threshold ($P < 5.0 \times 10^{-8}$). **c.** The ROC curve of SLALOM prediction for identifying suspicious loci in the simulations. Positive conditions were defined as whether a true causal variant in a locus is 1) a lead PIP variant, 2) in 95% CS, and 3) in 99% CS. AUROC values were shown in the labels. Black points represent the performance of our adopted metric, *i.e.*, whether a locus contains at least one outlier variant ($P_{\text{DENTIST-S}} < 1.0 \times 10^{-4}$ and $r^2 > 0.6$). **d.** Calibration plot in the simulations under different PIP thresholds. Calibration was measured as the mean PIP — fraction of true causal variants among variants above the threshold. Shadows around the lines represent 95% confidence intervals. **e.** The fraction of variants in predicted suspicious and non-suspicious loci under different PIP thresholds. Gray shadows in the panels **d,e** represent a PIP ≤ 0.1 region as we excluded loci with maximum PIP ≤ 0.1 in the actual SLALOM analysis based on these panels.

nonsynonymous variants are highly enriched for being causal, we tested the validity of our method by whether they achieve the highest PIP in the locus (*i.e.*, successful fine-mapping) in suspicious vs. non-suspicious loci (**3.4 Methods**). While 40% (1,557 / 3,860) of non-suspicious loci successfully fine-mapped nonsynonymous variants, only 17% (384 / 2,205) of suspicious loci did, demonstrating a significant depletion (2.3x) of successfully fine-mapped nonsynonymous variants in suspicious loci (Fisher's exact $P = 3.6 \times 10^{-79}$; **Fig. 3.4a**). We also tested whether nonsynonymous variants belonged to 95% and 99% CS and again observed significant depletion (1.4x and 1.3x, respectively; Fisher's exact $P < 4.6 \times 10^{-100}$). In addition, when we used a more stringent r2 threshold (> 0.8) for selecting loci that contain nonsynonymous variants, we also confirmed significant enrichment (Fisher's exact $P < 6.1 \times 10^{-65}$; **Supplementary Fig. C.6**). To quantify potential fine-mapping miscalibration in the GWAS Catalog, we investigated the difference between mean PIP for lead variants and fraction of lead variants that are nonsynonymous; assuming that nonsynonymous variants in these loci are truly causal, this difference equals the difference between the true and reported fraction of lead PIP variants that are causal. We observed differences between 26–51% and 10–18% under different PIP thresholds in suspicious and non-suspicious loci, respectively (**Fig. 3.4b**), marking 45% and 15% for high-PIP (> 0.9) variants.

We further assessed SLALOM performance in the GWAS Catalog meta-analyses by leveraging high-PIP (> 0.9) complex trait and *cis*-eQTL variants that were rigorously fine-mapped[67,68] in large-scale biobanks (Biobank Japan [BBJ][171], FinnGen[172], and UK Biobank [UKBB][76]) and eQTL resources (GTEx[103] v8 and eQTL Catalogue[191]). Among the 27,713 loci analyzed by SLALOM (maximum PIP > 0.1) that contain a lead variant that was included in biobank fine-mapping, 17% (3,266 / 19,692) of the non-suspicious loci successfully fine-mapped one of the high-PIP GWAS variants in biobank fine-mapping, whereas 7% (589 / 8,021) of suspicious loci did, showing a significant depletion (2.3x) of the high-PIP complex trait variants in suspicious loci (Fisher's exact $P = 4.6 \times 10^{-100}$; **Fig. 3.4c**). Similarly, among 26,901 loci analyzed by SLALOM that contain

a lead variant that was included in *cis*-eQTL fine-mapping, we found a significant depletion (1.9x) of the high-PIP *cis*-eQTL variants in suspicious loci, where 7% (1,247 / 18,976) of non-suspicious loci vs. 4% (281 / 7,925) of suspicious loci successfully fine-mapped one of the high-PIP *cis*-eQTL variants (Fisher's exact $P = 2.6 \times 10^{-24}$; **Fig. 3.4d**). We observed the same significant depletions of the high-PIP complex trait and *cis*-eQTL variants in suspicious loci that belonged to 95% and 99% CS set (**Fig. 3.4c,d**).

3.2.4 Suspicious loci for fine-mapping in the GBMI summary statistics

Next, we applied SLALOM to meta-analysis summary statistics of 14 disease endpoints from the GBMI[239]. These summary statistics were generated from a meta-analysis of 2.1 million individuals in total across 19 biobanks, representing six different genetic ancestry groups of approximately 33,000 African, 18,000 Admixed American, 31,000 Central and South Asian, 341,000 East Asian, 1.8 million European, and 1,600 Middle Eastern individuals (**Supplementary Table C.7**). Among 509 genome-wide significant loci across the 14 traits, we found that 87 loci (17%) showed maximum PIP < 0.1, thus not being further considered by SLALOM. Of the remaining 422 loci with maximum PIP > 0.1, SLALOM identified that 285 loci (68%) were suspicious loci for fine-mapping (**Fig. 3.5a**; **Supplementary Table C.8**). The fraction of suspicious loci and their maximum PIP varied by trait, reflecting different levels of statistical power (*e.g.*, sample sizes, heritability, and local LD structure) as well as inter-cohort heterogeneity (**Fig. 3.4b–o**).

While the fraction of suspicious loci (68%) in the GBMI meta-analyses is higher than in the GWAS Catalog (28%), there might be multiple reasons for this discrepancy, including association significance, sample size, ancestral diversity, and study-specific QC criteria. For example, the GBMI summary statistics were generated from multi-ancestry, large-scale meta-analyses of median sample size of 1.4 million individuals across six ancestries, while 63% of the 467 summary statistics from the GWAS Catalog were only in European-ancestry studies and 83% had less than 0.5 million discovery

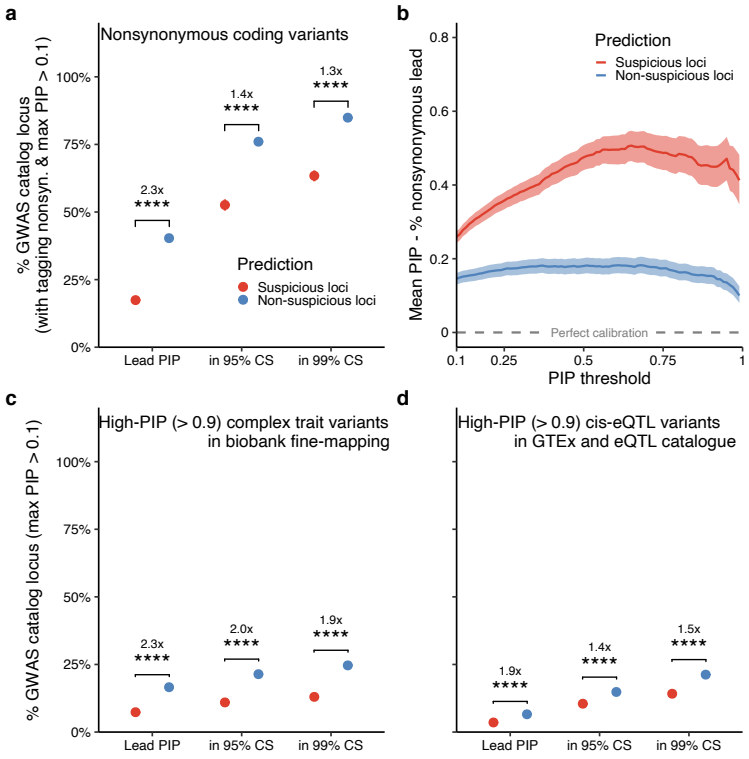


Figure 3.4: Evaluation of SLALOM performance in the GWAS Catalog summary statistics. a,c,d. Depletion of likely causal variants in predicted suspicious loci. We evaluated whether (**a**) nonsynonymous coding variants (pLoF and missense), (**c**) high-PIP (> 0.9) complex trait variants in biobank fine-mapping, and (**d**) high-PIP (> 0.9) *cis*-eQTL variants in GTEx v8 and eQTL Catalogue were lead PIP variants, in 95% CS, or in 99% CS in suspicious vs. non-suspicious loci. Depletion was calculated by relative risk (*i.e.* a ratio of proportions; **3.4 Methods**). Error bars correspond to 95% confidence intervals using bootstrapping. Significance represents a Fisher's exact test P-value (*, $P < 0.05$; **, < 0.01; ***, < 0.001; ***, $< 10^{-4}$). **b.** Plot of the estimated difference between true and reported proportion of causal variants in the loci tagging nonsynonymous variants (r^2 > 0.6 with the lead variants) in the GWAS Catalog under different PIP thresholds. Analogous to **Fig. 3.3b**, assuming nonsynonymous variants in these loci are truly causal, the mean PIP for lead variants minus the fraction of lead variants that are nonsynonymous above the threshold is equal to the difference between true and reported proportion of causal variants.

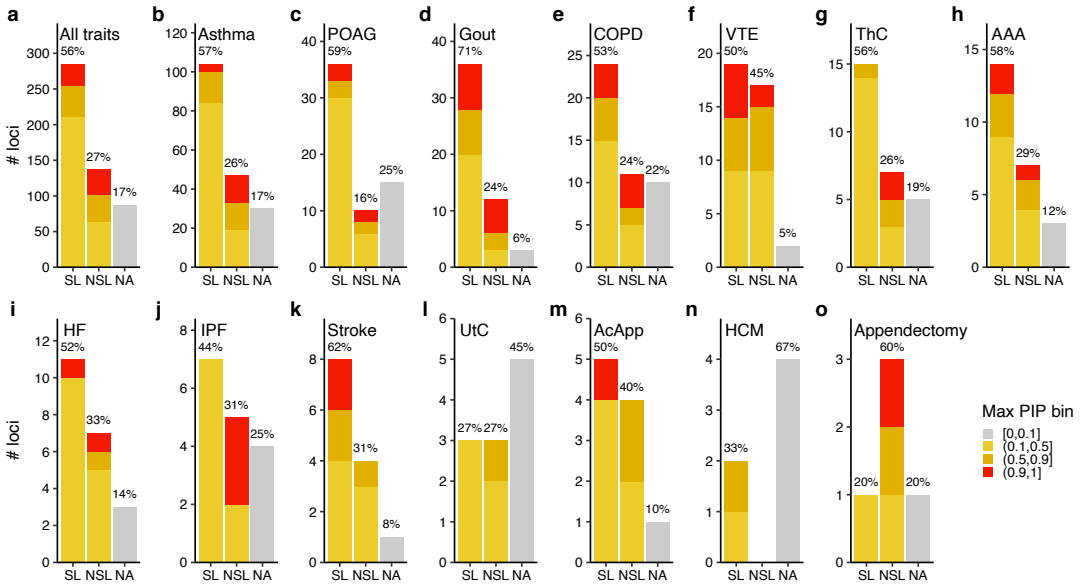


Figure 3.5: SLALOM prediction results in the GBMI summary statistics. For (**a**) all 14 traits and (**b–o**) individual traits, a number of predicted suspicious (SL), non-suspicious (NSL), and non-applicable (NA; maximum PIP < 0.1) loci were summarized. Individual traits are ordered by the total number of loci. Color represents the maximum PIP in a locus. Label represents the fraction of loci in each prediction category. AAA, abdominal aortic aneurysm. AcApp, acute appendicitis. COPD, chronic obstructive pulmonary disease. HCM, hypertrophic cardiomyopathy. HF, heart failure. IPF, idiopathic pulmonary fibrosis. POAG, primary open angle glaucoma. ThC, thyroid cancer. UtC, uterine cancer. VTE, venous thromboembolism.

samples. Nonetheless, predicted suspicious loci for fine-mapping were prevalent in both the GWAS Catalog and the GBMI.

Using nonsynonymous (pLoF and missense) and high-PIP (> 0.9) complex trait and *cis*-eQTL variants, we recapitulated a significant depletion of these likely causal variants in predicted suspicious loci (2.8x, 5.4x, and 5.2x for nonsynonymous, high-PIP complex trait, and high-PIP *cis*-eQTL variants being a lead PIP variant, respectively; Fisher's exact $P < 6.2 \times 10^{-4}$), confirming our observation in the GWAS Catalog analysis (**Fig. 3.6a–c**).

In 15/23 non-suspicious loci harboring a nonsynonymous variant, the nonsynonymous variant had the highest PIP. These included known missense variants such as rs116483731 (p.Arg20Gln)

in *SPDL1* for idiopathic pulmonary fibrosis (IPF)[256,257] and rs28929474 (p.Glu366Lys) in *SERPINA1* for chronic obstructive pulmonary disease (COPD)[258,259]. In addition, we observed successful fine-mapping in 2 novel loci for asthma, i) rs41286560 (p.Pro558Thr) in *RTL1*, a missense variant known for decreasing height[260,261] and ii) rs34187696 (p.Gly337Val) in *ZSCAN5A*, a known missense variant for increasing monocyte count[52].

To characterize fine-mapping failures in suspicious loci, we examined suspicious loci in which a nonsynonymous variant did not achieve the highest PIP. For example, the *FCGR2A/FCGR3A* (1q23.3) locus for COPD contained a genome-wide significant lead intergenic variant rs2099684 ($P = 1.7 \times 10^{-11}$) which is in LD ($r^2 = 0.92$) with a missense variant rs396991 (p.Phe176Val) of *FCGR3A* (**Fig. 3.6d**). This locus was not previously reported for COPD, but is known for associations with autoimmune diseases (*e.g.*, inflammatory bowel disease[73], rheumatoid arthritis[10], and systemic lupus erythematosus[262]) and encodes the low-affinity human FC-gamma receptors that bind to the Fc region of IgG and activate immune responses[263]. Notably, this locus contains copy number variations that contribute to the disease associations in addition to single-nucleotide variants, which makes genotyping challenging[263,264]. Despite strong LD with the lead variant, rs396991 did not achieve genome-wide significance ($P = 9.1 \times 10^{-3}$), showing a significant deviation from the expected association ($P_{\text{DENTIST-S}} = 5.3 \times 10^{-41}$; **Fig. 3.6e**). This is primarily due to missingness of rs396991 in 8 biobanks out of 17 (N_{eff} = 76,790 and 36,781 for rs2099684 and rs396991, respectively; **Fig. 3.6f**), which is caused by its absence from major imputation reference panels (*e.g.*, 1000GP[229], HRC[251], and UK10K[162]) despite having a high MAF in every population (MAF = 0.24–0.34 in African, admixed American, East Asian, European, and South Asian populations of gnomAD[214]).

Sample size imbalance across variants was pervasive in the GBMI meta-analyses[265], and was especially enriched in predicted suspicious loci—84% of suspicious loci vs. 24% of non-suspicious loci showed a maximum/minimum effective sample size ratio > 2 among variants in LD ($r^2 > 0.6$)

with lead variants (a median ratio = 4.2 and 1.2 in suspicious and non-suspicious loci, respectively; **Supplementary Fig. C.7**). These observations are consistent with our simulations, recapitulating that sample size imbalance results in miscalibration for meta-analysis fine-mapping. Notably, we observed a similar issue in other GBMI downstream analyses (*e.g.*, polygenic risk score [PRS][265] and drug discovery[266]), where predictive performance improved significantly after filtering out variants with maximum *N*eff < 50%. Although fine-mapping methods cannot simply take this approach because it inevitably reduces calibration and recall by removing true causal variants, other meta-analysis downstream analyses that primarily rely on polygenic signals rather than individual variants should consider this filtering as an extra QC step.

3.2.5 Comparison of fine-mapping results between the GBMI meta-analyses and individual biobanks

Motivated by successful validation of SLALOM performance, we investigated whether fine-mapping confidence and resolution were improved in the GBMI meta-analyses over individual biobanks. To this end, we used our fine-mapping results[67,68] of nine disease endpoints (asthma[259], COPD[259], gout, heart failure[267], IPF[257], primary open angle glaucoma[268], thyroid cancer, stroke[269], and venous thromboembolism[270]) in BBJ[171], FinnGen[172], and UKBB[76] Europeans that also contributed to the GBMI meta-analyses for the same traits.

To perform an unbiased comparison of PIP between the GBMI meta-analysis and individual biobanks, we investigated functional enrichment of fine-mapped variants based on top PIP rankings in the GBMI and individual biobanks (top 0.5%, 0.1%, and 0.05% PIP variants in the GBMI vs. maximum PIP across BBJ, FinnGen, and UKBB; **3.4 Methods**). Previous studies have shown that high-PIP (> 0.9) complex trait variants are significantly enriched for well-known functional categories, such as coding (pLoF, missense, and synonymou), 5'/3' UTR, promoter, and *cis*-regulatory element (CRE) regions (DNase I hypersensitive sites [DHS] and H3K27ac)[67,68]. Using these functional categories, we found no significant enrichment of variants in the top PIP rankings in the GBMI over individual biobanks (Fisher's exact $P > 0.05$; **Fig. 3.7a**) except for variants in the promoter region (1.8x; Fisher's exact $P = 3.1 \times 10^{-4}$ for the top 0.1% PIP variants). We observed similar trends regardless of whether variants were in suspicious or non-suspicious loci (**Fig. 3.7b,c**). To examine patterns of increased and decreased PIP for individual variants, we also calculated PIP difference between the GBMI and individual biobanks, defined as ΔPIP = PIP (GBMI) – maximum PIP across BBJ, FinnGen, and UKBB (**Supplementary Fig. C.8,C.9**). We investigated functional enrichment based on ΔPIP bins and observed inconsistent enrichment results using different ΔPIP thresholds (**Supplementary Fig. C.10**). Finally, to test whether fine-mapping resolution was im-

proved in the GBMI over individual biobanks, we compared the size of 95% CS after restricting them to cases where a GBMI CS overlapped with an individual biobank CS from BBJ, FinnGen, or UKBB (**3.4 Methods**). We observed the median 95% CS size of 2.5 and 2.5 in non-suspicious loci for the GBMI and individual biobanks, respectively, and 5 and 15 in suspicious loci, respectively (**Supplementary Fig. C.11**). The smaller credible set size in suspicious loci in GBMI could be due to improved resolution or to increased miscalibration. These results provide limited evidence of overall fine-mapping improvement in the GBMI meta-analyses over what is achievable by taking the best result from individual biobanks.

Individual examples, however, provide insights into the types of fine-mapping differences that can occur. To characterize the observed differences in fine-mapping confidence and resolution, we further examined non-suspicious loci with ΔPIP > 0.5 in asthma. In some cases, the increased power and/or ancestral diversity of GBMI led to improved fine-mapping: for example, an intergenic variant rs1888909 (~18 kb upstream of *IL33*) showed ΔPIP = 0.99 (PIP = 1.0 and 0.008 in GBMI and FinnGen, respectively; **Fig. 3.7d**), which was primarily owing to increased association significance in a meta-analysis ($P = 3.0 \times 10^{-86}$, 7.4×10^{-2}, 3.6×10^{-16}, and 1.9×10^{-53} in GBMI, BBJ, FinnGen, and UKBB Europeans, respectively) as well as a shorter LD length in the African population than in the European population (LD length = 4 kb vs. 41 kb for variants with $r^2 > 0.6$ with rs1888909 in the African and European populations, respectively; N_{eff} = 4,270 for Africans in the GBMI asthma meta-analysis; **Supplementary Fig. C.12**). This variant was also fine-mapped for eosinophil count in UKBB Europeans (PIP = 1.0; $P = 1.3 \times 10^{-314}$)[67] and was previously reported to regulate *IL33* gene expression in human airway epithelial cells via allele-specific transcription factor binding of OCT-1 (POU2F1)[271]. Likewise, we observed a missense variant rs16903574 (p.Phe319Leu) in OTULINL showed ΔPIP = 0.79 (PIP = 1.0 and 0.21 in GBMI and UKBB Europeans, respectively; **Fig. 3.7e**) owing to improved association significance ($P = 7.7 \times 10^{-15}$ and 4.7×10^{-12} in GBMI and UKBB Europeans, respectively).

However, we also observed very high ΔPIP for variants that are not likely causal. For example, we observed that an intronic variant rs1295686 in IL13 showed ΔPIP = 0.56 (PIP = 0.56 and 0.0002 in GBMI and UKBB Europeans, respectively; **Fig. 3.7f**), despite having strong LD with a nearby missense variant rs20541 (p.Gln144Arg; $r^2 = 0.96$ with rs1295686) which only showed ΔPIP = 0.13 (PIP = 0.13 and 0.0001 in GBMI and UKBB Europeans, respectively). The missense variant rs20541 showed PIP = 0.23 and 0.15 for a related allergic disease, atopic dermatitis, in BBJ and FinnGen, respectively[68], and was previously shown to induce STAT6 phosphorylation and upregulate CD23 expression in monocytes, promoting IgE synthesis[272]. Although the GBMI meta-analysis contributed to prioritizing these two variants (sum of PIP = 0.69 vs. 0.0003 in GBMI and UKBB Europeans, respectively), the observed ΔPIP was higher for rs1295686 than for rs20541.

While increasing sample size in meta-analysis improves association significance, we also found negative ΔPIP due to losing the ability to model multiple causal variants. A stop-gained variant rs61816761 (p.Arg501Ter) in FLG showed ΔPIP = −1.0 (PIP = 6.4×10^{-5} and 1.0 in GBMI and UKBB Europeans, respectively; **Fig. 3.7g**), which was primarily owing to a nearby lead variant rs12123821 (~17 kb downstream of *HRNR*; r^2 = 0.0 with rs61816761). This lead variant rs12123821 showed greater significance than rs61816761 in GBMI ($P = 9.3 \times 10^{-16}$ and 2.0×10^{-11} for rs12123821 and rs61816761, respectively) as well as in UKBB Europeans ($P = 7.1 \times 10^{-26}$ and 1.5×10^{-18}). While our biobank fine-mapping[67,68] assigned PIP = 1.0 for both variants based on multiple causal variant fine-mapping (*i.e.*, FINEMAP[28] and SuSiE[30]), our ABF fine-mapping in the GBMI meta-analysis was only able to assign PIP = 0.74 for the lead variant rs12123821 due to a single causal variant assumption. This recapitulates the importance of multiple causal variant fine-mapping in complex trait fine-mapping[67,68]—however, we note that multiple causal variant fine-mapping with an external LD reference is extremely error-prone as previously reported[67,87,118].

Figure 3.6 *(following page)*: Evaluation of SLALOM performance in the GBMI summary statistics. a–c. Similar to **Fig. 3.4**, we evaluated whether (**a**) nonsynonymous coding variants (pLoF and missense), (**b**) high-PIP (> 0.9) complex trait variants in biobank fine-mapping, and (**c**) high-PIP (> 0.9) *cis*-eQTL variants in GTEx v8 and eQTL Catalogue were lead PIP variants, in 95% CS, or in 99% CS in suspicious vs. non-suspicious loci. Depletion was calculated by relative risk (*i.e.* a ratio of proportions; **3.4 Methods**). Error bars correspond to 95% confidence intervals using bootstrapping. Significance represents a Fisher's exact test P-value (*, $P < 0.05$; **, < 0.01; ***, < 0.001; ***, $< 10^{-4}$). **d**. Locuszoom plot of the 1q23.3 locus for COPD. The top panel shows a Manhattan plot, where the lead variant rs2099684 (purple diamond) and a missense variant rs396991 (orange diamond) are highlighted. Color represents r^2 values to the lead variant. Horizontal line represents a genome-wide significance threshold ($P = 5.0 \times 10^{-8}$). The middle panel shows PIP from ABF fine-mapping. Color represents whether variants belong to a 95% CS. The bottom panel shows r^2 values with the lead variant in gnomAD populations. **e**. A diagnosis plot showing r^2 values to the lead variant vs. marginal χ^2. Color represents $-\log_{10} P_{\text{DENTIST-S}}$ values. Outlier variants with $P_{\text{DENTIST-S}} < 10^{-4}$ are depicted in red with a diamond shape. Diagonal line represents an expected marginal association. Horizontal line represents a genome-wide significance threshold. **f**. Z-scores of the lead variant (rs2099684) vs. the missense variant (rs396991) in the constituent cohorts of the meta-analysis. Open and closed circles represent whether both variants exist in a cohort or rs396991 is missing. Circle size corresponds to an effective sample size. Color represents genetic ancestry.

Figure 3.6: (continued)

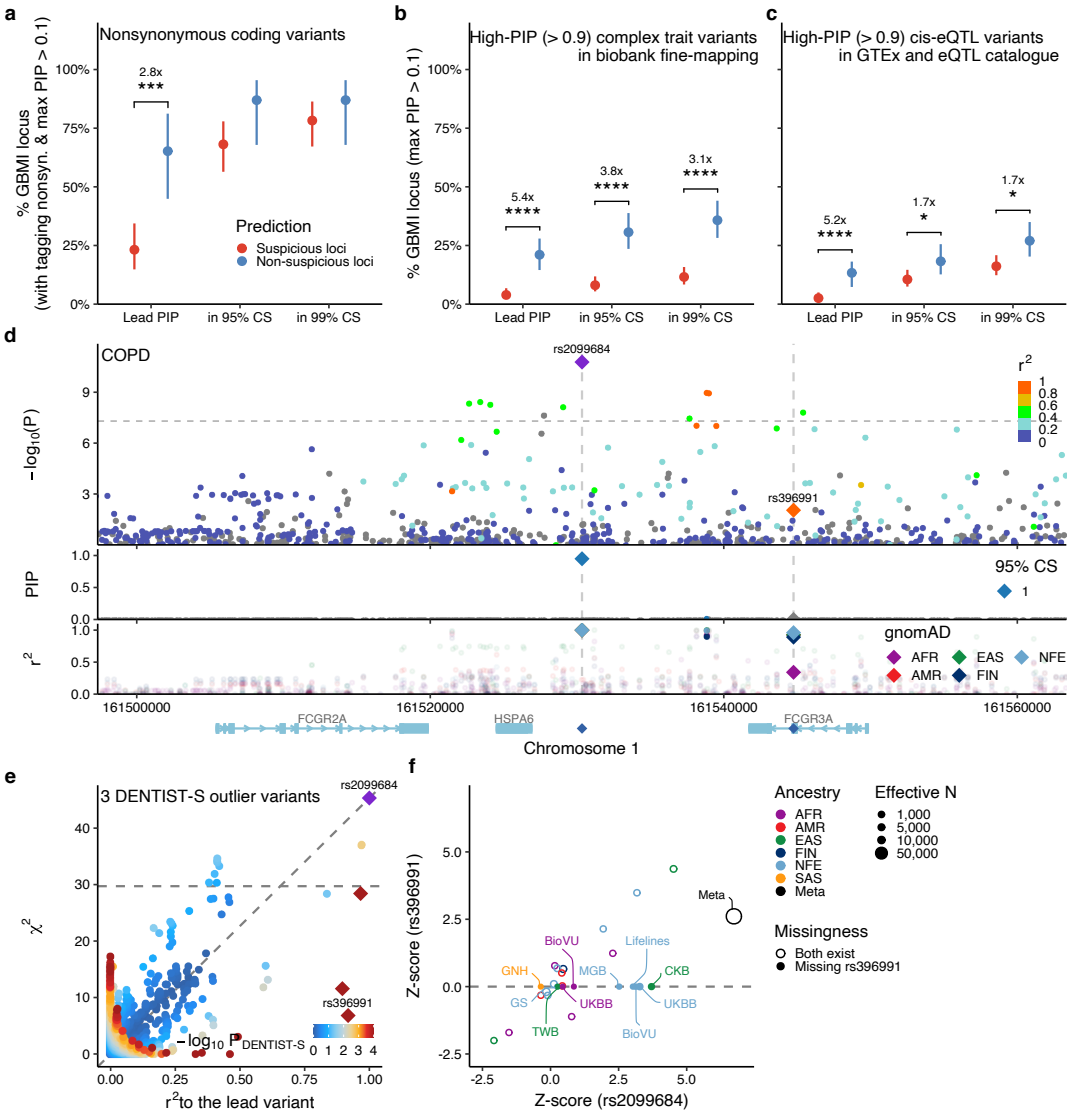
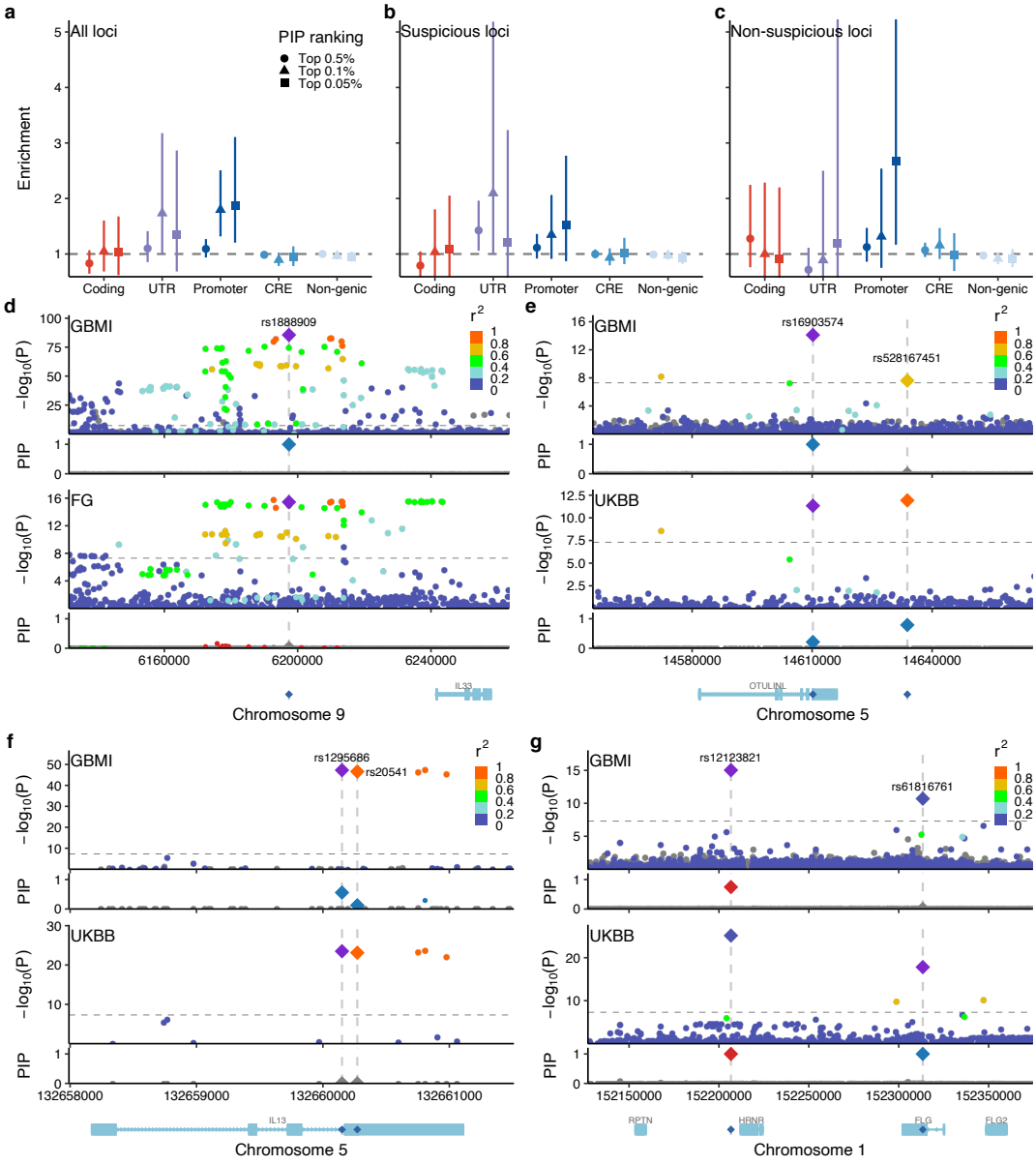


Figure 3.7 ***(following page)*****: Fine-mapping improvement and retrogression in the GBMI meta-analyses over individual biobanks. a–c**. Functional enrichment of variants in each functional category based on top PIP rankings in the GBMI and individual biobanks (maximum PIP of BBJ, FinnGen, and UKBB). Shape corresponds to top PIP ranking (top 0.5%, 0.1%, and 0.05%). Enrichment was calculated by a relative risk (*i.e.* a ratio of proportions; **3.4 Methods**). Error bars correspond to 95% confidence intervals using bootstrapping. **d–g**. Locuszoom plots for the same non-suspicious locus of asthma in the GBMI meta-analysis and an individual biobank (BBJ, FinnGen, or UKBB Europeans) that showed the highest PIP in our biobank fine-mapping. Colors in the Manhattan panels represent r^2 values to the lead variant. In the PIP panels, only fine-mapped variants in the 95% CS are colored, where the same colors are applied between the GBMI meta-analysis and an individual biobank based on merged CS as previously described. Horizontal line represents a genome-wide significance threshold ($P = 5.0 \times 10^{-8}$). **d**. rs1888909 for asthma in the GBMI and FinnGen. **e**. rs16903574 for asthma in the GBMI and UKBB Europeans. Nearby rs528167451 was also highlighted, which was in strong LD (r^2 = 0.86) and in the same 95% CS in UKBB Europeans, but not in the GBMI (r^2 = 0.67). **f**. rs1295686 for asthma in the GBMI and UKBB Europeans. A nearby missense, rs20541, showed lower PIP than rs1295686 despite having strong LD (r^2 = 0.96). **g**. rs12123821 for asthma in the GBMI and UKBB Europeans. Nearby stop-gained rs61816761 was independent of rs12123821 (r^2 = 0.0) and not fine-mapped in the GBMI due to a single causal variant assumption in the ABF fine-mapping.

Figure 3.7: (continued)



3.3 Discussion

In this study, we first demonstrated in simulations that meta-analysis fine-mapping is substantially miscalibrated when constituent cohorts are heterogeneous in phenotyping and imputation. To mitigate this issue, we developed SLALOM, a summary statistics-based QC method for identifying suspicious loci in meta-analysis fine-mapping. Applying SLALOM to 14 disease endpoints from the GBMI meta-analyses[239] as well as 467 summary statistics from the GWAS Catalog[18], we observed widespread suspicious loci in meta-analysis summary statistics, suggesting that meta-analysis fine-mapping is often miscalibrated in real data too. Indeed, we demonstrated that the predicted suspicious loci were significantly depleted for having likely causal variants as a lead PIP variant, such as nonsynonymous variants, high-PIP (> 0.9) GWAS and *cis*-eQTL fine-mapped variants from our previous fine-mapping studies[67,68]. Our method provides better calibration in non-suspicious loci for meta-analysis fine-mapping, generating a more reliable set of variants for further functional characterization.

We have found limited evidence of improved fine-mapping in the GBMI meta-analyses over individual biobanks. A few empirical examples in this study as well as other previous studies[10,52,176,178,238] suggested that multi-ancestry, large-scale meta-analysis could have potential to improve fine-mapping confidence and resolution owing to increased statistical power in associations and differential LD pattern across ancestries. However, we have highlighted that the observed improvement in PIP could be due to sample size imbalance in a locus, miscalibration, and technical confoundings too, which further emphasizes the importance of careful investigation of fine-mapped variants identified through meta-analysis fine-mapping.

As high-confidence fine-mapping results in large-scale biobanks and molecular QTLs continue to become available[67,68,191], we propose alternative approaches for prioritizing candidate causal variants in a meta-analysis. First, these high-confidence fine-mapped variants have been a valuable

resource to conduct a "PheWAS"[67] to match with associated variants in a meta-analysis, which provides a narrower list of candidate variants assuming they would equally be functional and causal in related complex traits or tissues/cell-types. Second, a traditional approach based on tagging variants (*e.g.*, $r^2 > 0.6$ with lead variants, or PICS[99] fine-mapping approach that only relies on a lead variant and LD) can be still highly effective, especially for known functional variants such as nonsynonymous coding variants. As we highlighted in this and previous[21] studies, potentially causal variants in strong LD with lead variants might not achieve genome-wide significance because of missingness and heterogeneity.

While using an external LD reference for fine-mapping has been shown to be extremely error-prone[67,87,118], we find here that it can be useful for flagging suspicious loci, even when it does not perfectly represent the in-sample LD structure of the meta-analyzed individuals. However, our use of external LD reference comes with several limitations. For example, due to the finite sample size of external LD reference, rare or low-frequency variants have larger uncertainties around r^2 than common variants. Moreover, our r^2 values in a multi-ancestry meta-analysis are currently approximated based on a sample-size-weighted average of r^2 across ancestries as previously suggested[63], but this can be different from actual r^2. These uncertainties around r^2 affect SLALOM prediction performance and should be modeled appropriately for further method development. On the other hand, we find it challenging to use a LD reference when true causal variants are located within a complex region (*e.g.*, major histocompatibility complex [MHC]), or are entirely missing from standard LD or imputation reference panels, especially for structural variants. These limitations are not specific to meta-analysis fine-mapping, and separate fine-mapping methods based on bespoke imputation references have been developed (*e.g.*, HLA[273], KIR[274], and variable numbers of tandem repeats [VNTR][275]).

In addition, there are several methodological limitations of SLALOM. First, our simulations only include one causal variant per locus. Although additional independent causal variants would not

affect SLALOM precision (but decrease recall), multiple correlated causal variants in a locus would violate SLALOM assumptions and could lead to some DENTIST-S outliers that are not due to heterogeneity or missingness but rather simply a product of tagging multiple causal variants in LD. In fact, our previous studies have illustrated infrequent but non-zero presence of such correlated causal variants in complex traits[67,68]. Second, SLALOM prediction is not perfect. Although fine-mapping calibration is certainly better in non-suspicious vs. suspicious loci, SLALOM has low precision, and we still observe some miscalibration in non-suspicious loci. Finally, SLALOM is a per-locus QC method and does not calibrate per-variant PIPs. Further methodological development that properly models heterogeneity, missingness, multiple causal variants, and LD uncertainty across multiple cohorts and ancestries is needed to refine per-variant calibration and recall in meta-analysis fine-mapping.

We have found evidence in our simulations and real data of severe miscalibration of fine-mapping results from GWAS meta-analysis; for example, we estimate that the difference between true and reported proportion of causal variants is 20% and 45% for high-PIP (> 0.9) variants in suspicious loci from the simulations and the GWAS Catalog, respectively. Our SLALOM method helps to exclude spurious results from meta-analysis fine-mapping; however, even fine-mapping results in SLALOM-predicted "non-suspicious" loci remain somewhat miscalibrated, showing estimated differences between true and reported proportion of causal variants of 4% and 15% for high-PIP variants in the simulations and the GWAS Catalog, respectively. We thus urge extreme caution when interpreting PIPs computed from meta-analyses until improved methods are available. We recommend that researchers looking to identify likely causal variants employ complete synchronization of study design, case/control ascertainment, genomic profiling, and analytical pipeline, or rely more heavily on functional annotations, biobank fine-mapping, or molecular QTLs.

3.4 Methods

3.4.1 Meta-analysis fine-mapping simulation

To benchmark fine-mapping performance in meta-analysis, we simulated a large-scale, realistic GWAS meta-analysis and performed fine-mapping under different scenarios. An overview of our simulation pipeline is summarized in **Supplementary Fig. C.2**.

Simulated true genotype

Using HAPGEN2 [276] with the 1000 Genomes Project Phase 3 reference [229], we simulated "true" genotypes of chromosome 3 for multiple independent cohorts from African, East Asian, and European ancestries. For each independent cohort from a given ancestry, we simulated 10,000 individuals each using the default parameters, with an ancestry-specific effective population size set to 17,469, 14,269, and 11,418 for Africans, East Asians, and Europeans, respectively, as recommended [276]. To mimic sample size imbalance of different ancestries in the current meta-analyses, we simulated 10 independent European cohorts, 1 African cohort, and 1 East Asian cohort.

To restrict our analysis to unrelated samples, we computed sample relatedness based on KING kinship coefficients [277] using PLINK 2.0 (ref. [278]) and removed monozygotic twins, duplicated individuals, or first-degree relatives with the coefficient threshold of 0.177. The detailed sample sizes of unrelated individuals for each cohort is summarized in **Supplementary Table C.1**.

Genotyping and imputation

To simulate realistic genotyping and imputation procedures, we first virtually genotyped each cohort by restricting variants to those that are available on different genotyping arrays. We selected three major genotyping arrays from Illumina, Inc. (Omni2.5, Multi-Ethnic Global Array

[MEGA], and Global Screening Array [GSA]) that have different densities of genotyping probes (**Supplementary Table C.2**). For each cohort, we created three virtually genotyped datasets by retaining variants that are genotyped on each array. For the sake of simplicity, we assumed no genotyping errors occurred between true genotypes and virtually genotyped data—however, in practice, genotyping error is one of the major sources of unexpected confounding (*e.g.*, see recent discussions here[279,280]) and should be treated carefully.

For each pair of cohort and genotyping array, we then imputed missing variants using different imputation reference panels. We used the Michigan Imputation Server (https://imputationserver.sph.umich.edu/)[228] and the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/)[252] with the default parameters, using three publicly available reference panels: the 1000 Genomes Project Phase 3 (version 5; n = 2,504; 1000GP3)[229], the Haplotype Reference Consortium (version r1.1; n = 32,470; HRC)[251], and the TOPMed (version R2; n = 97,256)[252]. Briefly, for each input, the imputation server created chunks of 20 Mb, applied the standard QC, pre-phased each chunk with Eagle2 (ref. [281]), and imputed non-genotyped variants using a specified reference panel with Minimac4 (https://genome.sph.umich.edu/wiki/Minimac4). The detailed documentation of the imputation pipeline is available on the Michigan and TOPMed websites and has been described elsewhere[228].

We applied post-imputation QC by only keeping variants with MAF > 0.001 and imputation Rsq > 0.6. Because the TOPMed panel is based on GRCh38 while the 1000GP3 and the HRC panels are on GRCh37, we lifted over TOPMed variants from GRCh38 to GRCh37 to meta-analyze with other cohorts. We excluded any variants which were lifted over to different chromosomes or for which the conversion failed. The number of virtually genotyped and imputed variants for each combination of cohort, genotyping array, and imputation panel is summarized in **Supplementary Table C.3**.

True phenotype

We simulated 300 true phenotypes that resemble observed complex trait genetic architecture and phenotypic heterogeneity across cohorts. Based on previous literature, we set parameters as follows: 1) 50% of 1 Mb loci contain a true causal variant[110]; 2) probability of being causal is proportional to functional enrichments of variant consequences (pLoF, missense, synonymous, 5'/3' UTR, promoter, *cis*-regulatory region, and non-genic) for fine-mapped variants as estimated in a previous complex trait fine-mapping study[68]; 3) per-allele causal effect sizes have a variance proportional to $[2p(1-p)]^{\alpha}$ where p represents a maximum MAF across the three ancestries (AFR, EAS, and EUR) and α is set to be -0.38 (ref. [111]); and 4) total SNP-heritability h^2_g for chromosome 3 equals 0.03 (ref. [112]). For the sake of simplicity, we randomly draw a single true causal variant per locus because ABF assumes a single causal variant[23,24]. We draw true causal variants from 1,150,893 non-ambiguous single-nucleotide variants in 1000GP3 that showed MAF > 0.01 in at least one of the three ancestries (AFR, EAS, or EUR) and were not located within conversion-unstable positions (CUP)[253] between the human genome builds GRCh37 and GRCh38. To mimic phenotypic heterogeneity across cohorts in real-world meta-analysis (due to *e.g.*, different ascertainment, measurement error, or true effect size heterogeneity), we introduced cross-cohort genetic correlation of true effect sizes r_g which is set to be one of 1, 0.9, or 0.5. For a true causal variant j, true causal effect sizes β_j across cohorts were randomly drawn from $\beta_j \sim \mathrm{MVN}(0, \Sigma)$ where diagonal elements of Σ were set to be $\sigma^2_g \cdot [2p(1-p)]^{\alpha}$. and off-diagonal elements of were set to be $r_g \cdot \sigma^2_g \cdot [2p(1-p)]^{\alpha}$. σ^2_g was determined by $\sigma^2_g = h^2_g / \sum_j [2p(1-p)]^{(1+\alpha)}$. For each cohort, true phenotype y was computed via $y = X\beta + \varepsilon$ where X is the above true genotype matrix from HAPGEN2 and $\varepsilon_i \sim N(0, 1-\sigma^2_g)$ i.i.d. We simulated 100 true phenotypes for each of r_g = 1, 0.9, and 0.5, respectively.

GWAS

For each combination of phenotype, cohort, genotyping chip, and imputation panel, we conducted GWAS via a standard linear regression as implemented in PLINK 2.0 using imputed dosages. For covariates, we included top 10 principal components that were calculated based on true genotypes after restricting to unrelated samples. We only used LD-pruned variants with MAF > 0.01 for PCA.

Meta-analysis

To simulate meta-analyses that resemble real-world settings, we generated multiple configurations of the above GWAS results to meta-analyze across 10 independent cohorts. Briefly, we chose configurations based on the following settings: 1) 10 EUR cohorts are genotyped and imputed using the same genotyping array (one of GSA, MEGA, or Omni2.5) and the same imputation panel (one of 1000GP3, HRC, TOPMed, or TOPMed-liftover); 2) 10 cohorts consisting of multiple ancestries (9 EUR + 1 AFR/EAS cohorts or 8 EUR + 1 AFR + 1 EAS cohorts), with all cohorts genotyped and imputed using the same array (Omni2.5) and the same panel (1000GP3); 3) 10 EUR or multi-ancestry cohorts are genotyped using the same array (Omni2.5) but imputed using different panels across cohorts; 4) 10 EUR or multi-ancestry cohorts are imputed using the same panel (1000GP3) but genotyped using different arrays across cohorts; 5) 10 EUR or multi-ancestry cohorts are genotyped and imputed using different arrays and panels across cohorts. For settings 3–5, we randomly draw a combination of a genotyping array and an imputation panel for each cohort five times each for 10 EUR and multi-ancestry cohorts. In total, we generated 45 configurations as summarized in **Supplementary Table C.4**.

For each configuration, we conducted a fixed-effect meta-analysis based on inverse-variance weighted betas and standard errors using a modified version of PLINK 1.9 (https://github.com/mkanai/plink-ng/tree/add_se_meta).

Fine-mapping

For each meta-analysis, we defined fine-mapping regions based on a 1 Mb window around each genome-wide significant lead variant and applied ABF[23,24] using prior effect size variance of $\sigma_0^2 = 0.04$. We set a prior variance of effect size to be 0.04 which was taken from Wakefield *et al.*[23] and is commonly used in meta-analysis fine-mapping studies[7,10]. We computed posterior inclusion probability (PIP) and 95% credible set (CS) for each locus and evaluated whether true causal variants were correctly fine-mapped.

3.4.2 The SLALOM method

SLALOM takes GWAS summary statistics and external LD reference as input and predicts whether a locus is suspicious for fine-mapping. SLALOM consists of the following three steps:

Locus definition

Consistent with common fine-mapping region definition, we defined loci based on a 1 Mb window around each genome-wide significant lead variant and merged them if they overlapped. We excluded the major histocompatibility complex (MHC) region (chr 6: 25-36 Mb) from analysis due to extensive LD structure in the region.

DENTIST-S outlier detection

For each variant in a locus, we computed DENTIST-S statistics using equation (1) based on the assumption of a single causal variant. DENTIST-S *P*-values ($P_{\text{DENTIST-S}}$) were computed using the χ^2 distribution with 1 degree of freedom. We applied ABF[23,24] using prior effect size variance of $\sigma_0^2 = 0.04$ and used the lead PIP variant (the variant with the highest PIP) as an approximation of

the causal variant in the locus. To retrieve correlation r among the variants, we used publicly available LD matrices from gnomAD[214] v2 as external LD reference for African, Admixed American, East Asian, Finnish, and non-Finnish European populations. When multiple populations exist, we computed a sample-size-weighted average of r^2 using per-variant sample sizes for each population as previously suggested[63]. We excluded variants without r^2 available in gnomAD from the analysis. Since gnomAD v2 LD matrices are based on the human genome assembly GRCh37, variants were lifted over to GRCh38 if the input summary statistics were based on GRCh38.

We determined DENTIST-S outlier variants using two thresholds: 1) $r^2 > \rho$ to the lead and 2) $P_{\text{DENTIST-S}} < \tau$. The thresholds ρ and τ were set to $\rho = 0.6$ and $\tau = 1.0 \times 10^{-4}$ based on the training in simulations as described below.

Suspicious loci prediction

We predicted whether a locus is suspicious or non-suspicious for fine-mapping based on the number of DENTIST-S outlier variants in the locus $> \kappa$. To determine the best-performing thresholds (ρ, τ, and κ), we used loci with maximum PIP > 0.9 in the simulations for training. Positive conditions were defined as whether a true causal variant in a locus is 1) a lead PIP variant, 2) in 95% CS, and 3) in 99% CS. We computed AUROC across different thresholds ($\rho = 0, 0.1, 0.2, , 0.9$; $-\log_{10}\tau = 0, 0.5, 1, , 10$; and $\kappa = 0, 1, 2,)$ and chose $\rho = 0.6$, $\tau = 1.0 \times 10^{-4}$, and $\kappa = 0$ that showed the highest AUROC for all the aforementioned positive conditions. Using all the loci in the simulations, we then evaluated fine-mapping miscalibration (defined as mean PIP − fraction of true causal variants) at different PIP thresholds in suspicious and non-suspicious loci and decided to only apply SLALOM to loci with maximum PIP > 0.1 owing to relatively lower miscalibration and specificity of SLALOM at lower PIP thresholds.

3.4.3 GWAS Catalog analysis

We retrieved full GWAS summary statistics publicly available on the GWAS Catalog[18]. Out of 33,052 studies from 5,553 publications registered at the GWAS Catalog (as of January 12, 2022), we selected 467 studies from 96 publications that have 1) full harmonized summary statistics pre-processed by the GWAS Catalog with non-missing variant ID, marginal beta, and standard error columns, 2) a discovery sample size of more than 10,000 individuals, 3) African (including African American, Afro-Caribbean, and Sub-Saharan African), admixed American (Hispanic and Latin American), East Asian, or European samples based on their broad ancestral category metadata, 4) at least one genome-wide significant association ($P < 5.0 \times 10^{-8}$), and 5) our manual annotation as a meta-analysis rather than a single-cohort study (**Supplementary Table C.5**). We applied SLALOM to the 467 summary statistics and identified 35,864 genome-wide significant loci (based on 1 Mb window around lead variants), of which 28,925 loci with maximum PIP > 0.1 were further classified into suspicious and non-suspicious loci. Since per-variant sample sizes were not available, we used overall sample sizes of each ancestry (African, Admixed American, East Asian, and European) to calculate the weighted-average of r2. All the variants were harmonized into the human genome assembly GRCh38 by the GWAS Catalog.

3.4.4 GBMI analysis

We used meta-analysis summary statistics of 14 disease endpoints from the GBMI (**Supplementary Table C.7**). These meta-analyses were conducted using up to 1.8 million individuals across 19 biobanks, representing six different genetic ancestry groups (approximately 33,000 African, 18,000 Admixed American, 31,000 Central and South Asian, 341,000 East Asian, 1.8 million European, and 1,600 Middle Eastern individuals). Detailed procedures of the GBMI meta-analyses were described in the GBMI flagship manuscript[239].

Across the 14 summary statistics, we defined 503 genome-wide significant loci ($P < 5.0 \times 10^{-8}$) based on a 1 Mb window around each lead variant and merged them if they overlapped. We applied SLALOM to 422 loci with maximum PIP > 0.1 based on the ABF fine-mapping and predicted whether they were suspicious or non-suspicious for fine-mapping. We used per-variant sample sizes of each ancestry (African, Admixed American, East Asian, Finnish, and non-Finnish European) to calculate the weighted-average of r^2. Since gnomAD LD matrices were not available for Central and South Asian and Middle Eastern, we did not use their sample sizes for the calculation. All the variants were processed on the human genome assembly GRCh38.

3.4.5 Fine-mapping results of complex traits and *cis*-eQTL

We retrieved our previous fine-mapping results for 1) complex traits in large-scale biobanks (BBJ [171], FinnGen [172], and UKBB [76] Europeans) [67,68] and 2) *cis*-eQTLs in GTEx [103] v8 and eQTL Catalogue [191]. Briefly, we conducted multiple-causal-variant fine-mapping (FINEMAP [28,29] and SuSiE [30]) of complex trait GWAS (# unique traits = 148) and *cis*-eQTL gene expression (# unique tissues/cell-types = 69) using summary statistics and in-sample LD. Detailed fine-mapping methods are described elsewhere [67,68].

In this study, we collected 1) high-PIP GWAS variants that achieved PIP > 0.9 for any traits in any biobank and 2) high-PIP *cis*-eQTL variants that acheived PIP > 0.9 for any gene expression in any tissues/cell-types. All the variants were originally processed on the human genome assembly GRCh37 and lifted over to the GRCh38 for comparison.

Additional fine-mapping results

To compare with the GBMI meta-analyses, we additionally conducted multi-causal-variant fine-mapping of four additional endpoints (gout, heart failure, thyroid cancer, and venous throm-

boembolism) that were not fine-mapped in our previous study[67,68]. We used exactly the same fine-mapping pipeline (FINEMAP[28,29] and SuSiE[30]) as described previously[67,68]. For UKBB Europeans, to use the exact same samples that contributed to the GBMI, we used individuals of European ancestry ($n = 420,531$) as defined in the Pan-UKBB project (https://pan.ukbb.broadinstitute.org), instead of those of "white British ancestry" ($n = 361,194$) used in our previous study[67,68].

3.4.6 Enrichment analysis of likely causal variants

To validate SLALOM performance, we asked whether suspicious and non-suspicious loci were enriched for having likely causal variants as a lead PIP variant, and for containing them in the 95% and 99% CS. We defined likely causal variants using 1) nonsynonymous coding variants, *i.e.*, pLoF and missense variants annotated[168] by the Ensembl Variant Effect Predictor (VEP) v101 (using GRCh38 and GENCODE v35), 2) the high-PIP (> 0.9) complex trait fine-mapped variants, and 3) the high-PIP (> 0.9) *cis*-eQTL fine-mapped variants from our previous studies as described above.

We estimated enrichment for suspicious and non-suspicious loci as a relative risk (*i.e.*, a ratio of proportion of variants) between being in suspicious/non-suspicious loci and having the annotated likely causal variants as a lead PIP variant (or containing them in the 95% or 99% CS). That is, a relative risk = (proportion of non-suspicious loci having the annotated variants as a lead PIP variant) / (proportion of suspicious loci having the annotated variants as a lead PIP variant). We computed 95% confidence intervals using bootstrapping.

3.4.7 Comparison of fine-mapping results between the GBMI and individual biobanks

To directly compare with fine-mapping results from the GBMI meta-analyses, we used our fine-mapping results of nine disease endpoints (asthma[259], COPD[259], gout, heart failure[267], IPF[257], primary open angle glaucoma[268], thyroid cancer, stroke[269], and venous thromboembolism[270]) in BBJ[171], FinnGen[172], and UKBB[76] Europeans that were also part of the GBMI meta-analyses for the same traits. For comparison, we computed the maximum PIP for each variant and the minimum size of 95% CS across BBJ, FinnGen, and UKBB. We restricted the 95% CS in biobanks to those that contain the lead variants from the GBMI. We defined the PIP difference between the GBMI and individual biobanks as ΔPIP = PIP (GBMI) the maximum PIP across the biobanks.

We conducted functional enrichment analysis to compare between the GBMI meta-analysis and individual biobanks because unbiased comparison of PIP requires conditioning on likely causal variants independent of the fine-mapping results, and functional annotations have been shown to be enriched for causal variants. Using functional categories (coding [pLoF, missense, and synonymous], 5'/3' UTR, promoter, and CRE) from our previous study[67,68], we estimated functional enrichments of variants in each functional category based on 1) top PIP rankings and 2) ΔPIP bins. Since fine-mapping PIP in the GBMI meta-analysis can be miscalibrated, we performed a comparison based on top PIP rankings to assess whether the ordering given by GBMI PIPs is more informative than the ordering given by the biobanks. For the top PIP rankings, we took the top 0.5%, 0.1%, and 0.05% variants based on the PIP rankings in the GBMI and individual biobanks. We computed enrichment as a relative risk = (proportion of top X% PIP variants in the GBMI that are in the annotation) / (proportion of top X% PIP variants in the individual biobanks that are in the annotation). For ΔPIP bins, we defined three bins using different thresholds (θ = 0.01, 0.05, and 0.1): 1) decreased PIP bin, $\Delta PIP < -\theta$, 2) null bin, $\theta \leq \Delta PIP \leq \theta$, and 3) increased PIP bin, $\theta < \Delta PIP$.

We computed enrichment as a relative risk = (proportion of variants in the decreased/increased PIP bin that are in the annotation) / (proportion of variants in the null PIP bin). We combined coding, UTR, and promoter categories for this analysis due to the limited number of variants for each bin.

3.5 Data availability

The GBMI summary statistics for the 14 endpoints are available at https://www.globalbiobankmeta.org/resources and are browserble at the GBMI PheWeb[235] website (http://results.globalbiobankmeta.org/).

3.6 Code availability

The SLALOM software is available at https://github.com/mkanai/slalom. Custom scripts to perform all the analyses and generate all the figures are available at https://github.com/mkanai/slalom-paper.

3.7 Acknowledgements

We acknowledge all the participants and researchers of the 19 biobanks that have contributed to the GBMI. Biobank-specific acknowledgements are available in the online version of the manuscript. We thank H. Huang, A.R. Martin, B.M. Neale, Y. Okada, K. Tsuo, J.C. Ulirsch, Y. Wang, and all the members of Finucane and Daly labs for their helpful feedback. M.K. was supported by a Nakajima Foundation Fellowship and the Masason Foundation. H.K.F. was funded by NIH grant DP5 OD024582.

3.8 Author contributions

M.K., M.J.D, and H.K.F. designed the study. M.K., R.E. and W.Z. performed analyses. H.K.F supervised this work. H.K.F. and M.K. obtained funding. M.K., R.E., M.J.D., and H.K.F. wrote the manuscript with input from all authors.

The work presented in this chapter was published as Weissbrod, O.*, Kanai, M.*, & Shi, H.* *et al.* (*Nat. Genet.*, 2022)[70].

4

Leveraging fine-mapping and multi-population training data to improve cross-population polygenic risk scores

Abstract

Polygenic risk scores (PRS) suffer reduced accuracy in non-European populations, exacerbating health disparities. We propose PolyPred, a method that improves cross-population PRS by combining two predictors: a new predictor that leverages functionally informed fine-mapping to estimate causal effects (instead of tagging effects), addressing LD differences; and BOLT-LMM, a published predictor. When a large training sample is available in the non-European target population, we propose PolyPred+, which further incorporates the non-European training data. We applied PolyPred to 49 diseases/traits in 4 UK Biobank populations using UK Biobank British training data, and observed relative improvements vs. BOLT-LMM ranging from +7% in South Asians to +32% in Africans, consistent with simulations. We applied PolyPred+ to 23 diseases/traits in UK Biobank East Asians using both UK Biobank British and Biobank Japan training data, and observed improvements of +24% vs. BOLT-LMM and +12% vs. PolyPred. Summary statistic-based analogues of PolyPred and PolyPred+ attained similar improvements.

4.1 Introduction

Polygenic risk scores (PRS) can identify individuals at elevated risk of complex diseases, providing opportunities for preventative action[31,32,282–285]. However, many studies have shown that PRS based on European training data attain lower accuracy when applied to populations of non-European ancestry[41–60]. This loss of accuracy is primarily driven by LD differences[46–49], allele frequency differences (including population-specific SNPs)[47,48,61], and causal effect size differences[46–48,62–65], though differences in heritability also play a minor role[47,48,66]. PRS based on non-European training data do not suffer from these limitations, but are currently limited by much

smaller training sample sizes [31,35,46–49,55] (however, lower non-European target sample sizes do not impact prediction accuracy). The development of new methods to reduce this gap in cross-population PRS accuracy has the potential to ameliorate health disparities [47].

Here, we propose PolyPred, which linearly combines two complementary predictors derived from European training data: (1) PolyFun-pred, a new predictor that circumvents LD differences by applying genome-wide functionally informed fine-mapping [13,118] to precisely estimate causal effects (instead of tagging effects); and (2) BOLT-LMM [109,159], a published predictors that analyzes all loci jointly and can capture all signals in extremely polygenic loci. BOLT-LMM requires individual-level training data. If individual-level training data is not available, we propose two analogous methods: (i) PolyPred-S, which linearly combines PolyFun-pred with SBayesR [36], and (ii) PolyPred-P, which linearly combines PolyFun-pred with PRS-CS [37]. Recommendations for when to use PolyPred, PolyPred-S, or PolyPred-P are provided below.

In the special case where there exists a large (*e.g.*, $N \geq 50$K) non-European training sample from the target population (or a closely related population), we propose PolyPred+, a polygenic prediction method that leverages both European and non-European training data. PolyPred+ linearly combines (1) PolyFun-pred; (2) BOLT-LMM; and (3) BOLT-LMM-pop, which is obtained by applying BOLT-LMM to the non-European training data, addressing MAF differences and causal effect size differences. If individual-level training data is not available, we propose the alternative methods PolyPred-S+ and PolyPred-P+, which replace BOLT-LMM with either SBayesR or PRS-CS, respectively. Recommendations for when to use PolyPred+, PolyPred-S+, or PolyPred-P+ are provided below.

We compared PolyPred and PolyPred+ (and their summary statistic-based analogues) to state-of-the-art polygenic prediction methods via simulations and analyses of 49 diseases and complex traits in 4 populations from the UK Biobank [76], additionally incorporating Biobank Japan [170] and Uganda-APCDR [286,287] to increase non-European training sample size and avoid cohort effects.

We conclude that PolyPred and its summary statistic-based analogues substantially increase cross-population polygenic prediction accuracy, and that PolyPred+ and its summary statistic-based analogues further increases cross-population prediction accuracy in the special case where non-European training data is available in large sample size.

4.2 Results

4.2.1 Overview of Methods

PolyPred combines two complementary predictors: PolyFun-pred and BOLT-LMM (**Table 4.1** and **Fig. 4.1a**). PolyFun-pred is a new predictor that leverages genome-wide functionally informed fine-mapping [13,118] to estimate posterior mean causal effects (instead of tagging effects; see **D.1 Supplementary Note**) for all SNPs with European MAF ≥ 0.1% (18 million SNPs in this study) by applying PolyFun + SuSiE [118] to European training data across 2,763 overlapping 3Mb loci. Leveraging fine-mapped posterior mean causal effects for cross-population polygenic prediction aims to address LD differences between populations. BOLT-LMM [109,159] is a published predictor that estimates posterior mean tagging effects of common SNPs (1.2 million HapMap 3 SNPs [288] in this study) using European individual-level training data. Combining PolyFun-pred with BOLT-LMM is advantageous because they have complementary advantages: PolyFun-pred estimates causal effects rather than tagging effects. BOLT-LMM estimates tagging effects, but it analyzes all loci jointly and it can potentially capture all signals in extremely polygenic loci (**4.4 Methods**).

In the special case where a large training sample is available in the target population (or a closely related population), we propose PolyPred+, which combines three complementary predictors: PolyFun-pred, BOLT-LMM, and BOLT-LMM-pop (**Table 4.1** and **Fig. 4.1b**); BOLT-LMM-pop refers to application of BOLT-LMM to common SNPs (1.2 million HapMap 3 SNPs in this study) using training data from the non-European target population, addressing MAF differences and

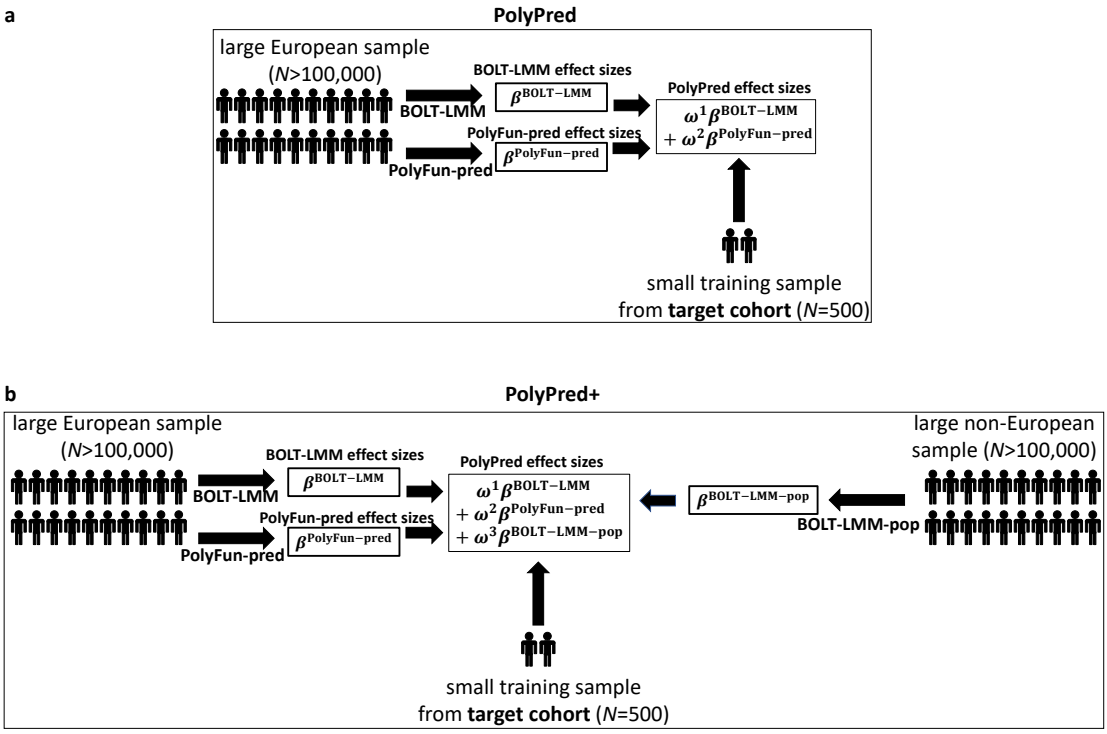


Figure 4.1: Overview of PolyPred and PolyPred+. a, Overview of PolyPred. PolyPred linearly combines the effect sizes of BOLT-LMM ($\hat{\beta}^{\text{BOLT-LMM}}$) and PolyFun-pred ($\hat{\beta}^{\text{PolyFun-pred}}$) (trained using European training data). It uses a small training sample from the target population to estimate mixing weights (ω^1, ω^2) for the constituent methods. **b**, Overview of PolyPred+. PolyPred+ linearly combines the effect sizes of BOLT-LMM ($\hat{\beta}^{\text{BOLT-LMM}}$), PolyFun-pred ($\hat{\beta}^{\text{PolyFun-pred}}$) (trained using European training data) and BOLT-LMM-pop ($\hat{\beta}^{\text{BOLT-LMM-pop}}$) (trained using non-European training data from the target population). It uses a small training sample from the target population to estimate mixing weights (ω^1, ω^2, ω^3) for the constituent methods. PolyPred-S and PolyPred-P (respectively, PolyPred-S+ and PolyPred-P+) replace all instances of BOLT-LMM with SBayesR or PRS-CS, respectively.

Table 4.1: Summary of main methods evaluated.

Method	Constituent methods	SNP set	Training data	Fine-mapped effect sizes	Summary statistics	Ref.
P+T	-	All (18 million)	Eur	No	Yes	289,290
BOLT-LMM	-	HapMap 3 (1.2 million)	Eur	No	No	109,159
SBayesR	-	HapMap 3 (1.2 million)	Eur	No	Yes	36
PRS-CS	-	HapMap 3 (1.2 million)	Eur	No	Yes	37
PolyPred	PolyFun-pred, BOLT-LMM	All (18 million)	Eur	Yes	No	This work
PolyPred-S	PolyFun-pred, SBayesR	All (18 million)	Eur	Yes	Yes	This work
PolyPred-P	PolyFun-pred, PRS-CS	All (18 million)	Eur	Yes	Yes	This work
PolyPred+	PolyFun-pred, BOLT-LMM, BOLT-LMM-pop	All (18 million)	Eur + target pop	Yes	No	This work
PolyPred-S+	PolyFun-pred, SBayesR, SBayesR-pop	All (18 million)	Eur + target pop	Yes	Yes	This work
PolyPred-P+	PolyFun-pred, PRS-CS, PRS-CS-pop	All (18 million)	Eur + target pop	Yes	Yes	This work

For each method we report its constituent methods (or "-" for individual methods), the set of SNPs analyzed in model training using UK Biobank training data (and its size when restricted to imputed UK Biobank SNPs with European MAF ≥ 0.1% and INFO score ≥ 0.6), the training data analyzed, whether it incorporates fine-mapped effect sizes (as opposed to tagging effect sizes), whether it can work with summary statistics, and the corresponding reference. Eur: European; target pop: non-European target population; Method-pop: Method applied to training data from non-European target population.

causal effect size differences.

PolyPred computes linear combinations of the estimated effect sizes of their constituent predictors:

$$\hat{\beta}_i^{\text{PolyPred(+)}} = \sum_j w^j \hat{\beta}_i^j, \tag{4.1}$$

where i indexes SNPs, j indexes the constituent predictors (PolyFun-pred and BOLT-LMM for PolyPred;PolyFun-pred, BOLT-LMM and BOLT-LMM-pop for PolyPred+), $\hat{\beta}_i^{\text{PolyPred(+)}}$ is the PolyPred (+) per-allele effect size of SNP i, w^j are method-specific weights, and $\hat{\beta}_i^j$ is the per-allele effect size of SNP i for method j (or 0 if SNP i was not considered by method j). Predicted phenotypes are computed by applying effect sizes to target genotypes:

$$\hat{y} = \sum_i x_i \hat{\beta}_i^{\text{PolyPred(+)}} \tag{4.2}$$

where $\hat{y}$ is the predicted phenotype of an individual from the target population and x_i is the number of minor alleles of SNP i carried by the individual. The mixing weights w^j in **Equation 4.1** are estimated via non-negative least squares regression using a small number of training individuals from

the target population (500 in this study), regressing true phenotypes on a linear combination of the constituent predictors (which are computed as in **Equation 4.2**).

PolyPred requires individual-level training data for its BOLT-LMM component. If only summary statistics (and summary LD information) are available, we propose two analogous methods (**Table 4.1**): (i) PolyPred-S, which linearly combines PolyFun-pred and SBayesR[36]; and (ii) PolyPred-P, which linearly combines PolyFun-pred and PRS-CS[37]. We also propose the analogous methods PolyPred-P+ and PolyPred-S+ (**Table 4.1**). Further details of PolyPred and PolyPred+ (and their summary statistic-based analogues) are provided in **4.4 Methods**; we have publicly released open-source software implementing these methods (see **4.6 Code availability**).

We evaluate prediction accuracy for each method and target population using relative-R^2, defined as the R^2 obtained in the target non-European population (after correcting for covariates and potential confounders; see **4.4 Methods**) divided by the R^2 obtained by BOLT-LMM in UK Biobank non-British Europeans (employing the same correction), using the same training data for the numerator and the denominator. This quotient transforms the prediction accuracies from an absolute scale to a scale of relative improvement (vs. the BOLT-LMM predictor in the UK Biobank non-British European target population), which is invariant to factors such as training sample size and trait heritability. For disease traits, we additionally evaluated the area under the receiving operating characteristic. We provide further details in the **4.4 Methods** section. We compare PolyPred and PolyPred+ (and their summary statistic-based analogues) to 4 published methods: LD-pruning + P-value thresholding (P+T)[289,290], BOLT-LMM[109,159], SBayesR[36], and PRS-CS[37] (**Table 4.1**).

Our recommendation for which version of PolyPred to use (**Table 4.1**) depends on three factors: (i) whether individual-level training data is available; (ii) the size and consistency of matched ancestry of the LD reference panel (if individual-level training data is not available); and (iii) whether non-European training data is available. Our results for the underlying constituent methods are summarized in **Table 4.2** (detailed below), and our recommendations are summarized in **Fig. 4.2**.

Table 4.2: Summary of the relative performance of constituent PRS methods.

LD	BOLT-LMM	SBayesR	PRS-CS	Figure(s)/Table(s)
Individual-level data (UKB, N=337K)	✓✓	✓	✓	**Fig. 4.3,4.4,4.4**
In-sample LD (UKB, N=337K)	—	✓✓	✓	**Fig. 4.3,4.4,4.4**
Very large unmatched LD (UKB, N=337K)	—	✓	✓✓	**Fig. 4.5**
Small unmatched LD (1000G, N=489)	—	✗	✓✓*	**Supplementary Tables D.4–D.6**

For each of three constituent PRS methods (BOLT-LMM, SBayesR, PRS-CS) we report its relative performance in prediction in UK Biobank non-British Europeans under various settings; we also provide links to the corresponding Figure(s)/Table(s). ✓✓: the method is significantly more accurate than the second best method in the same row, and combining this method with PolyFun-pred increases prediction accuracy; ✓✓*: the method is significantly more accurate than the second best method in the same row, and combining this method with PolyFun-pred does not increase prediction accuracy; ✓: the method is significantly less accurate than the best method in the same row, but is significantly more accurate than P+T; ✗: the method is not significantly more accurate than P+T; —: the method is not applicable, because it requires individual-level data. For very large unmatched LD (a likely scenario when analyzing summary statistics from a meta-analysis of many cohorts), we performed real trait analyses only, as simulations would have required another very large individual-level data set in addition to UK Biobank (see **D.1 Supplementary Note**). For individual-level data, the difference between BOLT-LMM and the second-best method was significant in simulations but non-significant in real trait analyses. For In-sample LD, the difference between SBayesR and PRS-CS was significant in simulations but non-significant in real traits analyses. For very large unmatched LD (a likely scenario when analyzing summary statistics from a meta-analysis of many cohorts), we performed real trait analyses only (see explanation in **D.1 Supplementary Note**). For small unmatched LD, we performed both simulations and real trait analyses but report results of real trait analyses, which we believe to be most reflective of real-life settings (in simulations, SBayesR was significantly more accurate than PRS-CS). Results for non-European target populations from UK Biobank were similar, though some of the differences were not statistically significant due to smaller prediction accuracies and sample sizes. We have facilitated the use of very large LD reference panels for European training data by publicly releasing summary LD information for N = 337K British-ancestry UK Biobank samples across 18 million SNPs (see **4.5 Data availability**).

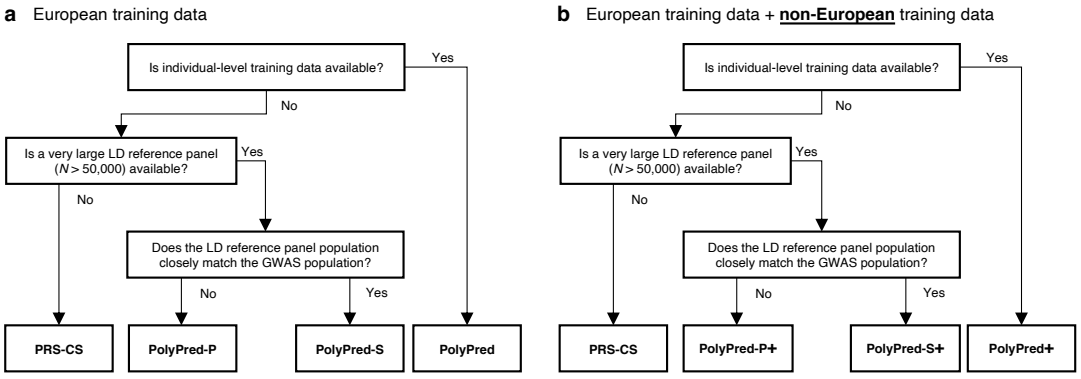


Figure 4.2: Recommendations for the application of PolyPred, PolyPred+ and related methods. a, Flowchart of recommendations when only European training data are available. **b**, Flowchart of recommendations when both European and non-European training data are available. We note that, when working with summary statistics from a meta-analysis of many cohorts, there is typically no LD reference panel that closely matches the GWAS population. Also, it is possible that the answers to the flowchart questions are different for European versus non-European training data, in which case the recommendation would be to use a hybrid method based on the answers to each flowchart in turn (for example, PolyFun-pred + BOLT-LMM + PRS-CS-pop; not listed in **Table 4.1**). For both a and b, we recommend training PolyFun-pred using a very large LD reference panel (for example, n = 337,000 UK Biobank British) with a dense SNP set (for example, 8 million SNPs). We have facilitated this by publicly releasing summary LD information for n = 337,000 British-ancestry UK Biobank samples across 18 million SNPs (**4.5 Data availability**).

4.2.2 Simulations with in-sample LD

We compared PolyPred, PolyPred-S and PolyPred-P to P+T, BOLT-LMM, SBayesR, and PRS-CS via simulations, using real genotypes or in-sample LD from the UK Biobank[76]. We trained each method using 337,491 unrelated British-ancestry individuals[76], and computed predictions in four target populations: non-British Europeans, South Asians, East Asians, and Africans. We estimated mixing weights for PolyPred, PolyPred-S and PolyPred-P using 500 individuals from the target population. We evaluated prediction accuracy using held-out individuals from each target population that were not included in the training sets: 42K non-British Europeans, 7.7K South Asians, 0.9K East Asians, and 6.2K Africans. We computed PRS using 250,963 MAF ≥ 0.1 % SNPs with INFO score ≥ 0.6 on chromosome 22.

Generative trait architectures were specified as follows. We simulated traits with polygenicity (genome-wide proportion of causal SNPs) equal to either 0.1% (less polygenic) or 0.3% (more polygenic) and heritability equal to 5%. We specified prior causal probabilities for each SNP in proportion to per-SNP heritabilities, which we generated for each SNP based on its British LD, MAF, and functional annotations, using the baseline-LF model[161]. For each causal SNP, we sampled ancestry-specific causal effect sizes from a multivariate normal distribution assuming cross-population genetic correlations of 0.8 (ref.[47,64]). Other parameter settings were explored in secondary analyses (see below).

We computed relative-R^2 for each method, target population, and trait architecture, averaged across 100 simulations. In addition to the simulations with in-sample LD described below, we also performed simulations with reference panel LD (**D.1 Supplementary Note**; also see **Table 4.2**). Further details of the simulation framework are provided in **4.4 Methods**.

The simulation results are reported in **Fig. 4.3** and **Supplementary Tables D.1** (also see **Table 4.2**). PolyPred was the most accurate method in each target population, with relative improve-

ments vs. BOLT-LMM (resp. P-values for improvement) ranging from +13% in non-British Europeans ($P < 10^{-16}$) to +65% in Africans ($P < 10^{-16}$) for the less polygenic architecture, and from +2% in non-British Europeans ($P = 0.0001$) to +17% in Africans ($P = 10^{-8}$) for the more polygenic architecture. PolyPred-S and PolyPred-P performed slightly worse than PolyPred, but were substantially and significantly more accurate than their corresponding constituent methods. Among the remaining methods, BOLT-LMM was consistently the most accurate and P+T was consistently the least accurate method, far underperforming the other methods (despite its widespread recent use[45,47–52,57,65,175,291–294]). We note that the higher accuracy of BOLT-LMM vs. SBayesR and PRS-CS does not imply that BOLT-LMM is a superior method, as BOLT-LMM analyzes individual-level training data whereas SBayesR and PRS-CS analyze summary statistics.

We additionally performed many secondary analyses to investigate the sensitivity of the results to the simulation parameters, the SNP set and the functional annotations, and to evaluate the computational cost and memory cost of each method (**D.1 Supplementary Note**, **Supplementary Tables D.1,D.2**).

We conclude that PolyPred and its summary statistic- based analogues are more accurate than BOLT-LMM, SBayesR, PRS-CS, and P+T, with small but significant improvements vs. BOLT-LMM in Europeans and substantial improvements in Africans.

4.2.3 PRS in 4 UK Biobank populations using British training data

We applied PolyPred and its summary statistic-based analogues to 49 diseases and complex traits from the UK Biobank, analyzing 4 target populations (**4.4 Methods**, **Supplementary Table D.3**). As in our simulations, we used UK Biobank British training data (average $N = 325$K) to estimate SNP effect sizes; used 500 additional individuals from the target population to estimate mixing weights; evaluated prediction accuracy using individuals from each of the 4 target populations that were not included in the training data: 42K non-British Europeans, 7.7K South Asians, 0.9K East

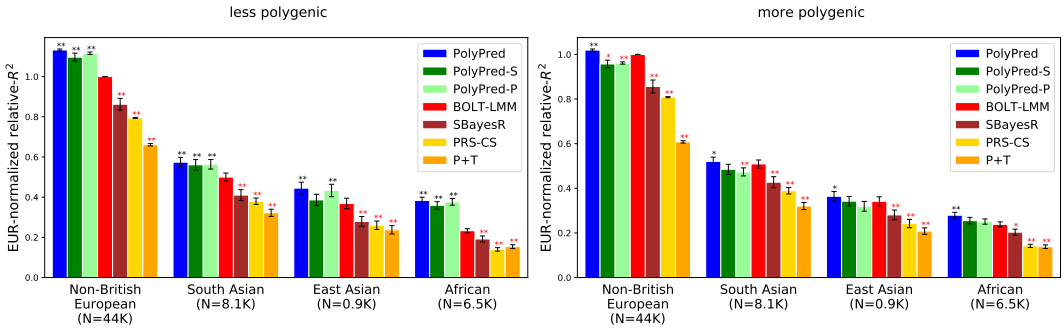


Figure 4.3: Cross-population PRS results for simulated UK Biobank traits using in-sample LD. We report average prediction accuracy (relative R^2; see text) for PRSs trained in UK Biobank British samples (n = 337,000) and applied to four UK Biobank target populations across 100 simulated traits with less polygenic (0.1% of SNPs causal; left panel) or more polygenic (0.3% of SNPs causal; right panel) architectures. Target population sample sizes are indicated in parentheses; PolyPred and its summary statistics-based analogs used 500 additional training samples from each target population to estimate mixing weights. Asterisks above each bar denote statistical significance of the difference versus BOLT-LMM, with black asterisks denoting an advantage and red asterisks a disadvantage ($^*P < 0.05$; $^{**}P < 0.001$). P values were computed using a two-sided Wald's test and were not adjusted for multiple comparisons. Error bars denote s.e. Numerical results, absolute prediction accuracies (R^2) and P values of relative improvements versus BOLT-LMM are reported in **Supplementary Table D.1**.

Asians, and 6.2K Africans; and compared PolyPred and its summary statistic-based analogues to P+T, BOLT-LMM, SBayesR, and PRS-CS. We meta-analyzed relative-R^2 across traits by restricting to 7 well-powered, independent complex traits from the UK Biobank[76] ($|r_g| < 0.3$; see **4.4 Methods** and **Supplementary Table D.3**) that were also available in Biobank Japan and in Uganda-APCDR (see below). We have publicly released SNP effect sizes used for prediction for each of the 4 methods (see **4.5 Data availability**).

We computed relative-R^2 for each method and target population. The results are summarized in **Fig. 4.4** and provided in **Supplementary Tables D.4–D.6** (also see **Table 4.2**). Among the published methods, BOLT-LMM attained the highest prediction accuracy in all target populations (differences between BOLT-LMM and SBayesR were small and not statistically significant). P+T was much less accurate than the other methods (despite its widespread recent use[45,47–52,57,65,175,291–294]), suffering relative losses of 37-50% vs. BOLT-LMM. We thus used BOLT-LMM as a benchmark.

Among all 7 methods, PolyPred attained the highest prediction accuracy in each target population. Improvements in average relative-R^2 of PolyPred vs. BOLT-LMM were equal to +7.5% in non-British Europeans ($P = 0.05$), +6.8% in South Asians ($P = 0.02$), +11% in East Asians ($P = 0.12$) and +32% in Africans ($P = 0.02$). The larger improvement in Africans reflects the larger LD differences vs. British training data, due to earlier divergence times[47,48,295]. The lack of statistical significance in East Asians reflects the low power to detect significant differences in very small target samples. PolyPred-S and PolyPred-P were consistently the second and third most accurate methods, respectively, with statistically significant improvements vs. their constituent methods. We additionally verified that PolyPred was well-calibrated (*i.e.*, regressing the true phenotype on the predicted phenotype yields a slope of 1) in all target populations, whereas the alternative methods were not always well-calibrated (**Supplementary Tables D.4–D.6**, **D.1 Supplementary Note**). Despite the improvements attained by PolyPred, the reductions in prediction accuracy in non-European populations remained significant ($P < 0.002$), with meta-analyzed absolute R^2 equal to 0.17 in non-British Europeans, 0.11 in South Asians, 0.093 in East Asians, and 0.053 in Africans (**4.4 Methods**, **Supplementary Tables D.4,D.5**).

As a secondary analysis, we meta-analyzed the results of each method across three independent diseases: type 2 diabetes, asthma, and all autoimmune disease (**4.4 Methods**); these diseases were not included in our primary meta-analyses due to low heritabilities. PolyPred attained the highest prediction accuracy for each target population and each disease, except for East Asians (where standard errors were large due to the small sample size) and for type 2 diabetes in non-British Europeans (where BOLT-LMM performed slightly but non-significantly better) (**Supplementary Table D.4**). We performed additional secondary analyses to evaluate the impact of the LD reference panel and the SNP set on prediction accuracy, to evaluate additional methods, and to evaluate the results when modifying the parameters of PolyPred and the other evaluated methods (**D.1 Supplementary Note**, **Supplementary Tables D.4–D.7**).

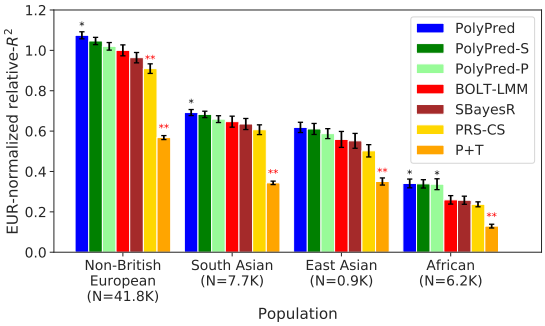


Figure 4.4: Cross-population PRS results for real UK Biobank traits. We report average prediction accuracy (relative R^2; see text), meta-analyzed across seven well-powered, independent traits, for PRSs trained in UK Biobank British samples (average n = 325,000) and applied to four UK Biobank target populations. Target population sample sizes are indicated in parentheses; PolyPred and its summary statistics-based analogs used 500 additional training samples from each target population to estimate mixing weights. Asterisks above each bar denote statistical significance of the difference versus BOLT-LMM, with black asterisks denoting an advantage ($^*P < 0.05$; $^{**}P < 0.001$). P values were computed using a two-sided Wald's test and were not adjusted for multiple comparisons. Error bars denote s.e. Numerical results, results for all 49 traits analyzed, absolute prediction accuracies (R^2) and P values of relative improvements versus BOLT-LMM are reported in **Supplementary Tables D.4–D.6**.

We conclude that PolyPred and its summary statistic-based analogues substantially increase cross-population polygenic prediction accuracy vs. published methods (with a particularly large improvement in Africans), consistent with simulations. However, there remains a large gap in cross-population polygenic prediction accuracy as compared to Europeans.

4.2.4 PRS USING ENGAGE META-ANALYSIS TRAINING DATA

We sought to analyze training data consisting of summary statistics for real traits from a meta-analysis of many European cohorts, for which in-sample LD is generally not available. We analyzed 8.1 million meta-analyzed summary statistics from the European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium [296–298] for four traits (BMI, waist-hip-ratio (adjusted for BMI), total cholesterol, and triglycerides; average $N = 61{,}365$), and evaluated the prediction accuracy using the same four UK Biobank populations analyzed previously. For each method, we used an LD reference panel based on UK Biobank British individuals; we emphasize that unlike the other

primary analyses, the LD reference panel was misspecified, because it was not based on in-sample LD. We excluded methods that require individual-level training data (BOLT-LMM and PolyPred) from this analysis.

The results are summarized in **Supplementary Fig. D.1** and reported in **Supplementary Tables D.5,D.8** (also see **Table 4.2**). Briefly, PolyPred-P was generally the most accurate method, and PRS-CS outperformed SBayesR (with a significant improvement for non-British Europeans and Africans), consistent with a previous study[299] (unlike our analysis of UK Biobank training data, where SBayesR outperformed PRS-CS; **Fig. 4.4**). However, differences between similarly performing methods were generally not statistically significant (due to moderately large standard errors), and thus caution should be exercised in their interpretation; for this reason, we did not perform secondary analyses to further assess differences between methods.

We conclude that PolyPred-P can increase cross-population polygenic prediction accuracy vs. published methods when analyzing summary statistics from a meta-analysis of many cohorts.

4.2.5 PRS in Biobank Japan and Uganda-APCDR cohorts

We applied PolyPred and its summary statistic-based analogues to predict 23 diseases and complex traits in Biobank Japan[170] and 7 complex traits in Uganda-APCDR, an African-ancestry cohort[286,287] (**4.4 Methods**, **Supplementary Table D.3**). We performed these experiments to avoid training effect sizes and testing predictions in the same cohort, which may produce inflated prediction accuracies[35,300–302]. We again used UK Biobank British training data (average N=325K) to estimate SNP effect sizes, and used 500 individuals from the target population to estimate mixing weights. We evaluated prediction accuracy using individuals from each of the 2 target cohorts that were not included in the training data: 5K Biobank Japan individuals and 1.3K Uganda-APCDR individuals. We again compared PolyPred and its summary statistic-based analogues to P+T, BOLT-LMM, SBayesR, and PRS-CS. We meta-analyzed relative-R^2 across the same 7 well-powered, inde-

pendent complex traits used in the UK Biobank analyses (**Supplementary Table D.3**).

The results are summarized in **Fig. 4.5** and reported in **Supplementary Tables D.5,D.9**. Among the published methods, we again observed that BOLT-LMM attained the highest prediction accuracy in each target population, and that P+T was substantially less accurate than the other methods. Among all 7 methods, PolyPred attained the highest prediction accuracy in Biobank Japan, and PolyPred-P attained the highest prediction accuracy in Uganda-APCDR (although the difference between PolyPred and PolyPred-P in Uganda-APCDR was not statistically significant). Improvements of PolyPred vs. BOLT-LMM in average relative-R^2 were equal to +13% in Biobank Japan ($P = 2 \times 10^{-6}$) and +22% in Uganda-APCDR ($P = 0.26$), similar to our UK Biobank results above. We observed similar improvements for PolyPred-S vs. SBayesR and PolyPred-P vs. PRS-CS (both of which were statistically significant in Biobank Japan). Prediction accuracy for each method was much smaller in Biobank Japan and Uganda-APCDR (*e.g.* 0.32 and 0.11 for PolyPred; **Fig. 4.5**) than in UK Biobank East Asians and UK Biobank Africans (0.62 and 0.34; **Fig. 4.4**), likely due to higher SNP-heritabilities in the UK Biobank (see below). We also applied PolyPred+ and its summary statistic-based analogues to Biobank Japan, incorporating additional Biobank Japan training data (average N = 124K), with the caveat that this analysis involved training and testing in the same cohort (**4.4 Methods**). PolyPred+ attained increased prediction accuracy, with a further +23% improvement vs. PolyPred ($P = 0.0004$), with similar results for PolyPred-S+ and PolyPred-P+ (**Supplementary Tables D.5,D.9**).

We performed additional experiments to investigate the above result of decreased prediction accuracy in Biobank Japan vs. UK Biobank East Asians. We matched the BOLT-LMM British training sample size to the Biobank Japan training sample size, and obtained a relative-R^2 in UK Biobank non-British Europeans (using UK Biobank British training samples) +108% larger than in Biobank Japan (using Biobank Japan training samples), consistent with the +104% increase expected from theory[61,62] based on the +67% higher SNP-heritabilities in UK Biobank (**Supplementary**

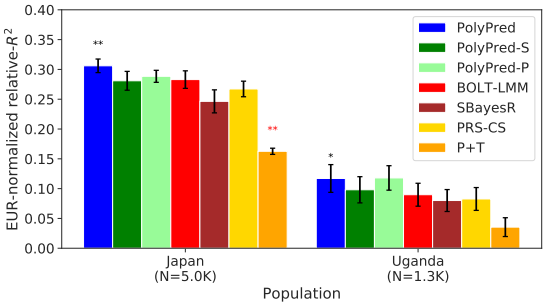


Figure 4.5: Cross-population PRS results for Biobank Japan and Uganda-APCDR traits. We report average prediction accuracy (relative R^2; see text), meta-analyzed across seven well-powered, independent traits, for PRSs trained in UK Biobank British samples (average n = 325,000) and applied to Biobank Japan and Uganda-APCDR target populations. Target population sample sizes are indicated in parentheses; PolyPred and its summary statistics-based analogs used 500 additional training samples from each target population to estimate mixing weights. Asterisks above each bar denote statistical significance of the difference versus BOLT-LMM, with black asterisks denoting an advantage and red asterisks a disadvantage ($^*P < 0.05$; $^{**}P < 0.001$). P values were computed using a two-sided Wald's test and were not adjusted for multiple comparisons. Error bars denote the s.e. Numerical results, results for all 23 traits analyzed, absolute prediction accuracies (R^2) and P values of relative improvements versus BOLT-LMM are reported in **Supplementary Table D.9**.

Table D.10, D.1 Supplementary Note). This suggests that differences in SNP-heritability due to ancestry or cohort differences may explain most of the differences in prediction accuracies observed between the UK Biobank and Biobank Japan. Further experiments and interpretation are provided in the **D.1 Supplementary Note**. We performed 6 additional secondary analyses to evaluate the sensitivity of the results to various factors (**D.1 Supplementary Note**, **Supplementary Tables D.5,D.9**).

We conclude that PolyPred and its summary statistic-based analogues substantially increase cross-population polygenic prediction accuracy vs. published methods when applied to target cohorts different from the training cohort.

4.2.6 PRS in East Asians using British and Japanese training data

We applied PolyPred+ and its summary statistic-based analogues to predict 23 diseases and complex traits in UK Biobank East Asians using UK Biobank British and Biobank Japan training data

(**Supplementary Table D.3**). We performed this experiment to explore the special case where non-European training data is available in large sample size from a population that is genetically similar to the target population, in a cohort that is distinct from the target cohort (previous studies considered only European training data or analyzed non-European training data from the target cohort[45,47–51]). We note that this experiment is still imperfect in that the European training data and non-European target data are from the same cohort (UK Biobank); however, we believe that cohort effects would deflate rather than inflate the relative improvement of PolyPred+ vs. other methods, since they would confer an advantage to the European training data but not the non-European training data. We used UK Biobank British training data (average $N = 325$K) and Biobank Japan training data (average $N = 124$K) to estimate SNP effect sizes. We again used 500 individuals from the target population to estimate mixing weights, and evaluated prediction accuracy using 900 UK Biobank East Asians that were not included in the training data. We compared PolyPred, PolyPred+, and their summary statistic-based analogues to P+T, BOLT-LMM, SBayesR, and PRS-CS (**4.4 Methods**). We meta-analyzed relative-R^2 across the same 7 well-powered, independent complex traits used in the previous analyses (**Supplementary Table D.3**).

The results are summarized in **Fig. 4.6** and reported in **Supplementary Tables D.4–D.6**. PolyPred+ attained the highest prediction accuracy, with a +24% improvement vs. BOLT-LMM ($P = 0.0009$) and a +12% improvement vs. PolyPred ($P = 0.0014$). This implies that incorporating non-European training data can provide a substantial advantage, if it is available in large sample size. Results for PolyPred-S+ (vs. SBayesR and PolyPred-S) and PolyPred-P+ (vs. PRS-CS and PolyPred-P) were similar. We emphasize that the +12% improvement for PolyPred+ vs. PolyPred should be viewed as a lower bound on the improvement that would be obtained in settings without cohort effects that may confer an advantage to the European training data. We performed additional secondary analyses to evaluate the sensitivity of the results to various factors (**D.1 Supplementary Note**, **Supplementary Tables D.4–D.6**).

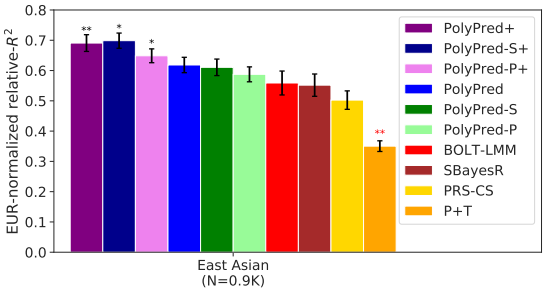


Figure 4.6: Cross-population PRS results for UK Biobank east Asians when incorporating both European and non-European training data. We report average prediction accuracy (relative R^2; see text), meta-analyzed across seven well-powered, independent traits, for PRSs trained in UK Biobank British (average n = 325,000) and Biobank Japan samples (average n = 124,000; used by PolyPred+ and its summary statistics-based analogs only) and applied to UK Biobank east Asians. The target population sample size is indicated in parentheses; PolyPred, PolyPred+ and their summary statistics-based analogs used 500 additional training samples from the target population to estimate mixing weights. Asterisks above each bar denote statistical significance of the difference versus BOLT-LMM, with black asterisks denoting an advantage and red asterisks a disadvantage ($^*P < 0.05$; $^{**}P < 0.001$). P values were computed using a two-sided Wald's test and were not adjusted for multiple comparisons. Error bars denote the s.e. Numerical results, results for all 23 traits analyzed, absolute prediction accuracies (R^2) and P values of relative improvements versus BOLT-LMM are reported in **Supplementary Tables D.4–D.6**.

We conclude that PolyPred+ and its summary statistic-based analogues further increase cross-population prediction accuracy in the special case where non-European training data from the target population (or a closely related population) is available in large sample size. We emphasize that efforts to assess the benefit of incorporating non-European training data should analyze non-European training data from a cohort that is distinct from the target cohort, otherwise results may be inflated due to cohort effects.

4.3 Discussion

We have introduced PolyPred, which improves cross-population polygenic risk prediction by incorporating causal effects in addition to tagging effects, addressing cross-population LD differences. Across seven well-powered independent traits, PolyPred significantly increased prediction accuracy over BOLT-LMM by 32% in UK Biobank Africans and by 13% in Biobank Japan (with similar re-

sults vs. SBayesR and PRS-CS). In the special case where a large training sample is available in the non-European target population (or a closely related population), we have introduced PolyPred+, which further incorporates the non-European training data, addressing MAF differences and causal effect size differences. PolyPred+ significantly increased prediction accuracy in UK Biobank East Asians over BOLT-LMM by 24% (and over PolyPred by 12%). PolyPred and PolyPred+ require individual-level training data (for their BOLT-LMM component), but we have also introduced summary statistic-based analogues of PolyPred and PolyPred+ in cases where individual-level training data is not available; specific recommendations are provided in **Fig. 4.2** (also see **Table 4.2**). In conclusion, PolyPred and its summary statistic-based analogues substantially improve cross-population polygenic prediction accuracy, ameliorating health disparities[47]. We have publicly released the PRS coefficients for all SNPs and traits analyzed under all evaluated methods (see **4.5 Data availability**).

Although we substantially improved cross-population PRS accuracy over the state of the art, prediction accuracy in non-Europeans is still substantially lower compared to Europeans, even within the UK Biobank. There are two reasons for the remaining accuracy gap. First, European sample sizes are still limited, which limits the ability of PolyFun-pred to estimate causal rather than tagging effects. Second, non-European sample sizes are limited, which limits the ability of BOLT-LMM applied to non-European samples to estimate tagging effects. Even with an infinite European training sample, which allows estimating causal effects perfectly (thus addressing LD differences), prediction accuracy could still be higher for Europeans vs. non-Europeans due to cross-population genetic correlations less than 1[47,64,244,303] and different allele frequencies (including population-specific SNPs) (**D.1 Supplementary Note**). Hence our theory and results confirm that larger non-European GWAS are the best way to further improve PRS accuracy in non- European populations[43,44,46,47,55].

Our work has several limitations, providing opportunities for future work. First, we did not evaluate a setting where the British training data, the non-British training data, and the target population are sampled from three different cohorts. Second, PolyPred requires a large number of imputed

SNPs (*e.g.* 8.1 million SNPs in the ENGAGE analysis) to perform fine-mapping, motivating the need for large cross-population imputation panels. Third, it may be possible to improve PRS accuracy for admixed individuals by using European effect sizes for European alleles and non-European effect sizes for non-European alleles [50,51]. Fourth, PolyPred and its summary statistic-based analogues are slower than alternative PRS methods (**D.1 Supplementary Note**). Fifth, PolyPred cannot use data from a fixed-effects meta-analysis of GWAS data of different populations (**D.1 Supplementary Note**). Sixth, PolyPred requires a small training sample from the target cohort to maintain calibrated predictions (**D.1 Supplementary Note**). Seventh, PolyPred prediction accuracy could in principle be improved if it were possible to decompose its constituent predictors into shared and non-shared components (**D.1 Supplementary Note**). Despite all these limitations, PolyPred and PolyPred+ and their summary statistic-based analogues provide a clear improvement for cross-population polygenic risk prediction.

4.4 Methods

4.4.1 PolyPred and its summary statistic-based analogues

All methods in this paper use a linear PRS, *i.e.*, $\hat{y} = \sum_i x_i \hat{\beta}_i$, where $\hat{y}$ is the PRS of an individual, x_i is the number of minor alleles of SNP i carried by that individual, and $\hat{\beta}_i$ is the estimated per-allele causal effect size of SNP i. The methods differ in the way they estimate $\hat{\beta}_i$.

PolyPred and PolyPred+ both combine the methods PolyFun-pred and BOLT-LMM; PolyPred-S and PolyPred-S+ both combine the methods PolyFun-pred and SBayesR; and PolyPred-P and PolyPred-P+ both combine the methods PolyFun-pred and PRS-CS. PolyFun-pred estimates $\hat{\beta}_i$ as the (approximate) data, using 187 functional annotations to specify prior causal probabilities (see below). BOLT-LMM (resp. SBayesR and PRS-CS) estimates tagging effects (**D.1 Supplementary Note**) of HapMap 3 SNPs by applying BOLT-LMM [109,159] (resp. SBayesR [36] and PRS-CS [37]) to

European training data. BOLT-LMM (resp. SBayesR) treats the effect of each SNP i as a random effect sampled from a mixture of two (resp. four) zero-mean normal distributions, whose variances and mixture weights are determined in a data-driven manner. PRS-CS treats the effect of each SNP i as a random effect sampled from a continuous shrinkage prior distribution.

PolyPred and its summary statistic-based analogues compute the effect size of each SNP i that is either in HapMap 3 or has a European MAF ≥ 0.1% and INFO score ≥ 0.6 as a weighted combination of (1) its PolyFun-pred effect size based on European training data; and (2) its BOLT-LMM (resp. SBayesR and PRS-CS) effect size based on European training data:

$$\hat{\beta}_i^{\text{PolyPred(-S)}} = w^{\text{PolyFun-pred}} \cdot \hat{\beta}_i^{\text{PolyFun-pred}} + w^{\text{BOLT-LMM/SBayesR/PRS-CS}} \cdot \hat{\beta}_i^{\text{BOLT-LMM/SBayesR/PRS-CS}}$$

where $\hat{\beta}_i^{\text{PolyFun-pred}}$ is the PolyFun-pred approximate posterior mean causal effect size of SNP i based on European training data, $\hat{\beta}_i^{\text{BOLT-LMM/SBayesR/PRS-CS}}$ is the approximate posterior mean tagging effect size of SNP i based on European training data using the indicated method (setting the effects of SNPs not in HapMap 3 to zero), and $w^{\text{PolyFun-pred}}$, $w^{\text{BOLT-LMM/SBayesR/PRS-CS}}$ are mixing weights. PolyPred estimates the mixing weights via non-negative least squares estimation (*i.e.*, least squares estimation constrained to produce to non-negative estimates) based on training individuals from the target cohort. Specifically, PolyPred (resp. PolyPred-S and PolyPred-P) estimates the mixing weights by computing the PRS corresponding to the PolyFun-pred effect sizes (given by $\hat{y}^{\text{PolyFun-pred}} = \sum_i x_i \hat{\beta}_i^{\text{PolyFun-pred}}$) and the PRS corresponding to the BOLT-LMM (resp. SBayesR and PRS-CS) effect sizes (given by $\hat{y}^{\text{BOLT-LMM}} = \sum_i x_i \hat{\beta}_i^{\text{BOLT-LMM}}$), and then fitting the mixing weights by regressing the true phenotypes y_i of of the training individuals in the target cohort on the PolyFun-pred and the BOLT-LMM (resp. SBayesR and PRS-CS) PRSs. The use of non-negative least squares estimation guarantees that the correlation of the predicted phenotype with the true phenotype is at least as large as the smallest of the correlations between each constituent predicted

phenotype and the true phenotype.

PolyPred+ and its summary statistic-based analogues compute the effect size of each SNP i that is either in HapMap 3 or has a European MAF ≥ 0.1% and INFO score ≥ 0.6 as a weighted combination of (1) its PolyFun-pred effect size based on European training data; (2) its BOLT-LMM (resp. SBayesR and PRS-CS) effect size based on European training data; and (3) its effect size as estimated by applying BOLT-LMM (resp. SBayesR and PRS-CS) to training data from the target population (or a closely related population):

$$\begin{aligned}\hat{\beta}_i^{\text{PolyPred+}} &= w^{\text{PolyFun-pred}} \cdot \hat{\beta}_i^{\text{PolyFun-pred}} \\ &+ w^{\text{BOLT-LMM/SBayesR/PRS-CS}} \cdot \hat{\beta}_i^{\text{BOLT-LMM/SBayesR/PRS-CS}} \\ &+ w^{\text{BOLT-LMM/SBayesR/PRS-CS-nonEur}} \cdot \hat{\beta}_i^{\text{BOLT-LMM/SBayesR/PRS-CS-nonEur}}\end{aligned}$$

where $\hat{\beta}_i^{\text{BOLT-LMM/SBayesR/PRS-CS-nonEur}}$ is the BOLT-LMM (resp. SBayesR or PRS-CS) approximate posterior mean tagging effect of SNP i based on training data from the non-European population (and set to zero for SNPs that are not in HapMap 3), and $w^{\text{BOLT-LMM/SBayesR/PRS-CS-nonEur}}$ is the mixing weight of $\hat{\beta}_i^{\text{BOLT-LMM/SBayesR/PRS-CS-nonEur}}$. The mixing weights are estimated as in PolyPred.

In practice, we apply PolyPred and its summary statistic-based analogues by linearly combining the PolyFun-pred PRS and the BOLT-LMM (or SBayesR or PRS-CS) PRS (rather than linearly combining the SNP effect sizes). The two procedures are almost mathematically identical, with the only difference being that a linear combination of PRSs can also accommodate an intercept, which explicitly bias-corrects the PRS to the target population.

We applied PolyFun-pred in the same way that we applied PolyFun + SuSiE in our previous work[118]. Briefly, we applied PolyFun-pred across 2,763 overlapping 3Mb loci (equally spaced starting at chromosome 1, position 0) spanning 18,212,157 European MAF > 0.1% imputed SNPs with INFO score > 0.6 (excluding the HLA and two other long-range LD regions)[118], assuming

10 causal SNPs per locus. We used summary statistics computed by BOLT-LMM, based on up to N = 337,491 unrelated British-ancestry UK Biobank individuals, and using summary LD information estimated directly from the target samples. Full details are provided in ref. [118]. We note that the use of BOLT-LMM summary statistics is mathematically equivalent to regressing the target phenotypes on BOLT-LMM off-chromosome PRS prior to applying PolyFun + SuSiE [109]. We also note that the use of 3Mb loci guarantees that for each SNP, the estimation of its causal effect size takes into account virtually all relevant SNPs that may be in LD with that SNP (because LD in European populations rarely ranges beyond 1Mb [229]), allowing to disentangle its causal effect size from its tagging effect size.

PRS methods that include non-common SNPs (MAF < 5%) may be sensitive to MAF-dependent and LD-dependent architectures [111,161,304]. Previous PRS methods have largely alleviated this concern by discarding non-common SNPs instead of explicitly modeling their lower per-SNP heritability [35–37,300,301,305–310]. In contrast, PolyFun-pred accounts for MAF-dependent and LD-dependent architectures by specifying SNP-specific prior causal probabilities based on the baseline-LF model [161] (**Supplementary Table D.11**). In detail, PolyFun-pred uses 187 overlapping functional annotations from the baseline-LF model (previously described in ref. [118]), including 10 common MAF bins (MAF ≥ 0.05); 10 low-frequency MAF bins (0.05 > MAF ≥ 0.001); 6 LD-related annotations for common SNPs; 5 LD-related annotations for low-frequency SNPs; 40 binary functional annotations for common SNPs; 7 continuous functional annotations for common SNPs; 40 binary functional annotations for low-frequency SNPs; 3 continuous functional annotations for low-frequency SNPs; and 66 annotations constructed via windows around other annotations [95] (**Supplementary Table D.11**).

4.4.2 Estimating relative-R^2 and its standard error

We measured prediction accuracy for each trait via a measure that we call relative-R^2, defined via the following computations:

1. Compute R^2-PRS: the R^2 obtained via a linear predictor that includes PRS, age, sex, age*sex (if the correlation with age was < 0.95), UK Biobank assessment center (defined via dummy binary variables), genotyping array, 10 principal components (computed separately for each ancestry; see below), and dilution factor (for biochemical traits only).

2. Compute R^2-noPRS, defined like R^2-PRS but omitting the PRS

3. Compute R^2-PRS-BOLT-EUR, computed by applying BOLT-LMM to UK Biobank non-British Europeans as in step 1

4. Compute R^2-noPRS-BOLT-EUR, computed by applying BOLT-LMM but omitting the PRS to non-British Europeans.

5. Compute relative-R^2 as (R^2-PRS $-$ R^2-noPRS) / (R^2-PRS-BOLT-EUR $-$ R^2-noPRS-BOLT-EUR).
   We note that fold improvement in relative-R^2 is the same as fold improvement in absolute difference in R^2, (*i.e.*, in R^2-PRS $-$ R^2-noPRS), because the denominator (R^2-PRS-BOLT-EUR $-$ R^2-noPRS-BOLT-EUR) is a trait-specific scaling factor.

We computed standard errors of relative-R^2, of differences in relative-R^2 (*e.g.*, vs. BOLT-LMM), of ancestry-specific regression slopes, and of the area under the receiver operating curve (for disease traits) via genomic block-jackknife, partitioning the genome into 200 equally-sized consecutive loci and omitting each one in turn. In secondary analyses, we computed standard errors by applying jackknife over individuals from the target population. These analyses yielded much smaller standard

errors in the UK Biobank, suggesting that genomic block-jackknife standard errors may be conservative, whereas individual-based jackknife estimates maty be anti-conservative. We emphasize that individual-based jackknife explicitly assumes a fixed training set.

We estimated statistics (*e.g.*, relative-R^2) for meta-analyzed traits via an inverse-variance weighted average, using weights inversely proportional to the standard error of the R^2 of BOLT-LMM in the target population (as estimated via genomic block-jackknife). We estimated the standard error of the meta-analyzed statistics as the square root of the weighted average of the trait-specific sampling variances (obtained via genomic block-jackknife), divided by the square root of the number of traits. We computed p-values of differences in relative-R^2 vs. BOLT-LMM via a Wald test.

We computed the statistical significance of the decrease in R^2 in non-European vs. European target samples via a Wald test for the difference in R^2, conservatively estimating the sampling variance of this difference as the sum of the sampling variances of the European R^2 and the non-European R^2 (this is a conservative estimate as long as the R^2 estimates in Europeans and non-Europeans are not negatively correlated, which is extremely unlikely).

4.4.3 Cohorts Analyzed

UK Biobank

The UK Biobank is a UK-based population cohort[76]. We used version 3 of the imputed genotypes, as described in our previous work[118]. We computed ancestry-specific PCs for UK Biobank Africans, UK Biobank East Asians, and UK Biobank South Asians via plink 1.9[311], restricting to SNPs that have ancestry-specific MAF > 5%, missingness < 10%, HWE p-value > 10^{-10}, and that were LD-pruned using the command `--indep-pairwise 1000 50 0.05`, and restricted to unrelated individuals (kinship coefficient <0.05) from the target ancestry with missingness < 10%. We used the UK Biobank provided PCs for UK Biobank Europeans.

We defined the 'autoimmune disease' trait in the UK Biobank as a union of the following UK Biobank codes: 1154 (irritable bowel syndrome); 1222 (type 1 diabetes); 1224 (thyroid problem); 1225 (hyperthyroidism/thyrotoxicosis); 1226 (hypothyroidism/myxoedema); 1256 (acute infective polyneuritis/guillain-barre syndrome); 1260 (myasthenia gravis); 1261 (multiple sclerosis); 1313 (ankylosing spondylitis); 1372 (vasculitis); 1377 (polymyalgia); 1378 (wegners granulmatosis); 1381 (systemic lupus erythematosis/sle); 1382 (sjogren's syndrome/sicca syndrome); 1384 (scleroderma/systemic sclerosis); 1437 (myasthenia gravis); 1453 (psoriasis); 1456 (malabsorption/coeliac disease); 1461 (inflammatory bowel disease); 1462 (Crohns disease); 1463 (ulcerative colitis); 1464 (rheumatoid arthritis); 1477 (psoriatic arthropathy); 1522 (grave's disease); 1661 (vitiligo); 1667 (alopecia / hair loss).

European Network for Genetic and Genomic Epidemiology

European Network for Genetic and Genomic Epidemiology (ENGAGE) is a consortium comprised of 24 cohorts to study the impact of genetic variations on medical phenotypes through GWAS[296]. The consortium has performed over 80,000 GWASs using genetic and phenotype samples from over 600,000 individuals, and made the GWAS summary statistics publicly available[296].

We obtained ENGAGE GWAS summary statistics, representing fixed-effect meta-analyses from 22 studies of European ancestry, for 2 lipid phenotypes[297] (triglyceride [N = 56,267] and total cholesterol [N = 58,327]), and 2 obesity-related phenotypes[298] (BMI [N = 80,938] and BMI-adjusted waist hip ratio [N = 49,877]). In each ENGAGE study, up to 37.4 million autosomal variants were imputed using the 1000 Genomes Project (we used 8.1 million variants which were also imputed in the UK Biobank); phenotypes were adjusted for age, age squared, genotype principal components, and other study-/trait-specific covariates, and were inverse rank normalized; GWASs were performed for each sex separately and combined using fixed-effect meta-analysis; a single genomic control correction was performed for each study prior to a cross-study meta-analysis[297,298].

Biobank Japan

Biobank Japan (BBJ) is a multi-institutional hospital-based biobank with DNA and serum samples from approximately 200,000 participants from 12 medical institutions in Japan [170]. The participants are mainly of Japanese ancestry and had been diagnosed with at least one of 47 diseases by physicians at the cooperating hospitals. Written informed consent was obtained from all the participants, as approved by the ethics committees of RIKEN Center for Integrative Medical Sciences and the Institute of Medical Sciences at the University of Tokyo.

We genotyped samples with either (i) the Illumina HumanOmniExpressExome BeadChip or (ii) a combination of the Illumina HumanOmniExpress and HumanExome BeadChips. We applied standard quality control criteria for both samples and variants as detailed elsewhere [230]. We then pre-phased genotypes with Eagle2 [281] and imputed dosages with Minimac3 [228] using 1000 Genomes project phase 3 (version 5) data (N = 2,504) and Japanese whole-genome sequencing (WGS) data (N = 1,037) as a reference [230]. We computed PCs using EIGENSOFT's smartpca [105].

For phenotypes, we retrieved clinical medical records from the participating hospitals through interviews and a standardized questionnaire. We used 23 diseases and complex traits in Biobank Japan which are also analyzed in UK Biobank (**Supplementary Table D.3**). We normalized quantitative phenotypes via inverse-rank normal transformation as described elsewhere [171]. We defined the 'autoimmune disease' trait in Biobank Japan as a union of Graves' disease and rheumatoid arthritis.

Uganda-APCDR

Uganda-APCDR is a population-based cohort from the General Population Cohort (GPC), Uganda. We retrieved genotype and phenotype data through the African Partnership for Chronic Disease Research (APCDR) initiative via the European Genome-Phenome Archive (EGA), using EGAD00010000965 to access genotype data. Phenotype data were accessed via sftp from EGA (reference: DD_PK_050716

gwas_phenotypes_28Oct14.txt). The participants are from nine ethno-linguistic groups in sub-Saharan Africa and had been recruited from the study area located in southwestern Uganda in Kyamulibwa subcounty of Kalungu district, approximately 120 km from Entebbe town. These ethno-linguistic groups have diverse population structure with varying degrees of admixture between Eurasian and East African Nilo-Saharan ancestries, which has been extensively characterized elsewhere[312]. The detailed cohort demographics, sample collection, and processing were described previously[286,287].

Briefly, the samples were genotyped using the Illumina HumanOmni 2.5M BeadChip at the Wellcome Trust Sanger Institute. We used the Ricopili pipeline to conduct pre-imputation QC and perform phasing and imputation[248]. Briefly, we phased the data using Eagle 2.3.5[281] and imputed variants using minimac3[228] in chunks ≥3Mb. The 1000 Genomes project phase 3 haplotypes[229] were used as the reference panel for phasing and imputation. As described previously, phenotypes were collected using a standard individual questionnaire, blood samples (laboratory tests), and biophysical measurements (height, weight, waist and hip circumferences and blood pressure)[286]. We normalized quantitative phenotypes via inverse-rank normal transformation.

4.4.4 UK Biobank Simulations

We simulated data based on real genotypes of UK Biobank individuals, using 250,963 MAF ≥ 0.1% SNPs with INFO score ≥ 0.6 on chromosome 22 (including short indels) (**D.1 Supplementary Note**). We trained all methods using 337,491 unrelated British-ancestry individuals[76], and we estimated the mixing weights of PolyPred and its summary statistic-based analogues using up to 1000 additional individuals from each of the four non-British ancestries. We computed summary statistics by applying linear regression via Plink 2.0. We did not evaluate PolyPred+ in the simulations because of the relatively small sample sizes of the UK Biobank non-European populations. We evaluated prediction accuracy via R^2, using held-out individuals that were not included in the training

sets and were unrelated to the training set individuals and to each other, using 42K non-British Europeans, 7.7K South Asians, 0.9K East Asians, and 6.2K Africans. We computed PRSs by applying plink 2.0 with the `--score` command, using imputed dosage data (rather than hard-called SNP values). We computed standard errors via a jackknife over simulations.

We trained BOLT-LMM by applying BOLT-LMM v2.3.4 to plink files of HapMap 3 SNPs (hard-coded from imputed dosages), using the same covariates specified in the "Estimating relative-R^2 and its standard error" Methods subsection, and specifying the flag `--predBetasFile` to report PRS coefficients.

We trained SBayesR using summary statistics from the infinitesimal version of BOLT-LMM (BOLT-LMM-inf[159]), which yielded far superior accuracy vs. using summary statistics from the non-infinitesimal version of BOLT-LMM. We ran SBayesR using 10,000 iterations, 4,000 burn-in iterations, using values from 10% of the iterations to compute posterior means, using the HapMap 3 LD files published the SBayesR authors. We attempted to run SBayesR using a mixture of four distributions (using $\pi = [0.95, 0.02, 0.02, 0.01]$ and $\gamma = [0, 0.01, 0.1, 1]$). In case SBayesR failed with these parameters, we iteratively shrank the last entry in the vector γ by 50% until it was smaller than 10^6, at which point we removed the last mixture component and redefined π such that the first entry was equal to 0.95 and all other entries had the same value such that all values sum to 1.0.

We trained PRS-CS using summary statistics from BOLT-LMM-inf (as in SBayesR) with the parameters $a = 1$, $b = 0.5$, thin $= 5$, n_iter $= 10,000$, n_burnin $= 500$, and without specifying the value of phi (corresponding to PRS-CS-auto). We used the UK Biobank LD reference panels made publicly available by the authors of PRS-CS (see **4.5 Data availability**).

We trained P+T by applying plink with the command `--clump-r2 0.5 --clump-kb 250` with various values of `--clump-p1` (following ref.[47]), and using 10,000 randomly selected unrelated UK Biobank British individuals to compute LD. We estimated LD using 10,000 individuals to balance between runtime and accuracy (noting that P+T is relatively insensitive to the LD reference panel

size compared to the other methods evaluated in this manuscript). We used summary statistics based on BOLT-LMM, using marginal effect sizes derived from reported χ^2 values (*i.e.*, the square root of χ^2 divided by the square root of the BOLT LMM effective sample size [118], and multiplied by the sign of the effect size estimated by the infinitesimal version of BOLT-LMM). We used the best value of `--clump-p1` (out of the evaluated values $10^{-2}, 10^{-3}, 10^{-4}, 10^{-6}, 5 \times 10^{-8}$) based on the target sample phenotypes, which leads to anti-conservative prediction accuracy estimates for P+T.

We used slightly different LD reference panels for PolyFun-pred, SBayesR, and PRS-CS, because (i) they use different algorithms to impose sparsity on LD matrices, and different file formats to store them; and (ii) we assume that naively running SBayesR or PRS-CS using summary LD from the 18 million SNPs used by PolyFun-pred would be computationally infeasible, based on information provided in the manuscripts describing these methods [36,37]. When modifying the training sample size, we kept the LD reference panel sample size fixed to alleviate computational costs.

4.4.5 Analysis of real data

We performed four sets of analyses: (i) Analysis of 4 UK Biobank populations using UK Biobank British training data; (ii) Analysis of 4 UK Biobank populations using ENGAGE meta-analysis training data; (iii) Analysis of Biobank Japan and Uganda-APCDR cohorts; and (iv) Analysis of UK Biobank East Asians using UK Biobank British and Biobank Japan training data. In analysis sets (i), (iii) and (iv), we evaluated PRSs generated by training all methods using unrelated UK Biobank British-ancestry individuals. In analysis set (ii), we evaluated PRSs generated by training all methods using summary statistics from 8.1 million meta-analyzed summary statistics from the ENGAGE consortium54–56. In a subset of analysis set (iii) and in analysis set (iv) we additionally evaluated PRSs generated by training BOLT-LMM-BBJ (BOLT-LMM trained on Biobank Japan individuals). In all analysis sets, the individuals in the target populations were unrelated to each other and to the individuals in the training set (when available).

In analysis sets (i), (iii) and (iv), we selected the 7 traits to meta-analyze by first restricting the set of 49 traits analyzed in ref. [118] to traits that are available in Biobank Japan and Uganda-APCDR and are well-powered across multiple ancestries, having $h^2 > 0.05$ in UK Biobank non-British Europeans, in UK Biobank South Asians, and in UK Biobank Africans (see below for details on ancestry-specific heritability estimation). We then iteratively greedily selected ranked traits according to their heritability in UK Biobank non-British Europeans (estimated as in ref. [118]), such that no selected trait had $|r_g| < 0.3$ with a previously selected trait.

We computed ancestry-specific SNP heritabilities in each UK Biobank ancestry by applying GCTA84 to unrelated sets of individuals using hard-called HapMap 3 SNPs (using a random set of 10,000 individuals for non-British Europeans to facilitate the computations). We did not use more advanced methods85 because of the relatively small sample sizes. We meta-analyzed ancestry-specific SNP heritabilities by averaging the estimated heritabilities, and we estimated the meta-analyzed standard error via the square root of the average sampling variance, divided by the square root of the number of traits.

In analysis sets (i), (iii) and (iv), We trained all PRS methods on UK Biobank unrelated British-ancestry individuals (average $N = 325$) as described in the Methods subsection "UK Biobank simulations", but using summary statistics generated by BOLT-LMM when applied to UK Biobank British-ancestry individuals, as described in our previous work [118]. We trained P+T separately for each non-UK Biobank cohort by restricting the set of SNPs considered to the set of SNPs available in both the UK Biobank and in the target cohort. We computed the contribution of PolyFun-pred (resp. BOLT-LMM) towards PolyPred via the ratio of the mixing weight of PolyFun-pred (resp. BOLT-LMM) to the sum of the mixing weights of PolyPred and of BOLT-LMM.

In analysis sets (i), (ii) and (iv), we computed a PRS for each UK Biobank individual using imputed dosage data as described in the "UK Biobank Simulations".

In analysis set (iii), we computed a PRS for each individual in Biobank Japan and in Uganda-

APCDR using imputed dosage data using Plink 2.0 (ref. [278]). In secondary analyses of analysis set (i) we also evaluated LDpred33. We trained LDpred using HapMap 3 SNPs and using two different LD reference panels: 1000 Genomes project [229] and UK10K [162]. We used summary statistics from the infinitesimal version of BOLT-LMM (as in SBayesR) and with default parameters, using the parameter `--ldr 400`. We used the value of "`--F`" (corresponding to the assumed proportion of causal SNPs, using all the default evaluated values) that yielded the best prediction accuracy in the target sample, yielding anti-conservative accuracy estimates as in P+T.

In analysis sets (iii) and (iv), we trained BOLT-LMM-BBJ, SBayesR-BBJ, and PRS-CS-BBJ (BOLT-LMM, SBayesR, and PRS-CS, respectively, trained using Biobank Japan training data) (average N = 124K). We selected individuals for training these methods as described in our previous work [47], but excluding a random subset of 5,000 individuals that were used for evaluating prediction accuracy. For SBayesR-BBJ, we used a subset of individuals (N = 50K) from Biobank Japan to compute in-sample LD, following the recommendations of the authors of SBayesR [36]. For PRS-CS-BBJ, we used the East Asian LD reference panels made publicly available by the authors of PRS-CS (see **4.5 Data availability**).

4.5 Data availability

Access to the UK Biobank resource is available via application (`http://www.ukbiobank.ac.uk`). PRS coefficients generated in this study are available for public download at `http://data.broadinstitute.org/alkesgroup/polypred_results`. Summary LD information of N = 337K British-ancestry UK Biobank individuals for 2,763 overlapping 3Mb loci is available at: `https://data.broadinstitute.org/alkesgroup/UKBB_LD`. Summary LD information of N = 50K UK Biobank individuals for SBayesR is available at: `https://zenodo.org/record/3350914`. Summary LD information used by PRS-CS is available at: `https://github.com/getian107/PRScs`. Baseline-LF v2.2.UKB annota-

tions and LD-scores for UK Biobank SNPs are available at: https://data.broadinstitute.org/alkesgroup/LDSCORE/baselineLF_v2.2.UKB.tar.gz

4.6 Code availability

PolyPred and PolyPred+ are provided as part of the open-source software package PolyFun, which is freely available at https://doi.org/10.5281/zenodo.613967989 and https://github.com/omerwe/polyfun. BOLT-LMM is available at https://data.broadinstitute.org/alkesgroup/BOLT-LMM. SBayesR is available at https://cnsgenomics.com/software/gctb. PRS-CS is available at https://github.com/getian107/PRScs.

4.7 Acknowledgements

We thank Armin Schoech and Carla Márquez-Luna for helpful discussions. This research was conducted using the UK Biobank Resource under Application #16549 and was funded by NIH grants U01 HG009379, U01 HG012009, R37 MH107649, R01 MH101244 and R01 HG006399. MK is supported by a Nakajima Foundation Fellowship and the Masason Foundation. WJP is supported by an NWO Veni grant (91619152). ARM is supported by NIMH K99/R00MH117229. HKF is supported by Eric and Wendy Schmidt. AVK is supported by grants 1K08HG010155 and 1U01HG011719 from the National Human Genome Research Institute and a sponsored research agreement from IBM Research. YO is supported by JSPS KAKENHI (19H01021, 20K21834), and AMED (JP21km0405211, JP21ek0109413, JP21ek0410075, JP21gm4010006, JP21km0405217), JST Moonshot R&D (JPMJMS2021, JPMJMS2024). Computational analyses were performed on the O2 High-Performance Compute Cluster at Harvard Medical School.

4.8 Author contributions

O.W., M.K., H.S. and A.L.P. designed the study; O.W., M.K., H.S. and S.G. analyzed data; O.W., M.K., H.S. and A.L.P. wrote the manuscript with assistance from S.G., W.J.P., A.V.K, Y.O., A.R.M. and H.F.

5

Conclusion

In this dissertation, I described a series of fine-mapping analyses in large-scale biobanks across diverse populations. The work presented here provides insights into candidate causal variants of human complex traits, along with potential applications for further functional characterization and polygenic prediction. While **Chapters 1–4** represent key advances in fine-mapping complex traits across diverse populations, I foresee several remaining challenges and opportunities for future studies.

First, the current fine-mapping methods have many limitations. Model misspecification and data heterogeneity are major (but commonly overlooked) sources of miscalibration, including mismatched LD, misspecified number of causal variants and effect size distribution, missing causal variants (*e.g.*, structural variants), uncontrolled confounding factors (*e.g.*, population stratification), and heterogeneity in phenotyping and genotyping (especially for meta-analysis, **Chapter 3**). Further methodical development that properly models these factors is required in addition to improved study design at the outset that is tailored to post-GWAS variant prioritization.

Second, the vast majority of candidate causal variants in our study (**Chapters 1,2**) remains unannotated for biological mechanisms and functional consequences, even when they have been definitively resolved and replicated across multiple independent cohorts. Future systematic variant-to-function (V2F) efforts will require the development of high-throughput assays as well as recruitment of samples from diverse cell types, conditions, and genetic backgrounds. Besides V2F, locus-to-gene (L2G) mapping remains particularly challenging for non-coding loci. As the recent studies have demonstrated emerging convergence between rare and common variant associations[11,313,314], I envision that more biobank-scale resource generation in the future would help us learn L2G principles by leveraging rare and common fine-mapped coding variants to link genes with regulatory variants.

Third, the utility and portability of PRS remain largely limited, especially for non-European populations. Despite the continuous method development in the field (including our PolyPred

method, **Chapter 4**), the most fundamental solution for equitable polygenic prediction is only made possible by recruiting more diverse samples from different backgrounds, including genetic ancestry, geographic location, time points, and other environmental factors.

Finally, increasing the diversity of study participants is also crucial for further variant discovery and replication. As demonstrated in our allelic series examples (**Chapter 2**) and many other studies, aggregating data across multiple populations enables identification of population-enriched variants and their convergence on the same gene. While this dissertation describes an incredible opportunity of identifying Finnish- and Japanese-enriched putative causal variants using FinnGen and BioBank Japan, I envision that fast-evolving biobanks worldwide will flourish in the next decade and provide novel insights into the biology of human complex diseases.

A

Supplementary Materials for Chapter 1

A.1 Supplementary Tables

The following Supplementary Tables will be made available in the online version of the manuscript.

Table A.1: Overview of traits included in study

Table A.2: LD score regression estimates

Table A.3: Genetic correlation between traits

Table A.4: Merged SuSiE 95% credible sets

Table A.5: Baseline model annotation enrichments

Table A.6: Fine-mapped variants with weak p-values

Table A.7: Fine-mapped variants with marginal and posterior effect sign disagreements

Table A.8: Fine-mapped pleiotropic variants across 3 or more domains

Table A.9: Fine-mapped pleiotropic variants where effect directions disagree with polygenic expectation

Table A.10: Phenome-wide association study of fine-mapped UKBB variants

Table A.11: Likely causal variants affecting gene expression

Table A.12: Fine-mapped eQTL variant enrichments for fine-mapped complex traits

Table A.13: Disjoint genomic annotation enrichments

Table A.14: Single variant colocalization results

Table A.15: Credible set colocalization results

Table A.16: Colocalization for gene prioritization results

Table A.17: Functional enrichment of fine-mapping variants in accessible chromatin across datasets

Table A.18: Fine-mapped variants in accessible chromatin for trait-specific enriched cell-types

Table A.19: Transcription factor features selected for inclusion

Table A.20: Enrichment of molecular mechanisms for CRE single nucleotide variants

Table A.21: Putative mechanistic annotations of fine-mapped complex trait variants

A.2 Supplementary Figures

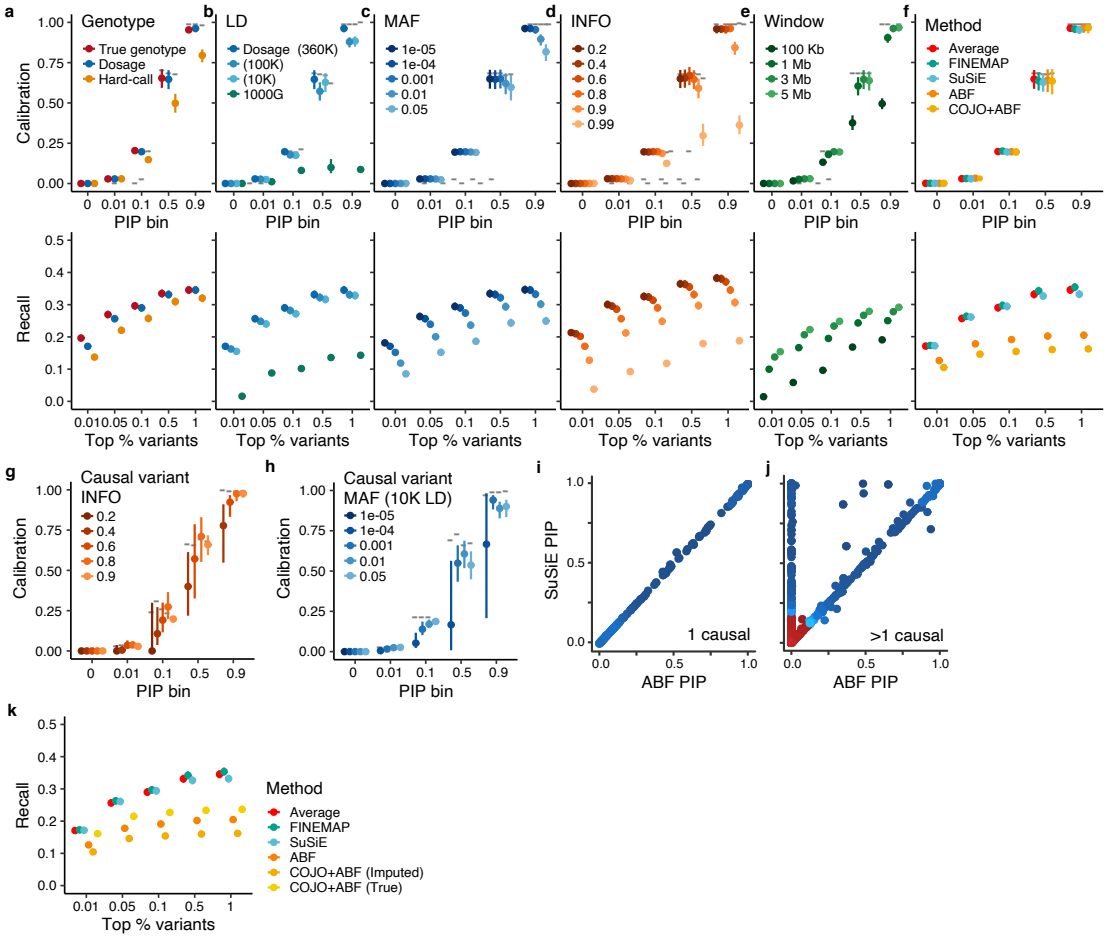


Figure A.1: Evaluation of fine-mapping approaches in realistic biobank simulations. Main results of fine-mapping simulations are shown in panels **a–f**. In the top panel, calibration, the difference between the observed and expected proportion of causal variants in a range of PIP values is shown. In the bottom panel, recall, the number of true causal variants detected in the top ranked variants by PIP is shown. We evaluate the performance of fine-mapping using true, hard-called, or dosage genotypes (**a**), using different LD reference panels (**b**), varying QC thresholds for variant inclusion in fine-mapping based upon allele frequency and imputation quality (**c,d**), varying window size around sentinel variants (**e**), and across different fine-mapped methods (**f**). Error bars represent 95% CIs. In **g,h**, calibration is shown for causal variants with varying imputation confidence (**g**) and with varying MAF (**h**) using an LD reference panel of 10,000 randomly sampled individuals from the UK Biobank. Error bars represent 95% CIs. **i,j**. Comparison between single causal variant (ABF) and multiple causal variant (SuSiE) fine-mapping PIPs. When there is one simulated causal variant, the methods agree (**d**), but when there are two causal variants, ABF underperforms (**j**). **k** Similar to (**f**) bottom panel, except with a focus on the differences in recall, the number of true causal variants detected in the top ranked variants by PIP, between approximate conditional analysis followed by ABF fine-mapping (COJO+ABF) when using the (typically unobserved) true genotype matrix or the (observed) imputation genotype matrix (which GCTA can currently only hard call rather than using the dosage values). Methods had similar precision in the simulation framework.

Figure A.2 ***(following page)*****: Narrow-sense heritability and genetic correlations of UK Biobank traits. a.** S-LDSC estimated narrow-sense common (h^2_c; MAF > 0.05) and low-frequency (h^2_{lf}; 0.05 > MAF > 0.005) heritability. Traits are grouped by phenotypic domain and arranged by decreasing total heritability ($h^2_c + h^2_{lf}$). **b.** Genetic correlation estimated using S-LDSC for traits in **a**. Square size is proportional to the Bonferroni adjusted P-value. LD scores were derived from the UK 10K cohort.

Figure A.2: (continued)

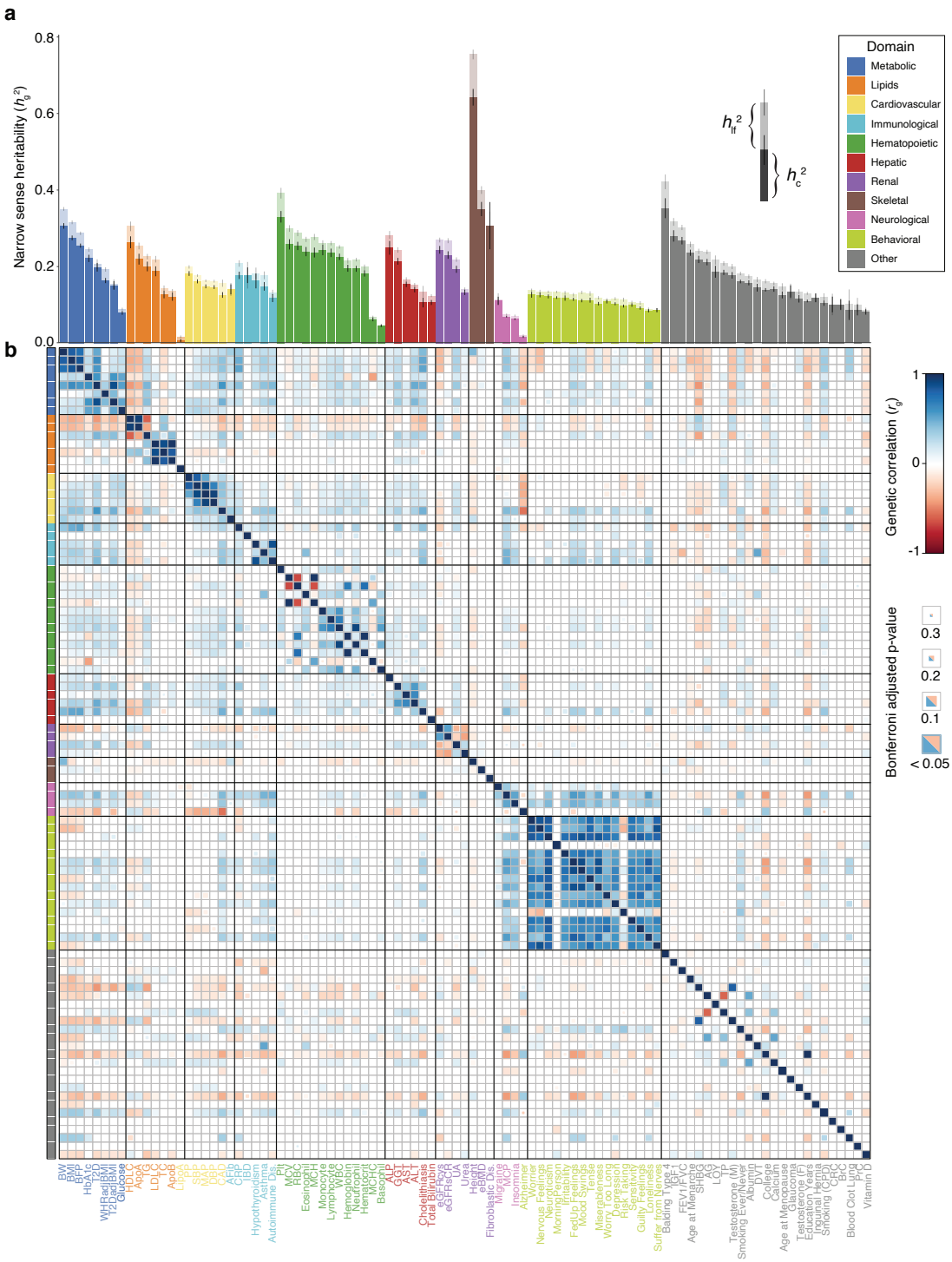
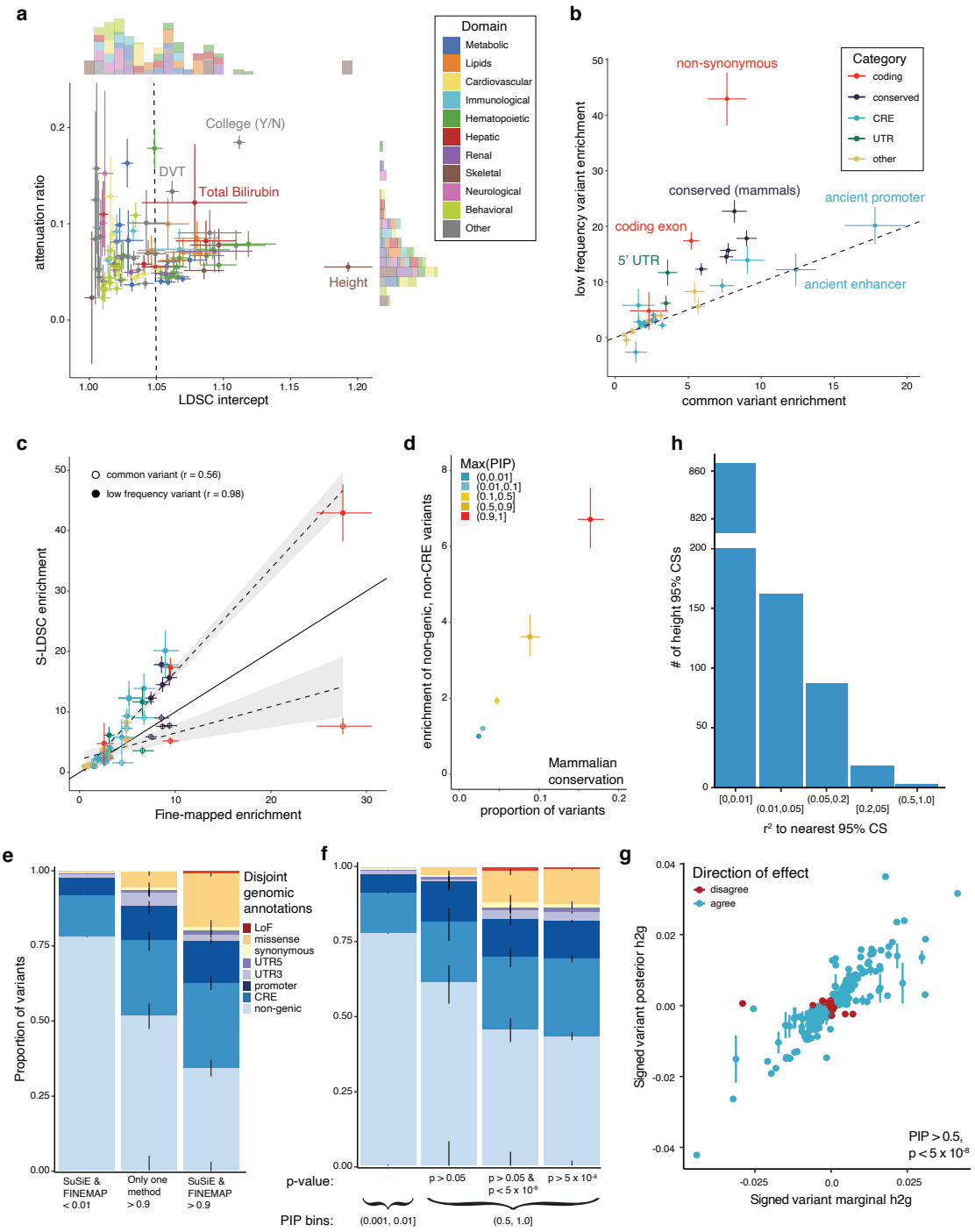


Figure A.3 *(following page)*: Additional characterization of genetic fine-mapping of complex traits in UK Biobank. a. Comparison of the LD score regression intercept and the attenuation ratio [(LDSC intercept – 1) / (mean χ^2 – 1)] for different traits colored by phenotypic domain. Similar to Loh *et al.*[109], we highlight a subset of traits with extreme values in either, or particularly both, dimensions. A dashed line is drawn at an LDSC intercept of 1.05. Traits are colored by phenotypic domain. Error bars represent 95% CIs. **b**. Comparison of common variant (MAF > 0.05) enrichment and low-frequency (0.05 > MAF > 0.005) variant enrichment. Enrichment is defined as the proportion of (common or low frequency) heritability divided by the proportion of (common or low frequency) variants in that genomic annotation. Results are consistent with those reported in Gazal *et al.*[161]. Genomic annotations obtained from the LDSC baseline v2.2 are colored by broad category. **c**. Common and low frequency variant enrichment from b. compared to fine-mapped variant enrichment (defined as proportion of variants in annotation with PIP > 0.9 vs. proportion of those with PIP < 0.01). A solid line represents $y = x$. Dotted lines represent linear regression fits to the model $y = xB$ where y is the S-LDSC enrichment point estimates and x is the fine-mapping enrichment point estimates. Pearson correlations are also estimated and provided. Error bars and grey ribbons represent 95% CIs. **d**. Enrichment of non-genic (excluding variants in CREs, defined as in **Fig. 1.1g**) variants in each indicated PIP bin (vs. similar variants with PIP < 0.01) and proportion of those variants that are evolutionarily conserved across mammals. **e**. Comparison of 8 distinct genomic enrichments depending upon agreement of FINEMAP and SuSiE, similar to **Fig. 1.1g**. Error bars represent 95% CIs. **f**. Comparison of 8 distinct genomic enrichments across selected PIP-bins and marginal p-value thresholds, similar to **Fig. 1.1g**. Error bars represent 95% CIs. **g**. Comparison of marginal and SuSiE posterior effect sizes for variants with PIP > 0.5. Variants are colored according to whether the effect direction agrees between marginal and posterior estimates. Error bars represent 95% CIs. **h**. Using the best fine-mapped variant in each 95% CS for height, we calculate the LD (r^2) between all 95% CS pairs and report the nearest 95% CS (highest LD) for each unique 95% CS. An axis break between 200 and 750 on the y-axis is indicated.

Figure A.3: (continued)



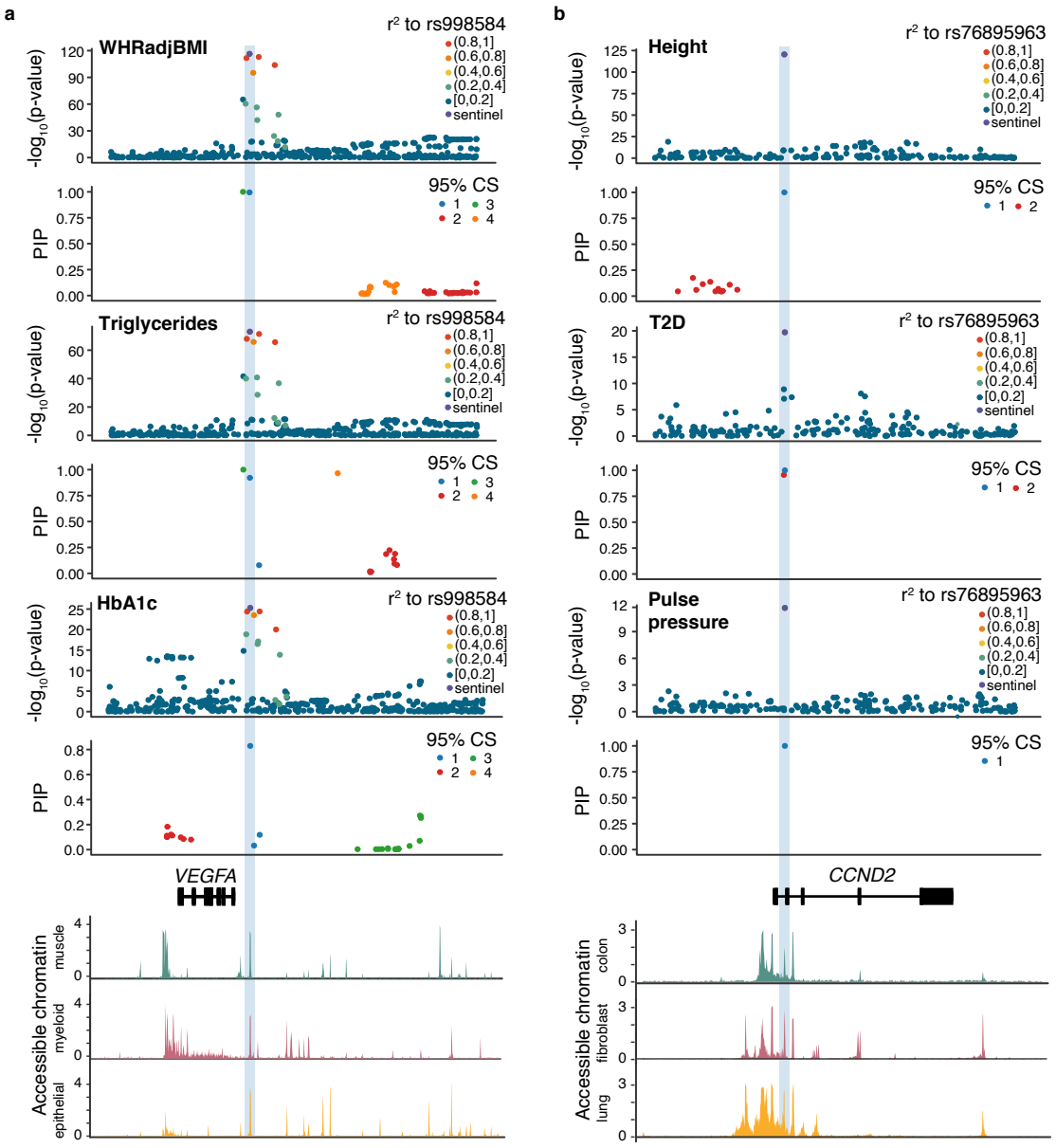


Figure A.4: Examples of fine-mapped pleiotropic variants. a. Locus-zoom plots for WHR adjusted for BMI, triglyceride levels, and HbA1c around *VEGFA*. Fine-mapping pinpoints the pleiotropic variant rs998584 as the likely causal variant in the region. Overlap with chromatin occupancy for a subset of investigated cell types suggests putative cell-types of action. **b**. Locus-zoom plots for height, T2D, and pulse pressure levels around *CCND2*. Fine-mapping pinpoints the pleiotropic variant rs76895963 as the likely causal variant in the region. Overlap with chromatin occupancy for a subset of investigated cell types suggests a broad range of action across cell-types for this variant. For both loci, there are additional 95% CSs that vary between traits.

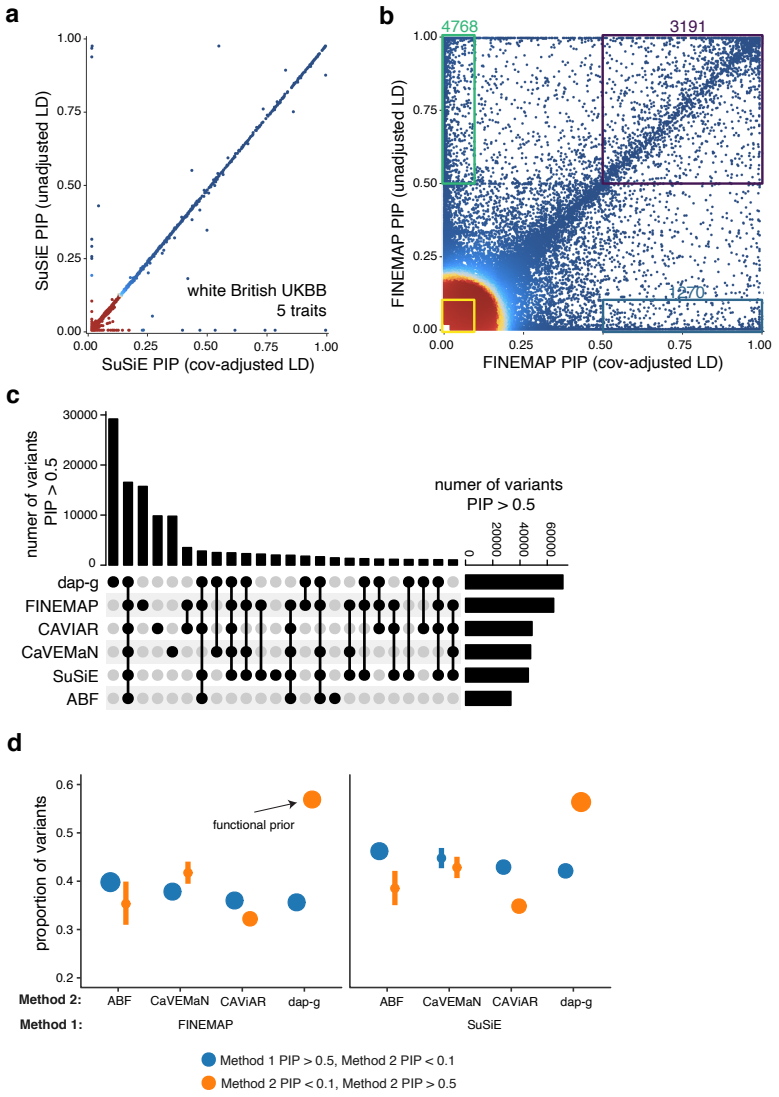
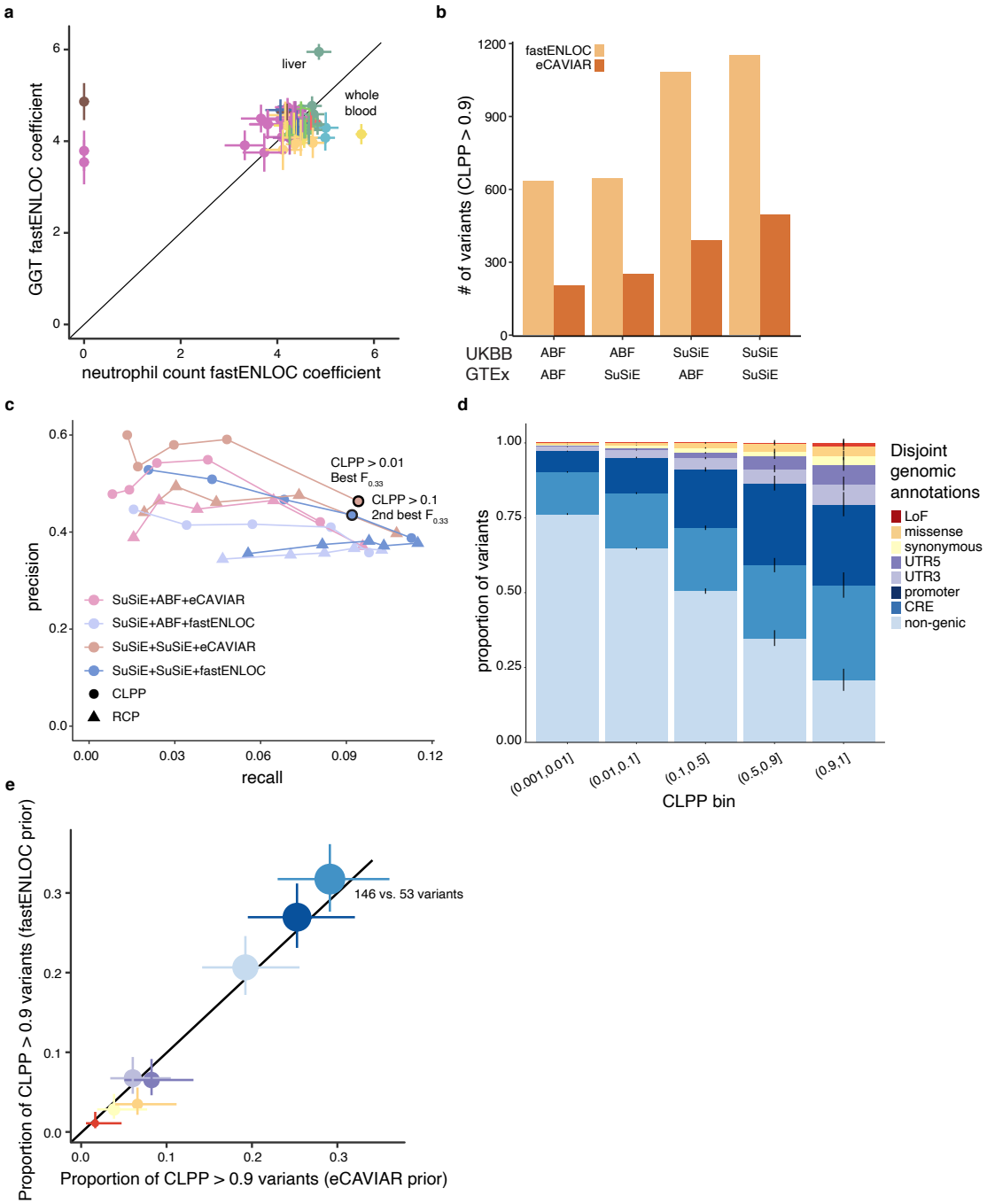


Figure A.5: Additional comparison of GTEx fine-mapping approaches. a. Comparison of PIPs from SuSiE fine-mapping using unadjusted in-sample LD or covariate-adjusted in-sample LD for 5 traits in UK Biobank. Red indicates higher density of variants, and blue indicates lower. **b**. Comparison of PIPs from FINEMAP fine-mapping using unadjusted in-sample LD or covariate-adjusted in-sample LD for Muscle tissue in GTEx v8. Red indicates higher density of variants, and blue indicates lower. **c**. Upset plot showing overlap of fine-mapped variants (max PIP > 0.5) across all methods. **d**. Proportion of fine-mapped variants overlapping one of 7 distinct genic or regulatory genomic annotations from **Fig. 1.1g**. In the left panel, blue points represent SuSiE PIP > 0.5 and a corresponding method with PIP < 0.1. Orange points represent SuSiE PIP < 0.1 and a corresponding method with PIP > 0.5. The right panel is the same with FINEMAP instead of SuSiE. SuSiE-specific variants outperform 3/3 uniform prior methods and FINEMAP outperforms 2/3 uniform prior methods. dap-g shows the highest enrichment but this is likely due to its functional prior. Variant PIPs are taken as max across all genes and tissues. Point sizes are proportional to the number of variants in the analysis. Error bars represent 95% CIs.

Figure A.6 ***(following page):*** **Additional characterization of colocalization of complex and molecular traits.** **a**. Comparison of informative fastENLOC priors (coefficient estimates from the regularized logistic model) between neutrophil count and gamma-glutamyl transferase (GGT) levels. Trait-specific outliers are labeled. Colors indicate the specific physiological system of the tissue as shown in **Fig. 1.4a**. Error bars represent 95% CIs. **b**. Number of variant sets with regional colocalization probability (RCP) > 0.1 for each colocalization method. Fine-mapping using a single causal variant (ABF) and allowing for multiple causal variants (SuSiE) for both complex traits and eQTLs is varied as well as priors assuming independence between complex and molecular traits (eCAVIAR) or informative empirical priors (fastENLOC). Similar to **Fig. 1.4e**. **c**. Using a validation set of fine-mapped coding variants with PIP > 0.5, precision and recall estimated for non-coding 95% CSs within 500 kb across different colocalization methods varying single or multiple causal variant assumptions, independent or informative priors, and estimand (RCP vs. CLPP), similar to **Fig. 1.4h**. Estimates are shown for posterior probability values equal to 0.9, 0.5, 0.25, 0.1, and 0.01. Weighting our preference for precision to recall at a ratio of 3:1, we highlight the best approaches based upon top $F_{0.33}$ values. **d**. Overlap of fine-mapped variants in each CLPP bin and 8 disjoint genomic annotations. The max CLPP for each variant across traits is used. Error bars represent 95% CIs. Similar to **Fig. 1.1g**. **e**. Comparison of the proportion of colocalized (CLPP > 0.9) variants when using an eCAVIAR or fastENLOC prior. The size of points corresponds to the increase in number of variants detected for each genomic annotation when switching from an eCAVIAR to fastENLOC prior. Colors indicate genomic annotation from **Fig. 1.1g** and from **d**. Error bars represent 95% CIs.

Figure A.6: (continued)



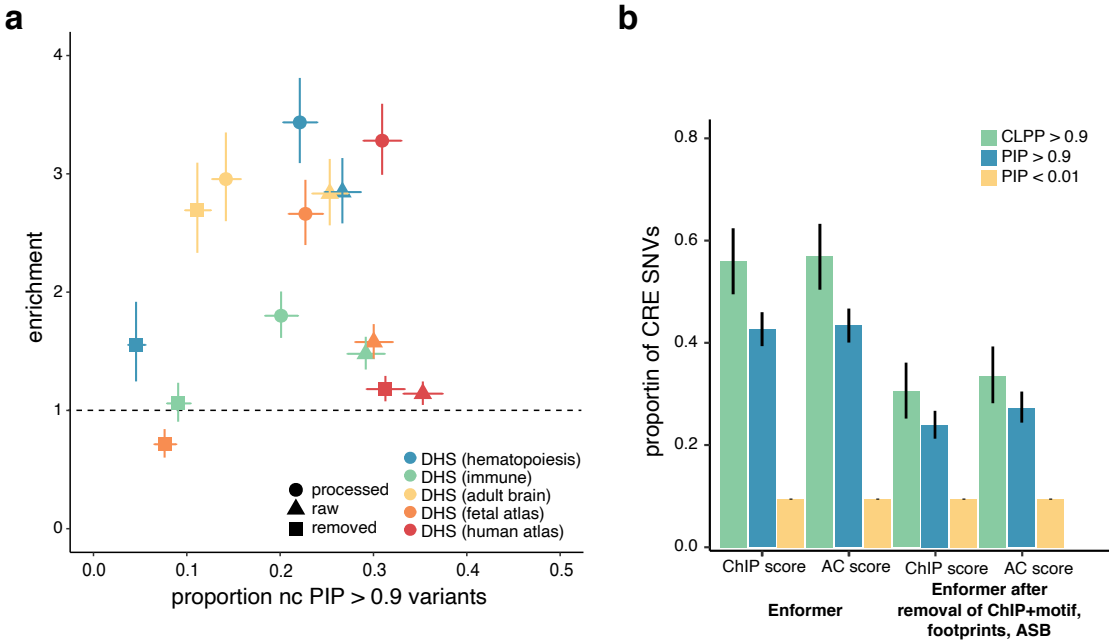


Figure A.7: Characterization and examples of accessible chromatin datasets and fine-mapped regulatory variants. a. Proportion and enrichment (PIP > 0.9 vs. < 0.1) of fine-mapped variants across different accessible chromatin datasets, similar to **Fig. 1.4a**. Datasets with quantitative accessible chromatin values for each peak (excluding ChIP-Atlas and Roadmap) are divided into the raw downloaded peaks, the processed peaks, and the peaks removed by our QC procedure. Error bars represent 95% CIs. **b.** Proportion of CRE variants with TF occupancy or accessible chromatin predicted changes from deep neural network models (Enformer) that are not annotated as occupied by the corresponding TF (ChIP + motif), residing within an accessible chromatin footprint (AC footprint), exhibiting allele specific TF binding (TF ASB), or exhibiting allele specific accessible chromatin (AC ASB), similar to **Fig. 1.4c**.

B

Supplementary Materials for Chapter 2

B.1 Supplementary Note

B.1.1 Novel genes implicated by population-enriched variants

Fine-mapping of complex traits in FinnGen led to the identification of Finnish-enriched missense variants in genes novel for several traits. Many of these genes possess compelling biological rationales linking them to these traits.

THBS3

On chromosome 1, rs199935580 corresponds to an Arginine to Tryptophan substitution (p.Arg520Trp) in the gene *THBS3*. This rare missense variant (MAF = 1.0×10^{-3} in gnomAD Finnish; 1.2×10^{-5} in gnomAD Non-Finnish-Swedish-Estonian Europeans [NFSEE]) is fine-mapped for increased risk of carpal tunnel syndrome ($P = 7.4 \times 10^{-10}$; $\beta = 1.44$; PIP = 1.0). In carpal tunnel syndrome (CTS), the median nerve is pinched as it traverses through the wrist. A previous GWAS on CTS revealed an important role for the extracellular matrix in its etiology[315]. *THBS3* encodes a member of the thrombospondin family, a group of proteins known for binding to various extracellular matrix proteins[316]. No specific role for *THBS3* in CTS has been noted previously. However, its closest homolog, *COMP*, is causal for familial carpal tunnel syndrome 2 (ref. [317]).

LUM

On chromosome 12, the rare variant rs191692991 corresponds to an Arginine to Cysteine substitution (p.Arg310Cys) in the gene *LUM*. This missense variant (MAF = 5.3×10^{-3} in gnomAD Finnish; 1.2×10^{-5} in gnomAD NFSEE) is strongly associated with "fibroblast disorders" such as Dupuytren's contracture, a condition in which the skin under the palm of the hand becomes thick and fibrous ($P = 6.9 \times 10^{-9}$; $\beta = 1.02$; PIP = 1.0). LUM encodes lumican, a leucine-rich repeat

glycoprotein important for regulation of collagen fibril formation[318]. One hypothesis is that this mutation reduces stability or activity of lumican, leading to misregulation of collagen and accumulation of fibrils in the hand and elsewhere.

POF1B

On the X chromosome, the rare variant rs200939713 corresponds to an Arginine to Tryptophan substitution (p.Arg339Trp) in *POF1B* (MAF = 1.6×10^{-3} in gnomAD Finnish; monomorphic in gnomAD NFSEE). This missense variant is associated with risk of varicose veins ($P = 3.4 \times 10^{-11}$; $\beta = 0.84$; PIP = 0.99), a condition in which veins just below the skin become enlarged and prominent. *POF1B* encodes a protein important for epithelial structural integrity through desmosomes[319]. It's possible this mutation reduces *POF1B* stability, reducing epidermal integrity increasing occurrence or appearance of varicose veins.

B.1.2 "Missing" variants from summary statistics in other populations

After restricting to the 26 traits available in every population (BBJ, FinnGen, and UKBB), we found 301 high-PIP variants (PIP > 0.9) fine-mapped in a discovery population that are missing from summary statistics in other populations (**Fig. 2.2a–c**). We characterized reasons for the missingness using the following criteria:

1. Variants do not exist in imputation reference panels used in each cohort, *i.e.*, BBJ[230]: the 1000 Genomes Phase 3 (n = 2,504) + Japanese WGS (n = 1,037); FinnGen: Finnish WGS (n = 3,775); and UKBB[76]: the Haplotype Reference Consortium (n = 64,976) + the 1000 Genomes Phase 3 + UK10K (n = 3,781).

2. Low MAF (MAF < 0.005) in a population based on the GEM-J WGS[215] for BBJ and the gnomAD[214] v2 for FinnGen and UKBB.

3. Low imputation INFO score (INFO < 0.7 for BBJ and INFO < 0.8 for FinnGen/UKBB)

4. Hardy-Weinberg equilibrium (HWE) outlier (HWE test P-value $< 1 \times 10^{-10}$) only in UKBB.

We confirmed that the missingness are primarily due to low frequency in other populations (**Supplementary Fig. B.11**). Note that since BBJ and UKBB included the 1000 Genomes Project in their reference panels, there are variants that exist in the reference but showed very low MAF in the Japanese or White British populations; This is in contrast to the FinnGen which only used a Finnish-specific reference panel.

Low INFO variants

We found six high-PIP variant-trait pairs are missing from summary statistics in other populations due to low INFO score despite having a high MAF in a population:

- rs138381300 (frameshift, *FLG*:p.Ser761CysfsTer36) fine-mapped for atopic dermatitis (PIP = 1.0) in FinnGen, but missing from UKBB (INFO = 0.60 in UKBB; MAF = 0.02 in non-Finnish Europeans). This variant is monomorphic in Japanese.

- rs6874142 (intron variant of *STC2*) fine-mapped for height (PIP = 1.0) in UKBB, but missing from BBJ (INFO = 0.64 in BBJ; MAF = 0.03 in Japanese). This variant is significantly associated in FinnGen ($P = 1.6 \times 10^{-10}$) but not fine-mapped (PIP = 0.008).

- rs117137535 (intron variant of *ARRDC1*) fine-mapped for atopic dermatitis (PIP = 1.0) in FinnGen, but missing from BBJ (INFO = 0.55 in BBJ; MAF = 0.10 in Japanese). This variant is not associated in UKBB ($P = 0.75$).

- rs9893867 (intron variant of *SLC43A2*) fine-mapped for type 2 diabetes (PIP = 0.97) in FinnGen, but missing from BBJ (INFO = 0.62 in BBJ; MAF = 0.08 in Japanese). This variant is not associated in UKBB ($P = 0.15$).

- rs11653578 (intron variant of *ANKFN1*) fine-mapped for type 2 diabetes (PIP = 0.98) in UKBB, but missing from BBJ (INFO = 0.61 in BBJ; MAF = 0.23 in Japanese). This variant is not associated in FinnGen ($P = 0.20$).

- rs3810291 (3' UTR variant of *ZC3H4*) fine-mapped for body mass index (PIP = 0.96) in UKBB, but missing from BBJ (INFO = 0.61 in BBJ; MAF = 0.23 in Japanese). This variant is significantly associated in FinnGen ($P = 1.4 \times 10^{-10}$) but not fine-mapped (PIP = 0.009).

There are a few potential reasons for relatively low INFO scores of these variants. First, rs138381300 is a frameshift variant of *FLG* which is known to have a highly repetitive coding sequence. This makes short-read next-generation sequencing (NGS) extremely challenging; and indeed, the region is registered as NCBI GeT-RM NGS Dead Zone[320]. Second, we found three variants are located near the telomere regions (rs6874142: 5q35.1, rs117137535: 9q34.3, and rs9893867: 17p13.3) which are difficult to impute. Lastly, although we did not find any simple reason for rs11653578 and rs3810291, we speculate that the low INFO scores of these variants are due to a combination of several factors including reference panel and genotyping array quality, given that their INFO scores are just borderline below the threshold (INFO = 0.61 < 0.7).

Of the six pairs, we are confident that rs138381300 is a putative causal variant for atopic dermatitis (PIP = 1.0 in FinnGen) since it is a frameshift variant for the known pathogenic gene FLG. The rest of the variant-trait pairs show varying evidence of association, emphasizing the critical needs for replication in fine-mapping studies.

HWE outlier variants in UKBB

In addition, we observed that five high-PIP variant-trait pairs (three unique variants) are missing from UKBB summary statistics due to HWE outlier, namely:

- rs2237897 (intron variant of *KCNQ1*) fine-mapped for body mass index, body weight, and type 2 diabetes (PIP =1.0) in BBJ (MAF = 0.04, HWE *P*-value = 6.3×10^{-134} in UKBB).
- rs4765138 (intergenic variant) fine-mapped for height (PIP = 0.99) in FinnGen (MAF = 0.31, HWE *P*-value = 5.7×10^{-32} in UKBB).
- rs117952254 (intergenic variant) fine-mapped for myocardial infarction (PIP = 0.98) in FinnGen (MAF = 0.027, HWE *P*-value = 2.9×10^{-44} in UKBB).

We previously reported that UKBB imputed data contain genotyped SNPs failing the HWE test (http://www.nealelab.is/blog/2019/9/17/genotyped-snps-in-uk-biobank-failing-hardy-weinberg-equilibrium-test). Briefly, we observed 15,069 genotyped variants that are retained in the imputed bgen files with INFO = 1 and HWE *P*-value $< 1 \times 10^{-12}$, due to UKBB's QC criteria relying on a per-batch HWE test[76] (**Supplementary Fig. B.12**). This observation was particularly concerning given that we found that 3,987 of these variants have no homozygous alternative genotypes despite having a MAF > 1%. To mitigate this issue, we applied an additional post-hoc filtering that excludes any imputed variants with HWE test *P*-value $< 1 \times 10^{-10}$; however, this might exclude a potential causal variant too.

Having said that, we are confident rs2237897 is a putative causal variant (PIP = 1.0 and 0.31 in BBJ and FinnGen, respectively) that confers a risk for type 2 diabetes as previously reported[8]. However, rs4765138 and rs117952254 are not well-characterized in the current literature, with lack of fine-mapping replication in BBJ (rs4765138: $P = 4.1 \times 10^{-19}$ and PIP = 1.1×10^{-5} for height; rs117952254: missing in BBJ), suggesting that further replication effort should be warranted.

B.2 Supplementary Tables

The following Supplementary Tables are available in the online version of the manuscript.

Table B.1: Overview of the studied cohorts

Table B.2: Overview of the studied traits

Table B.3: High-PIP (> 0.9) variant-trait pairs

Table B.4: Merged credible set summary

Table B.5: Functional enrichment of fine-mapped variants for the seven main distinct functional categories

Table B.6: Functional enrichment of fine-mapped variants for 35 binary annotations from the baselineLD v2.2 model

Table B.7: Colocalization

Table B.8: High-confidence fine-mapped coding variant-trait pairs

Table B.9: High-confidence fine-mapped non-coding variant-trait paris

Table B.10: High-confidence fine-mapped intergenic variants that are more than 250 kb away from the closest gene

Table B.11: Extremely population-enriched (> 10-fold) high-PIP non-coding variants

Table B.12: Nonsynonymous (pLoF/missense) coding variants with the best PIP > 0.1

Table B.13: Allelic series of nonsynonymous coding variants (PIP > 0.1)

Table B.14: Allelic series of nonsynonymous coding variants and proximal non-coding variants (< 100 kb)

B.3 Supplementary Figures

Figure B.1 ***(following page)*****: Functional enrichments of fine-mapped variants. a–d**. Proportion of variants for the seven main functional categories (Methods), stratified by the best PIP bin for a variant in BBJ, FinnGen, UKBB, and all cohorts combined. Labels above each bar represent the number of variants in each bin. **e–h**. Enrichments of fine-mapped variants (PIP > 0.9) in each functional category compared to non-fine-mapped variants (PIP ≥ 0.01). **i**. Enrichments in 35 binary annotations from the baselineLD v2.2 model [188]. Enrichment was calculated as a relative risk (*i.e.*, a ratio of proportion of variants) between being in an annotation and fine-mapped (PIP ≥ 0.01 or PIP > 0.9; **2.4 Methods**. Error bars correspond to 95% confidence intervals using bootstrapping. Numerical results are available in **Supplementary Tables B.5,B.6**.

Figure B.1: (continued)

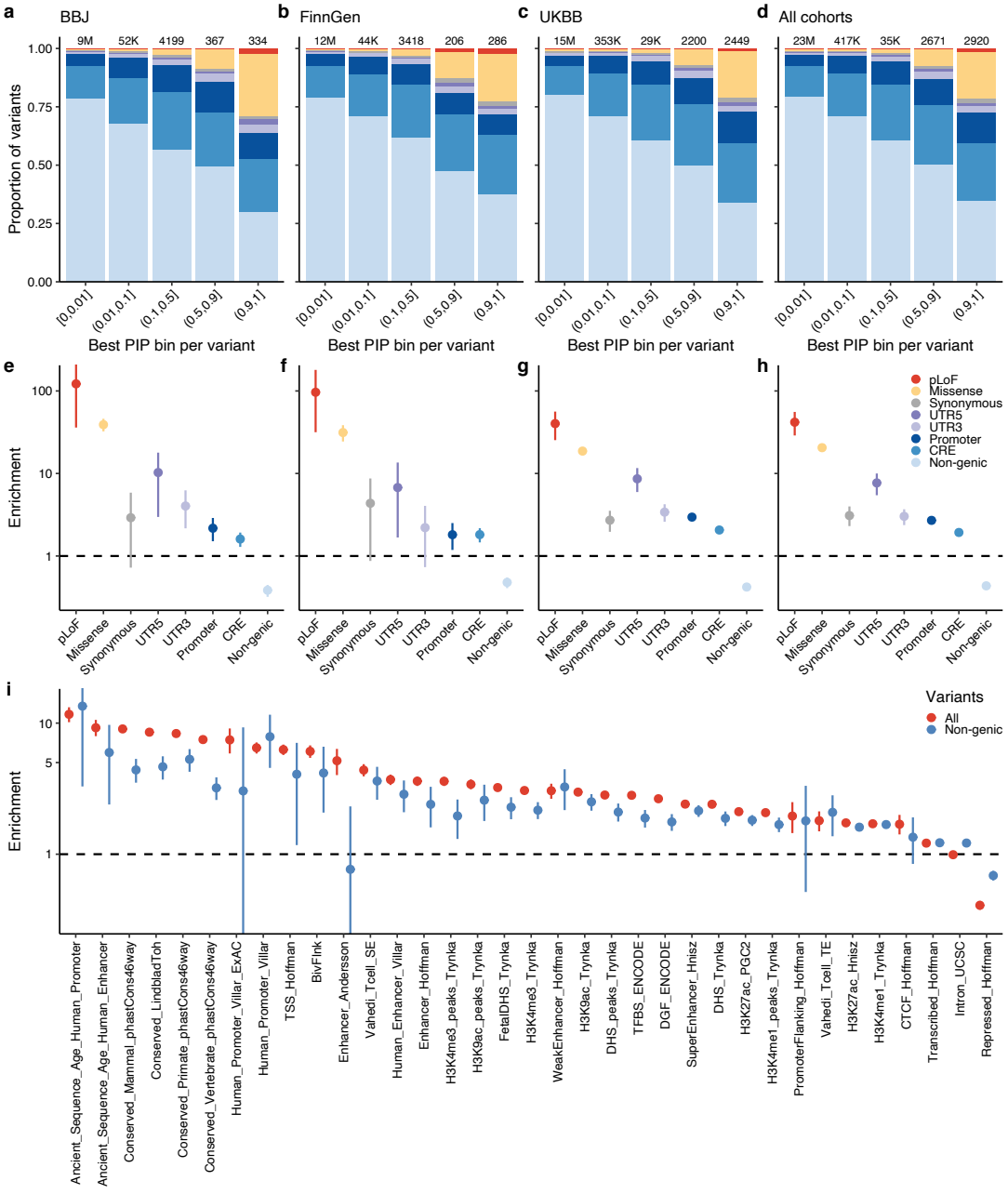


Figure B.2 ***(following page):*** **Additional details of fine-mapping replication status across populations. a,b**. Breakdowns for the genome-wide significant variant-trait pairs ($P_{\mathrm{GWAS}} < 5.0 \times 10^{-8}$) in a secondary population, using distinct fine-mapping replication criteria (**a**. PIP > 0.05 and **b**. in 95% CS) different from **Fig. 2.2** (PIP > 0.1). **c–e**. PIP distributions of the genome-wide significant variant-trait pairs in a secondary population, stratified by PIP bins in a discovery population. Half-sided violin plots represent PIP distributions for each secondary population. Points represent mean PIP in the secondary population for each PIP bin in a discovery population. **f**. Comparison of PIP distributions in a secondary population for real data and simulations, stratified by discovery cohorts which true causal variants (PIP > 0.9) were taken from for simulations (**2.4 Methods**). Left-sided plots represent PIP distributions from real data. Right-sided plots represent PIP distributions from simulations. **g**. Barplots showing a fraction of the high-PIP (> 0.9) variant-trait pairs in real data and simulations. The top bar represents the result from the real data which is slightly different from **Fig. 2.2** to make an apple-to-apple comparison with simulations (**2.4 Methods**). The remaining bars represent the results from simulations. We categorized the genome-wide significant variant-trait pairs ($P_{\mathrm{GWAS}} < 5.0 \times 10^{-8}$ in a simulation) into whether they showed simulated PIP > 0.1 in a secondary population or not, which is stratified by discovery cohorts which true causal variants (PIP > 0.9) were taken from for simulations (**2.4 Methods**). **h**. Proportion of variants for the seven main functional categories, stratified by the replication status shown in **Fig. 2.2** (**2.4 Methods**). Labels above each bar represent the number of variants in each status.

Figure B.2: (continued)

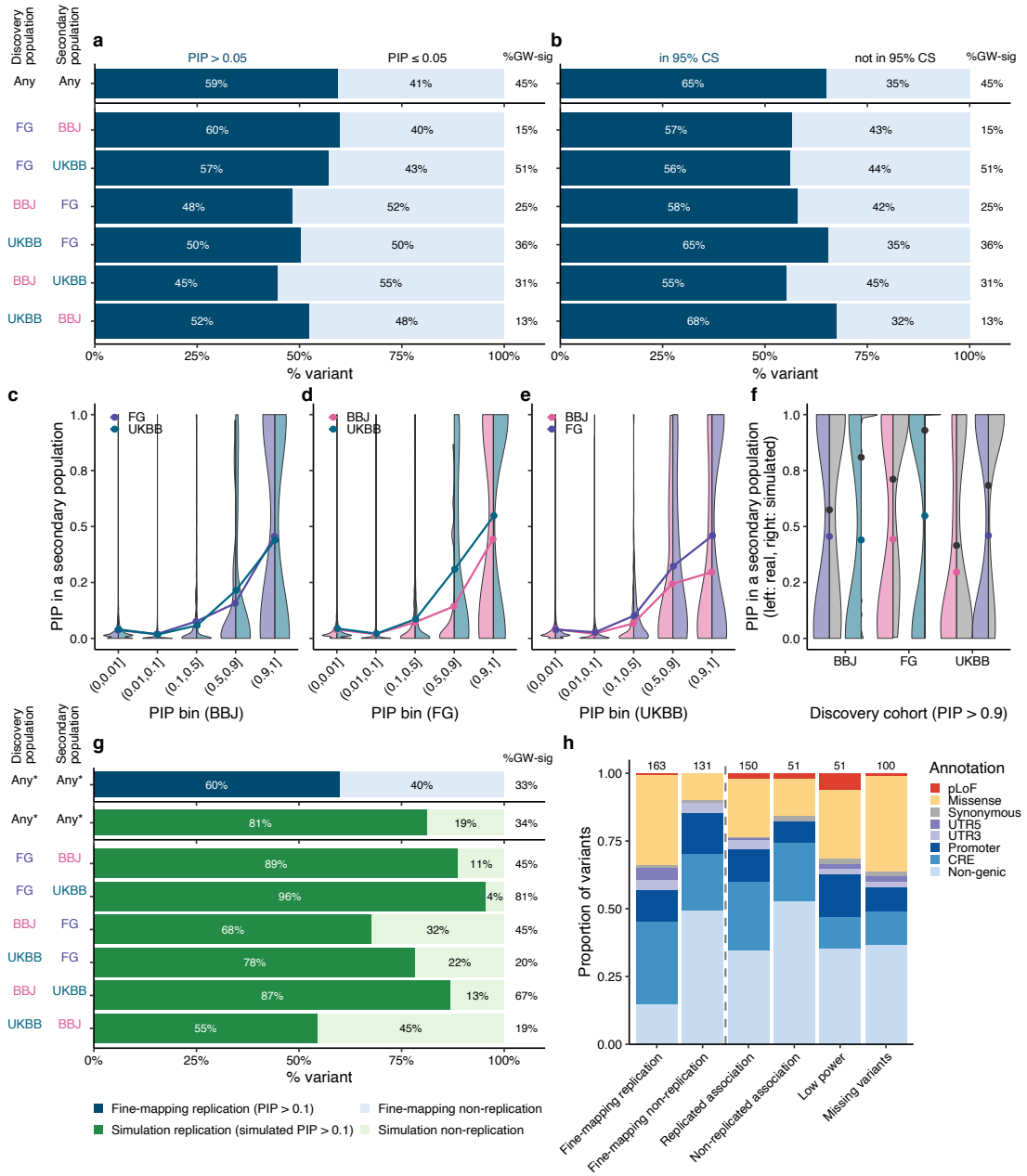
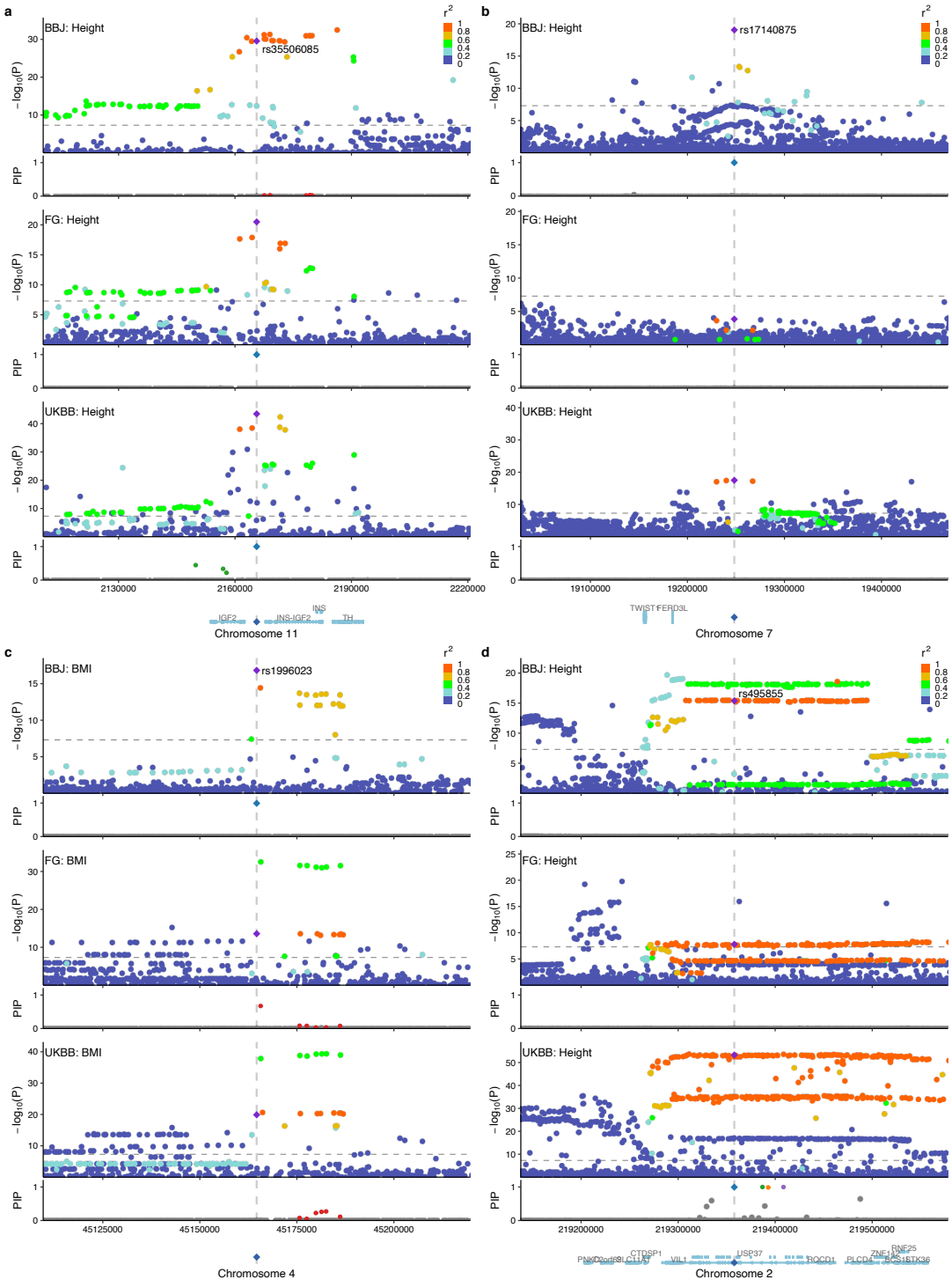


Figure B.3 ***(following page):*** **Illustrative examples of fine-mapping non-replication across populations.** Locuszoom plots for the same locus across populations. Colors in the locuszoom panels represent r^2 values to the lead variant. In the PIP panels, only fine-mapped variants in SuSiE 95% CS are colored, where the same colors are applied across populations based on the merged CS (**2.4 Methods**). **a**. rs35506085 for height that was fine-mapped in FinnGen and UKBB (PIP = 1.0), but not in BBJ (PIP ~ 0) likely due to extensive LD. **b**. rs17140875 for height that was fine-mapped in BBJ (PIP = 1.0) but not in FinnGen or UKBB (PIP ~ 0). The variant is more common in BBJ (MAF = 0.08) than in FinnGen or UKBB (MAF = 0.04 and 0.05, respectively) and has more LD neighbors in Europeans. **c**. rs1996023 for BMI that was fine-mapped in BBJ (PIP = 0.99), but not in FinnGen or UKBB (PIP ~ 0). Instead, we found other CS in FinnGen and UKBB that showed modest LD with rs1996023 in Europeans (r^2 ~ 0.5) but high LD in BBJ (r^2 ~ 0.8). **d**. rs495855 for height that was fine-mapped only in UKBB (PIP = 1.0). This seems very likely a false positive given extensive LD in every population.

Figure B.3: (continued)



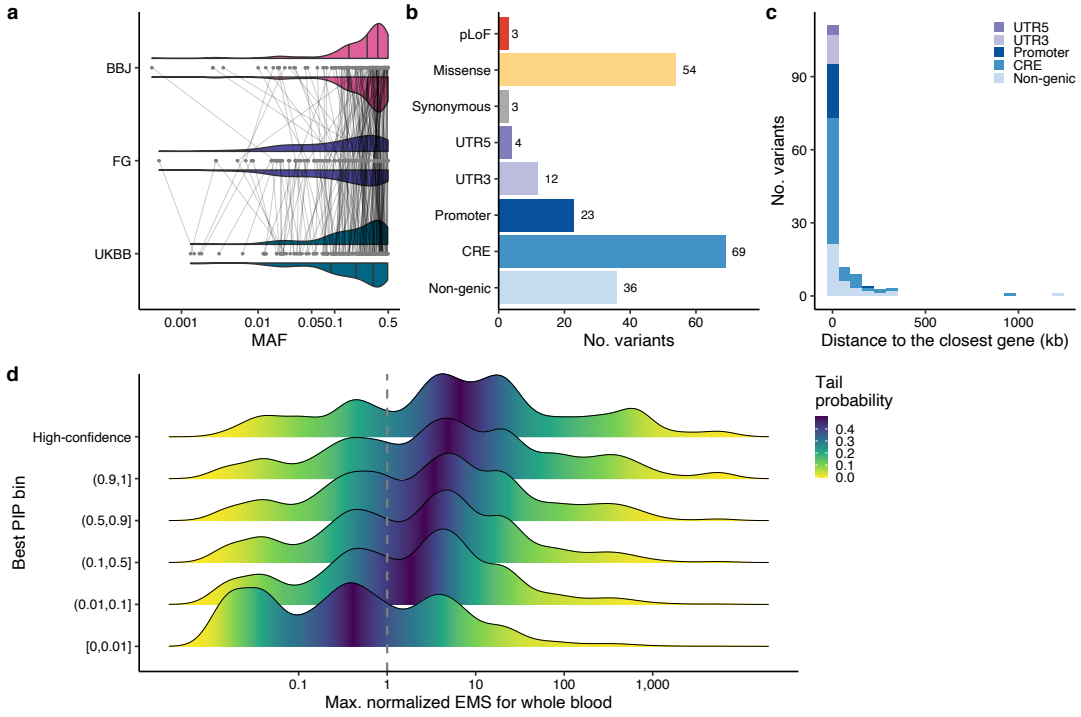


Figure B.4: Overview of high-confidence fine-mapped variants. a. Distribution of minor allele frequencies (MAF) in each cohort. Violin plots represent the distribution. Each point represents a high-confidence fine-mapped variant and each line connects the same variant across cohorts. **b**. Consequences annotated by VEP (see **2.4 Methods**). **c**. Histogram of distance to the closest gene for high-confidence fine-mapped non-coding variants. Color represents non-coding consequences same as **b**. **d**. Distribution of predicted expression modifier score (EMS)[138] for fine-mapped non-coding variants, stratified by the best PIP bins. The highest bin ($0.9 < \text{PIP} \leq 1$) was further stratified into the high-confidence variants or not based on replication across populations (see **2.4 Methods**). Maximum normalized EMS score over genes was calculated for each fine-mapped variant using the whole blood tissue. Details of the high-confidence fine-mapped variants are summarized in **Supplementary Tables B.8,B.9**.

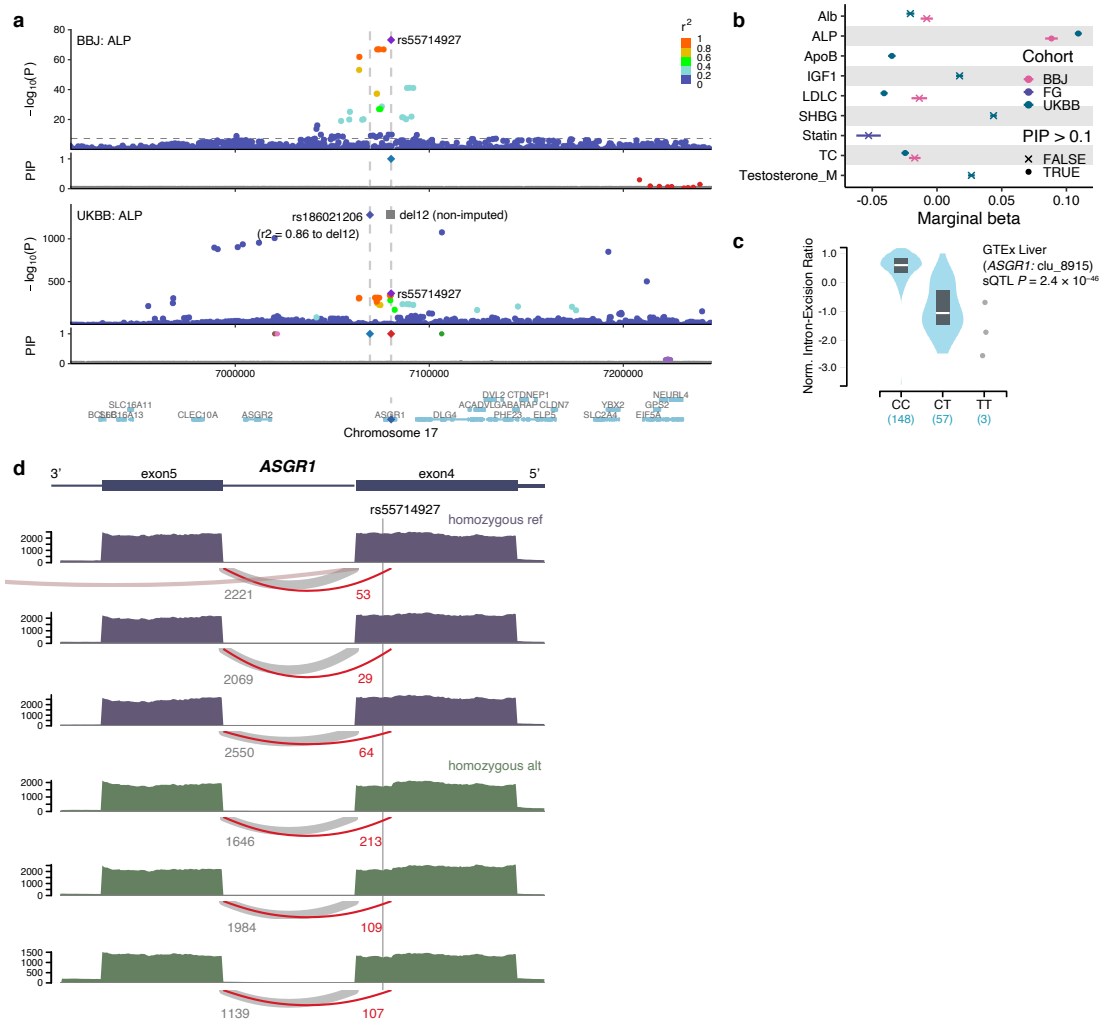
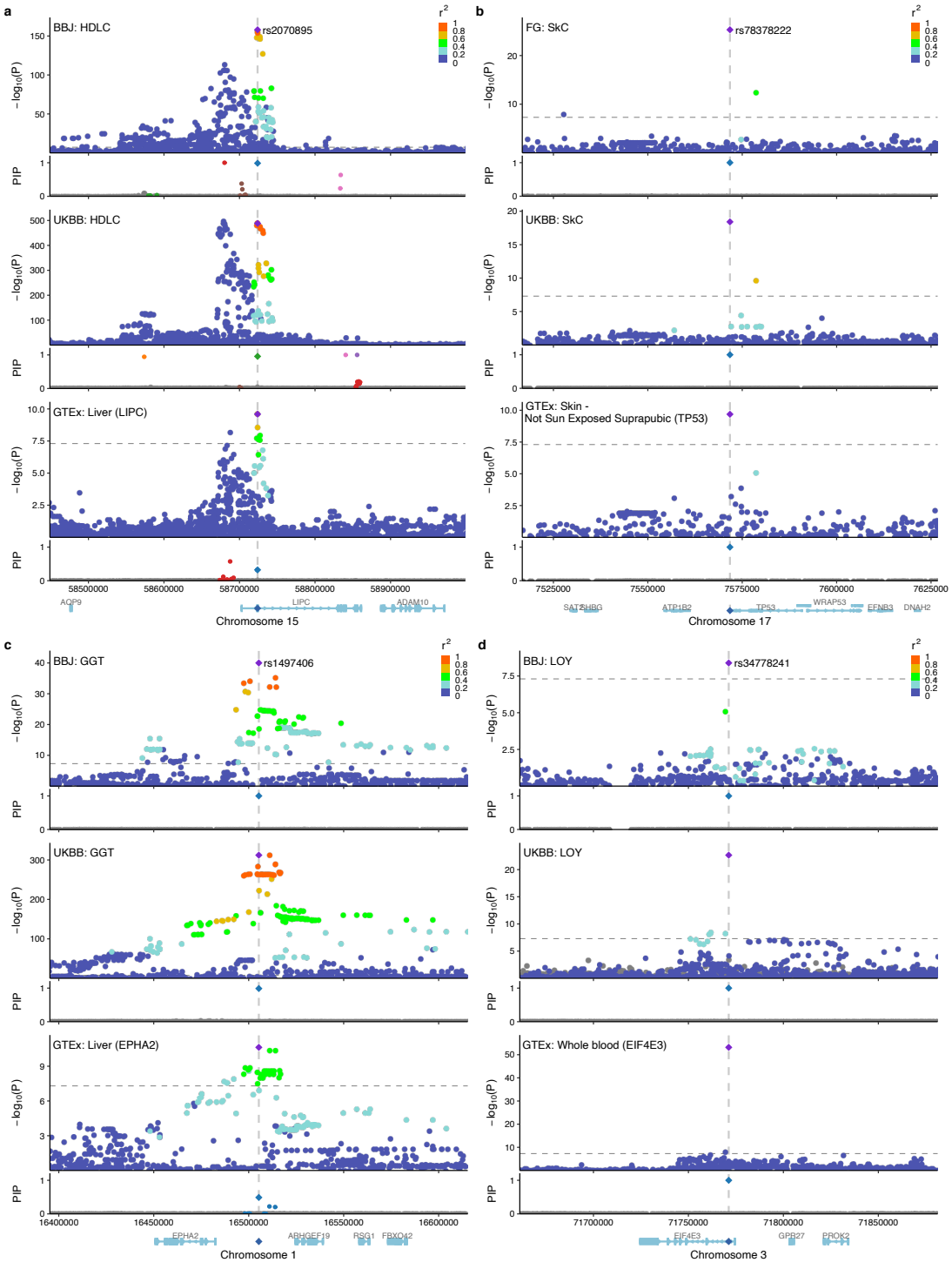


Figure B.5: Synonymous variant rs55714927 shows splicing effect in *ASGR1*. **a**. Locuszoom plots for alkaline phosphatase (ALP) in BBJ and UKBB. **b**. Phenome-wide association study (PheWAS) of rs55714927 across all the traits analyzed in this study. Only phenotypes that showed $P < 5.0 \times 10^{-8}$ in any cohort are displayed. Each point represents a marginal beta for a given trait in a cohort, with an error bar representing the standard error. Shape of each point represents whether each variant showed PIP > 0.1. **c**. sQTL effect of rs55714927 in GTEx liver. **d**. Sashimi plot showing splicing effects of rs55714927 in three homozygous reference allele carriers vs. three homozygous alternative allele carriers that were randomly chosen.

Figure B.6 *(following page)*: Colocalization between high-confidence fine-mapped non-coding variants for complex traits and cis-eQTL associations in trait-relevant tissues. Locuszoom plots for the same locus of complex traits across populations and of cis-eQTL associations in trait-relevant tissues. Colors in the locuszoom panels represent r^2 values to the lead variant. In the PIP panels, only fine-mapped variants in SuSiE 95% CS are colored, where the same colors are applied across populations based on the merged CS (**2.4 Methods**). **a**. rs2070895 for HDL cholesterol in BBJ and UKBB and for *LIPC* expression in GTEx liver. **b**. rs78378222 for skin cancer in FinnGen and UKBB and for *TP53* expression in GTEx skin. **c**. rs1497406 for γ-glutamyl transferase (GGT) in BBJ and UKBB and for *EPHA2* expression in GTEx liver. **d**. rs34778241 for loss of chromosome Y (LOY) in BBJ and UKBB and for *EIF4E3* expression in GTEx whole blood.

Figure B.6: (continued)



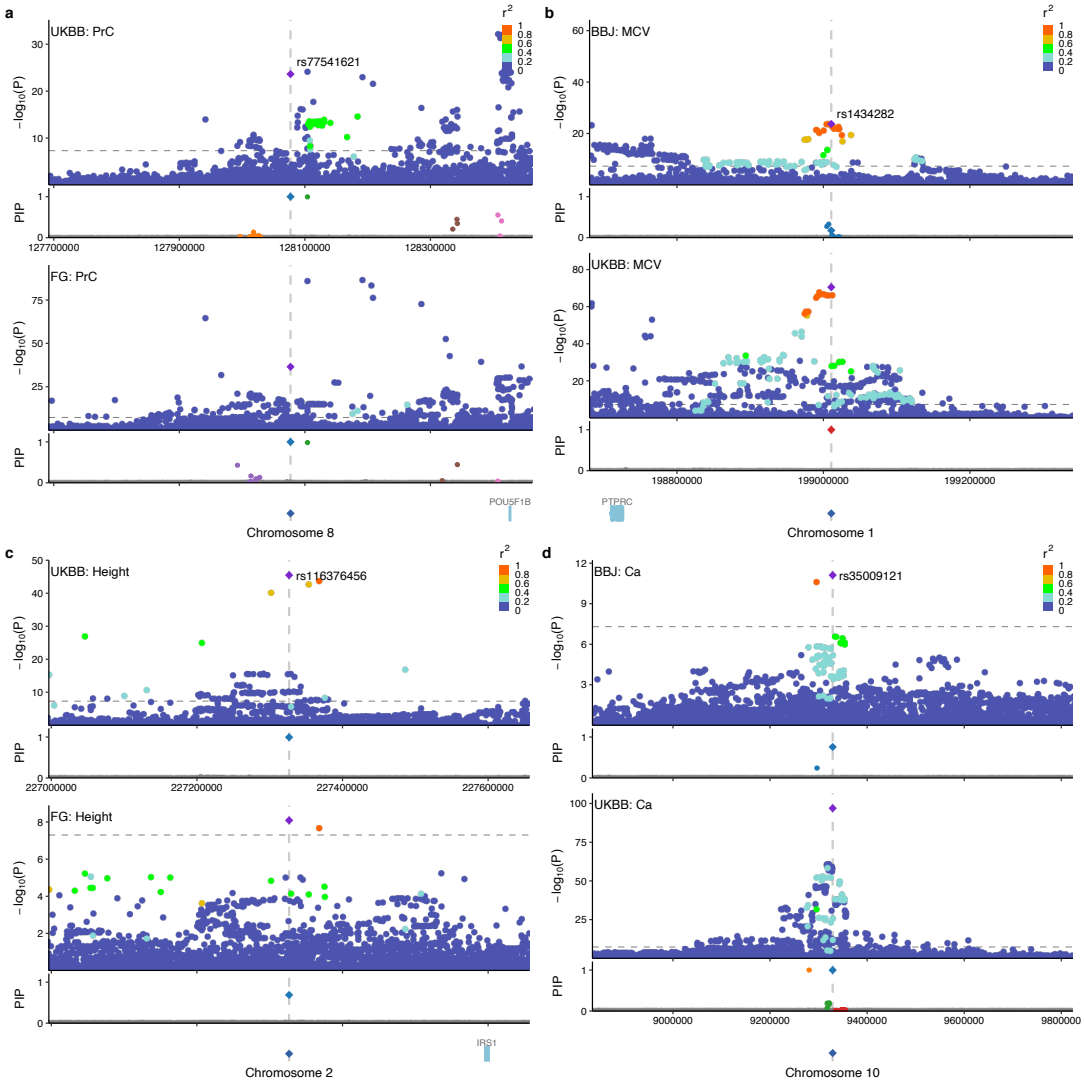


Figure B.7: High-confidence fine-mapped intergeneric variants in a gene desert. Locuszoom plots for the same locus across populations. Colors in the locuszoom panels represent r^2 values to the lead variant. In the PIP panels, only fine-mapped variants in SuSiE 95% CS are colored, where the same colors are applied across populations based on the merged CS (**2.4 Methods**). **a**. rs77541621 in the 8q24 locus for prostate cancer in UKBB and FinnGen. **b**. rs1434282 in the 1q32 locus for mean corpuscular volume (MCV) in BBJ and UKBB. **c**. rs116376456 in the 2q36 locus for height in UKBB and FinnGen. **d**. rs35009121 in the 10p14 locus for calcium levels in BBJ and UKBB.

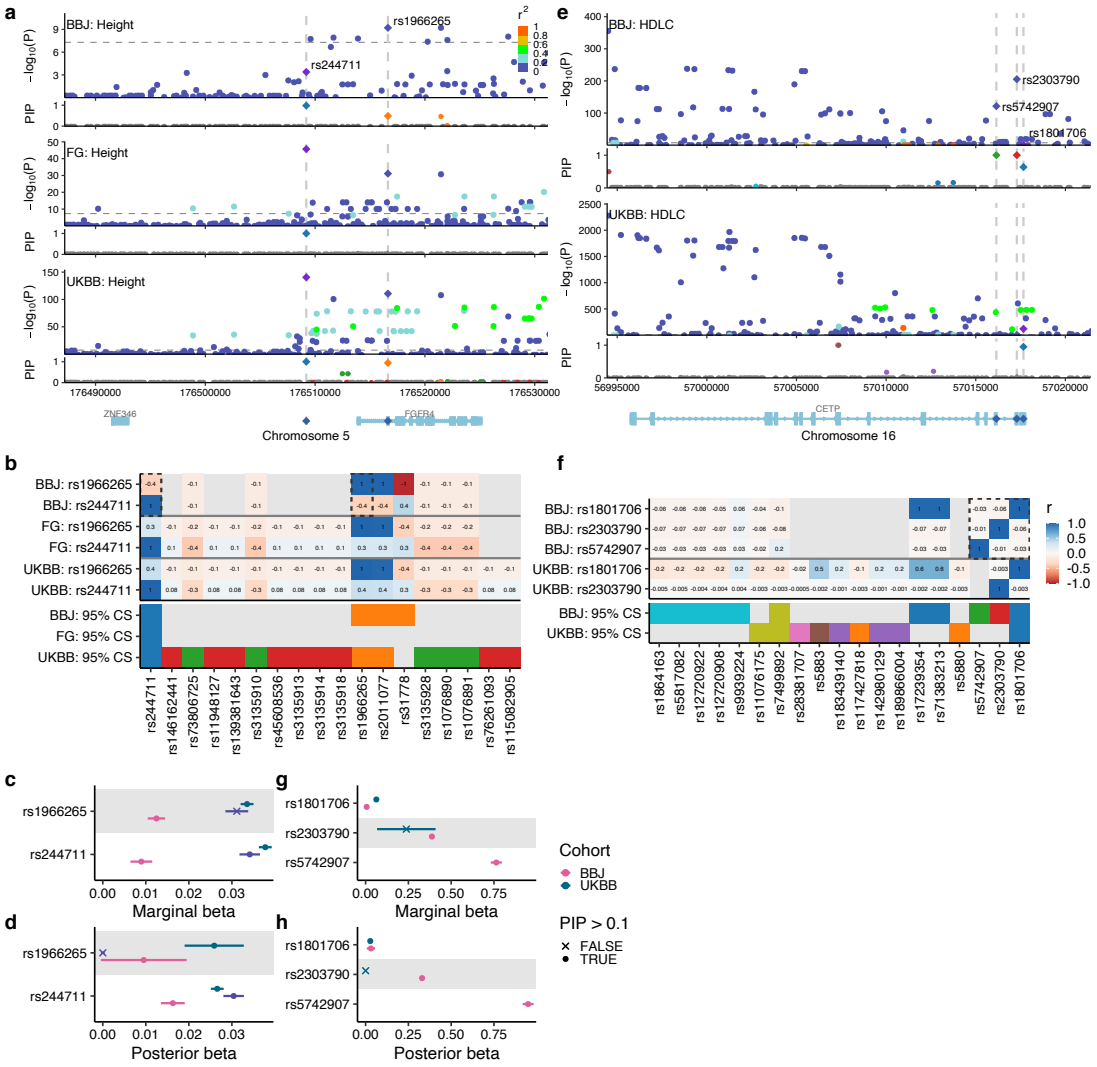


Figure B.8: Putative causal variants are negatively correlated in a locus. a–d. rs244711 and rs1966265 for height in BBJ, FinnGen, and UKBB. **e–h**. rs1801706, rs5742907 and rs2303790 for HDL cholesterol in BBJ and UKBB. **a, e**. Locuszoom plots for the same locus across populations. Colors in the manhattan panels represent r^2 values to the lead variant. In the PIP panels, only fine-mapped variants in SuSiE 95% CS are colored, where the same colors are applied across populations based on the merged CS (**2.4 Methods**). **b, f**. Heatmaps showing r values between the highlighted variants and the other variants in 95% CS in each population. In a CS panel, variants are colored by the same colors in the locuszoom plots (**a, e**). **c, d, g, h**. Forest plots showing marginal and posterior betas of fine-mapped variants. Point color represents each cohort and shape represents whether the variant showed PIP > 0.1 in each cohort.

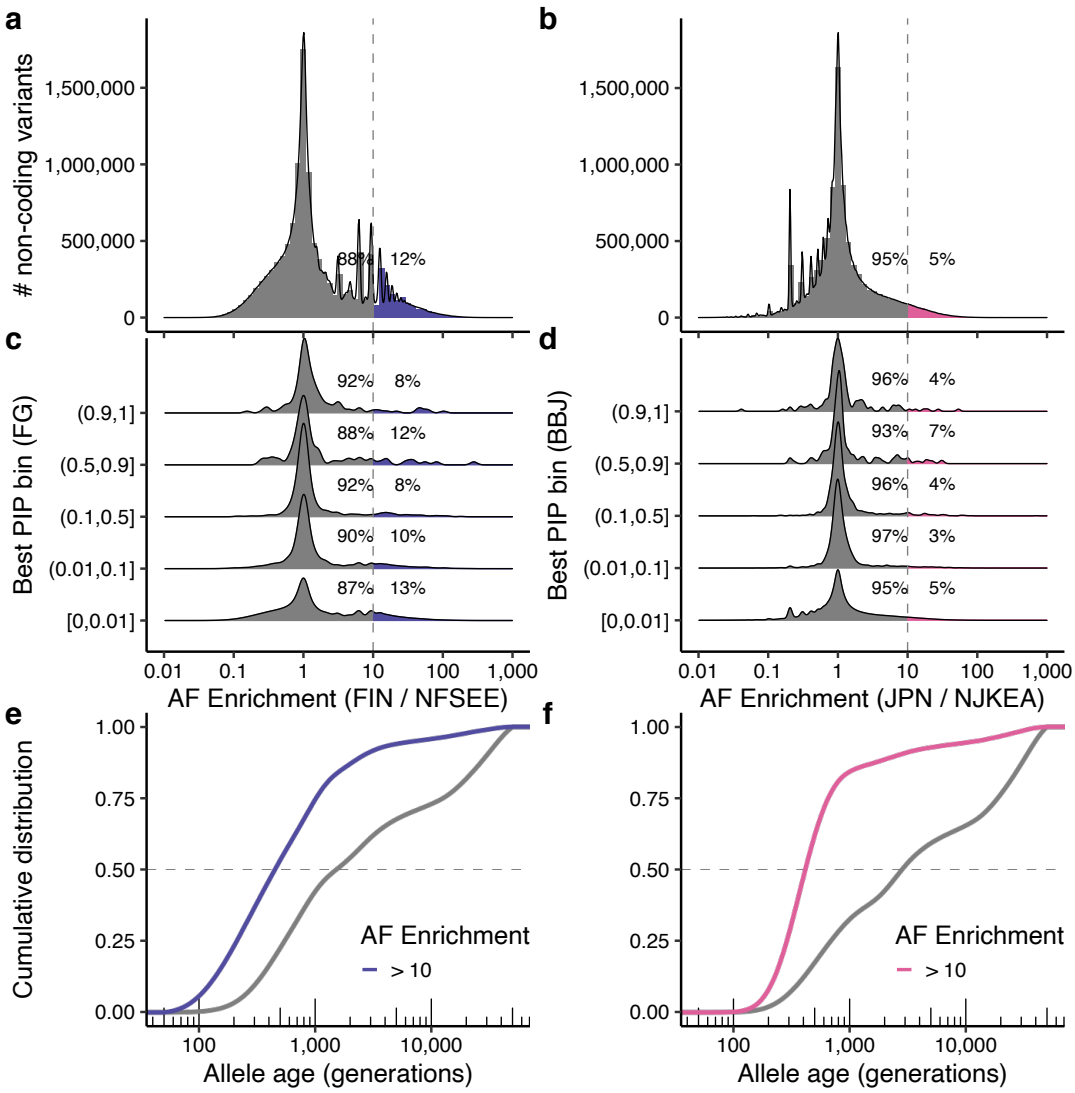
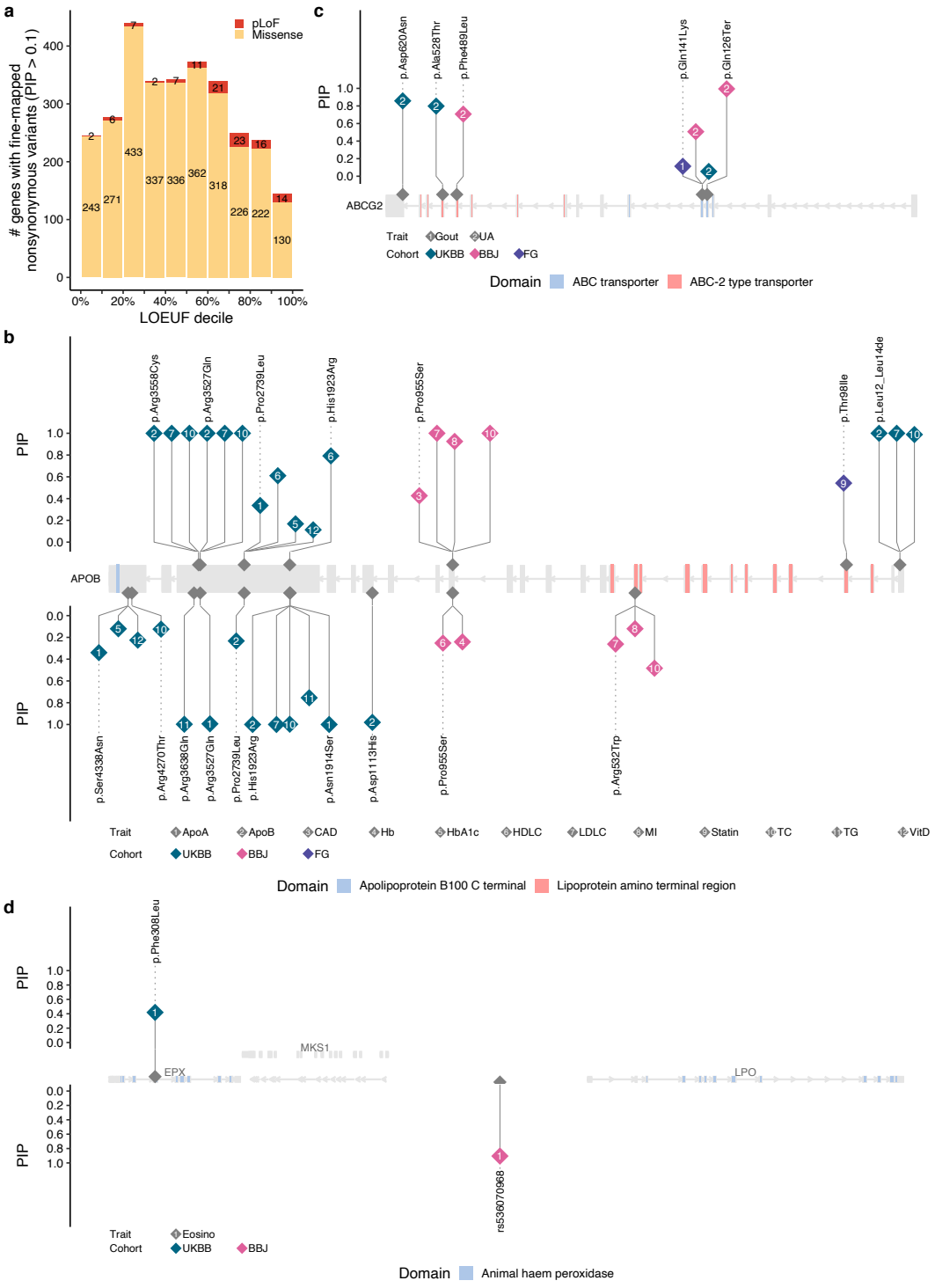


Figure B.9: Population-enriched non-coding variants. a–d. Histograms showing a distribution of allele frequency (AF) enrichment metric in (**a**) Finnish (n = 1,738) and (**b**) Japanese (n = 7,609) populations. A ratio of AF was computed against NFSEE (n = 5,421) and NJKEA (n = 780) for non-coding variants analyzed in FinnGen or BBJ GWAS that exist in gnomAD WGS or GEM-J WGS, respectively. For a subset of variants that are fine-mapped in our analysis (see **2.4 Methods**), we show AF enrichment distribution across maximum PIP bins computed in (**c**) FinnGen or (**d**) BBJ. **e,f.** Cumulative distribution of estimated allele age for non-coding variants, stratified by AF enrichment in (**e**) Finnish or (**f**) Japanese. FIN: Finnish, JPN: Japanese, NFSEE: Non-Finnish-Swedish-Estonian European, NJKEA: Non-Japanese-Korean East Asian.

Figure B.10 *(following page)*: Allelic series of putative causal variants across populations. a. Number of genes with fine-mapped nonsynonymous variants (pLoF and missense) with best PIP > 0.1 for each LOEUF decile[214]. Genes without fine-mapped nonsynonymous variants are not plotted. Colors represent the consequence of each variant. When multiple nonsynonymous variants are found, the most deleterious consequence is colored. **b–d**. Lollipop plots of allelic series for (**b**) *APOB*, (**c**) *ABCG2*, and (**d**) *EPX*. Each point represents a fine-mapped variant from a single trait and cohort. Point color represents discovery cohort and number label represents a fine-mapped trait. Points above the gene body correspond to those with positive effect sizes, whereas points below the gene body correspond to those with negative effect sizes. Coding variants are labeled with the HGVS protein nomenclature and non-coding variants (in **d**) are labeled with rsids. Protein domains are annotated based on the Pfam database.

Figure B.10: (continued)



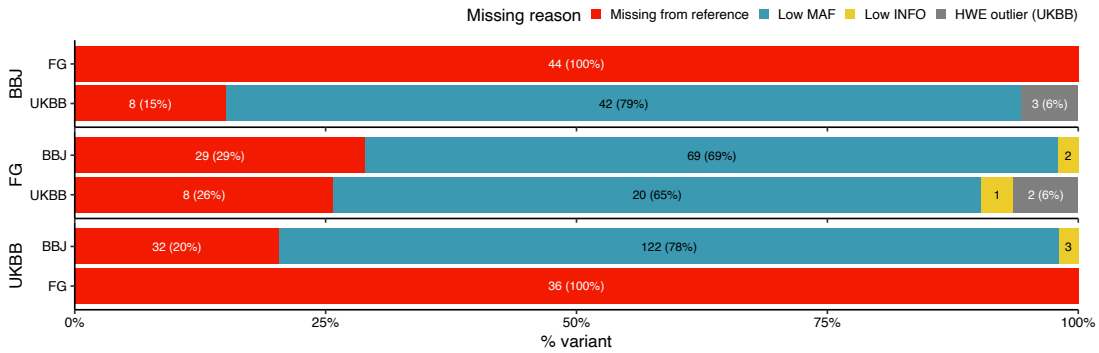


Figure B.11: Overview of "missing" variants from summary statistics. We characterized the reasons that high-PIP variants (PIP > 0.9) in a single population are missing from summary statistics in other populations. The following criteria represent imputation and quality control procedures adopted in each cohort (**2.4 Methods**). Missing from reference: Variants do not exist in imputation reference panels used in each cohort, *i.e.*, BBJ: the 1000 Genomes Phase 3 (n = 2,504) + Japanese WGS (n = 1,037); FinnGen: Finnish WGS (n = 3,775); and UKBB: the Haplotype Reference Consortium (n = 64,976) + the 1000 Genomes Phase 3 + UK10K (n = 3,781). Low MAF: MAF < 0.005 in a population based on the GEM-J WGS for BBJ and the gnomAD v2 for FinnGen and UKBB. Low INFO: INFO < 0.7 for BBJ and INFO < 0.8 in FinnGen/UKBB. HWE outlier (UKBB): HWE P-value < 1.0×10^{-10} (only in UKBB).

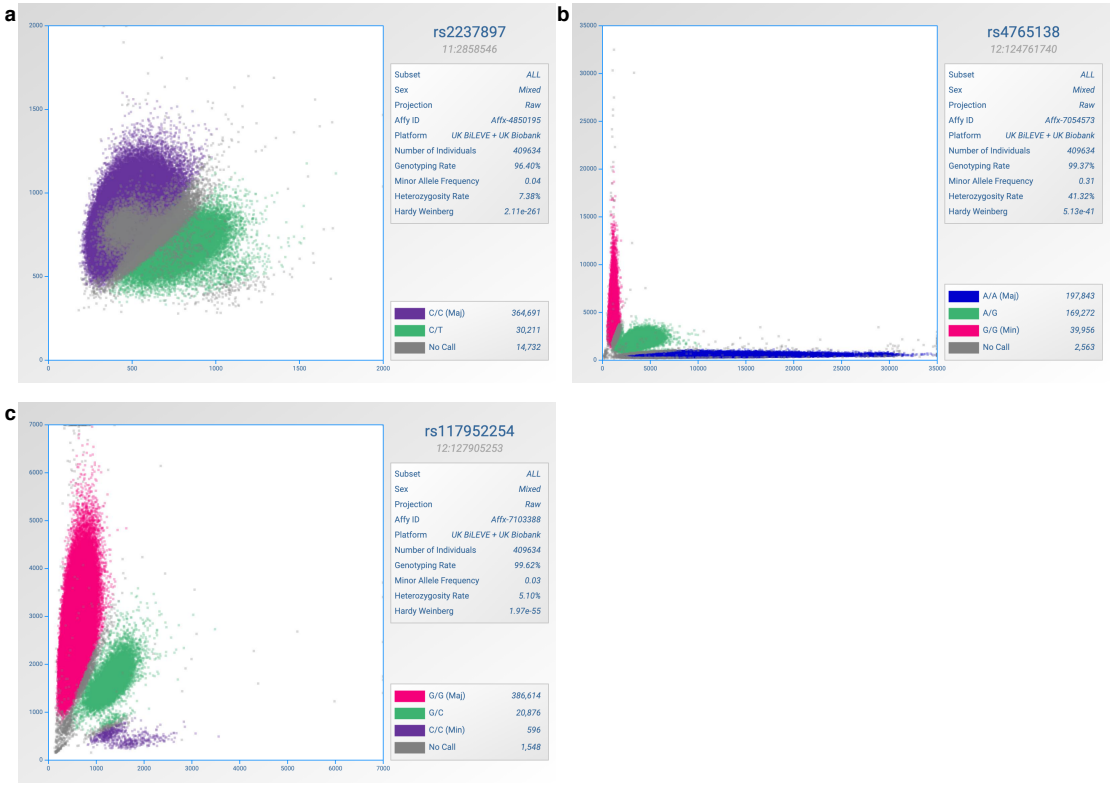


Figure B.12: Genotype cluster plots in UKBB "white British" individuals. Cluster plots of genotyped SNP intensity are shown for three SNPs: **a**. rs2237897, **b**. rs4765138 and **c**. rs117952254. Colors correspond to called genotypes. All the variants passed a per-batch QC but showed HWE test P-value < 1×10^{-10} in aggregate. The cluster plots were generated by ScatterShot (http://mccarthy.well.ox.ac.uk/static/software/scattershot/).

C

Supplementary Materials for Chapter 3

C.1 Supplementary Note

C.1.1 Standard variant-level QC procedures for meta-analysis summary statistics

Various QC practices on GWAS summary statistics have been adopted for meta-analysis, which were extensively described elsewhere[249]. Here, we summarize pre- and post-meta-analysis QC procedures that primarily act on individual variants, and thus affect calibration and recall for meta-analysis fine-mapping at single-variant resolution. Study-level QC should be conducted separately, such as integrity check of submitted files (*e.g.*, nonsense/missing values), population stratification (*e.g.*, LD score regression intercept[321]), and ancestry composition (*e.g.*, PCA projection to the unified reference[239]).

Pre-meta-analysis QC (per-cohort):

- Minor allele frequency or count filtering (*e.g.*, MAF > 0.1% or MAC > 20)
- Imputation quality score filtering (*e.g.*, INFO > 0.3)
- Variant normalization
  - Variants should have the same ID nomenclature (*e.g.*, chromosome:position:ref:alt) that represents a unique genomic position and alleles on the forward strand (*i.e.*, rsid is not recommended)
  - Indels should be normalized (left-aligned and trimmed to be parsimonious)[322]
- Allele frequency consistency with external public reference (*e.g.*, gnomAD)
  - Extra care should be taken for palindromic single-nucleotide variants (with A/T or G/C alleles) and indels.

  - Please refer to the GBMI flagship paper[239] for our filtering criteria based on the Mahalanobis distance.

- If multiple genome builds exist, lifting over variants to the major genome build
  - Variants that are located within conversion-unstable positions (CUP)[253] between genome builds should be removed, including those that fail at conversion, map to different chromosomes, and do not map back to the original position when lifting back to the original genome build.
  - Effect alleles should be consistent between genome builds because reference and alternative alleles could be flipped during liftover.

Post-meta-analysis QC:

- Effect size heterogeneity test (*e.g.*, Cochran's Q-test)
- Rffective sample size filtering (*e.g.*, maximum $N_{\text{eff}} > 50$

C.1.2 Approximate Bayes factors as a function of sample size

We assume a model

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{e}$$

$$\mathbf{e} \sim N_n(0, \sigma^2 I_n)$$

$$\beta = b\gamma$$

where

$$b \sim N_1(0, \sigma_0^2)$$

and γ is a unit basis vector chosen uniformly at random. Assuming a single causal variant, the posterior inclusion probability (PIP) for variant i is the posterior probability that variant i is causal, and can be calculated as

$$\begin{aligned}\text{PIP}(i) &= \frac{\mathbb{P}(y \mid x, \sigma^2, \sigma_0^2, \gamma = e_i)}{\sum_j \mathbb{P}(y \mid x, \sigma^2, \sigma_0^2, \beta_j \neq 0)} \\ &= \frac{\text{BF}(i)}{\sum_j \text{BF}(j)}\end{aligned}$$

where

$$\text{BF}(i) = \frac{\mathbb{P}(y \mid x, \sigma^2, \sigma_0^2, \gamma = e_i)}{\mathbb{P}(y \mid x, \sigma^2, \beta = 0)}.$$

Let x_i denote the genotype at the i-th SNP. When no genotypes are missing, we have

$$\mathbb{P}(y \mid x, \sigma^2, \sigma_0^2, \gamma = e_i) = \mathbb{P}(y \mid x_i, \sigma^2, \sigma_0^2, \gamma = e_i)$$

and

$$\mathbb{P}(y \mid x, \sigma^2, \sigma_0^2, \gamma = 0) = \mathbb{P}(y \mid x_i, \sigma^2, \sigma_0^2, \gamma = 0).$$

Thus, $\text{BF}(i)$ can be computed by considering only variant i. However, when there are missing genotypes, these equalities no longer hold. Meta-analysis fine-mapping typically proceeds by ignoring this fact and continuing to use the standard approximation to $\text{BF}(i)$ introduced by Wakefield[6]

To concretize the lack of accuracy of ABF-based PIPs in the presence of missing data, we show how sample size differences between variants in a locus affect their respective ABF PIPs. Specifically, we show that for every marginal effect size $\hat{\beta}$ and prior effect size variance σ_0^2, there exists a sample size n above which the approximate Bayes Factor for a variant, and thus the PIP, increases monotonically with increasing sample size (and below which, the PIP decreases monotonically with increasing sample size). For a variant i, we have the following linear regression model, where $\mathbf{y}$ and $\mathbf{x_i}$ are vec-

tors of length n representing the standardized phenotype and standardized genotype at one locus, respectively:

$$\mathbf{y} = \mathbf{x_i}\beta_i + \mathbf{e}$$

$$\mathbf{e} \sim N_n(0, \sigma^2 I_n)$$

$$\beta_i \sim N_1(0, \sigma_0^2)$$

with $n, \sigma_0^2 > 0$. Approximate Bayes factors are computed as follows:

$$\mathrm{BF}(i) = \frac{\mathbb{P}(y \mid x_i, \sigma^2, \sigma_0^2, \beta_i \neq 0)}{\mathbb{P}(y \mid x_i, \sigma^2, \sigma_0^2, \beta_i = 0)} = \sqrt{\frac{s^2}{\sigma_0^2 + s^2}} \exp \frac{z^2}{2} \times \frac{\sigma_0^2}{\sigma_0^2 + s^2}$$

where $z = \hat{\beta}_i / s, \hat{\beta}_i = (\mathbf{x_i}^T \mathbf{x_i})^{-1} \mathbf{x_i}^T \mathbf{y}$ is the least squares estimate of β, and its variance $s^2 = \sigma^2 / (\mathbf{x_i}^T \mathbf{x_i})$. Since $\mathbf{x_i}$ and $\mathbf{y}$ are mean-centered and scaled to unit variance $s \approx \sqrt{1/n}$ and thus $z^2 \approx n\hat{\beta}_i^{\,2}$.

To see how changes in n for fixed $\hat{\beta}_i$, σ_0^2, and σ^2 impact the value of the Bayes factor, we take it's derivative with respect to n and set it equal to zero. After some algebra, we get:

$$\sigma_0^2 n^2 + (2 - \frac{\sigma_0^2}{\hat{\beta}_i^{\,2}})n - \frac{1}{\hat{\beta}_i^{\,2}} = 0$$

which is a parabola with one positive root at $n = \frac{-1}{\sigma_0^2} + \frac{1}{2\hat{\beta}_i^{\,2}} + \sqrt{\frac{1}{\sigma_0^4} + \frac{1}{4\hat{\beta}_i^{\,4}}}$. Above this value of n, the Bayes factor increases monotonically with increasing sample size. For example, when the estimated effect $\hat{\beta}_i = 0.01$ and the prior effect size variance $\sigma_0^2 = 0.04$, the Bayes factor will increase monotonically when the sample size $n > 9{,}976$. Thus, a causal variant that is not imputed in some studies and thus has a lower sample size than nearby variants will typically have a relatively lower Bayes Factor.

C.1.3 Global Biobank Meta-analysis Initiative

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

Sebastian Zoellner[32], Biobank Japan, BioMe, BioVU, Canadian Partnership for Tomorrow, China Kadoorie Biobank Collaborative Group, Colorado Center for Personalized Medicine, deCODE Genetics, Estonian Biobank, FinnGen, Generation Scotland, Genes & Health, LifeLines, Mass General Brigham Biobank, Michigan Genomics Initiative, QIMR Berghofer Biobank, Taiwan Biobank, The HUNT Study, UCLA ATLAS Community Health Initiative, UK Biobank, Alicia R Martin[1,2,3], Cristen J Willer[6,14,74*], Mark J Daly[1,2,3,21*], Benjamin M Neale[1,2,3*]

*These authors jointly supervised this work

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA, [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, [5]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan, [6]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA, [7]K. G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health, NTNU, Norwegian University of Science and Technology, Trondheim, Norway, [8]Division of Medicine and Laboratory Sciences, University of Oslo, Norway, [9]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK, [10]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA, [11]Vanderbilt Genetic Institute, Vanderbilt University Medical Center, Nashville, TN, USA, [12]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [13]MRC Integrative Epidemiology Unit (IEU), Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, UK, [14]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA, [15]University of Groningen, UMCG, Department of Ophthalmology, Groningen, the Netherlands,

[16]Department of Clinical Genetics, Amsterdam University Medical Center (AMC), Amsterdam, the Netherlands, [17]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden, [18]University of Groningen, UMCG, Department of Genetics, Groningen, the Netherlands, [19]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia, [20]Ontario Institute for Cancer Research, Toronto, ON, Canada, [21]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland, [22]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim, 7030, Norway, [23]University of Colorado - Anschutz Medical Campus, Aurora, CO, USA, [24]Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan, [25]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA, [26]Division of Biostatistics, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taiwan, [27]QIMR Berghofer Medical Research Institute, Brisbane, Australia, [28]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA, [29]Nuffield Department of Population Health, University of Oxford, Oxford, UK, [30]Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China, [31]MRC Population Health Research Unit, University of Oxford, Oxford, UK, [32]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA, [33]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [34]deCODE Genetics/Amgen inc., 101, Reykjavik, Iceland, [35]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashviile, TN, USA, [36]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK, [37]University of Groningen, UMCG, Department of Epidemiology, Groningen, the Netherlands, [38]Department of Anesthesiology, Michigan Medicine, Ann Arbor, MI, USA, [39]Institute of Healthcare Policy & Innovation, University of Michigan, Ann

Arbor, MI, USA, [40]Wolfson Institute of Population Health, Queen Mary University of London, London, UK, [41]Chinese Academy of Medical Sciences, Beijing, China, [42]Blizard Institute, Queen Mary University of London, London, UK, [43]Central Pharmaceutical Research Institute, JAPAN TOBACCO INC., Takatsuki 569-1125, Japan, [44]Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK, [45]Department of Biostatistics, Vanderbilt University Medical Center, Nashviile, TN, USA, [46]Biogen, Cambridge, MA, USA, [47]NIHR Bristol Biomedical Research Centre, Bristol, UK, [48]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK, [49]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, [50]Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA, [51]Centre for Population Genomics, VA Boston Healthcare System, Boston, MA, USA, [52]Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA, USA, [53]Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [54]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK, [55]Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan, [56]University of Michigan, Department of Pediatrics, Ann Arbor MI 48109, [57]School of Basic and Medical Biosciences, Faculty of Life Sciences and Medicine, King's College London, London, UK, [58]Bradford Institute for Health Research, Bradford Teaching Hospitals National Health Service (NHS) Foundation Trust, Bradford, UK, [59]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, [60]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA, [61]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [62]Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA, [63]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [64]Department of Public

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

C.2 Supplementary Tables

The following Supplementary Tables are available in the online version of the manuscript.

Table C.1: Number of unrelated samples for simulated cohorts
Table C.2: Number of chromosome 3 variants in Illumina manifest and those extracted from 1000GP African, East Asian, and European populations
Table C.3: Number of imputed and QC-passing variants (MAF > 0.001 and Rsq > 0.6)
Table C.4: List of configurations for meta-analysis simulation
Table C.5: List of studies used in the GWAS Catalog analysis
Table C.6: SLALOM prediction for the GWAS Catalog loci
Table C.7: Overview of the GBMI meta-analyses
Table C.8: SLALOM prediction for the GBMI loci

C.3 Supplementary Figures

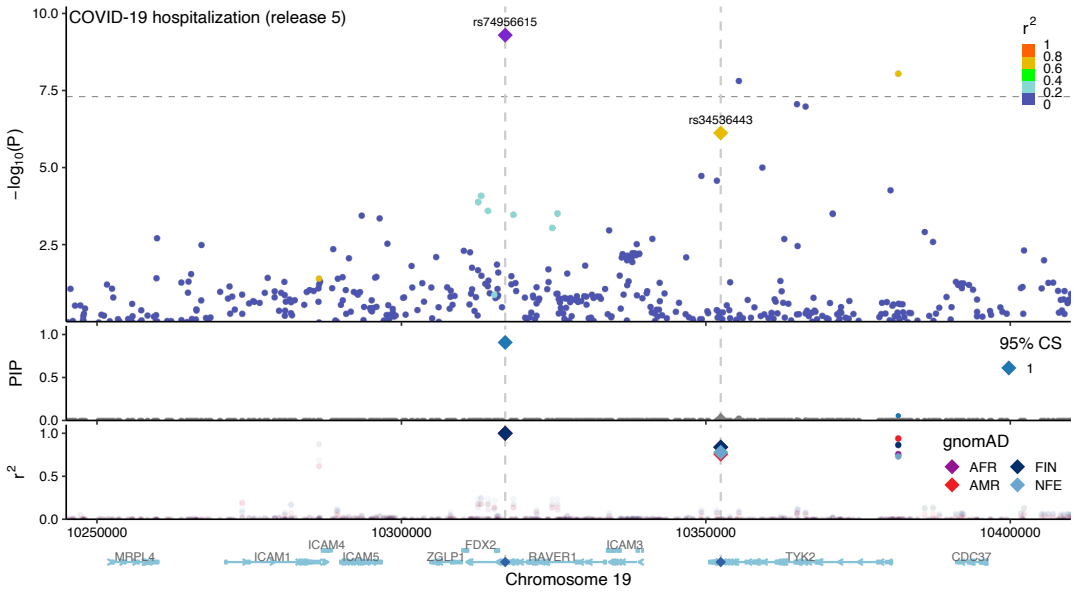


Figure C.1: Locuszoom plot of the *TYK2* locus (19p13.2) for COVID-19 hospitalization in the COVID-19 Host Genetics Initiative meta-analysis (release 5)[21]**.** The top panel shows a Manhattan plot, where the lead variant rs74956615 (purple diamond) and a missense variant rs34536443 (gold diamond) are highlighted. Color represents r2 values to the lead variant. Horizontal line represents a genome-wide significance threshold ($P = 5.0 \times 10^{-8}$). The middle panel shows PIP from ABF fine-mapping. Color represents whether variants belong to a 95% CS. The bottom panel shows r^2 values to the lead variant in gnomAD populations.

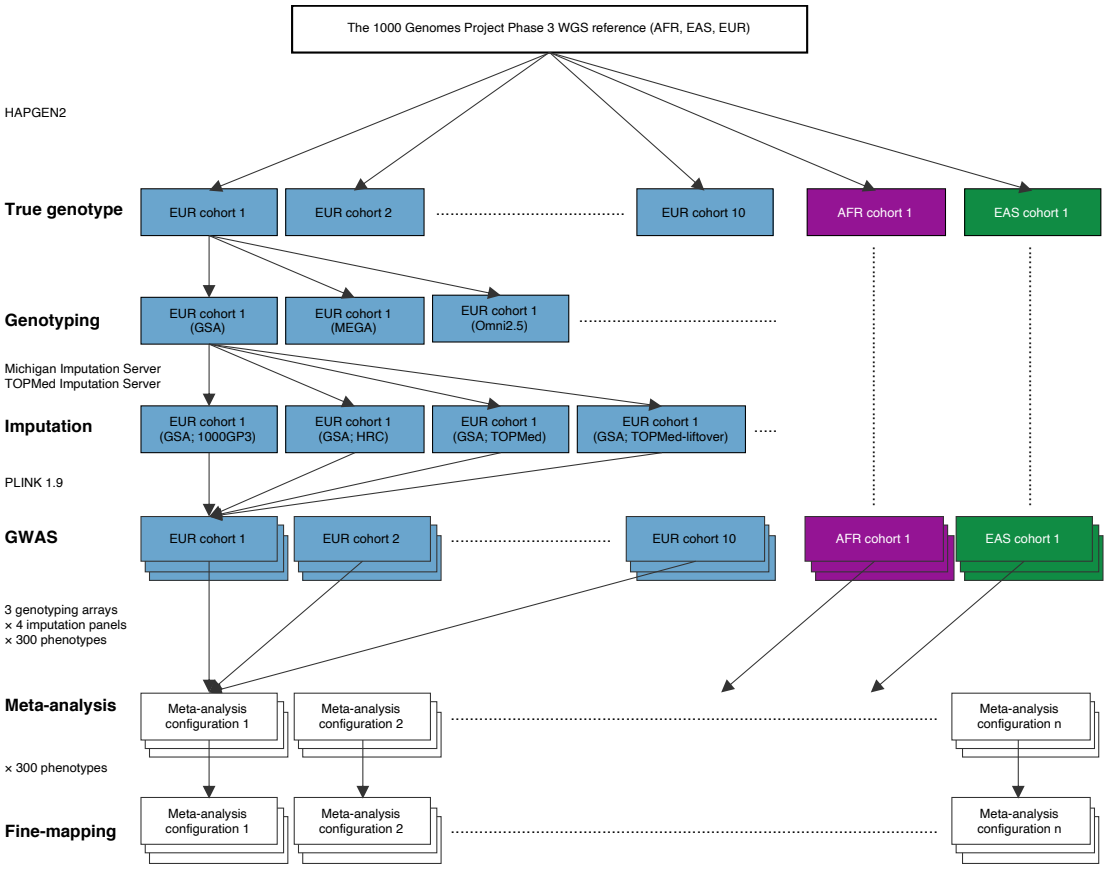


Figure C.2: Overview of our simulation pipeline. The flow of major simulation steps are summarized in the order of 1) simulating true genotypes, 2) genotyping, 3) imputation, 4) GWAS, 5) meta-analysis, and 6) fine-mapping (**3.4 Methods**). Arrows represent how each single simulated cohort (AFR, EAS, or EUR) is differentiated based on genotyping arrays and imputation panels and then combined for meta-analysis and fine-mapping.

Figure C.3 *(following page)*: UpSet plots of QC-passing GWAS variants across simulated GWAS cohorts under different MAF thresholds. Each simulated cohort has a different combination of ancestry, genotyping array, and imputation panel. For each MAF threshold, the top panel shows a number of QC-passing GWAS variants (Rsq > 0.6 and gnomAD or per-combination MAF) across different intersections of the simulated cohorts. Color represents the number of different imputation panels included in the intersection. Shape represents the number of genotyping arrays included in the interaction. The bottom panel shows a combination of the simulated cohorts for each intersection. Point color represents an imputation panel, while background color represents ancestry. Point shape represents a genotyping array. Dotted vertical lines split intersections by the number of ancestries. Horizontal lines split the simulated cohorts by a combination of ancestry and an imputation panel. Only the top 30 intersections for each MAF threshold are shown, ordered by the number of ancestries, and then the number of QC-passing variants for each intersection. For European simulated cohorts, we only used one cohort out of the 10 cohorts that we simulated.

Figure C.3: (continued)

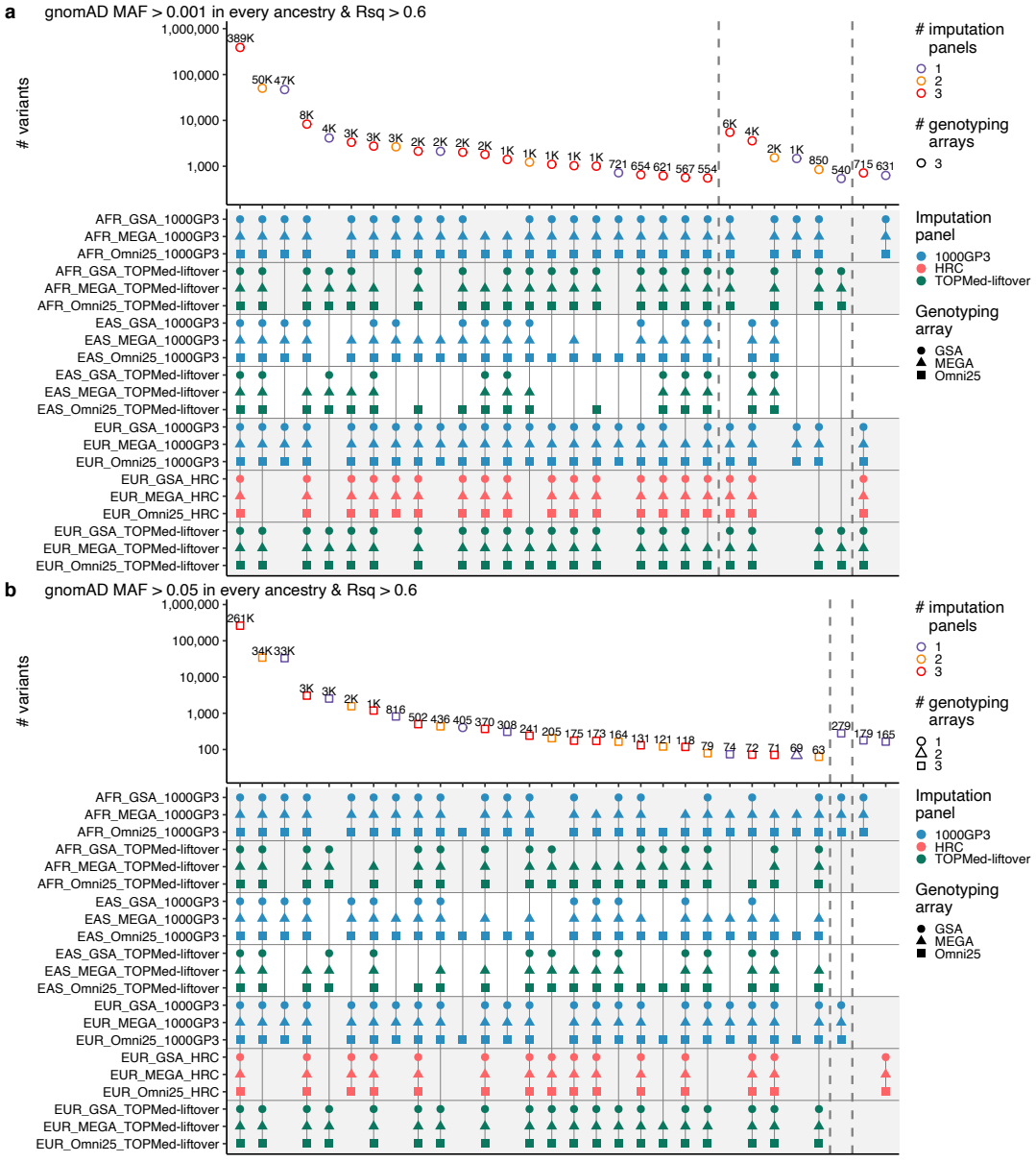
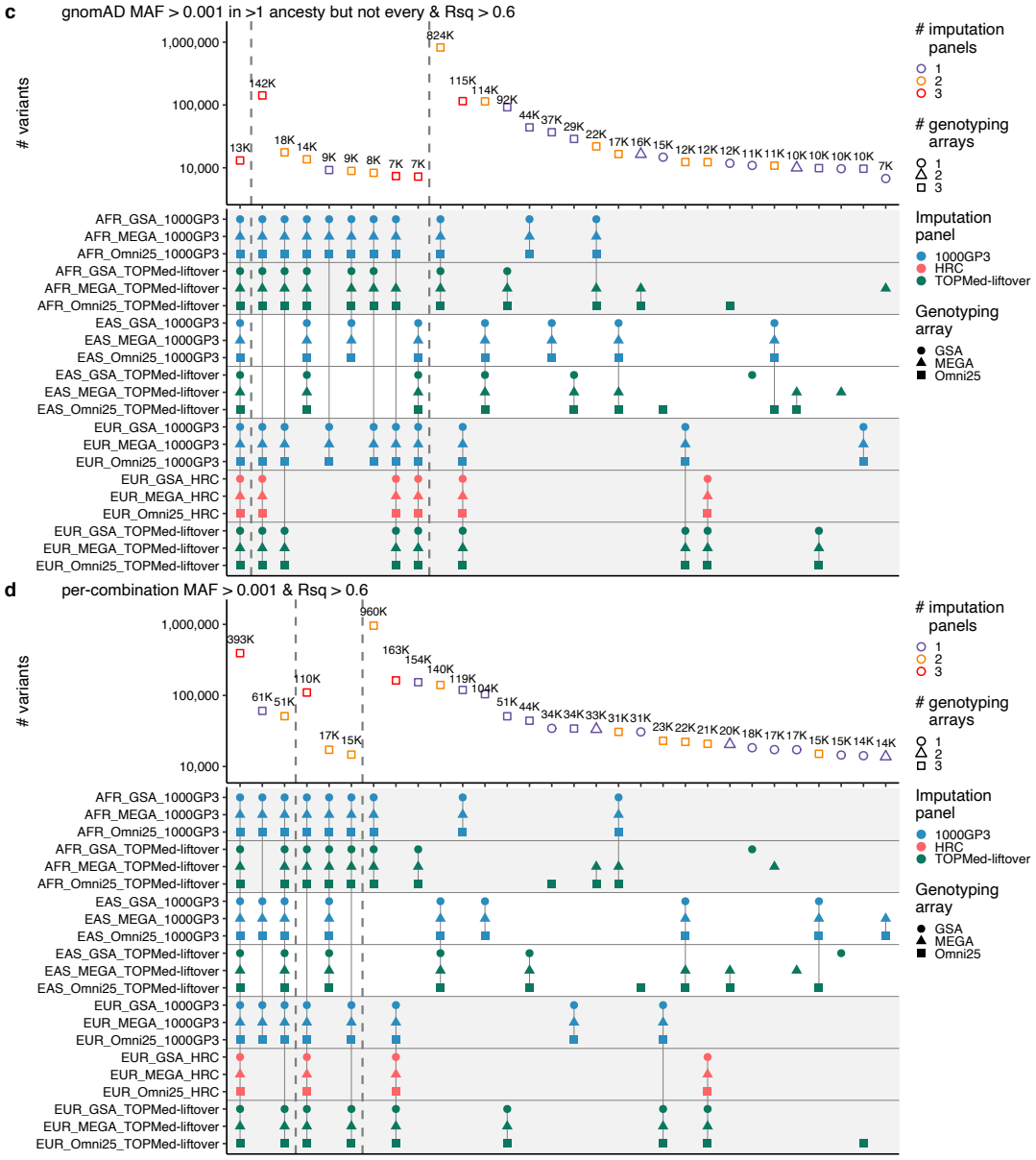


Figure C.3: (continued)



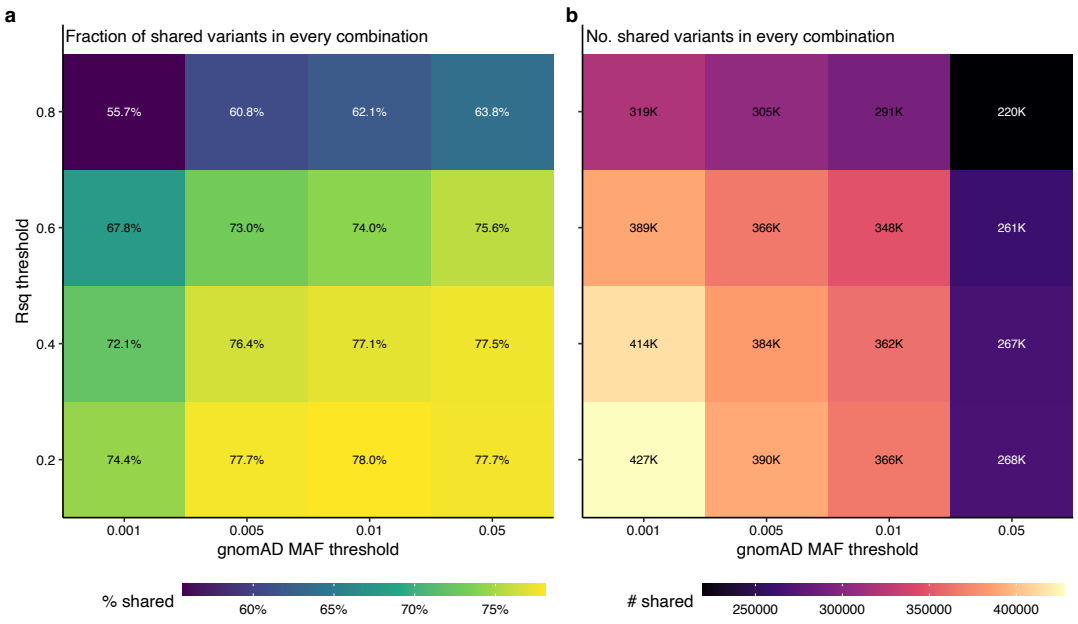


Figure C.4: QC-passing shared variants across simulated GWAS cohorts. Heatmaps under different combinations of gnomAD MAF and Rsq thresholds are shown for (**a**) the fraction of shared variants in every combination of ancestry, genotyping array, and imputation panel, and (**b**) the number of shared variants in every combination.

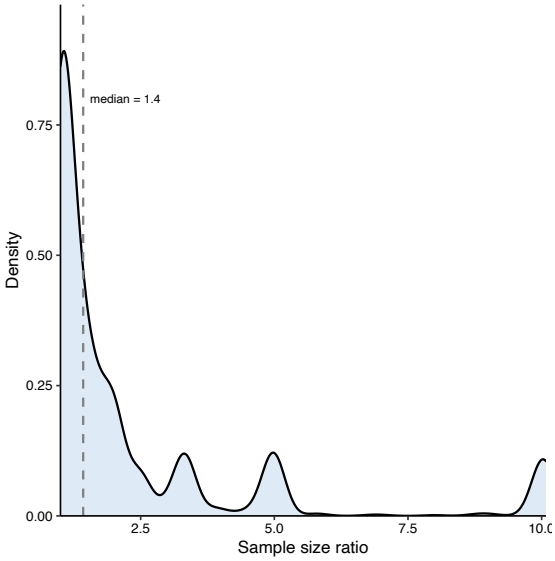


Figure C.5: Sample size ratio between true causal and false positive variants in simulations. Sample size ratio is defined as a ratio of maximum and minimum sample sizes for a true causal variant and a false positive variant (non-causal variant with PIP > 0.9) in the same locus. Dotted vertical line represents the median of 1.4. We used the simulation results for the most heterogeneous configurations (**3.4 Methods**).

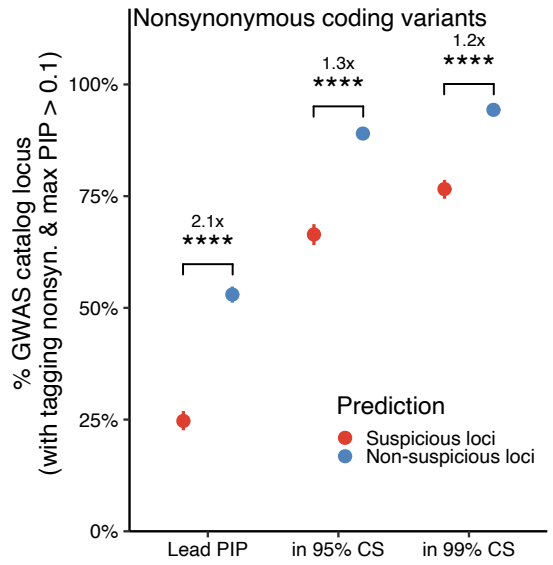


Figure C.6: Evaluation of SLALOM performance in the GWAS Catalog summary statistics using a more stringent r^2 threshold (> 0.8) for loci tagging nonsynonymous variants. Similar to **Fig. 3.4a**, we evaluated whether nonsynonymous coding variants (pLoF and missense) were lead PIP variants, in 95% CS, or in 99% CS in suspicious vs. non-suspicious loci. Depletion was calculated by relative risk (*i.e.* a ratio of proportions; Methods). Error bars correspond to 95% confidence intervals using bootstrapping. Significance represents a Fisher's exact test P-value (*, $P < 0.05$; **, < 0.01; ***, < 0.001; ***, $< 10^{-4}$).

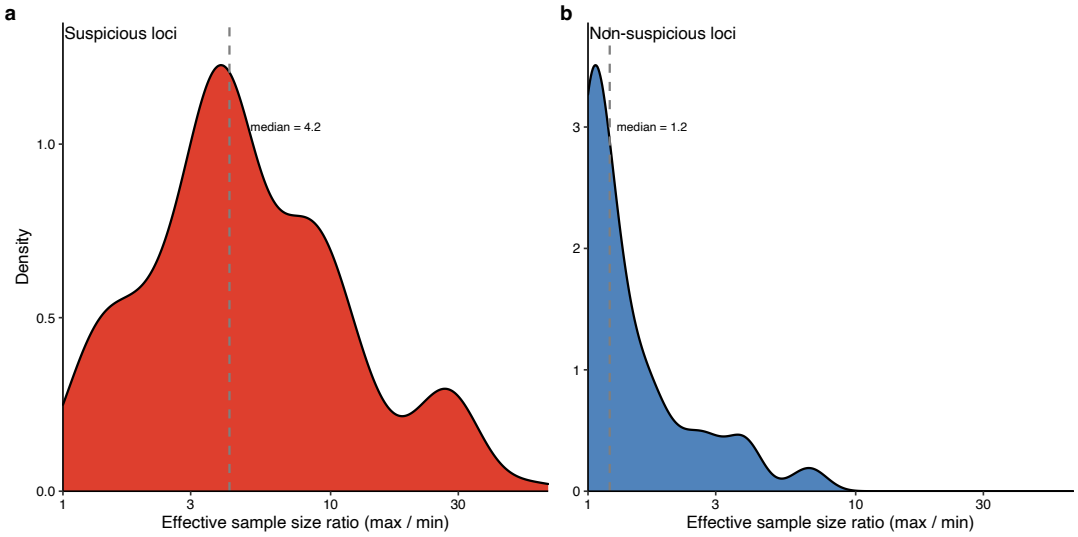


Figure C.7: Effective sample size ratio in the GBMI meta-analyses. Effective sample size ratio is defined as a ratio of maximum and minimum effective sample sizes among variants in LD (r^2 > 0.6) with a lead variant in a locus. Each panel represents (**a**) suspicious and (**b**) non-suspicious loc in the GBMI meta-analyses predicted by SLALOM. Dotted vertical lines represent the median values for each panel.

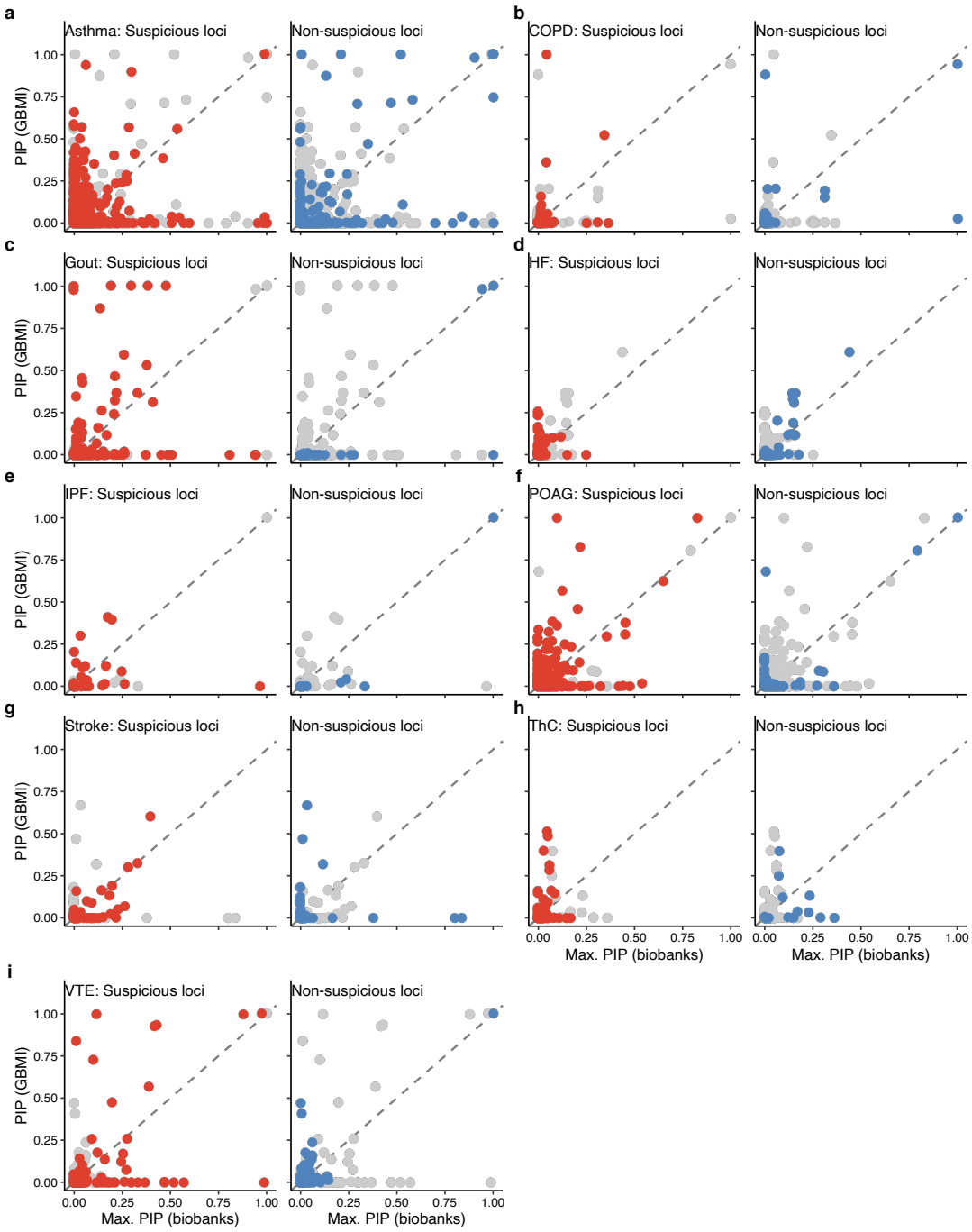


Figure C.8: Scatter plot of PIP in the GBMI and individual biobanks. For each variant, PIP in the GBMI vs. maximum PIP across BBJ, FinnGen, and UKBB are plotted, stratified by traits and suspiciou/non-suspiciou loci. Colored points represent variants in either suspicious or non-suspicious loci, while gray points represent variants in the other loci.

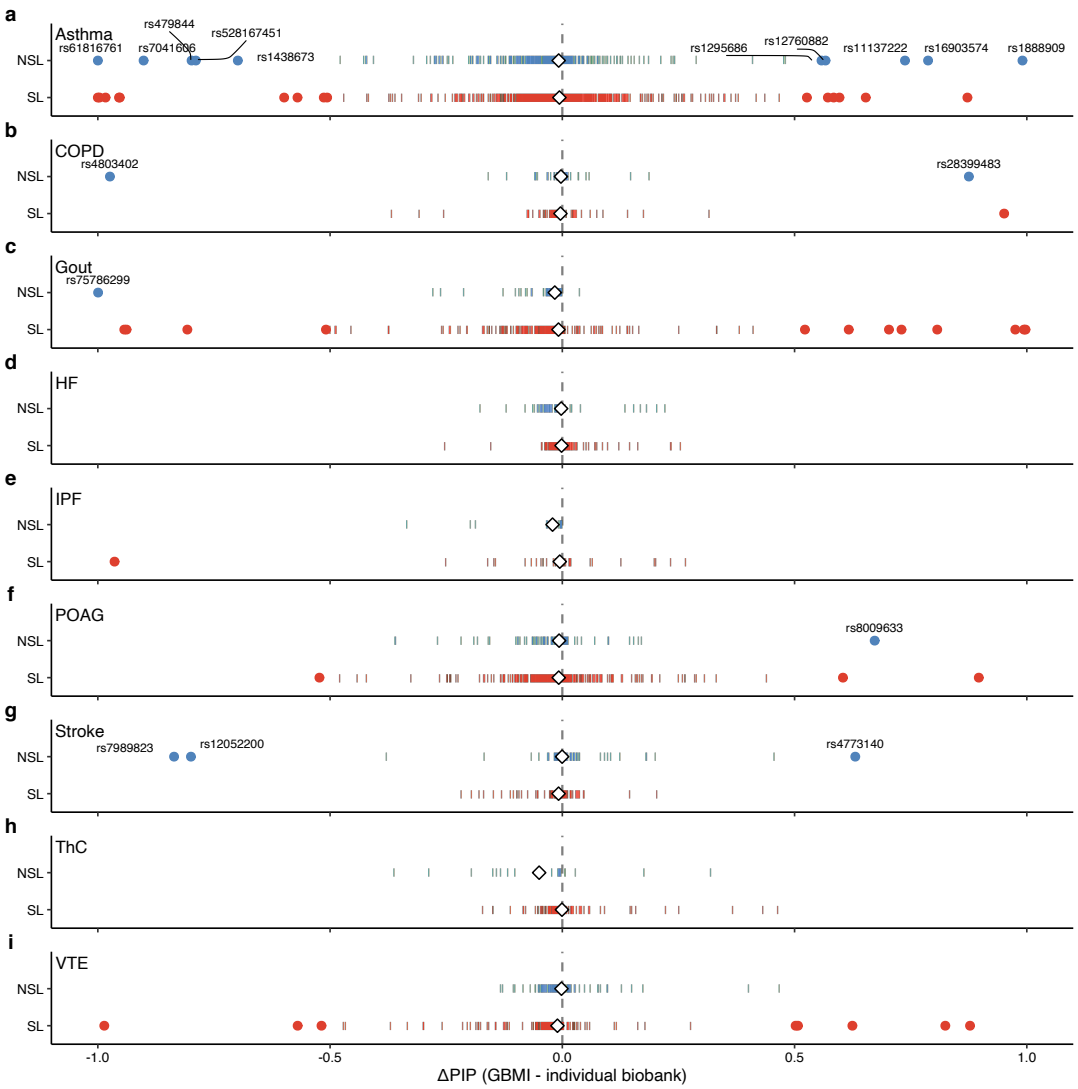


Figure C.9: Distribution of ΔPIP between the GBMI and individual biobanks. For each trait, distributions of ΔPIP in suspicious and non-suspicious loci are plotted. Circular points represent variants with |ΔPIP| > 0.5 while vertical bars represent those |ΔPIP| ≤ 0.5 in a similar fashion to a rug plot. Variants with |ΔPIP| > 0.5 in non-suspicious loci are labeled with rsids. White diamonds represent median values of ΔPIP. Dotted vertical line represents ΔPIP = 0.

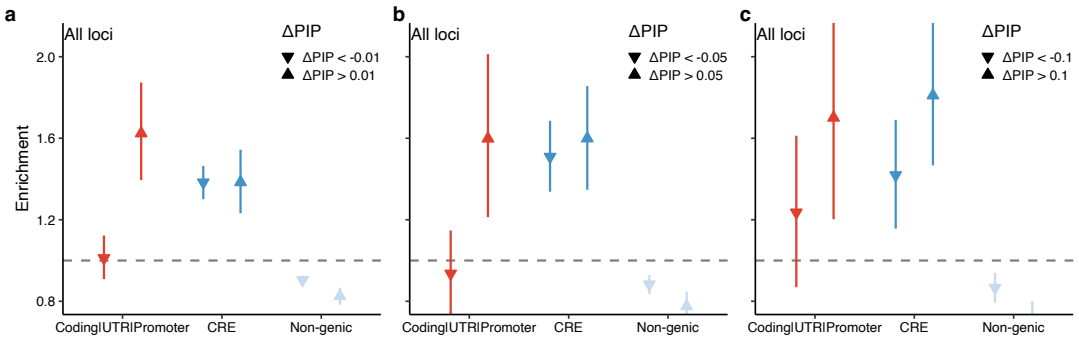


Figure C.10: Functional enrichment of variants with PIP difference using a different threshold. We computed functional enrichment of variants with PIP difference ($\Delta\text{PIP} > \theta$ or $< \theta$) in each functional category compared to variants with no substantial PIP difference ($\theta \leq \Delta\text{PIP} \leq \theta$) using thresholds of $\theta = 0.01$, 0.05, and 0.1. A regular triangle represents $\Delta\text{PIP} > 0.01$ while an upside-down triangle represents $\Delta\text{PIP} < -0.01$. Enrichment was calculated by a relative risk (*i.e*, a ratio of proportions; **3.4 Methods**). Error bars correspond to 95% confidence intervals using bootstrapping.

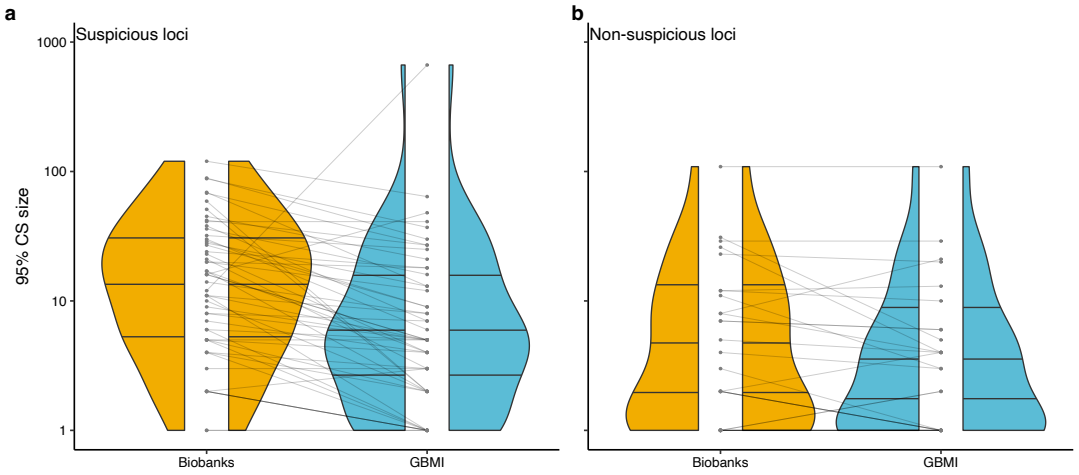


Figure C.11: Distribution of the 95% CS size in the GBMI and individual biobanks. Violin plots represent the distribution in (**a**) suspicious loci and (**b**) non-suspicious loci. Each point represents each 95% CS in the GBMI and individual biobanks. Each line connects the overlapping CS that contains the same lead variants from the GBMI.

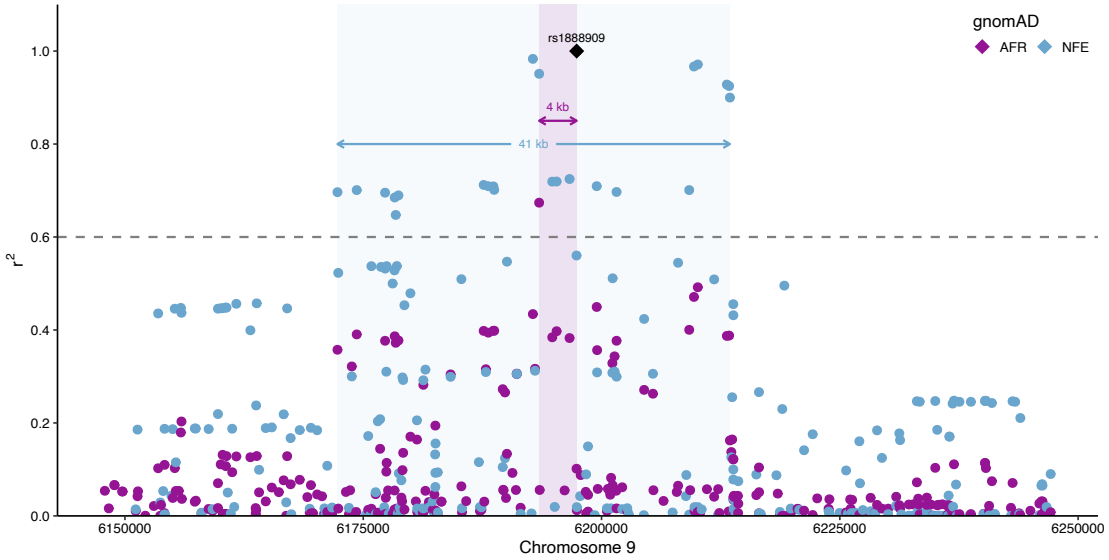


Figure C.12: LD structure around rs1888909 in the African and European populations. r^2 values with rs1888909 in the gnomAD African (AFR) and non-Finnish European (NFE) populations are plotted. Black diamond corresponds to rs1888909. Purple and light blue areas represent LD regions that showed $r^2 > 0.6$ with rs1888909 in AFR and NFE, respectively. Dotted horizontal corresponds to $r^2 = 0.6$.

D

Supplementary Materials for Chapter 4

D.1 Supplementary Note

D.1.1 Secondary analyses for simulations with in-sample LD

We performed 5 secondary analyses to investigate the sensitivity of the results to the simulation parameters. First, we performed simulations for much less polygenic (0.05%) and much more polygenic (0.5%) architectures. PolyPred remained the most accurate method, attaining the largest relative improvements vs. BOLT-LMM for the much less polygenic architecture, with slightly worse results for PolyPred-S and PolyPred-P (**Supplementary Table D.1**); we conservatively restricted the remaining secondary analyses to the more polygenic (0.3%) architecture (for which PolyPred attains smaller relative improvements among the two main architectures simulated) and omitted PolyPred-S and PolyPred-P (due to their close similarity to PolyPred), unless otherwise indicated. Second, we performed simulations with lower (3%) or higher (7%) chromosome 22 heritability. PolyPred remained the most accurate method, with relative improvements vs. BOLT-LMM increasing with heritability (**Supplementary Table D.1**). Third, we performed simulations with cross-population genetic correlations increased from 0.8 to 1.0. PolyPred remained the most accurate method, with relative improvements vs. BOLT-LMM remaining broadly similar (**Supplementary Table D.1**). Fourth, we modified the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to various values from 100–1,000. PolyPred remained the most accurate method in all these experiments, with relative improvements vs. BOLT-LMM increasing with N_{mix} but limited improvement above N_{mix} = 500 (**Supplementary Table D.1**). Fifth, we decreased the number of British-ancestry training samples (N) from N = 337K to N = 100K or N = 10K. Prediction accuracies decreased with decreasing training sample size for all methods, and the relative improvements of PolyPred vs. BOLT-LMM (and other methods) were substantially decreased for N = 10K, though they remained statistically significant in Africans under 0.1% polygenicity (**Supplementary Table D.1**).

We performed two secondary analyses to investigate the sensitivity of the results to the SNP set and functional annotations. First, we evaluated a modified version of PolyPred that uses only 1.2 million HapMap 3 SNPs (matching the SNP sets of BOLT-LMM, SBayesR, and PRS-CS) instead of 18 million SNPs. PolyPred suffered a substantial loss of accuracy in this setting, demonstrating the importance of using a dense SNP set for fine-mapping based PRS (**Supplementary Table D.1**). Second, we evaluated a non- functionally informed method (PolyPred-NoFun) that linearly combines PolyNoFun-pred (a modification of PolyFun-pred that is not functionally-informed; see Methods) and BOLT-LMM, precluding the need for functional annotations. PolyPred-NoFun was slightly less accurate than PolyPred, but still more accurate than BOLT-LMM (**Supplementary Table D.1**).

We performed two secondary analyses to evaluate the computational cost and memory cost of each method. First, we evaluated the computational cost of each method (for PolyPred, PolyPred-S, and PolyPred-P, we included the time cost of each constituent method); we focused on the time cost to compute SNP effect sizes used for prediction, as the time cost to compute predictions in target samples using these SNP effect sizes is approximately the same for each method. SBayesR was the fastest method (2.8 minutes), P+T was the second fastest method (7.4 minutes), PRS-CS was the third fastest method (113 minutes), BOLT-LMM was the fourth fastest method (224 minutes), PolyPred-S was the fifth fastest method (447 minutes), PolyPred-P was sixth fastest method (557 minutes), and PolyPred was the slowest method (668 minutes) (**Supplementary Table D.2**). Second, we evaluated the memory cost of each method (for PolyPred, we computed the maximum memory cost of each constituent method). We performed this analysis using chromosome 1 instead of chromosome 22 because memory cost can increase with the number of SNPs in the analysis (but the memory cost of PolyFun-pred is fixed because it analyzes each 3Mb-locus separately). P+T used the least memory (1.5GB), PRS-CS used the second smallest amount of memory (1.8GB), SBayesR used the third smallest amount of memory (2.6GB), BOLT-LMM used the fourth smallest amount

of memory (11GB), and PolyPred, PolyPred-S, and PolyPred-P all used the most memory (57GB) (**Supplementary Table D.2**). The larger computational cost of PolyPred and its summary statistic-based analogues is dominated by the PolyFun-pred component, which is computationally intensive because (i) it performs fine-mapping and (ii) it analyses a large number of SNPs (see the **D.1 Supplementary Note** subsection Limitations of PolyPred and PolyPred+).

D.1.2 Simulations with reference LD

The simulations described in the main text use in-sample LD (*i.e.*, LD summary data based on the UK Biobank GWAS sample). However, researchers often do not have access to in-sample LD, necessitating external LD reference panels. We thus evaluated modified versions of PolyFun-pred, SBayesR and PRS-CS that use summary LD estimated from 1000 Genomes project Europeans4 ($N = 489$). We note that this LD reference panel is both smaller than the UK Biobank British LD reference panel ($N = 337$K) and less well-matched to the GWAS sample, because it consists of pan-European ancestries rather than only British-ancestry individuals. We excluded BOLT-LMM from these analyses because it requires individual-level data.

The results of simulations with reference LD are reported in **Supplementary Table D.1**. All methods became less accurate when using 1000 Genomes project Europeans LD summary data. The loss of accuracy was modest for SBayesR (−5% R^2 for non-British Europeans vs. using in-sample LD) but severe for PRS-CS (−42% R^2 for non-British Europeans vs. using in-sample LD) and PolyFun-pred (-90% for non-British Europeans vs. using in-sample LD). We caution that the differences observed in real trait analysis for SBayesR and PRS-CS were substantially different from those observed in our simulations (large loss of accuracy for SBayesR, no significant loss of accuracy for PRS-CS), suggesting that the effect of LD mismatch on PRS accuracy may be sensitive to the underlying genetic architecture.

We performed 3 secondary analyses. First, we evaluated a modified version of PolyFun-pred that

uses summary LD from UK10K5 ($N = 3{,}567$). We observed only a moderate loss of accuracy in PolyFun-pred vs. using in-sample LD (–8% R^2 in non-British Europeans) (**Supplementary Table D.1**). However, we caution that using UK10K led to substantial and statistically significant loss of accuracy in real trait analysis, suggesting that the results may be sensitive to the underlying genetic architecture. Second, we evaluated modified versions of PolyFun-pred using subsets of UK10K as an LD reference panel, ranging from $N = 3{,}000$ to $N = 489$ (matching the 1000 Genomes project Europeans reference LD sample size). The accuracy of PolyFun-pred decreased with the LD reference panel sample size, with the loss in accuracy vs. using in-sample LD (for non-British Europeans) ranging from –8% for $N = 3{,}000$, to –90% for $N = 489$ (**Supplementary Table D.1**). Finally, we evaluated a modified version of PolyFun-pred (PolyFun-pred1) that assumes a single causal variant per locus, precluding the need for a reference LD panel (because fine-mapping under a single causal variant assumption does not require any LD information1). PolyFun-pred1 was substantially less accurate than all other methods (including P+T) and is thus not recommended for polygenic prediction (**Supplementary Table D.1**).

We conclude that the accuracy of all methods increases with the size of the LD reference panel and its concordance with the GWAS sample population, but that the relationship may depend on the underlying genetic architecture. Hence, it may be best to assess the accuracy obtained under various LD reference panels using real trait analysis rather than simulations. Specifically, the simulation results do not support the use of PolyPred-S or PolyPred-P in the specific scenarios considered in these simulations. However, real data results with very large LD reference panels do support the use of PolyPred-S or PolyPred-P (**Supplementary Fig. D.1**). We did not perform simulations with very large unmatched LD (analogous to **Supplementary Fig. D.1**), as this would have required another very large individual-level data set in addition to UK Biobank.

D.1.3 Evaluating method calibration for PRS in 4 UK Biobank populations using British training data

We assessed the calibration of each prediction method. A predictor is correctly calibrated if a regression of the true phenotype vs. the predictor yields a slope of 1, and is miscalibrated otherwise[35]. Regression slopes are reported in **Supplementary Table D.4**. In non-British Europeans, PolyPred was well-calibrated (regression slope = 1.01), BOLT-LMM and SBayesR were approximately well-calibrated (0.96–1.08), PRS-CS was slightly miscalibrated (1.26), and P+T was poorly calibrated (0.08). In non-European populations, PRS-CS was approximately well-calibrated (0.85–1.11), but BOLT-LMM and SBayesR suffered reduced regression slopes (0.57–0.90), consistent with reduced prediction accuracy. In contrast, PolyPred and its summary statistic-based analogues remained well-calibrated (0.95–1.17), as expected due to their extra training step to estimate mixing weights in the target population.

D.1.4 Secondary analyses for PRS in 4 UK Biobank populations using British training data

We performed 5 secondary analyses to evaluate the impact of the LD reference panel and the SNP set on prediction accuracy (we note that analyses of summary statistics from a meta-analysis of many cohorts generally require using an LD reference panel instead of in-sample LD). First, we evaluated a modified version of PolyFun-pred using a reference panel based on UK10K (N = 3,567) and observed a substantial and statistically significant reduction in accuracy, to a far greater degree that observed in simulations (**Supplementary Table D.4–D.6**). Second, we evaluated a modified version of PRS-CS that uses an LD reference panel from 1000 Genomes project Europeans (N = 489) and observed statistically indistinguishable results from those obtained using in-sample LD (unlike in simulations, where we observed significantly reduced accuracy when using an LD reference

panel from 1000 Genomes project Europeans) (**Supplementary Table D.4–D.6**). Third, we evaluated modified versions of SBayesR that use (i) an LD reference panel using UK10K ($N = 3{,}567$); (ii) an LD reference panel using 1000 Genomes project Europeans ($N = 489$); or (iii) an LD reference panel using a subset of UK10K ($N = 489$). We observed (i) very similar and statistically indistinguishable accuracy when using UK10K, (ii) severely reduced accuracy ($P < 4 \times 10^{-6}$) when using 1000 Genomes project Europeans, and (iii) moderately reduced accuracy ($P = 0.07$ in East-Asians, $P < 7 \times 10^{-6}$ in other target populations) when using a subset of UK10K, suggesting that the loss of accuracy primarily stems from LD mismatch rather than reduced sample size (**Supplementary Table D.4–D.6**). Fourth, we evaluated a modified version of SBayesR (SBayesR-2.8M) that uses 2.8M common SNPs specified by the authors of SBayesR[36] instead of 1.2 million HapMap 3 SNPs. SBayesR-2.8M was less accurate than SBayesR (significantly so for Africans) (**Supplementary Table D.4–D.6**). Thus, our use of SBayesR (using 1.2 million HapMap 3 SNPs) instead of SBayesR-2.8M in all primary comparisons is a conservative choice, since SBayesR outperforms SBayesR-2.8M (we note that naively scaling SBayesR and PRS-CS to use 18 million SNPs as in PolyFun-pred would be computationally infeasible[36,37]). Fifth, we evaluated a modified version of BOLT-LMM (BOLT-LMM-727K) that estimates effect sizes using only 727K genotyped SNPs (instead of 1.2 million imputed HapMap 3 SNPs). BOLT-LMM-727K was substantially and significantly less accurate than BOLT-LMM (**Supplementary Table D.4**).

We performed 9 additional secondary analyses. First, we evaluated LDpred6 using 1000 Genomes project Europeans4 or UK10K5 as the LD reference panel (**4.4 Methods**). Both versions of LDpred were consistently less accurate than BOLT-LMM (**Supplementary Table D.4**). Second, we evaluated modified versions of PolyPred that specify fixed mixing weights instead of estimating mixing weights in the target populations. We considered mixing weights for PolyFun-pred/BOLT-LMM equal to 0%/100%, 25%/75%, 50%/50%, 75%/25%, and 100%/0%. The 25%/75% and 50%/50% methods performed very similarly to PolyPred, with no statistically significant differences (**Supplementary**

Table D.6). Third, we restricted the PolyFun-pred component of PolyPred to only include SNPs with posterior causal probability greater than a fixed threshold (0.05, 0.50 or 0.95). This restriction decreased prediction accuracy (**Supplementary Tables D.4,D.6**), implying that estimating causal effect sizes is beneficial for prediction even at loci that cannot be confidently fine-mapped. Fourth, we evaluated a non-functionally informed method (PolyPred-NoFun) that linearly combines PolyNoFun-pred (a modification of PolyFun-pred that is not functionally-informed; see Methods) and BOLT-LMM. PolyPred-NoFun was slightly less accurate than PolyPred, but still more accurate than BOLT-LMM (**Supplementary Tables D.4,D.6**). The difference between PolyPred-NoFun vs. PolyPred was not statistically significant, in contrast to previous studies reporting a large and statistically significant increase in prediction accuracy from incorporating functional annotations [305–307]. Fifth, we reduced the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to 100. PolyPred suffered slightly reduced accuracy but remained the most accurate method, although relative improvements vs. BOLT-LMM were no longer statistically significant due to larger standard errors (**Supplementary Table D.4**). Sixth, we computed standard errors of relative-R^2 using a jackknife over individuals [305] (instead of a genomic block-jackknife over SNPs; see Methods). Standard errors computed using a jackknife over individuals were generally smaller, increasing the statistical significance of relative improvements of PolyPred vs. BOLT-LMM (**Supplementary Table D.4**). Seventh, we observed very similar results when down-sampling the non-British European target sample size to match the African target sample size, demonstrating that the reduced accuracy in Africans vs. Europeans is not due to the lower target sample size (**Supplementary Table D.4**). Eighth, we evaluated two versions of PRS-CS that use pre-specified values of its global shrinkage parameter (0.01 and 0.001, following the recommendations of the authors of PRS-CS [37]). Both versions were less accurate than the default version of PRS-CS (which automatically adjusts the value of this parameter), justifying the use of the default version of PRS-CS in this work (**Supplementary Tables D.4,D.5**). Finally, we assessed the potential

contribution of ancestry-specific heritability to reductions in cross-population prediction accuracy[48], by applying GCTA[323] to estimate the SNP-heritability explained by HapMap 3 SNPs[324,325] in each target population. SNP-heritabilities were largest in non-British Europeans and smallest in Africans (**Supplementary Table D.7**) (these differences could be due to SNP ascertainment[326], sample ascertainment, and/or ancestry-specific architectures), likely contributing to reductions in cross-population prediction accuracy.

D.1.5 Secondary analyses for PRS in Biobank Japan and Uganda-APCDR cohorts

We performed 6 secondary analyses. First, we assessed the calibration of each method by computing regression slopes, which are reported in **Supplementary Table D.9**. Similar to our analyses of non-European UK Biobank target populations, PolyPred and its summary statistic-based analogues were the only approximately well-calibrated methods, as expected due to their extra training step to estimate mixing weights in the target population. We restricted the remaining secondary analyses to PolyPred (as PolyPred-S and PolyPred-P are analogous to PolyPred with respect to these analyses). Second, we evaluated a modification of PolyPred that estimates mixing weights using 500 UK Biobank individuals from the genetically closest target population (UK Biobank East Asians for Biobank Japan, UK Biobank Africans for Uganda-APCDR) instead of 500 individuals from the target cohort. The differences between the original and modified versions of PolyPred were small and not statistically significant (**Supplementary Table D.9** indicating that PolyPred mixing weights can be estimated using 500 individuals from any cohort with the same continental ancestry as the target population. Third, we evaluated modified versions of PolyPred that specify fixed mixing weights instead of estimating mixing weights in the target populations. We considered mixing weights for PolyFun-pred/BOLT-LMM equal to 0%/100%, 25%/75%, 50%/50%, 75%/25%, and 100%/0%. The 25%/75% and 50%/50% methods performed very similarly to PolyPred, with no statistically signif-

icant differences (**Supplementary Table D.9**). Fourth, we reduced the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to 100. PolyPred suffered slightly reduced accuracy but remained the most accurate method, with the improvement vs. BOLT-LMM in Biobank Japan remaining statistically significant (**Supplementary Table D.9**). Fifth, we computed standard errors of relative-R^2 using a jackknife over individuals9 jackknife over SNPs). We obtained standard errors that were almost identical to those obtained using a (instead of a genomic block-genomic block-jackknife (unlike the above results for UK Biobank), suggesting that Biobank Japan may be more heterogeneous across samples, possibly due to its hospital-based recruitment (**Supplementary Table D.9**). Finally, we meta-analyzed the results of each method across three independent diseases in Biobank Japan: type 2 diabetes, asthma, and all autoimmune disease. Similar to our UK Biobank analyses above, PolyPred attained the highest prediction accuracy in each disease, though relative improvements were not statistically significant due to lower power (**Supplementary Table D.9**).

D.1.6 Secondary analyses for PRS in East Asians using British and Japanese training data

We performed 6 secondary analyses. We restricted these secondary analyses to PolyPred+ (as PolyPred-S+ and PolyPred-P+ are analogous to PolyPred+ with respect to these analyses). First, we verified that PolyPred+ using European and East Asian training data does not outperform PolyPred in UK Biobank populations other than East Asians; differences between PolyPred+ and PolyPred were very small and not statistically significant (**Supplementary Table D.6**). Second, we verified that PolyPred+ was well-calibrated (**Supplementary Table D.4**; results for other methods are described above), as expected due to its extra training step to estimate mixing weights in the target population. Third, we evaluated a modified version of PolyPred+ that estimates mixing weights using 500 Biobank Japan individuals instead of 500 UK Biobank East Asians. The modified ver-

sion of PolyPred+ was far less accurate than the original version (52% lower relative-R^2; **Supplementary Table D.6**). The mixing weights estimated in Biobank Japan assign much higher weight to the Biobank Japan training data (**Supplementary Table D.6**), perhaps due to cohort effects; thus, it may be important to estimate PolyPred+ mixing weights using the target cohort (as opposed to the training cohort) if cohort effects are present. Fourth, we reduced the number of training samples from the target population used to estimate mixing weights (N_{mix}) from 500 to 100. PolyPred+ suffered slightly reduced accuracy, though the difference was not statistically significant (**Supplementary Table D.6**). Fifth, we evaluated a prediction method using only the $N = 124$K Biobank Japan individuals to train effect sizes (BOLT-LMM-BBJ). BOLT-LMM-BBJ substantially underperformed methods that use UK Biobank British training data (−27% vs. BOLT-LMM, −34% vs. PolyPred, −41% vs. PolyPred+; **Supplementary Table D.4**). Finally, we computed standard errors of relative-R^2 using a jackknife over individuals [305] (instead of a genomic block-jackknife over SNPs). Standard errors computed using a jackknife over individuals were smaller, increasing the statistical significance of relative improvements of PolyPred+ vs. other methods (**Supplementary Table D.6**).

D.1.7 Loss of PRS accuracy under an infinite European training sample

Under an infinite European training sample, the ratio between R^2_{EUR} and $R^2_{\text{non-EUR}}$, which denote R^2 in a European sample and in a non-European sample, respectively, is approximately given by:

$$\rho_g^2 \times \frac{h^2_{\text{non-EUR}}}{h^2_{\text{EUR}}} \times \left(\sum_k \sqrt{\frac{p_{k,\text{non-EUR}}(1-p_{k,\text{non-EUR}})}{p_{k,\text{non-EUR}}(1-p_{k,\text{non-EUR}})}} \right)^2 \times \frac{\text{var}(\text{PGS}_{\text{EUR}})}{\text{var}(\text{PGS}_{\text{non-EUR}})}$$

Here, ρ_g is the cross-population genetic correlation, $h^2_{\text{non-EUR}}$, h^2_{EUR} are the heritabilities in the non-European and the European populations, respectively, k iterates over causal SNPs, $p_{k,\text{non-EUR}}$, $p_{k,\text{EUR}}$ are minor allele frequencies in the non-European and the European population, respec-

tively, and var(PGS_{EUR}), var($PGS_{non\text{-}EUR}$) are the variances of the polygenic risk scores in the non-European and the European populations, respectively. This equation is directly derived from Equation 1 in ref. 12, after assuming that causal SNPs are approximately not in LD with each other, and that the predictor SNPs are the causal SNPs under an infinite sample size.

D.1.8 Limitations of PolyPred and PolyPred+

Polypred training time is slower than alternative PRS methods

PolyPred and its summary statistic-based analogues are slower than alternative PRS methods, requiring over 1,000 hours of computation time for training, vs. less than 100 hours for BOLT-LMM (**D.1 Supplementary Note**). This is dominated by the PolyFun-pred component, which is computationally intensive because (i) PolyFun-pred performs fine-mapping, which is a more computationally intensive task than other approaches to computing PRS coefficients (*e.g.* computing posterior mean tagging effect sizes, as in SBayesR); and (ii) PolyFun-pred analyzes a large number of SNPs, *e.g.* 18 million SNPs in UK Biobank training data and 8.1 million SNPs in ENGAGE training data (vs. 1.2 million SNPs for SBayesR). We do not foresee the larger computation time for training as a major limitation in real-world settings, because training only needs to be performed once, can be parallelized, and provides genome-wide fine-mapping results of direct interest1.

PolyPred cannot use data from a fixed-effects meta-analysis of GWAS data from different ancestry groups

One of the main conclusions of our work is that leveraging training data from different ancestry groups (*e.g.* different continental ancestries) improves PRS in diverse populations. However, we recommend against using training data consisting of a traditional fixed-effect meta-analysis of GWAS data from different ancestry groups, for two reasons: (i) fixed-effect meta-analysis implies that Eu-

ropean training samples and training samples from the non-European target population would receive equal weight, whereas our work shows that the latter should receive higher weight in order to maximize PRS accuracy; and (ii) it may be challenging to construct an LD reference panel whose ancestry matches the ancestry of the meta-analysis of different ancestry groups. When possible, it would be preferable to separately incorporate European training data and training data from the non-European target population, with appropriate LD reference panels. Although there is no single optimal way to choose a training cohort, training sample size should be a primary consideration, as it is a critical factor impacting PRS accuracy.

PolyPred requires a small training sample from the target cohort to maintain calibrated predictions

PolyPred ideally requires a small training sample from the target cohort to estimate mixing weights. Our results suggest that it is possible to improve cross-population PRS accuracy even without such a training sample, by linearly combining PolyFun-pred and BOLT using mixing weights of either 25%/75% or 50%/50%, respectively. However, we caution that PRS linearly combined using fixed mixing weights may not always be well-calibrated.

D.1.9 Causal vs. tagging effects

We consider a linear model $y = \sum_i x_i \beta_i + \varepsilon$ where y is a trait, x_i is the number of minor alleles at SNP i, β_i is the (true) causal effect sizes of SNP i, and ε is a residual term sampled from a normal distribution. We consider a method (such as PolyFun-pred) that estimates β_i. If the generative model holds and all SNPs i are considered in the estimation procedure, then any consistent estimator $\hat{\beta}_i$ of β_i represents a causal effect. In contrast, if only a subset of the SNPs, such as HapMap3 SNPs, are considered in the estimation procedure (*i.e.* if we incorrectly assume the generative model

$y = \sum_{i \in S} x_i \beta_i + \varepsilon$, where S is a subset of SNPs) then the estimated value $\hat{\beta}_i$, represents a linear combination of β_i and of the effect sizes of other SNPs.

The exact value estimated by $\hat{\beta}_i$ depends on the estimation procedure. For example, assuming an ordinary least squares estimator for simplicity, the vector $\hat{\beta}_S$ of estimated coefficients is a consistent estimator of $[I_{m-k} R_{SS}^{-1} R_{S\bar{S}}]\beta$, where m is the total number of SNPs, k is the number of SNPs in the set S, R_{SS} is the LD matrix of the SNPs in the set S, $R_{S\bar{S}}$ is a matrix wherein each entry i,j is the correlation between SNP i in the set S and SNP j in the set of SNPs that are not in S, and β is the vector of true effect sizes, assuming without loss of generality that the set S includes the first k SNPs (out on m SNPs considered). It is easy to derive this quantity by writing down the conditional expectation of $\hat{\beta}_S$ under an ordinary least squares estimator, given by $\mathrm{E}[\hat{\beta}_S \mid \beta] = \mathrm{E}[(\boldsymbol{X}_S^T \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^T \boldsymbol{y} \mid \beta]$, where $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$ is a vector of observed phenotypes and $\boldsymbol{X}$ is the corresponding matrix of SNPs, $\boldsymbol{X}_S$ is the submatrix of $\boldsymbol{X}$ consisting of columns of SNPs in the set S, and we assume that ε is independent of $\boldsymbol{X}$.

D.1.10 Investigating if off-cohort loss of accuracy is driven by SNP heritability differences

We investigated if lower prediction accuracies in Biobank Japan vs. the UK Biobank can be largely explained by SNP heritability differences. We began by comparing trait heritabilities across the UK Biobank and Biobank Japan, using BOLT-REML16 applied to UK Biobank British-ancestry individuals (average $N = 325$K) and to Biobank Japan (average $N = 124$K), restricting to HapMap 3 SNPs. The average heritability in the UK Biobank was 67% larger (**Supplementary Table D.10**), indicating differences in either trait measurement, cohort ascertainment, the ability of HapMap 3 SNPs to tag East Asian causal SNPs[64], or in the true underlying heritabilities (we could not perform a similar analysis with UK Biobank East Asian individuals due to small sample sizes leading to large standard errors). We next asked if the observed differences in PRS accuracy between Biobank

Japan and the UK Biobank can be explained by the 67% increased average SNP heritability in the UK Biobank. To this end, we computed the expected R^2 within each cohort as function of SNP heritability, sample size, and the effective number of independent SNPs[327,328]:

$$\mathrm{E}[R^2] = h^2 \frac{h^2}{h^2 + \frac{m}{n}}$$

Here, h^2 is SNP heritability, n is sample size, and m is the effective number of independent SNPs (which we specified as 55,000, determined by dividing the number of HapMap 3 SNPs by their average within-HapMap 3 LD-score). We used the smaller Biobank Japan sample size in both cohorts to eliminate differences due to sample size differences (by choosing a random subset of UK Biobank British individuals as a training set). The average expected R^2 in the UK Biobank was 104% larger than in Biobank Japan (**Supplementary Table D.10**). We then trained BOLT-LMM using subsets of the UK Biobank British sample (matching the Biobank Japan sample size for each trait) and applied the predictions to UK Biobank non-British Europeans. The average R^2 in UK Biobank non-British Europeans (when training BOLT-LMM using the reduced British training sample) was 108% larger than the average R^2 in Biobank Japan (when training BOLT-LMM using the Biobank Japan training sample) (**Supplementary Table D.10**), strongly consistent with the 104% increase expected from theory. Finally, we determined that when training BOLT-LMM using the full UK Biobank British training set (average N = 325K), the average R^2 in UK Biobank East Asians across the 7 independent traits is 93% larger than in Biobank Japan (**Supplementary Tables D.4,D.9**), broadly consistent with the previous results. Assuming that the main factor differentiating the UK Biobank East Asian sample from the Biobank Japan sample is SNP heritability differences (rather than differences in MAF, LD, or causal effect sizes), these findings suggest that the main factor leading to lower prediction accuracies in Biobank Japan vs. the UK Biobank is SNP heritability differences.

To further investigate if off-cohort loss of accuracy is driven by SNP heritability differences, we

compared prediction accuracies in UK Biobank East Asians and in Biobank Japan, when training BOLT-LMM using the Biobank Japan training sample. The average relative-R^2 in UK Biobank East Asians across the 7 independent traits was 9.0% larger (**Supplementary Tables D.4,D.10**), though the difference was not statistically significant (P =0.18), possibly owing to the small UK Biobank East Asian sample size.

Although these results are not conclusive, they suggest that heritability differences drive most of the differences in prediction accuracies observed between the UK Biobank and Biobank Japan. Surprisingly, these results are consistent with a model in which HapMap 3 SNPs in Biobank Japan tag approximately 50% of the causal SNPs that they tag in the UK Biobank, rather than a model in which SNP heritabilities in Biobank Japan are smaller due to smaller causal effect sizes. This is because under the second model, we would expect to see large increase in prediction accuracy in UK Biobank East Asians vs. Biobank Japan when training BOLT-LMM using Biobank Japan (compared with only a 9.0% increase observed in practice). A partial explanation is that the HapMap 3 SNP set consists of a combination of two genotyping chips, one of which is explicitly designed to optimize tagging in Europeans[288].

Overall, these results suggest that differences in SNP-heritability due to ancestry differences (*e.g.* SNP ascertainment[326], sample ascertainment, and/or ancestry-specific architectures[64]) or due to cohort differences (*e.g.* differences in phenotype definitions[47], different recruiting strategies[47], or assay artifacts) may explain most of the differences in prediction accuracies observed between the UK Biobank and Biobank Japan. Our results are consistent with recent results showing almost no loss of accuracy when applying PRS based on UK Biobank training data to other European-ancestry cohorts[36]. Importantly, our results suggest that factors that inflate within-cohort PRS accuracy[329] (such as cohort-specific GxE, cohort-specific indirect effects[330], cohort-specific population structure, or cohort-specific assortative mating) are unlikely explanations for the observed accuracy differences between the UK Biobank and Biobank Japan.

D.1.11 Decomposing the PolyFun-pred and BOLT-LMM predictors into shared and non-shared components

A linear combination of PRS predictors is not necessarily suboptimal, even if the methods are correlated. (As an extreme example, a linear combination of two perfectly correlated predictors is optimal.) However, a linear combination could be suboptimal if the correlation between the (effect sizes underlying the) two predictors varies across the genome. As an extreme example, consider a scenario where one predictor is perfectly accurate across the first half of a chromosome but uninformative across the second half, whereas the second predictor is uninformative across the first half but perfectly accurate across the second half. Clearly, the optimal combination would use only the (effect sizes of the) first predictor for the first half of the chromosome, and only the (effect sizes of the) second predictor for the second half of the chromosome. However, a simple linear combination assigns only a single mixing weight to each predictor, and will thus assign equal weights to both predictors, resulting in a suboptimal predictor.

We performed several attempts to improve upon a simple linear combination of PRS predictors by partitioning the genome into segments and estimating different linear mixing weights in different segments. However, this more complex approach did not outperform the simple approach of assigning a simple mixing weight to each predictor (results not shown), and we thus did not pursue it further.

D.1.12 Generating data for UK Biobank simulations

To simulate data, we first computed the variance of per-standardized-genotype effect η_i, for every SNP i with annotations $\boldsymbol{a}_i$ using the baseline-LF (version 2.2.UKB) model, $\mathrm{var}[\eta_i \mid \boldsymbol{a}_i] = \sum_c \tau^c a_i^c$, where c are annotations and τ^c estimates are taken from a fixed-effects meta-analysis of 16 well-powered genetically uncorrelated ($|r_g| < 0.2$) UK Biobank traits (age of menarche, BMI, balding,

bone mineral density, eosinophil count, FEV1/FVC ratio, forced vital capacity, hair color, height, platelet count, red blood cell distribution width, red blood cell count, systolic blood pressure, tanning, waist-hip ratio adjusted for BMI, white blood count), scaled such that $\sum_i \text{var}[\eta_i \mid \boldsymbol{a}_i]$ is the same across all traits (as detailed in ref. [118]). Each SNP was specified to be causal with probability proportional to $\text{var}[\eta_i \mid \boldsymbol{a}_i]$, such that the average causal probability was equal to the desired proportion of causal SNPs (0.1% or 0.3% in most simulations).

We generated ancestry-specific effect sizes as follows. First, we generated a British per-allele causal effect size for each SNP i via $\beta_i^{\text{British}} = \gamma_i / \sqrt{2f_i(1-f_i)}$, where $\gamma_i \sim N(0, h^2/m)$, m is the number of causal SNPs, and f_i is the maximal MAF of SNP i among British, non-British European, South Asian, East Asian, or African UK Biobank individuals. Afterwards, for each of the main UK Biobank non-European ancestries (South Asian, East Asian, and African) a we generated an ancestry-specific per-allele effect size β_i^a via $\beta_i^a = r_g \cdot \beta_i^{\text{British}} + \sqrt{1-r_g^2} z_i^a$, where r_g is the cross-population genetic correlation (set to 0.8 by default, following previous works [64,244,303]), and $z_i^a \sim N(0,1)$. The use of f_i bounds the per-allele causal effect sizes by the MAF of the ancestry in which the SNP is most common, which guarantees that SNPs that are infrequent in Europeans but are common in other ancestries do not explain a very large proportion of heritability. After generating ancestry-specific per-allele causal effect sizes, we generated a phenotype y for every UK Biobank individual in each ancestry a via $y = \sum_i x_i \beta_i^a + \varepsilon$, where x_i is the number of minor alleles of SNP i carried by that individuals, β_i^a is the ancestry-specific per-allele causal effect size of SNP i, and $\varepsilon \sim N(0, 1-h^2)$ is the environmental variance of the generated trait. We generated phenotypes based on dosage data from imputed genotypes, using Plink 2.0 (ref. [278]. We used self-reported ancestry based on UK Biobank data field 21000 (Ethnic background). We considered Irish-ancestry as a non-British European ancestry.

D.2 Supplementary Tables

The following Supplementary Tables are available in the online version of the manuscript.

Table D.1: Detailed simulation results

Table D.2: Detailed simulation runtime analysis

Table D.3: List of 49 diseases and complex traits

Table D.4: Detailed results of analyses using UKB British training individuals applied to other UKB populations, compared vs. BOLT-LMM

Table D.5: Comparisons between pairs of methods in analyses of real UK Biobank and Biobank Japan traits

Table D.6: Detailed results of analyses using UKB British training individuals applied to other UKB populations, compared vs. PolyPred

Table D.7: Ancestry-specific SNP heritability estimates in the UK Biobank, across 7 independent complex traits

Table D.8: Prediction accuracy using summary statistics from the from the European Network for Genetic and Genomic Epidemiology

Table D.9: Detailed results of analyses applied to Biobank Japan and to Uganda-APCDR

Table D.10: Comparing prediction accuracy in UK Biobank Non-British Europeans and in Biobank Japan when using equal training set sample sizes

Table D.11: Description of 187 baseline-LF model annotations used by PolyFun-pred

D.3 Supplementary Figures

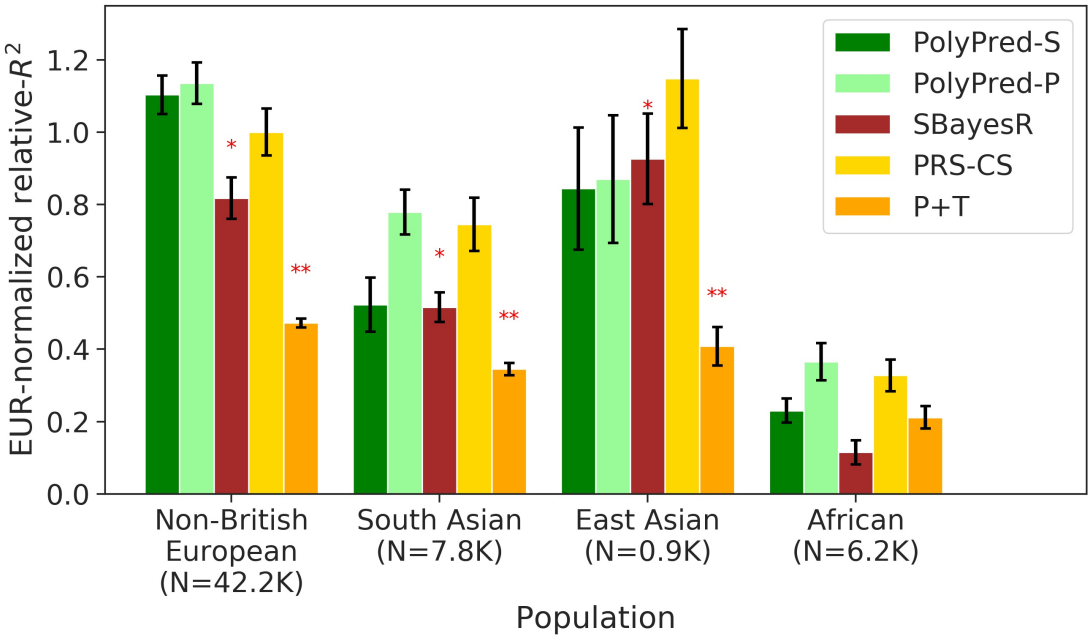


Figure D.1: Cross-population PRS results for real UK Biobank traits, using summary statistics from a meta-analysis of many cohorts. We report average prediction accuracy (relative-R^2, but computed with respect to PRS-CS instead of BOLT-LMM; see main text), meta-analyzed across 4 well-powered, approximately independent traits, for PRS trained in European Network for Genetic and Genomic Epidemiology (ENGAGE) samples (average N = 61,365) and applied to four UK Biobank populations. Target population sample sizes are indicated in parentheses; PolyPred and its summary statistic-based analogues used 500 additional training samples from each target population to estimate mixing weights. Asterisks above each bar denote statistical significance of the difference vs. PRS-CS, with red asterisks denoting a disadvantage ($^*P < 0.05$; $^{**}P < 0.001$). P-values were computed using a two-sided Wald test and were not adjusted for multiple comparisons. Errors bars denote standard errors. Numerical results, results for all 4 traits analyzed, absolute prediction accuracies (R^2), and P-values of relative improvements vs. PRS-CS are reported in **Supplementary Tables D.5,D.8**.

References

[1] Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era–concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

[2] Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2018).

[3] O'Connor, L. J. The distribution of common-variant effect sizes. *Nat. Genet.* **53**, 1243–1249 (2021).

[4] Chong, J. X. *et al.* The genetic basis of mendelian phenotypes: Discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).

[5] MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. the huntington's disease collaborative research group. *Cell* **72**, 971–983 (1993).

[6] Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).

[7] Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

[8] Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 east asian individuals. *Nature* **582**, 240–245 (2020).

[9] Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

[10] Ishigaki, K. *et al.* Trans-ancestry genome-wide association study identifies novel genetic mechanisms in rheumatoid arthritis. *medRxiv* (2021).

[11] Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

[12] Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

[13] Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

[14] Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine-Progress, pitfalls, and promise. *Cell* **177**, 45–57 (2019).

[15] Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).

[16] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).

[17] Kanai, M., Tanaka, T. & Okada, Y. Empirical estimation of genome-wide significance thresholds based on the 1000 genomes project data set. *J. Hum. Genet.* **61**, 861–866 (2016).

[18] Buniello, A. *et al.* The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

[19] Claussnitzer, M. *et al.* FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).

[20] Grant, S. F. A. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).

[21] COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).

[22] Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 229–232 (2001).

[23] Wakefield, J. A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).

[24] Wakefield, J. Bayes factors for genome-wide association studies: comparison with p-values. *Genet. Epidemiol.* **33**, 79–86 (2009).

[25] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).

[26] Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).

[27] Kichaev, G. & Pasaniuc, B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).

[28] Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

[29] Benner, C., Havulinna, A. S., Salomaa, V., Ripatti, S. & Pirinen, M. Refining fine-mapping: effect sizes and regional heritability. *bioRxiv* 318618 (2018).

[30] Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

[31] Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

[32] Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

[33] Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat. Med.* **27**, 1876–1884 (2021).

[34] International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

[35] Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

[36] Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

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

[38] Maas, P. *et al.* Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the united states. *JAMA Oncol* **2**, 1295–1302 (2016).

[39] Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

[40] Sharp, S. A. *et al.* Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care* **42**, 200–207 (2019).

[41] Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

[42] Grinde, K. E. *et al.* Generalizing polygenic risk scores from europeans to Hispanics/Latinos. *Genet. Epidemiol.* **43**, 50–62 (2019).

[43] Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).

[44] Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).

[45] Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).

[46] Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).

[47] Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

[48] Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).

[49] Amariuta, T. *et al.* Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat. Genet.* **52**, 1346–1354 (2020).

[50] Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.* **11**, 1628 (2020).

[51] Bitarello, B. D. & Mathieson, I. Polygenic scores for height in admixed populations. *G3* **10**, 4027–4036 (2020).

[52] Chen, M.-H. *et al.* Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198–1213.e14 (2020).

[53] Mahajan, A. *et al.* Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *medRxiv* (2020).

[54] Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *HGG Adv* **2** (2021).

[55] Mills, M. C. & Rahal, C. The GWAS diversity monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).

[56] Lehmann, B. C. L., Mackintosh, M., McVean, G. & Holmes, C. C. High trait variability in optimal polygenic prediction strategy within multiple-ancestry cohorts. *bioRxiv* (2021).

[57] Ji, Y. *et al.* Incorporating european GWAS findings improve polygenic risk prediction accuracy of breast cancer among east asians. *Genet. Epidemiol.* **45**, 471–484 (2021).

[58] Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* (2022).

[59] Cai, M. *et al.* A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am. J. Hum. Genet.* **108**, 632–655 (2021).

[60] Huang, Q. Q. *et al.* Transferability of genetic loci and polygenic scores for cardiometabolic traits in british pakistanis and bangladeshis. *medRxiv* (2021).

[61] Durvasula, A. & Lohmueller, K. E. Negative selection on complex traits limits phenotype prediction accuracy between populations. *Am. J. Hum. Genet.* **108**, 620–631 (2021).

[62] Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. & Tang, H. Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *Am. J. Hum. Genet.* **101**, 218–226 (2017).

[63] Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

[64] Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1098 (2021).

[65] Kuchenbaecker, K. *et al.* The transferability of lipid loci across african, asian and european cohorts. *Nat. Commun.* **10**, 4330 (2019).

[66] Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* **9** (2020).

[67] Ulirsch, J. C. *et al.* An annotated atlas of causal variants for complex human traits. *In revision* .

[68] Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv* (2021).

[69] Kanai, M. *et al.* Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *medRxiv* (2022).

[70] Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).

[71] Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* (2005).

[72] Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).

[73] Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

[74] Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).

[75] Lee, Y., Luca, F., Pique-Regi, R. & Wen, X. Bayesian multi-snp genetic association analysis: Control of FDR and use of summary statistics. *bioRxiv* 316471 (2018).

[76] Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[77] Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

[78] Hukku, A. *et al.* Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Hum. Genet.* **108**, 25–35 (2021).

[79] Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

[80] Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).

[81] Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).

[82] Corces, M. R. *et al.* Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for alzheimer's and parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).

[83] Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).

[84] Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* **19** (2018).

[85] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

[86] Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370** (2020).

[87] Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).

[88] Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

[89] Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

[90] Abramov, S. *et al.* Landscape of allele-specific transcription factor binding in the human genome. *Nat. Commun.* **12**, 1–15 (2021).

[91] Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

[92] Carvalho-Silva, D. *et al.* Open targets platform: new developments and updates two years on. *Nucleic Acids Res.* **47**, D1056–D1065 (2019).

[93] Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).

[94] Raychaudhuri, S. Mapping rare and common causal alleles for complex human diseases. *Cell* **147**, 57–69 (2011).

[95] Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

[96] O'Connor, L. J. *et al.* Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).

[97] Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

[98] Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

[99] Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

[100] Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).

[101] Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2018).

[102] Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).

[103] The GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

[104] Lovell, M. C. A simple proof of the FWL theorem. *J. Econ. Educ.* **39**, 88–91 (2008).

[105] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

[106] Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* **21**, 233 (2020).

[107] Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

[108] Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

[109] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

[110] Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).

[111] Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 UK biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).

[112] Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).

[113] Benner, C. *et al.* Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).

[114] Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75 (2012).

[115] Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

[116] Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).

[117] Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).

[118] Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

[119] Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

[120] Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).

[121] Matschinsky, F. M. Glucokinase, glucose homeostasis, and diabetes mellitus. *Curr. Diab. Rep.* **5**, 171–176 (2005).

[122] Gery, S. & Koeffler, H. P. Role of the adaptor protein LNK in normal and malignant hematopoiesis. *Oncogene* **32**, 3111–3118 (2013).

[123] Mahley, R. W. Apolipoprotein e: cholesterol transport protein with expanding role in cell biology. *Science* **240**, 622–630 (1988).

[124] Takeshita, T., Mao, X. Q. & Morimoto, K. The contribution of polymorphism in the alcohol dehydrogenase beta subunit to alcohol sensitivity in a japanese population. *Hum. Genet.* **97**, 409–413 (1996).

[125] Carmeliet, P. & Jain, R. K. Molecular mechanisms and clinical applications of angiogenesis. *Nature* **473**, 298–307 (2011).

[126] Bouchard, C. *et al.* Direct induction of cyclin D2 by myc contributes to cell cycle progression and sequestration of p27. *EMBO J.* **18**, 5321–5333 (1999).

[127] Pattaro, C. *et al.* Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat. Commun.* **7**, 10023 (2016).

[128] Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).

[129] Prins, B. P. *et al.* Genome-wide analysis of health-related biomarkers in the UK household longitudinal study reveals novel associations. *Sci. Rep.* **7**, 11008 (2017).

[130] Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).

[131] Hao, Y. *et al.* Genome-wide association study: Functional variant rs2076295 regulates desmoplakin expression in airway epithelial cells. *Am. J. Respir. Crit. Care Med.* **202**, 1225–1236 (2020).

[132] Thom, C. S. & Voight, B. F. Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes. *BMC Med. Genomics* **13**, 89 (2020).

[133] Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).

[134] Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative multi-snp association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).

[135] Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* **49**, 1747–1751 (2017).

[136] Luo, Y. *et al.* Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Hum. Mol. Genet.* **30**, 1521–1534 (2021).

[137] Wen, X. Molecular QTL discovery incorporating genomic annotations using bayesian false discovery rate control. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).

[138] Wang, Q. S. *et al.* Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 1–11 (2021).

[139] Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

[140] Klos, K. *et al.* APOE/C1/C4/C2 hepatic control region polymorphism influences plasma apoe and LDL cholesterol levels. *Hum. Mol. Genet.* **17**, 2039–2046 (2008).

[141] Sobreira, D. R. *et al.* Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of IRX3 and IRX5. *Science* **372**, 1085–1091 (2021).

[142] Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv* (2020).

[143] Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

[144] Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).

[145] Tucker, N. R. *et al.* Transcriptional and cellular diversity of the human heart. *Circulation* **142**, 466–482 (2020).

[146] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

[147] Novikova, G. *et al.* Integration of alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat. Commun.* **12**, 1–14 (2021).

[148] Ruth, K. S. *et al.* Using human genetics to understand the disease impacts of testosterone in men and women. *Nat. Med.* **26**, 252–258 (2020).

[149] Ferkingstad, E. *et al.* Genome-wide association meta-analysis yields 20 loci associated with gallstone disease. *Nat. Commun.* **9**, 1–11 (2018).

[150] Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).

[151] Tournamille, C., Colin, Y., Cartron, J. P. & Le Van Kim, C. Disruption of a GATA motif in the duffy gene promoter abolishes erythroid gene expression in duffy-negative individuals. *Nat. Genet.* **10**, 224–228 (1995).

[152] Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS Genet.* **10**, e1004633 (2014).

[153] Nielson, C. M. *et al.* Novel genetic variants associated with increased vertebral volumetric BMD, reduced vertebral fracture risk, and increased expression of SLC1A3 and EPHB2. *J. Bone Miner. Res.* **31**, 2085–2097 (2016).

[154] Griesemer, D. *et al.* Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247–5260.e19 (2021).

[155] Connally, N. *et al.* The missing link between genetic association and regulatory function. *medRxiv* (2021).

[156] Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

[157] Bastarache, L. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).

[158] Stegle, O., Parts, L., Durbin, R. & Winn, J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).

[159] Loh, P.-R. *et al.* Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

[160] Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).

[161] Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).

[162] UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

[163] Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).

[164] Tsukiyama, S., Ide, M., Ariyoshi, H. & Shirakawa, I. A new algorithm for generating all the maximal independent sets. *SIAM J. Comput.* **6**, 505–517 (1977).

[165] Ghoussaini, M. *et al.* Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, D1311–D1320 (2021).

[166] Malone, J. *et al.* Modeling sample variables with an experimental factor ontology. *Bioinformatics* **26**, 1112–1118 (2010).

[167] The Hail Team. Hail. https://github.com/hail-is/hail.

[168] McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

[169] Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakr: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).

[170] Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

[171] Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).

[172] Kurki, M. I. *et al.* FinnGen: Unique genetic insights from combining isolated population and national health register data. *medRxiv* (2022).

[173] Chen, W. *et al.* Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015).

[174] Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).

[175] Lam, M. *et al.* Comparative genetic architectures of schizophrenia in east asian and european populations. *Nat. Genet.* **51**, 1670–1678 (2019).

[176] Gharahkhani, P. *et al.* Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. *Nat. Commun.* **12**, 1258 (2021).

[177] Levey, D. F. *et al.* Bi-ancestral depression GWAS in the million veteran program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat. Neurosci.* **24**, 954–963 (2021).

[178] Chen, J. *et al.* The trans-ancestral genomic architecture of glycemic traits. *Nat. Genet.* **53**, 840–860 (2021).

[179] Robertson, C. C. *et al.* Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. *Nat. Genet.* **53**, 962–971 (2021).

[180] Stanzick, K. J. *et al.* Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat. Commun.* **12**, 4350 (2021).

[181] Novembre, J. *et al.* Genes mirror geography within europe. *Nature* **456**, 98–101 (2008).

[182] Martin, A. R. *et al.* Haplotype sharing provides insights into Fine-Scale population history and disease in finland. *Am. J. Hum. Genet.* **102**, 760–775 (2018).

[183] Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).

[184] Chheda, H. *et al.* Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from finland and united kingdom. *Eur. J. Hum. Genet.* **25**, 477–484 (2017).

[185] Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated european populations. *Nat. Commun.* **8**, 15927 (2017).

[186] Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).

[187] Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

[188] Hujoel, M. L. A., Gazal, S., Hormozdiari, F., van de Geijn, B. & Price, A. L. Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *Am. J. Hum. Genet.* **104**, 611–624 (2019).

[189] Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675–1678 (2012).

[190] Marnetto, D. *et al.* Evolutionary rewiring of human regulatory networks by waves of genome expansion. *Am. J. Hum. Genet.* **102**, 207–218 (2018).

[191] Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).

[192] Zollner, S. & Pritchard, J. K. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80**, 605–615 (2007).

[193] Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C. & Bu, G. Apolipoprotein E and alzheimer disease: pathobiology and targeting strategies. *Nat. Rev. Neurol.* **15**, 501–518 (2019).

[194] Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to crohn's disease. *Nature* **411**, 599–603 (2001).

[195] Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to crohn's disease. *Nature* **411**, 603–606 (2001).

[196] Benyamin, B. *et al.* Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. *Nat. Genet.* **41**, 1173–1175 (2009).

[197] Luukkonen, P. K. *et al.* MARC1 variant rs2642438 increases hepatic phosphatidylcholines and decreases severity of non-alcoholic fatty liver disease in humans. *J. Hepatol.* **73**, 725–726 (2020).

[198] Emdin, C. A. *et al.* A missense variant in mitochondrial amidoxime reducing component 1 gene and protection against liver disease. *PLoS Genet.* **16**, e1008629 (2020).

[199] Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin levels. *Hum. Mol. Genet.* **18**, 2700–2710 (2009).

[200] Nioi, P. *et al.* Variant ASGR1 associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.* **374**, 2131–2141 (2016).

[201] Wang Xiao & Musunuru Kiran. Confirmation of causal rs9349379-PHACTR1 expression quantitative trait locus in Human-Induced pluripotent stem cell endothelial cells. *Circulation: Genomic and Precision Medicine* **11**, e002327 (2018).

[202] Gupta, R. M. *et al.* A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell* **170**, 522–533.e15 (2017).

[203] Adlam, D. *et al.* Association of the PHACTR1/EDN1 genetic locus with spontaneous coronary artery dissection. *J. Am. Coll. Cardiol.* **73**, 58–66 (2019).

[204] Ford, T. J. *et al.* Genetic dysregulation of endothelin-1 is implicated in coronary microvascular dysfunction. *Eur. Heart J.* **41**, 3239–3252 (2020).

[205] Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).

[206] Grisanzio, C. & Freedman, M. L. Chromosome 8q24-associated cancers and MYC. *Genes Cancer* **1**, 555–559 (2010).

[207] Huppi, K., Pitt, J. J., Wahlberg, B. M. & Caplen, N. J. The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Front. Genet.* **3**, 69 (2012).

[208] Matejcic, M. *et al.* Germline variation at 8q24 and prostate cancer risk in men of european ancestry. *Nat. Commun.* **9**, 4616 (2018).

[209] Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

[210] Sajantila, A. *et al.* Paternal and maternal DNA lineages reveal a bottleneck in the founding of the finnish population. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 12035–12039 (1996).

[211] Kittles, R. A. *et al.* Dual origins of finns revealed by Y chromosome haplotype variation. *Am. J. Hum. Genet.* **62**, 1171–1179 (1998).

[212] Jinam, T. *et al.* The history of human populations in the japanese archipelago inferred from genome-wide SNP data with a special reference to the ainu and the ryukyuan populations. *J. Hum. Genet.* **57**, 787–795 (2012).

[213] Takeuchi, F. *et al.* The fine-scale genetic structure and evolution of the japanese population. *PLoS One* **12**, e0185487 (2017).

[214] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

[215] GEnome Medical alliance Japan Project (GEM-J). GEM japan whole genome aggregation (GEM-J WGA) panel. https://togovar.biosciencedbc.jp/doc/datasets/gem_j_wga.

[216] Rivas, M. A. *et al.* Insights into the genetic epidemiology of crohn's and rare diseases in the ashkenazi jewish population. *PLoS Genet.* **14**, e1007329 (2018).

[217] Casals, F. *et al.* Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet.* **9**, e1003815 (2013).

[218] Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the icelandic population. *Nat. Genet.* **47**, 435–444 (2015).

[219] Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).

[220] Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).

[221] Henn, B. M. *et al.* Distance from sub-saharan africa predicts mutational load in diverse human genomes. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E440–9 (2016).

[222] Kashiwagi, H. *et al.* Molecular basis of CD36 deficiency. evidence that a 478C–>T substitution (proline90–>serine) in CD36 cDNA accounts for CD36 deficiency. *J. Clin. Invest.* **95**, 1040–1046 (1995).

[223] Lenkkeri, U. *et al.* Structure of the gene for congenital nephrotic syndrome of the finnish type (NPHS1) and characterization of mutations. *Am. J. Hum. Genet.* **64**, 51–61 (1999).

[224] van der Slot, A. J. *et al.* Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis *. *J. Biol. Chem.* **278**, 40967–40972 (2003).

[225] Hirata, M. *et al.* Cross-sectional analysis of BioBank japan clinical data: A large cohort of 200,000 patients with 47 common diseases. *Journal of Epidemiology* **27**, S9–S21 (2017).

[226] Ishigaki, K. *et al.* Large-scale genome-wide association study in a japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).

[227] Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).

[228] Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

[229] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[230] Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the japanese population. *Nat. Commun.* **10**, 4393 (2019).

[231] Terao, C. *et al.* GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat. Commun.* **10**, 4719 (2019).

[232] Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

[233] Buenrostro, J. D. *et al.* Integrated Single-Cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548.e16 (2018).

[234] Vukcevic, D., Hechter, E., Spencer, C. & Donnelly, P. Disease model distortion in association studies. *Genet. Epidemiol.* **35**, 278–290 (2011).

[235] Gagliano Taliun, S. A. *et al.* Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).

[236] Evangelou, E. & Ioannidis, J. P. a. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).

[237] Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

[238] Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

[239] Zhou, W. *et al.* Global biobank meta-analysis initiative: powering genetic discovery across human diseases. *medRxiv* (2021).

[240] Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).

[241] Zhou, W. *et al.* GWAS of thyroid stimulating hormone highlights pleiotropic effects and inverse association with thyroid cancer. *Nat. Commun.* **11**, 1–13 (2020).

[242] Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).

[243] Li, D., Zhao, H. & Gelernter, J. Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (*2) allele against alcoholism and alcohol-induced medical diseases in asians. *Hum. Genet.* **131**, 725–737 (2012).

[244] Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

[245] Dendrou, C. A. *et al.* Resolving *tyk2* locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* **8**, 363ra149 (2016).

[246] Couturier, N. *et al.* Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. *Brain* **134**, 693–703 (2011).

[247] Li, Z. *et al.* Two rare disease-associated tyk2 variants are catalytically impaired but signaling competent. *J. Immunol.* **190**, 2335–2344 (2013).

[248] Lam, M. *et al.* RICOPILI: Rapid imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2020).

[249] Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).

[250] Chen, W. *et al.* Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. *Nat. Commun.* **12**, 7117 (2021).

[251] McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

[252] Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* **590**, 290–299 (2021).

[253] Ormond, C., Ryan, N. M., Corvin, A. & Heron, E. A. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Brief. Bioinform.* **22** (2021).

[254] Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).

[255] Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

[256] Koskela, J. T. *et al.* Genetic variant in SPDL1 reveals novel mechanism linking pulmonary fibrosis risk and cancer protection. *medRxiv* (2021).

[257] Partanen, J. J. *et al.* Leveraging global multi-ancestry meta-analysis in the study of idiopathic pulmonary fibrosis genetics. *medRxiv* (2021).

[258] Foreman, M. G. *et al.* Alpha-1 antitrypsin PiMZ genotype is associated with chronic obstructive pulmonary disease in two racial groups. *Ann. Am. Thorac. Soc.* **14**, 1280–1287 (2017).

[259] Tsuo, K. *et al.* Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. *medRxiv* (2021).

[260] Benonisdottir, S. *et al.* Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).

[261] Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).

[262] Langefeld, C. D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).

[263] Hargreaves, C. E. *et al.* Fcγ receptors: genetic variation, function, and disease. *Immunol. Rev.* **268**, 6–24 (2015).

[264] Franke, L. *et al.* Association analysis of copy numbers of FC-gamma receptor genes for rheumatoid arthritis and other immune-mediated phenotypes. *Eur. J. Hum. Genet.* **24**, 263–270 (2016).

[265] Wang, Y. *et al.* Global biobank analyses provide lessons for computing polygenic risk scores across diverse cohorts. *medRxiv* (2021).

[266] Namba, S. *et al.* A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis. *medRxiv* (2021).

[267] Wu, K.-H. H. *et al.* Polygenic risk score from a multi-ancestry GWAS uncovers susceptibility of heart failure. *medRxiv* (2021).

[268] Lo Faro, V. *et al.* Genome-wide association meta-analysis identifies novel ancestry-specific primary open-angle glaucoma loci and shared biology with vascular mechanisms and cell proliferation. *medRxiv* (2021).

[269] Surakka, I. *et al.* Multi-ancestry meta-analysis identifies 2 novel loci associated with ischemic stroke and reveals heterogeneity of effects between sexes and ancestries. *medRxiv* (2022).

[270] Wolford, B. *et al.* Multi-ancestry GWAS for venous thromboembolism identifies novel loci followed by experimental validation. *In preparation* .

[271] Aneas, I. *et al.* Asthma-associated genetic variants induce IL33 differential expression through an enhancer-blocking regulatory region. *Nat. Commun.* **12**, 6115 (2021).

[272] Vladich, F. D. *et al.* IL-13 R130Q, a common variant associated with allergy and asthma, enhances effector mechanisms essential for human allergic inflammation. *J. Clin. Invest.* **115**, 747–754 (2005).

[273] Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat. Genet.* **53**, 1504–1516 (2021).

[274] Sakaue, S. *et al.* Decoding the diversity of killer immunoglobulin-like receptors by deep sequencing and a high-resolution imputation method. *Cell Genomics* **2** (2022).

[275] Mukamel, R. E. *et al.* Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).

[276] Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).

[277] Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

[278] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

[279] Wei, X. & Nielsen, R. CCR5-Δ32 is deleterious in the homozygous state in humans. *Nat. Med.* **25**, 909–910 (2019).

[280] Maier, R. *et al.* No statistical evidence for an effect of CCR5-Δ32 on lifespan in the UK biobank cohort. *Nat. Med.* **26**, 178–180 (2020).

[281] Loh, P.-R. *et al.* Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

[282] Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

[283] Khera, A. V. *et al.* Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* **177**, 587–596.e9 (2019).

[284] Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).

[285] Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.* (2020).

[286] Asiki, G. *et al.* The general population cohort in rural south-western uganda: a platform for communicable and non-communicable disease studies. *Int. J. Epidemiol.* **42**, 129–141 (2013).

[287] Heckerman, D. *et al.* Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 7377–7382 (2016).

[288] Duan, S., Zhang, W., Cox, N. J. & Dolan, M. E. FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformation* **3**, 139–141 (2008).

[289] Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

[290] Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).

[291] Nievergelt, C. M. *et al.* International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. *Nat. Commun.* **10**, 4558 (2019).

[292] Sakaue, S. *et al.* Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. *Nat. Med.* **26**, 542–548 (2020).

[293] Vuckovic, D. *et al.* The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231.e11 (2020).

[294] Guo, J. *et al.* Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat. Commun.* **9**, 1865 (2018).

[295] Sved, J. A., McRae, A. F. & Visscher, P. M. Divergence between human populations estimated from linkage disequilibrium. *Am. J. Hum. Genet.* **83**, 737–743 (2008).

[296] Budin-Ljøsne, I. *et al.* Data sharing in large research consortia: experiences and recommendations from ENGAGE. *Eur. J. Hum. Genet.* **22**, 317–321 (2014).

[297] Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).

[298] Horikoshi, M. *et al.* Discovery and Fine-Mapping of glycaemic and Obesity-Related trait loci using High-Density imputation. *PLoS Genet.* **11**, e1005230 (2015).

[299] Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* **17**, e1009021 (2021).

[300] Chung, W. *et al.* Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. *Nat. Commun.* **10**, 569 (2019).

[301] Chun, S. *et al.* Non-parametric polygenic risk prediction via partitioned GWAS summary statistics. *Am. J. Hum. Genet.* **107**, 46–59 (2020).

[302] Im, C. *et al.* Generalizability of "GWAS hits" in clinical populations: Lessons from childhood cancer survivors. *Am. J. Hum. Genet.* **107**, 636–653 (2020).

[303] Galinsky, K. J. *et al.* Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).

[304] Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746 (2018).

[305] Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).

[306] Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* **13**, e1005589 (2017).

[307] Márquez-Luna, C. *et al.* Incorporating functional priors improves polygenic prediction accuracy in UK biobank and 23andme data sets. *Nat. Commun.* **12**, 6052 (2021).

[308] Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).

[309] Yang, S. & Zhou, X. Accurate and scalable construction of polygenic scores in large biobank data sets. *Am. J. Hum. Genet.* **106**, 679–693 (2020).

[310] Qian, J. *et al.* A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK biobank. *PLoS Genet.* **16**, e1009141 (2020).

[311] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

[312] Gurdasani, D. *et al.* Uganda genome resource enables insights into population history and genomic discovery in africa. *Cell* **179**, 984–1002.e36 (2019).

[313] Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

[314] Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509–516 (2022).

[315] Wiberg, A. *et al.* A genome-wide association analysis identifies 16 novel susceptibility loci for carpal tunnel syndrome. *Nat. Commun.* **10**, 1030 (2019).

[316] Tan, K. & Lawler, J. The interaction of thrombospondins with extracellular matrix proteins. *J. Cell Commun. Signal.* **3**, 177–187 (2009).

[317] Li, C. *et al.* Mutations in COMP cause familial carpal tunnel syndrome. *Nat. Commun.* **11**, 3642 (2020).

[318] Svensson, L., Närlid, I. & Oldberg, A. Fibromodulin and lumican bind to the same region on collagen type I fibrils. *FEBS Lett.* **470**, 178–182 (2000).

[319] Crespi, A. *et al.* POF1B localizes to desmosomes and regulates cell adhesion in human intestinal and keratinocyte cell lines. *J. Invest. Dermatol.* **135**, 192–201 (2015).

[320] Mandelker, D. *et al.* Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med.* **18**, 1282–1289 (2016).

[321] Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

[322] Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).

[323] Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

[324] HapMap3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).

[325] Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

[326] Bhangale, T. R., Rieder, M. J. & Nickerson, D. A. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat. Genet.* **40**, 841–843 (2008).

[327] Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3**, e3395 (2008).

[328] Visscher, P. M. & Hill, W. G. The limits of individual identification from sample allele frequencies: Theory and statistical analysis. *PLoS Genet.* **5**, e1000628 (2009).

[329] Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

[330] Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424–428 (2018).

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.