Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity

Kristin Tsuo[1,2,3,*], Wei Zhou[1,2], Ying Wang[1,2,3], Masahiro Kanai[1,2,3,4,5], Shinichi Namba[5], Rahul Gupta[1,2,3,6], Lerato Majara[7], Lethukuthula L. Nkambule[1,2,3], Yukinori Okada[5,8,9,10,11], Takayuki Morisaki[13], Benjamin M. Neale[1,2,3], Global Biobank Meta-analysis Initiative, Mark J. Daly[1,2,3,12,*], Alicia R. Martin[1,2,3,*]

Affiliations

1. Analytic and Translational Genetics Unit, 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, Japan
6. Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, MA USA
7. Department of Psychiatry and Mental Health, Faculty of Health Sciences, University of Cape Town, South Africa
8. Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
9. Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan
10. Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan
11. Center for Infectious Disease Education and Research (CiDER), Osaka University, Suita 565-0871, Japan
12. Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland
13. Division of Molecular Pathology, The Institute of Medical Science, The University of Tokyo, Minato-ku, Japan

Lead Contact: Kristin Tsuo (ktsuo@broadinstitute.org)
*Correspondence: ktsuo@broadinstitute.org, mjdaly@broadinstitute.org, armartin@broadinstitute.org.

Keywords

asthma, GWAS, heterogeneity, multi-ancestry, polygenic risk prediction, cross-trait

Summary

Asthma is a complex disease that affects millions of people and varies in prevalence by an order of magnitude across geographic regions and populations. However, the extent to which genetic variation contributes to these disparities is unclear, as studies probing the genetics of asthma have been primarily limited to populations of European (EUR) descent. As part of the Global Biobank Meta-analysis Initiative (GBMI), we conducted the largest genome-wide association study of asthma to date (153,763 cases and 1,647,022 controls) via meta-analysis across 18 biobanks spanning multiple countries and ancestries. Altogether, we discovered 180 genome-wide significant loci ($p < 5x10^{-8}$) associated with asthma, 49 of which are not previously reported. We replicate well-known associations such as *IL1RL1* and *STAT6*, and find that overall the novel associations have smaller effects than previously-discovered loci, highlighting our substantial increase in statistical power. Despite the considerable range in prevalence among biobanks, from 3% to 24%, the genetic effects of associated loci are largely consistent across the biobanks and ancestries. To further investigate the polygenic architecture of asthma, we construct polygenic risk scores (PRS) using a multi-ancestry approach, which yields higher predictive power for asthma in non-EUR populations compared to PRS derived from previous asthma meta-analyses and using other methods. Additionally, we find considerable genetic overlap between asthma and chronic obstructive pulmonary disease (COPD) across ancestries but minimal overlap in enriched biological pathways. Our work underscores the multifactorial nature of asthma development and offers insight into the shared genetic architecture of asthma that may be differentially perturbed by environmental factors and contribute to variation in prevalence.

Introduction

Asthma is a complex and multifactorial disease that affects millions of people worldwide, yet much of its genetic architecture has eluded discovery. Genetic factors contribute substantially to asthma risk, with heritability estimates ranging from 35% to 95% (Ober and Yao, 2011; 2015). The most recent meta-analysis of asthma discovered 212 asthma-associated loci across the genome, confirming the polygenic nature of asthma (Han et al., 2020). However, these risk loci only account for a small proportion of the total heritability for asthma. Furthermore, the discovery GWAS, like the majority of previous asthma GWAS, were primarily conducted in populations of European ancestry. A major exception is the Trans-National Asthma Genetic Consortium (TAGC) which included modest sample sizes from populations of African, Japanese and Latino ancestries (Demenais et al., 2018). Surveys of asthma worldwide have found that prevalence can vary by as much as 21-fold among countries (Asher et al., 2020; Sembajwe et al., 2010). Prevalence varies greatly within countries as well (Akinbami et al., 2012) and this variation cannot be attributed to any single known risk factor such as air pollution. Rather, genetic and environmental contributions to the wide variation in asthma prevalence are complex. Therefore, assessing the genetic architecture of asthma in diverse cohorts is particularly important.

This heterogeneity in prevalence is mirrored by, and may be a consequence of, the heterogeneity of the disease itself. Asthma is commonly viewed as not a single disease, but rather a syndrome composed of overlapping phenotypes and endotypes ((Borish and Culp, 2008; Dharmage et al., 2019)). These are driven by a myriad of risk factors, both genetic and non-genetic (Borish and Culp, 2008; Dharmage et al., 2019). Asthma also shares genetic components with various comorbid diseases, including other respiratory diseases like chronic obstructive pulmonary disease (COPD), allergic diseases, obesity, and neuropsychiatric disorders (Maselli and Hanania, 2018; Postma and Rabe, 2015; Zhu et al., 2018, 2019, 2020, 2021). The complex etiology and clinical presentations of asthma in turn complicate standards for defining the phenotype; for example, nearly 60 different definitions of “childhood asthma” were used across more than 100 studies in the literature (Van Wonderen et al., 2010). This likely also contributes to the observed variation in asthma prevalence across populations.

High quality clinical models of asthma are necessary to comprehensively and quantitatively investigate risk factors within and among populations. However, with relatively few large and diverse datasets of asthma, genetic risk predictors for asthma have been rare in the literature. For a highly heterogeneous disease like asthma, polygenic risk scores (PRS) may be particularly useful tools for predicting subtypes, disease severity, and the development of comorbidities in the clinical setting, and for investigating potential gene-environment interactions in the research setting. Existing PRS for asthma were generated primarily from UK Biobank (UKBB), a large population-based cohort study, and BioBank Japan (BBJ), a similarly large-scale hospital-based study (Amariuta et al., 2020), or cohorts of smaller sample sizes (Song et al., 2020; Sordillo et al., 2021). The PRS had limited predictive ability in these studies and lower performance compared to PRS for other common complex diseases. This

underscores the genetic complexity of asthma and highlights the need for more large-scale, genomic studies of asthma.

To more deeply interrogate the genetic architecture of asthma across different populations, we analyzed paired phenotypic and genetic data from the Global Biobank Meta-analysis (GBMI). Participating biobanks shared summary statistics for the meta-analyses of 14 phenotypes, including asthma (Zhou et al., 2021). Compared to previous asthma resources and studies, this collaborative effort increased both the sample size and diversity of asthma cases by many folds, covered more variants with higher imputation quality, and harmonized phenotypes using consistent electronic health record definitions for asthma across datasets. Harnessing this resource, we identified 49 loci not previously associated with asthma. Despite prevalence differences of nearly an order of magnitude, we also demonstrated remarkable consistency of genetic effects across the biobanks and ancestries in GBMI. Further, we showed that the increased diversity of data from GBMI improves genetic risk prediction accuracy in multiple populations. Finally, we provided additional evidence for shared genetic architectures between asthma and known coexisting diseases such as COPD and hay fever. Our findings highlight the need for future investigations into how genetic effects shared with different diseases contribute to the heterogeneity of asthma.

Results

Meta-analysis for asthma across 18 biobanks in GBMI

To identify novel loci associated with asthma, we performed fixed-effects inverse-variance weighted meta-analysis using the harmonized GWAS summary statistics for asthma from 18 biobanks participating in GBMI. The combined sample size from all studies was 153,763 cases and 1,647,022 controls, spanning individuals of European (EUR), African (AFR), Admixed American (AMR), East Asian (EAS), Middle Eastern (MID), and Central and South Asian (CSA) ancestry (**Fig. 1**). Despite the standardized phenotype definitions used by each biobank, which included the asthma PheCode and/or self-reported data (**Supplementary Table 3**), the prevalence of asthma varies widely across these biobanks, ranging from 3% in the Taiwan Biobank to 24% in the Mass General Brigham Biobank. We applied pre- and post-GWAS quality control filters that resulted in 70.8 million SNPs for meta-analysis; for downstream analyses we analyzed SNPs present in at least 2 biobanks (Zhou et al., 2021). The meta-analysis identified 180 loci of genome-wide significance ($p < 5x10^{-8}$), 49 of which have not been previously reported to be associated with asthma (**Fig. 2A**). The potentially novel associations had smaller effect sizes on average compared to the previously reported loci, across the allele frequency spectrum (**Fig. 2B**). This illustrates that with the increased power and effective sample size of GBMI, we were able to uncover SNPs with more modest effects on asthma.

Because the GBMI meta-analysis includes data from UKBB, we compared our results to the TAGC meta-analysis results that did not include the UKBB GWAS to facilitate analyses that require independent samples (Demenais et al., 2018). Of the 180 lead variants in GBMI, 122 were in the TAGC meta-analysis or had a tagging variant in high LD ($r^2 > 0.8$) in the TAGC study; 76 of these had $p < 0.05$ in the TAGC results. We compared the effect sizes of these 76 SNPs in the GBMI and the TAGC meta-analyses using a previously proposed Deming regression method that considers standard errors in both effect size estimates (Deming, 1943). We found that all 76 SNPs were directionally consistent and aligned across the studies (**Supplementary Table 4, Supplementary Fig. 1**).

Among the 49 novel loci, six were missense variants or in high LD ($r^2 > 0.8$) with a missense variant (**Supplementary Table 2**). One of these SNPs, chr10:94279840:G:C ($p_{meta\text{-}analysis} = 2.5\times10^{-9}$), resides in *PLCE1,* an autosomal recessive nephrotic syndrome gene (Jefferson and Shankland, 2007); high prevalence of atopic disorders, like asthma, among children with nephrotic syndrome has long been observed in the clinic, suggesting potential shared pathways underlying asthma and nephrotic syndrome (Riar et al., 2019). The asthma risk allele has also been previously linked to lower blood pressure (UK Biobank — Neale lab). The lead SNPs chr14:100883117:G:T ($p_{meta\text{-}analysis} = 2.6\times10^{-8}$) and chr19:56222056:C:A ($p_{meta\text{-}analysis} = 2.4\times10^{-8}$) also implicate novel genes, *RTL1* and *ZSCAN5A* respectively. *RTL1* has been found to play a role in muscle regeneration (Loo et al., 2019), and *ZSCAN5A* has been linked to monocyte count (Chen et al., 2020). Additionally, three of the novel lead SNPs co-localized with a fine-mapped cis-eQTL (**Supplementary Table 2**). One of these, chr19:49513502:C:T ($p_{meta\text{-}analysis} = 7.98\times10^{-9}$), implicates *FCGRT,* which regulates IgG recycling and is a potential drug target for autoimmune neurological disease therapies (Dalakas and Spaeth, 2021). The other previously-reported missense variants replicated previous findings; among these, chr4:102267552:C:T (p.Ala391Thr, $p = 2.5\times10^{-12}$) is a highly pleiotropic variant in *SLC39A8* that has been associated with many psychiatric, neurologic, inflammatory and metabolic diseases (Huang et al., 2017; International Consortium for Blood Pressure Genome-Wide Association Studies et al., 2011; Li et al., 2016; Nebert and Liu, 2019; Pickrell et al., 2016; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Speliotes et al., 2010) and has been demonstrated to disrupt manganese homeostasis (Nakata et al., 2020). Variants implicating well-known asthma-associated genes with large effects, like *IL1RL1, IL2RA, STAT6, IL33, GSDMB,* and *TSLP,* were replicated in the meta-analysis as well.

Shared genetic architecture of asthma across biobanks and ancestry groups

Given that sample size, disease prevalence, ancestry, and sampling approaches differed across the 18 biobanks, we next investigated the consistency of the asthma-associated loci across the biobanks and their attributes. We first implemented an approach that estimates the correlation (r_b) between the effects of the 180 lead variants in each biobank GWAS and the corresponding meta-analysis excluding that biobank (Qi et al., 2018). Most of the r_b estimates were highly

correlated (i.e. did not differ significantly from 1) (**Supplementary Table 5**). To further interrogate the consistency of the lead variants in all biobanks, we computed the ratio of the effect size of these SNPs in the biobank-specific GWAS over that in the corresponding leave-that-biobank-out meta-analysis. We found that the average per-biobank ratios were almost evenly split between those greater than and less than 1 (**Supplementary Fig. 2**). This indicates that any significant difference in effects likely does not reflect technical artifacts in the meta-analysis or GWAS procedures. We also applied Deming regression to assess the alignment of the lead SNP effects in each biobank-specific GWAS with the effects in the corresponding leave-that-biobank-out meta-analysis (Deming, 1943) and observed that the effect sizes were comparable across the biobanks (**Fig. 3**). Furthermore, the genome-wide genetic correlations between the biobanks with non-zero heritability estimates and the respective leave-that-biobank-out meta-analyses were all close to 1 (Zhou et al., 2021). Taken together, these analyses indicate that the genetic architecture of asthma is largely shared across diverse cohorts, even when cohort characteristics like sample size and disease prevalence differ.

We also found little evidence of heterogeneity in the ancestry-specific effect sizes for the lead SNPs. One SNP, chr10:9010779:G:A, was significantly heterogeneous (p-value for Cochran's Q test < 0.0003, the Bonferroni-corrected p-value threshold) across the ancestry-specific meta-analyses of AFR, AMR, CSA, EAS, and EUR individuals (**Fig. 4A, Supplementary Table 6**). One known SNP that nearly reached the Bonferroni-corrected p-value threshold for heterogeneity, chr16:27344041:G:A, displayed different effects in the EUR and EAS cohorts. This SNP lies within an intron of *IL4R* (**Fig. 4B**), which has known associations with asthma (El-Husseini et al., 2020; Massoud et al., 2016). Previous studies have investigated the association of *IL4R* with asthma in different populations, with inconsistent results, so future studies on the population-specific effects of this gene will be needed (Battle et al., 2007; Kousha et al., 2020; Nie et al., 2013). Our findings demonstrate that despite broad consistency of effect sizes across ancestries among the lead variants, the increased power and diversity of GBMI enabled the detection of SNPs with significantly different effects across ancestries.

Polygenic risk score analyses

We next explored the impact of the increased sample sizes and diversity in GBMI on genome-wide risk prediction of asthma. To establish a baseline understanding of PRS performance for the 14 phenotypes analyzed in GBMI, we implemented PRS-CS (Ge et al., 2019) to construct PRS using the leave-one-biobank-out meta-analyses as discovery data (Wang et al., 2021). The PRS derived from the GBMI leave-one-biobank-out meta-analyses of asthma had higher predictive accuracy, as measured by R^2 on the liability scale ($R^2_{liability}$), compared to the PRS constructed from the TAGC meta-analysis (Demenais et al., 2018) across all biobanks tested (**Fig. 5**).

To expand on these analyses, we tested a recently-developed extension of PRS-CS, PRS-CSx (Ruan et al., 2021). This method jointly models multiple summary statistics from different ancestries to enable more accurate effect size estimation for prediction. For input to PRS-CSx, we used the AFR, AMR, EAS, CSA, and EUR ancestry-specific meta-analyses from GBMI; the discovery meta-analysis that matched the ancestry of the target cohort excluded the target cohort (**Supplementary Fig. 5**). With the posterior SNP effect size estimates from PRS-CSx, we tested the multi-ancestry PRS in the following target populations: AFR ancestry individuals in UKBB, CSA ancestry individuals in UKBB, a holdout set of EAS ancestry individuals in BBJ, and a holdout set of EUR ancestry individuals in UKBB. The final prediction models tested in these target populations were the optimal linear combinations of the population-specific PRS. In both the EAS and EUR target cohorts, the $R^2_{liability}$ approached the SNP-based heritability (h^2_{SNP}), estimated to be 0.085 for asthma using the all-biobank meta-analysis (Wang et al., 2021) (**Fig. 5**). However, we acknowledge that h^2_{SNP} estimates may differ across biobanks and ancestries given differences in disease prevalence, environmental exposures, phenotype definitions, and other factors; these differences may contribute to the PRS in EAS individuals performing similarly to PRS in EUR individuals in our analyses, despite the smaller sample size of the EAS discovery cohort. The $R^2_{liability}$ across the target populations for the PRS-CSx scores were roughly the same as the $R^2_{liability}$ of the PRS derived from the PRS-CS analyses. It is important to note that the discovery data used in the PRS-CS analyses differed slightly in sample size and composition, since the leave-one-biobank-out approach was used for PRS-CS, but the target cohorts in which the PRS were evaluated were the same (**Supplementary Table 14**). To investigate why improvement in performance using PRS-CSx was only incremental in most of the target cohorts, we examined the performances of each population-specific PRS. We found that across all target cohorts, PRS derived from either the EUR or EAS set of posterior effect size estimates outperformed the linear combination, and the $R^2_{liability}$ of these PRS were also higher compared to that of the PRS-CS scores (**Supplementary Fig. 6**). This suggests that the addition of more discovery GWAS to PRS-CSx can improve the accuracy of PRS based on a single set of posterior effect size estimates, but the linear combination of PRS from multiple GWAS does not necessarily yield higher accuracy. This may be due to the considerably smaller sample sizes of some of the input discovery meta-analyses in our analyses and thus varying signal to noise ratios.

Genetic correlation between asthma and significantly heritable phenotypes across biobanks

Previous studies have shown that asthma is genetically correlated with a wide range of diseases and traits (Maselli and Hanania, 2018; Postma and Rabe, 2015; Zhu et al., 2018, 2019, 2020, 2021). We aimed to do a comprehensive genetic correlation analysis and identify all phenotypes that are significantly correlated with asthma from the GBMI biobanks. Among the 14

phenotypes analyzed in GBMI, COPD, a late-onset respiratory disease, had the highest genetic correlation with asthma (r_g (se) = 0.67 (0.021), p = 1.55×10^{-226}). This genetic correlation estimate is higher than estimates from previous studies, which ranged from 0.38-0.42 (Hobbs et al., 2017; Sakornsakolpat et al., 2019). This may be a result of greater power in the GBMI meta-analysis of COPD and different phenotype definitions used. Of note, only 6 of the 180 asthma-associated index variants (3%) had a genome-wide significant p-value in the COPD meta-analysis. Conversely, 12 of the 46 COPD-associated index variants (26%) had a genome-wide significant p-value in the asthma meta-analysis (**Supplementary Table 7**); 8 of these 12 variants were previously reported COPD associations. This shows that despite the strong genetic overlap between the COPD and asthma meta-analyses, the COPD meta-analysis largely identified variants with COPD-specific effects.

We next expanded these genetic correlation analyses beyond GBMI to measure correlations between asthma and the full breadth of phenotypes in UKBB. Of the 7,142 phenotypes for which GWAS were conducted in the EUR ancestry cohort in UKBB, 1,008 were significantly heritable (heritability Z score > 4) (Pan UKBB). Applying linkage-disequilibrium score correlation (LDSC) to these GWAS and the GBMI leave-UKBB-out meta-analysis of asthma, we obtained pairwise genetic correlation estimates between the heritable UKBB phenotypes and asthma, and observed strong correlations ($|r_g| > 0.4$) with 95 of these phenotypes, which spanned prescriptions, PheCodes, and other categories (**Supplementary Table 8**). Many of these replicated previously-found correlations with asthma. Digestive system disorders, including gastritis and gastroesophageal reflux disease (GERD), emerged as a disease category with significant and strong genetic correlations with asthma. Although the association between asthma and digestive disorders has not been as well studied, this does reinforce a previous finding of shared genetics between asthma and diseases of the digestive system (Demenais et al., 2018), indicating that the commonly-observed co-presentation of asthma and gastroesophageal disease in the clinic may be partially due to pleiotropic genetic effects. Our results also showed moderate and significant correlations (r_g ranging from 0.2-0.3) between asthma and neuropsychiatric diseases, like anxiety and depression, and obesity-related traits, like body mass index, which is similarly consistent with previous findings (Zhu et al., 2019, 2020).

To assess the consistency of the correlations in another population, we computed genetic correlation estimates between the GBMI EAS meta-analysis of asthma and other phenotypes in BBJ (**Supplementary Table 9, Supplementary Fig. 4**). Of the 48 available diseases in BBJ, 8 were significantly heritable in both BBJ and UKBB. Among these phenotypes, COPD showed the strongest and most significant correlation with asthma in BBJ (r_g = 0.29, p = 6.41×10^{-6}), although this was notably lower than across the full GBMI meta-analysis; differences in phenotype definition and curation may potentially contribute to variation in correlation estimates. Among all significantly heritable phenotypes in BBJ, pollinosis, also known as hay fever, showed moderate correlation with asthma as well (r_g = 0.28, p = 0.0004). These were directionally consistent with the correlation results from UKBB, which showed strong and significant

correlation with COPD (r_g = 0.71, p = 3.88x10^{-57}) and moderate correlation with pollinosis (r_g = 0.39, p = 4.60x10^{-3}).

Gene and pathway enrichment analyses for asthma

To further evaluate the extent of genetic overlap between asthma and COPD, we applied a gene prioritization method, MAGMA, to the GBMI EUR, AFR, EAS, and CSA meta-analyses of asthma as well as the GBMI EUR, AFR, and EAS meta-analyses of COPD (de Leeuw et al., 2015). After Bonferroni correction, we found that 442, 149, and 6 genes were significantly associated with asthma in the EUR (p < 2.50x10^{-6}), EAS (p < 2.50x10^{-6}), and CSA (p < 2.52x10^{-6}) populations, respectively, with no significantly associated genes in the AFR cohort (all p > 2.51x10^{-6}) (**Supplementary Table 10**). The majority of the genes associated with asthma identified in the EAS meta-analysis overlapped with the genes from the EUR meta-analysis (126 out of 149 genes), and all 6 genes associated with asthma as identified in the CSA meta-analysis were also significantly associated in the EUR and EAS meta-analyses. We identified 46 and 33 genes significantly associated with COPD in the EUR (p < 2.50x10^{-6}) and EAS (p < 2.50x10^{-6}) cohorts, respectively, and similarly to asthma, no significantly associated genes from the AFR meta-analysis (all p > 2.51x10^{-6}) (**Supplementary Table 11**). Of the 75 genes associated with COPD across the EUR and EAS meta-analyses, 24 overlapped with the asthma-associated genes.

Utilizing these sets of genes, we also adopted MAGMA for gene-set enrichment based on the curated and ontology gene sets from the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005). We found hundreds of gene sets that were significantly enriched (FDR < 0.05) by the asthma-associated genes discovered in the EUR and EAS meta-analyses (**Supplementary Table 12**). In contrast, only a handful of gene sets were significantly enriched among COPD-associated genes discovered in the AFR meta-analysis, likely reflecting the smaller overall sample size of COPD (**Supplementary Table 13**). The top-ranked asthma pathways from the EUR meta-analysis included cytokine and interleukin signaling and T-cell activation. Consistently biologically, the EAS meta-analysis identified autoimmune thyroid disease and graft vs. host disease pathways. The top-ranked COPD pathways from the EUR meta-analysis, although not significant, included several pathways related to nicotine receptor activity. These results reinforce that despite the substantial genetic overlap, asthma and COPD are governed by distinct biological processes as well. Future investigations will be required to fully parse out the etiology and comorbidities of asthma, like COPD, that develop later on in adulthood.

Discussion

Assembling the largest and most diverse collection of asthma cohorts to date, we conducted a GWAS meta-analysis of 18 biobanks around the world and identified 49 novel associations. Despite the substantial sample sizes of previous meta-analyses of asthma (Demenais et al.,

2018), our results indicated that the heterogeneity and complexity of asthma necessitate even larger sample sizes for genomic discovery. We interrogated the overall consistency of genetic effects across the cohorts and found that in spite of variability in recruitment continent, sampling strategy, health system design, and disease prevalence, the effects of the loci discovered in the meta-analysis were mostly concordant across the biobanks. Additionally, genetic correlation estimates across ancestries, which ranged from 0.65 to 0.99 for the well-powered ancestry groups, as well as specific genes prioritized using MAGMA, strongly supported the finding that the genetic architecture of asthma is largely shared across the ancestry groups studied. This study provided further evidence for substantial genetic overlap between asthma and well-known, immune-related comorbidities like COPD and allergic diseases. We also identified genetic correlations between asthma and less well-studied comorbidities like digestive system disorders, while highlighting additional complexity in the etiology and comorbidities of asthma. For example, gene set enrichment analyses using MAGMA did not yield many shared pathways for asthma and COPD despite the strong genetic correlation.

We also showed that applying novel Bayesian PRS construction methods like PRS-CSx and PRS-CS to association data from larger and more diverse cohorts can improve prediction in asthma, particularly for populations of non-European ancestry. However, we found that imbalances in the sample sizes of the discovery cohorts may need to be taken into careful consideration when using these methods. Previous studies have shown that imbalanced sample sizes across ancestries contribute somewhat unpredictably to varying prediction performances, with a low signal-to-noise ratio in ancestry-matched target populations reducing prediction performance (Majara et al., 2021). Therefore, further investigations are needed to fully understand the interplay between sample size and ancestry in the context of polygenic prediction. Ultimately, these analyses highlight the pressing need for more well-powered and ancestrally-diverse resources that will help reduce these imbalances.

We have highlighted the harmonization of the phenotype definitions across biobanks, but it is important to acknowledge that the criteria used, which allowed for both self-reported and PheCode information, are vulnerable to imprecision and variability in the data collected. Self-reported data for asthma is particularly susceptible to imprecision, since it relies on personal recollection of asthma diagnoses that are often given in childhood. On the other hand, PheCodes, which are based on ICD codes, may fail to capture diagnoses made earlier in the lifetime of individuals in hospital-based cohorts. Therefore, including both self-reported and PheCode data -- an approach adopted by some but not all biobanks -- may be optimal for association analyses for asthma. In the UKBB we found that the genetic correlation between the EUR GWAS based on the asthma PheCode and the EUR GWAS based on either PheCode or self-reported data was nearly 1 (r_g = 0.97). So, while we have demonstrated that the variation of phenotype definition used does not significantly influence the genetic association results in this case, we cannot confirm the same pattern for all biobanks in GBMI and especially for other diseases. However, given the relative alignment of genetic effects across the biobanks, we would expect that minor differences in phenotype definition would not substantially change the association results.

Additionally, we acknowledge that since the definitions used here for asthma and COPD do not exclude individuals with concurrent diagnoses, we are not able to fully distinguish the distinct biological pathways affecting asthma and COPD. Comorbidity rates of asthma and COPD reported in the literature range across studies but population-based estimates generally are low, around 2-3% (Hosseini et al., 2019; Kumbhare et al., 2016), while hospital-based prevalence estimates tend to be higher, around 13% (Akmatov et al., 2020). Among biobanks participating in GBMI, for example, 15.5% of all individuals with asthma in UKBB have a concurrent COPD diagnosis, 21% in BioVU, and 7.4% in BBJ. A previous study found that using stricter definitions of asthma, such as excluding subjects with COPD, resulted in stronger association signals for some of the asthma-associated loci (Han et al., 2020). However, it is important to note that this case exclusion would introduce an additional source of ascertainment bias. We also note that estimates of genetic correlation by LDSC are not biased by sample overlap (Bulik-Sullivan et al., 2015). In fact, this has been explored in the context of asthma and allergic diseases, where r_g estimates from LDSC were shown to be robust to overlapping cases and/or controls (Zhu et al., 2018).

We also recognize the importance of analyzing environmental factors in conjunction with genetic factors for a disease that is heavily influenced by the environment. Our genetic analyses offer insight into the potential shared biological pathways that may be differentially affected by non-genetic factors, but we were not able to explicitly investigate environmental effects given the lack of available environmental exposure data among the biobanks. The high degree of alignment among genetic associations, coupled with the large variability in asthma prevalence, points to a particularly important role of the environment for asthma risk across populations. Gaining a greater understanding of the specific non-genetic factors that contribute to asthma development in different environments may help guide more accurate disease prediction across populations.

This study, and importantly the data sharing across biobanks facilitated by this initiative, have laid the groundwork for deeper dives into the shared and distinct genetic signatures of asthma subtypes. Previous studies have categorized the UKBB individuals with asthma into childhood- versus adult-onset subtypes based on their ages at first diagnosis, discovering partly distinct genetic architectures (Ferreira et al., 2019; Zhu et al., 2020). Data from other biobanks in GBMI make it possible to perform similar stratifications and enable multiple downstream analyses. For example, future studies can evaluate the genetic overlap between subtypes, further validate previously reported subtype-specific variants in different populations, and test the power of PRS to discern different subtypes, empowered by the meta-analysis conducted here. Additionally, with access to biobanks with a wide range of phenotypes beyond the 14 analyzed in this initiative, we can begin to tease out the underlying biological relationships between asthma subtypes and other correlated phenotypes, particularly immune-related pulmonary diseases. Harnessing these cross-trait correlations for prediction may also be a fruitful approach to improving the accuracy of polygenic prediction models for asthma.

Methods

Asthma phenotype definitions for association analyses

The phenotype definition guidelines that were developed by GBMI and shared with all participating biobanks can be found in Zhou et al. (Zhou et al., 2021). Disease endpoints, including asthma, were defined following the PheCode maps, which maps ICD-9 or ICD-10 codes to PheCodes (Denny et al., 2013). Asthma cases were all study participants with the asthma PheCode, and controls were all study participants without the asthma PheCode (or asthma-related PheCodes). Biobanks that did not have ICD codes primarily used self-reported data (**Supplementary Table 3**).

Meta-analysis for asthma in GBMI

We performed fixed-effects meta-analysis with inverse variance weighting for 18 biobanks in GBMI: China Kadoorie Biobank (CKB), Generation Scotland (GS), Lifelines, QSKIN, East London Genes & Health (GNH), HUNT, UCLA Precision Health Biobank (UCLA), Colorado Center for Personalized Medicine (CCPM), Mass General Brigham (MGB), BioVU, BioMe, Michigan Genomics Initiative (MGI), BioBank Japan (BBJ), Estonian Biobank (ESTBB), deCODE Genetics (DECODE), FinnGen, Taiwan Biobank (TWB), and UK Biobank (UKBB). Basic information on the biobanks are described in Zhou et al., as well as details on the genotyping, imputation, GWAS, post-GWAS quality control, and meta-analysis procedures (Zhou et al., 2021). Genetic variants with minor allele count (MAC) < 20 and imputation score < 0.3 were excluded from the analyses. Altogether, these cohorts had a total sample size of 153,763 cases and 1,647,022 controls (**Supplementary Table 1**). GWAS meta-analyses were first conducted within continental ancestry groups to control for population stratification. 5,051 cases and 27,607 controls were of African (AFR) ancestry; 4,069 cases and 14,104 controls were of Admixed American (AMR) ancestry; 18,549 cases and 322,655 controls were of East Asian (EAS) ancestry; 121,940 cases and 1,254,131 were of European (EUR) ancestry; 139 cases and 1,434 controls were of Middle Eastern (MID) ancestry; and 4,015 cases and 27,091 controls were of Central and South Asian (CSA) ancestry.

Lead SNP and locus definitions

We used a threshold of $p < 5e-8$ to identify SNPs with a genome-wide significant effect. To identify lead variants, we used a window size of 500 kb upstream and downstream of the SNPs with the strongest evidence of association in the meta-analysis, and merged overlapping regions until no genome-wide significant variants were detected within the ± 500 kb region. To designate lead SNPs as previously discovered or potentially novel, we compiled a list of known

asthma-associated SNPs ($p < 5x10^{-8}$) from the associations collected by El-Husseini et al. (El-Husseini et al., 2020) and listed in the GWAS catalog (as of 11/14/2021) (Buniello et al., 2019). We extended 500 kb upstream and downstream of each of these variants to define a locus, and intersected these regions with the loci defined by the lead SNPs in our meta-analysis to identify any overlaps. We annotated genetic variants with the nearest genes using ANNOVAR (Wang et al., 2010) and putative loss-of-function using VEP (McLaren et al., 2016) with the LOFTEE plug (Karczewski et al., 2021) as implemented in Hail (Zhou et al., 2021). We also annotated whether the lead or tagging variants ($r^2 > 0.8$) of asthma were fine-mapped in any of the cis-eQTL fine-mapping resources. We retrieved cis-eQTL fine-mapped variants with posterior inclusion probability (PIP) > 0.9 in any tissues and cell types from the GTEx v8 (Aguet et al., 2020) and eQTL catalogue release 4 (Kerimov et al., 2021). Fine-mapping was conducted using SuSiE (Wang et al., 2020) with summary statistics and covariate-adjusted in-sample LD matrix as described previously (Kanai et al., 2021; Ulirsch and Kanai) (**Supplementary Table 2**).

Lead SNP effects across biobanks

To estimate the correlation of SNP effects for the 180 top-associated SNPs between one specific biobank and the leave-that-biobank-out meta-analysis, we used the method proposed by Qi et al. using GWAS summary statistics (Qi et al., 2018) (**Supplementary Table 5**). Specifically, the method directly calculates SNP effect correlation as:

$$\hat{r}_b = \frac{\hat{cov}(\hat{b}_{biobank}, \hat{b}_{leave-biobank})}{\sqrt{\left[\hat{var}\left(\hat{b}_{biobank}\right) - \hat{var}(e_{biobank})\right]\left[\hat{var}\left(\hat{b}_{leave-biobank}\right) - \hat{var}(e_{leave-biobank})\right]}},$$

where $\hat{b}_{biobank}$ and $\hat{b}_{leave-biobank}$ denote the estimated SNP effects from GWAS conducted in one specific biobank and from GWAS performed in the leave-that-biobank-out meta-analysis, respectively. The $\hat{cov}(\hat{b}_{biobank}, \hat{b}_{leave-biobank})$ is calculated as the sampling covariance between $\hat{b}_{biobank}$ and $\hat{b}_{leave-biobank}$. The $\hat{var}\left(\hat{b}_{biobank}\right)$ and $\hat{var}\left(\hat{b}_{leave-biobank}\right)$ are the estimated variances of $\hat{b}_{biobank}$ and $\hat{b}_{leave-biobank}$, separately. The $\hat{var}(e_{biobank})$ and $\hat{var}(e_{leave-biobank})$ are the estimated variance of the estimation errors of $\hat{b}_{biobank}$ and $\hat{b}_{leave-biobank}$, which are approximated as the mean of the squared standard errors of estimated SNP effect ($\hat{b}_{biobank}$ and $\hat{b}_{leave-biobank}$) across all the top-associated SNPs, respectively. The standard error of $\hat{r}_b$ is obtained through the jackknife approach by leaving one SNP out each time. SNPs with large standard errors in CKB and HUNT (chr12:123241280:T:C and chr17:7878812:T:C, respectively) were excluded from these analyses.

Then, for the lead SNPs present in each biobank, we computed:

$$\frac{biobank\ meta-analysis\ effect\ size}{leave-that-biobank-out\ meta-analysis\ effect\ size}$$

for the biobank and leave-that-biobank-out pair. We took the average ratio across the lead SNPs for each biobank and leave-that-biobank-out pair. We then used the regression method introduced in Deming et al., which considers the errors in both the *X*- and *Y*-variables, to compare the effect sizes of these SNPs in each biobank GWAS with their effects in the leave-that-biobank-out meta-analysis (Deming, 1943). We set the intercept equal to 0 for these analyses.

Ancestry-specific heterogeneity analyses

To assess heterogeneity of per-SNP effect sizes for the 180 lead SNPs across ancestries in GBMI, we conducted ancestry-specific meta-analyses of the five most well-powered ancestry groups in GBMI (EUR, AFR, AMR, EAS, and CSA). We applied the Cochran's Q test to the SNP effects in the ancestry-specific meta-analyses and identified SNPs with significant heterogeneity based on a Bonferroni-corrected p-value cut-off of 0.05/169 = 0.0003, accounting for the number of SNPs present in all studies (**Supplementary Table 6**). Regions displaying heterogeneity in effects across ancestry groups were visualized using the LocalZoom tool (Boughton et al., 2021).

Polygenic risk score analyses

A description of the PRS analyses conducted using PRS-CS (Ge et al., 2019), as well as the leave-one-biobank-out meta-analysis strategy applied, is provided in Wang et al. (Wang et al., 2021).

We used PRS-CSx, which jointly models GWAS summary statistics from populations of different ancestries and returns posterior SNP effect size estimates for each input population (Ruan et al., 2021). We applied this method to the AMR, AFR, CSA, EAS, and EUR ancestry-specific meta-analyses, which served as the discovery data for PRS construction. For the ancestry-specific meta-analysis that matched the ancestry of the target cohort, we excluded the target cohort. We evaluated the predictive performance of the PRS in 4 target cohorts: 1) AFR ancestry individuals in UKBB (849 cases, 5190 controls), 2) CSA ancestry individuals in UKBB (1232 cases, 6744 controls), 3) EAS ancestry individuals in BBJ that were part of a randomly-selected 1k holdout set (500 cases, 500 controls), and 4) EUR ancestry individuals in UKBB that were part of a randomly-selected 10k holdout set (1164 cases, 7577 controls). As an example, for the AFR ancestry individuals, the full set of discovery data for PRS construction consisted of the AMR, CSA, EAS, and EUR ancestry-specific meta-analyses, as well as the AFR ancestry-specific meta-analysis excluding the AFR ancestry individuals in UKBB. The same strategy was applied to the other 3 target cohorts (**Supplementary Fig. 5**). We used ancestry-matched LD reference panels from UKBB data and the default PRS-CSx settings for all input parameters. We evenly and randomly split cases and controls in the target cohorts into validation and testing subsets. Using the posterior SNP effect size estimates from PRS-CSx, we computed one PRS per discovery population for the validation subsets to learn the optimal

linear combination of the ancestry-specific PRS using PLINK v1.9 (Chang et al., 2015; Purcell and Chang). Then, with these weights, we evaluated the prediction accuracy of this linear combination of PRS in the testing subset. We reported the average prediction accuracy, measured by variance explained on the liability scale ($R^2_{liability}$), estimated using the prevalence of asthma in the BBJ biobank for the EAS target cohort and in the UKBB biobank for the other target cohorts, across 100 random splits.

Genetic correlation analyses in UKBB and BBJ

We used linkage-disequilibrium score correlation (LDSC) to compute pairwise genetic correlations (r_g) (Bulik-Sullivan et al., 2015). We estimated r_g between all EUR-ancestry UKBB phenotypes with heritability Z-score > 4 and (1) the GBMI leave-UKBB-out meta-analysis for asthma and (2) the UKBB EUR-ancestry GWAS of asthma (PheCode ID 495 in UKBB). The heritability Z-scores were obtained from the stratified-LDSC (S-LDSC) computations of heritability reported by the Pan-UK Biobank team (Finucane et al., 2015, 2018; Pan UKBB). Summary statistics from the UKBB EUR GWAS were obtained from the Pan-UK Biobank team as well (Pan UKBB).

We also used LDSC (Bulik-Sullivan et al., 2015) to compute r_g between 48 phenotypes in BioBank Japan (BBJ) and (1) the GBMI leave-BBJ-out meta-analysis for asthma and (2) the BBJ GWAS of asthma. We used publicly available GWAS summary statistics for all traits (Akiyama et al., 2017; Ishigaki et al., 2020; Kanai et al., 2018). Genetic correlation results were visualized using the R corrplot package (Wei and Simko, 2021).

Gene- and pathway-based enrichment analyses for asthma and COPD

Fixed-effects meta-analysis with inverse variance weighting was also performed for 16 biobanks in GBMI with COPD data: BBJ, BioMe, BioVU, CCPM, CKB, ESTBB, FinnGen, GNH, GS, HUNT, Lifelines, MGB, MGI, TWB, UCLA, and UKBB. The same processing and methods were applied here as for the asthma meta-analysis. These cohorts had a total sample size of 81,568 cases and 1,310,798 controls. COPD cases were defined based on the COPD PheCode, and controls were all study participants without the COPD or COPD-related PheCodes. Biobanks that did not have ICD codes available used spirometry data (Lifelines) or self-reported data (TWB). Details can be found in Zhou et al. (Zhou et al., 2021). Meta-analyses were also conducted within continental ancestry groups: 19,044 cases and 310,689 controls of EAS ancestry, 1,978 cases and 27,704 controls of AFR ancestry, and 58,559 cases and 937,358 controls of EUR ancestry.

We used MAGMA v1.09b for gene prioritization and gene-set enrichment analyses, applying this method to the GBMI asthma EUR, AFR, EAS, and CSA ancestry-specific meta-analyses and the GBMI COPD EUR, AFR, and EAS ancestry-specific meta-analyses (de Leeuw et al., 2015).

For the gene-level analyses in MAGMA, we first mapped the SNPs to the provided list of genes using a window size of 20kb, and then performed gene analysis using the ancestry-matched 1000G LD reference panels to account for LD structure. Gene-set enrichment was performed using the default settings to correct for gene length, gene density, and the inverse mean minor allele count. The gene sets used were the c2, “curated gene sets,” and c5, “ontology gene sets,” obtained from the Molecular Signatures Database v7.4 (Subramanian et al., 2005).

Data and Code Availability

The all-biobank GWAS summary statistics are publicly available for downloading at https://www.globalbiobankmeta.org/resources and can be browsed at the PheWeb Browser (http://results.globalbiobankmeta.org). Custom scripts used for quality control, meta-analysis, and loci definition are available at https://github.com/globalbiobankmeta. Other analyses utilized publicly available tools: the R deming package for Deming regression (Therneau, 2018), PRS-CSx for polygenic prediction (https://github.com/getian107/PRScsx), LDSC for genetic correlation (https://github.com/bulik/ldsc), and MAGMA v1.09b for gene-set enrichment (https://ctg.cncr.nl/software/magma).

Acknowledgements

We acknowledge helpful comments from Cristen Willer, Hailiang Huang, Yunfeng Ruan, and Tian Ge. A.R.M is funded by the K99/R00MH117229. S.N. is supported by Takeda Science Foundation. Y.O. is 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. R.G. is supported by the T32AG000222.

Author Contributions

Study design: K.T., A.R.M., M.J.D., B.M.N.
Data collection/contribution: W.Z., Y.O., T.M.
Data analysis: K.T., W.Z., Y.W., M.K., S.N., R.G., L.M., L.N.
Writing: K.T., A.R.M., S.N., R.G.
Revision: K.T., A.R.M., Y.W., R.G., M.J.D.

Declaration of Interests

M.J.D. is a founder of Maze Therapeutics. B.M.N. is a member of the scientific advisory board at Deep Genomics and consultant for Camp4 Therapeutics, Takeda Pharmaceutical, and Biogen.

Figures

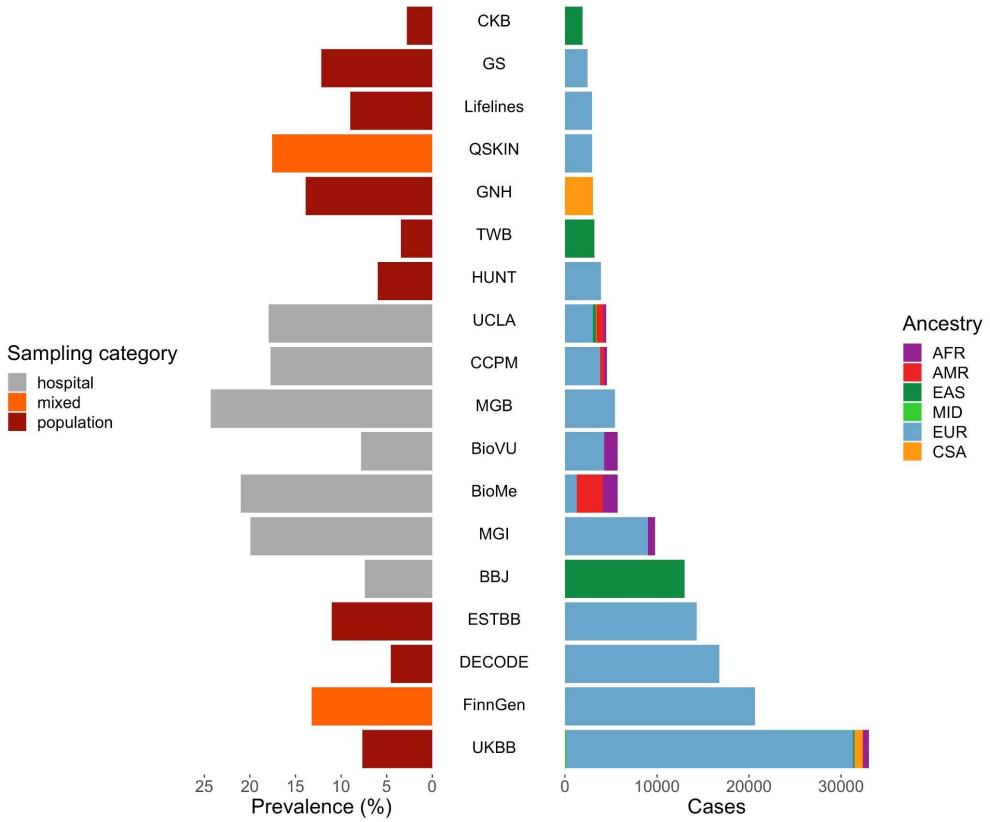


Figure 1 | 18 biobanks in GBMI contributing GWAS of asthma. Distribution of prevalence of asthma on left and number of cases of asthma on right across biobanks in GBMI. Biobanks span different sampling approaches and ancestries (AFR = African; AMR = Admixed American; EAS = East Asian; MID = Middle Eastern; EUR = European; CSA = Central and South Asian).

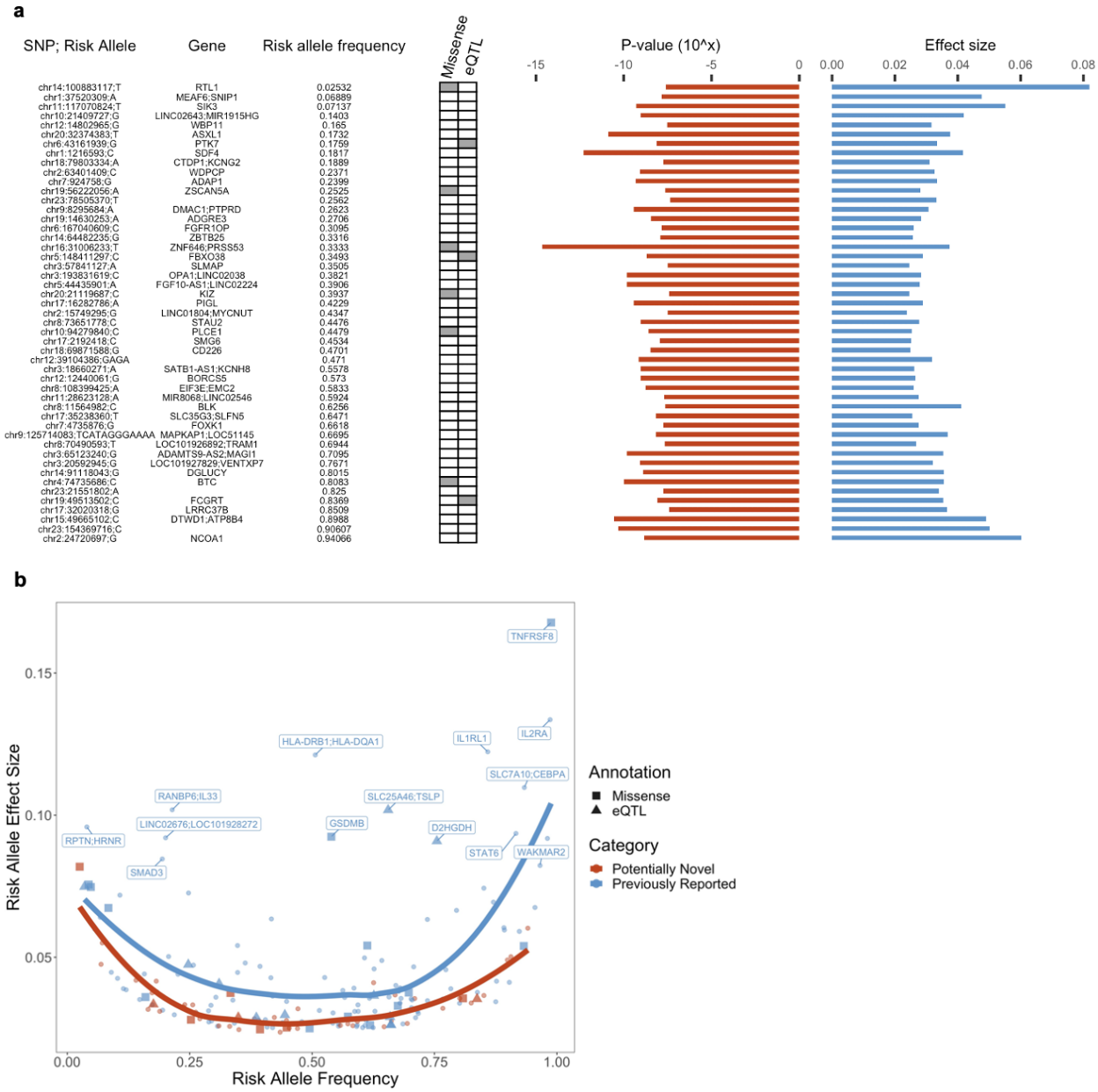


Figure 2 | Lead variants associated with asthma. a, 49 asthma lead variants that are potentially novel. VEP was used to identify missense variants that overlapped with an index or tagging variant (r2 > 0.8). cis-eQTLs fine-mapped with PIP > 0.9 that overlapped with an index or tagging variant are also annotated here. Frequency of risk allele and effect size estimate in GBMI meta-analysis are shown on the right. **b,** Frequency and effect size of risk alleles of all 180 lead variants, categorized by whether they are potentially novel or previously reported. Missense variants and those overlapping with a fine-mapped cis-eQTL are highlighted, as well as previously reported genes with large effect sizes.

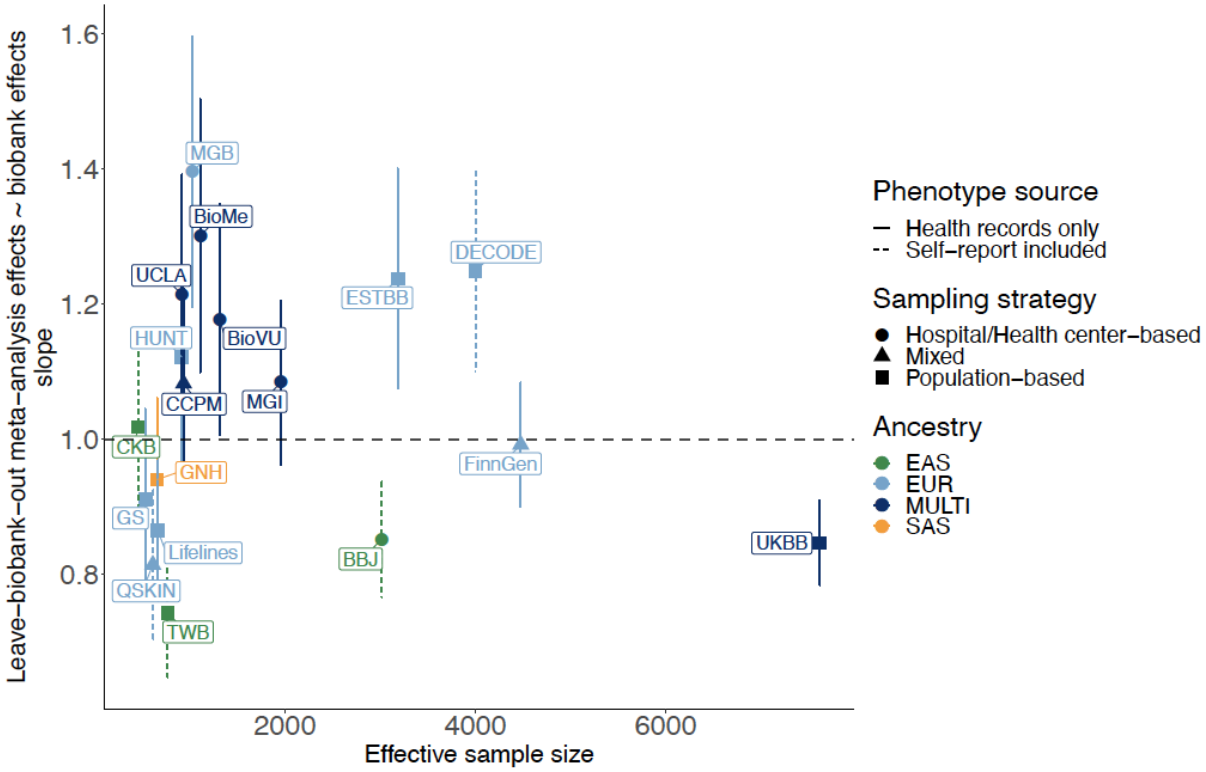


Figure 3 | Consistency of lead variants across biobanks. Regression slopes computed using the Deming regression method, which compared effects of index variants in each biobank GWAS against their effects in the corresponding leave-that-biobank-out meta-analysis (Deming 1943). The x-axis shows the effective sample size of each biobank, computed as 4/(1/cases + 1/controls).

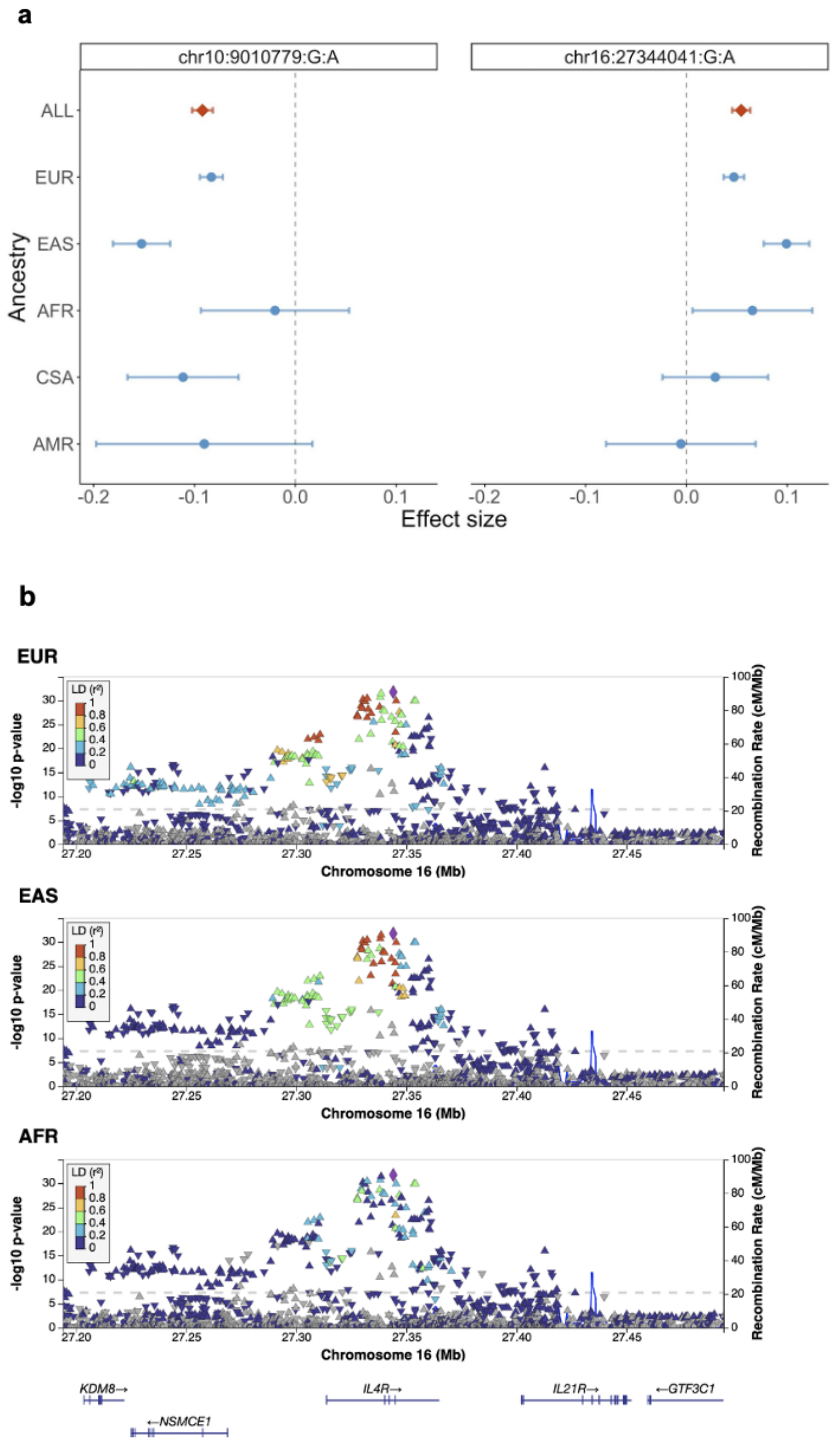


Figure 4 | Lead variants showing heterogeneity in ancestry-specific effect sizes. a, The index variants with the most significant $p_{\text{Cochran's Q}}$. Effect sizes of these variants in each ancestry-specific meta-analysis are shown here. **b,** LocusZoom plots showing the association of chr16:27344041:G:A (purple symbol) and variants within 150kb upstream and downstream of this variant with asthma. Color

coding of other SNPs indicates LD with this SNP. EUR, EAS, and AFR indicate the population from which LD information was estimated.

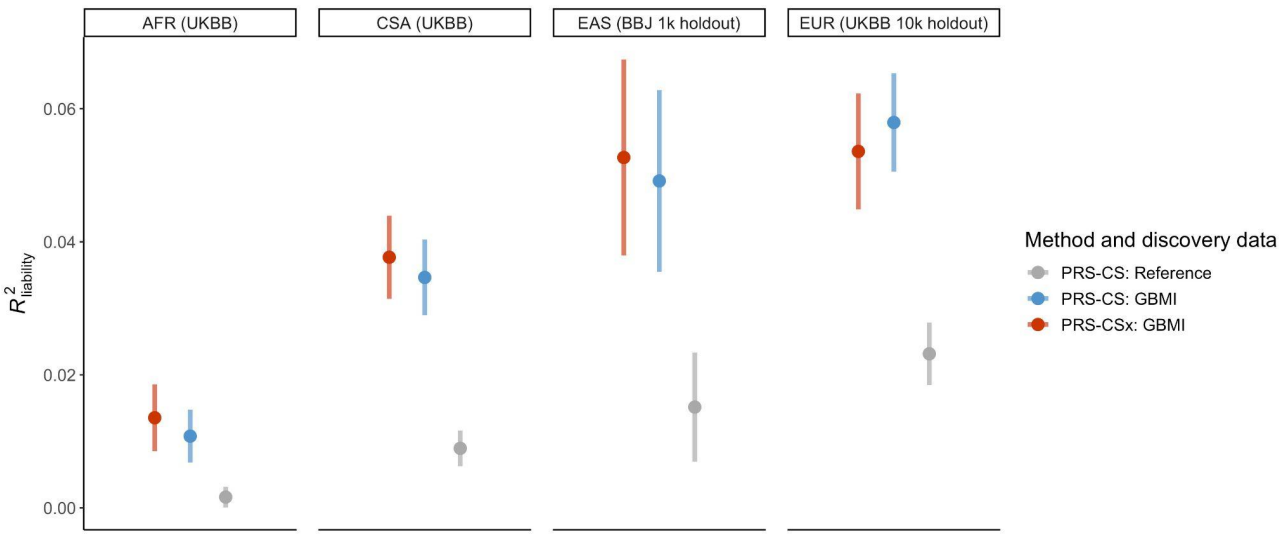
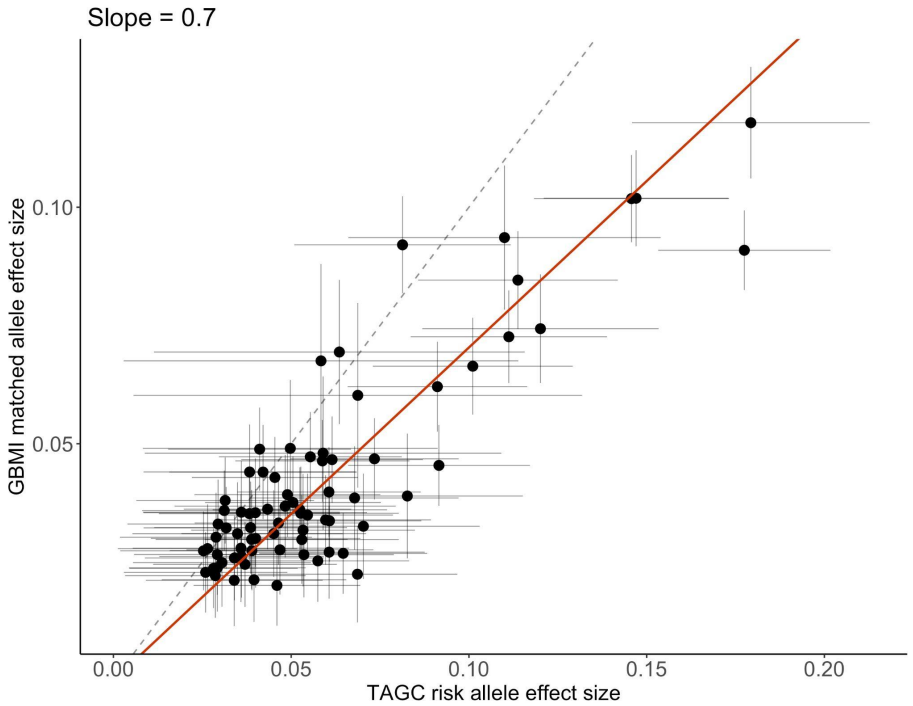


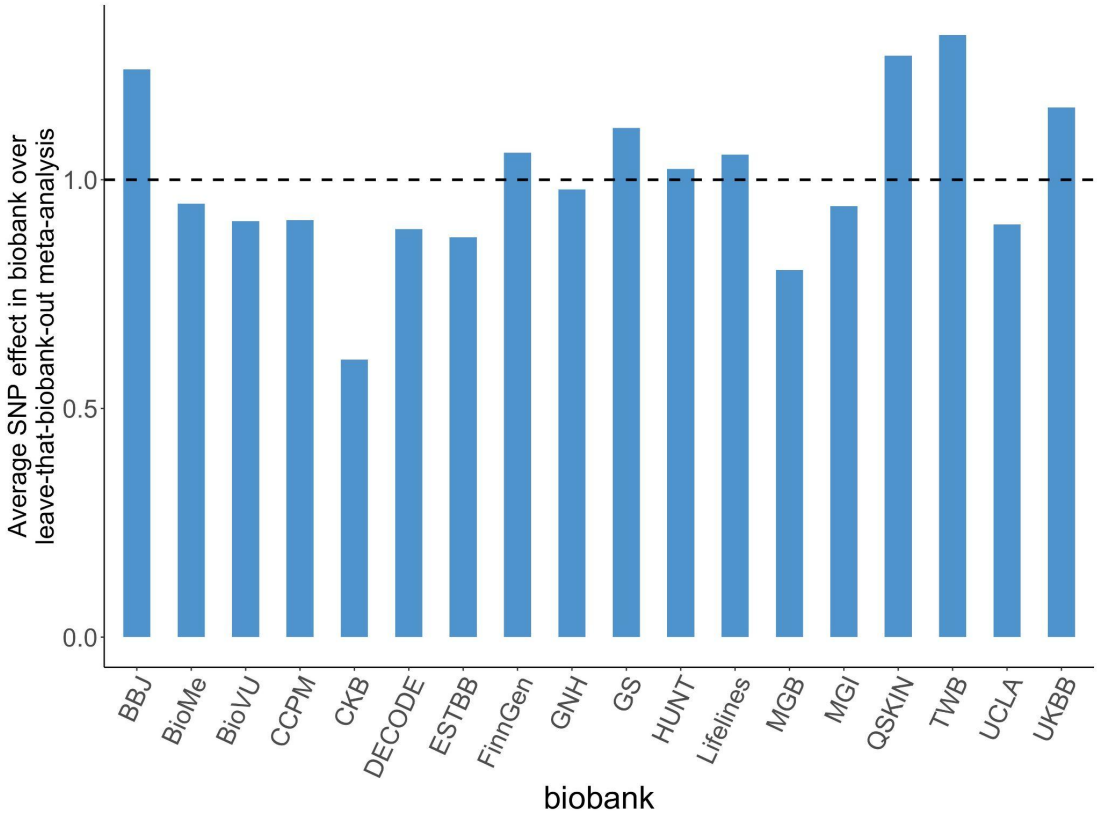
Figure 5 | PRS performance across ancestries. Each panel represents a target cohort in which PRS constructed using PRS-CSx and PRS-CS were evaluated. PRS-CS results are from Wang et al. (Wang et al. 2021), which used the GBMI leave-BBJ-out meta-analysis and GBMI leave-UKBB-out meta-analysis as discovery data for the BBJ and all UKBB target cohorts, respectively (**Supplementary Table 14**). The reference dataset used in the PRS-CS analyses was the TAGC meta-analysis (Demenais et al. 2018). Sample sizes for the target cohorts are: cases=849 and controls=5190 for AFR; cases=500 and controls=500 for EAS; cases=1164 and controls=7577 for EUR; cases=1232 and controls=6744 for SAS. Error bars represent standard deviation of R^2 on the liability scale across 100 replicates.

Supplementary Information

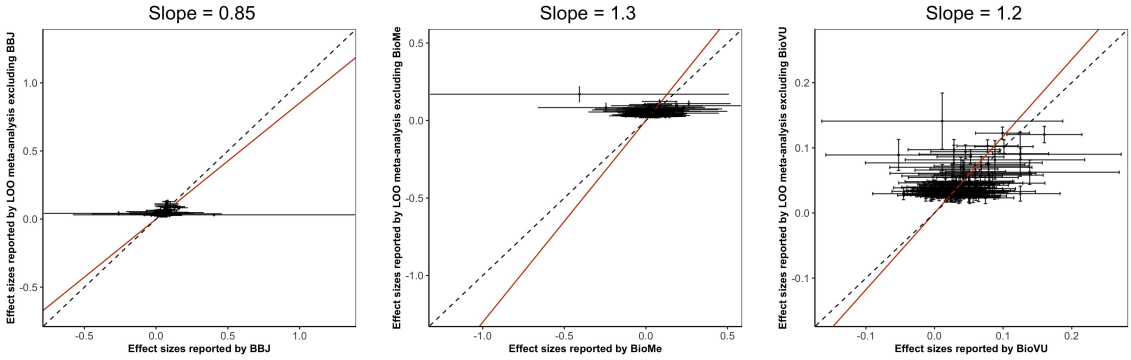
Supplementary Figures

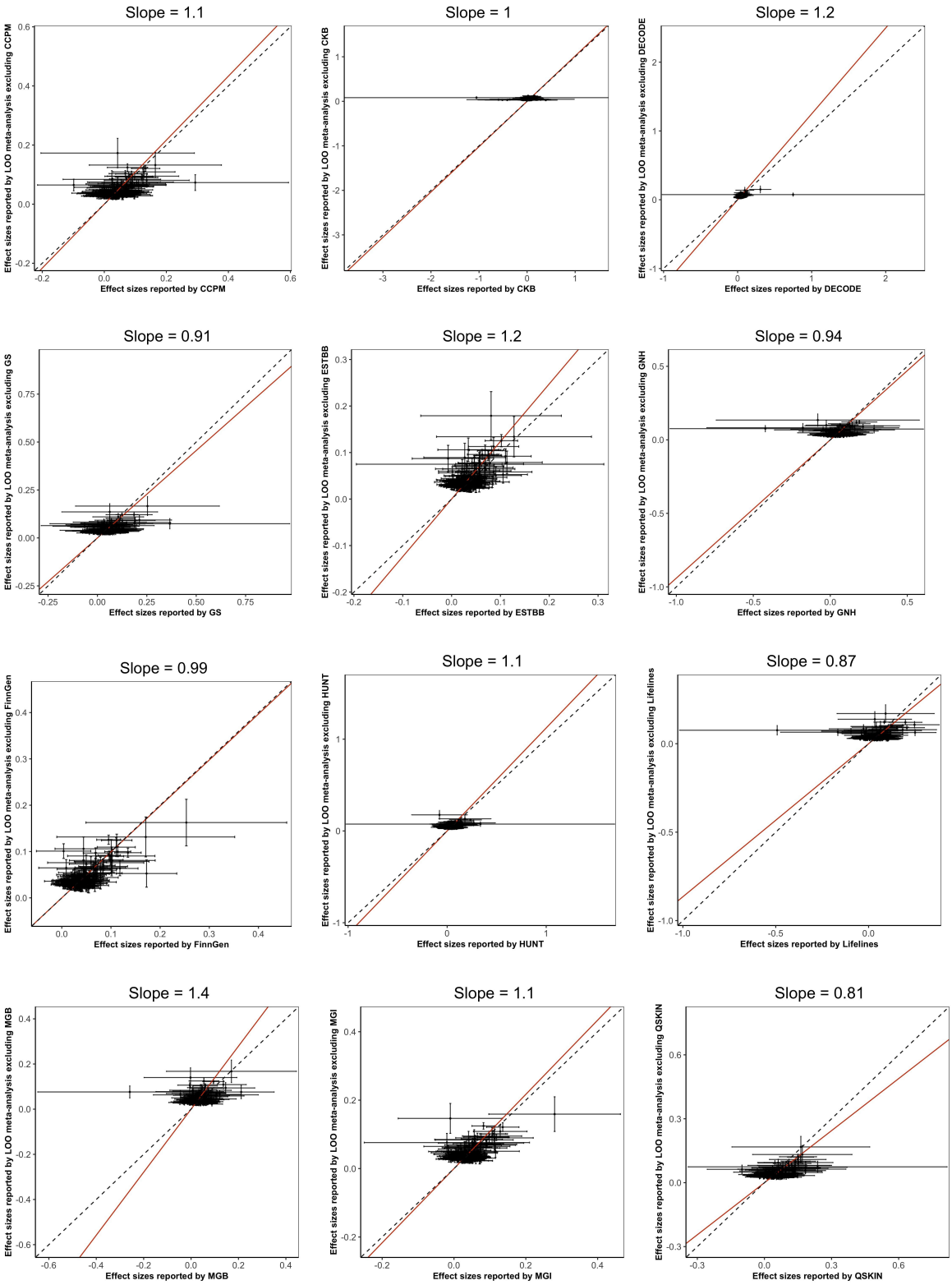


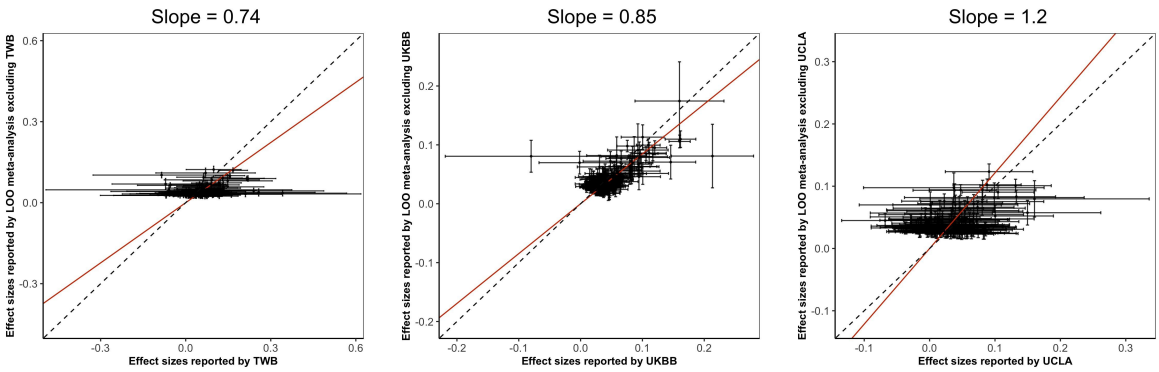
Supplementary Figure 1 | GBMI lead variants in TAGC. 76 of the 180 lead variants associated with asthma discovered in the GBMI meta-analysis were found in the TAGC meta-analysis of asthma, or had a tagging variant (r2 > 0.8) in the TAGC study, with a p-value < 0.05 (Demenais et al., 2018). The effect sizes of these 76 variants as estimated in the TAGC vs. GBMI meta-analyses were compared using the Deming regression method (Deming 1943). The intercept was set to be 0; the slope estimated from the regression analysis is reported here.



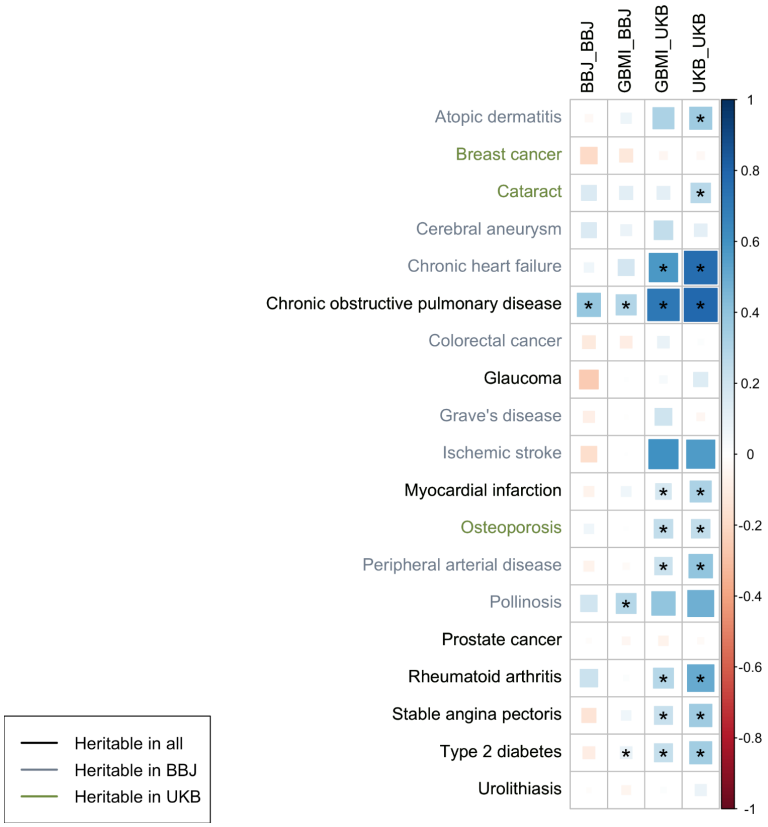
Supplementary Figure 2 | Consistency of asthma lead variants across biobanks. For each biobank shown on the x-axis, we computed the average ratio of effect sizes of the index variants in the biobank vs. in the corresponding leave-that-biobank-out meta-analysis.



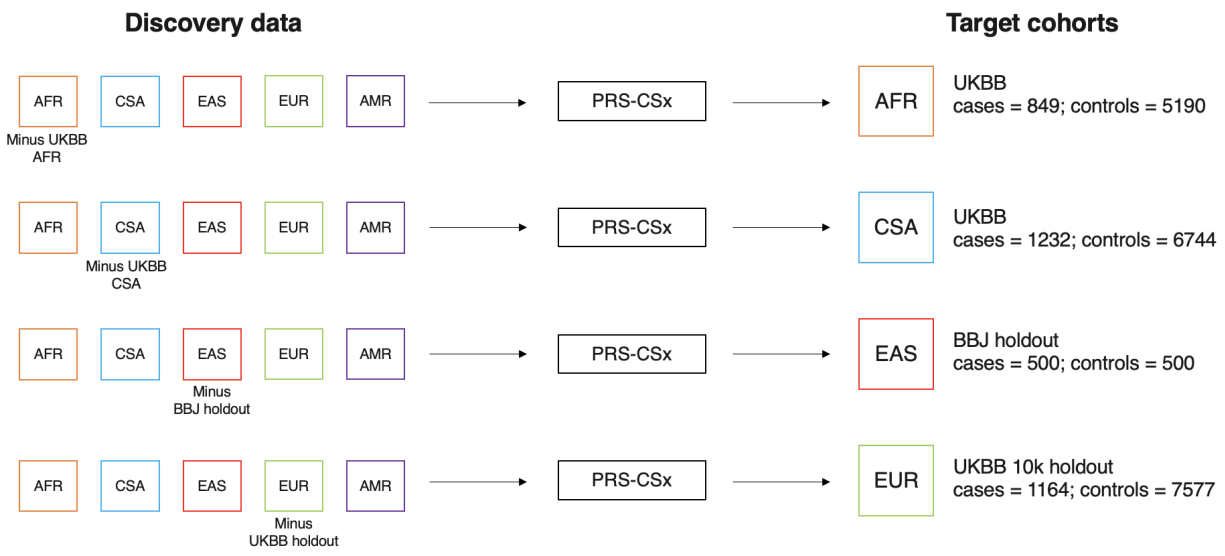

Slope = 1.1
Effect sizes reported by LOO meta-analysis excluding CCPM
Effect sizes reported by CCPM
Slope = 1
Effect sizes reported by LOO meta-analysis excluding CKB
Effect sizes reported by CKB
Slope = 1.2
Effect sizes reported by LOO meta-analysis excluding DECODE
Effect sizes reported by DECODE
Slope = 0.91
Effect sizes reported by LOO meta-analysis excluding GS
Effect sizes reported by GS
Slope = 1.2
Effect sizes reported by LOO meta-analysis excluding ESTBB
Effect sizes reported by ESTBB
Slope = 0.94
Effect sizes reported by LOO meta-analysis excluding GNH
Effect sizes reported by GNH
Slope = 0.99
Effect sizes reported by LOO meta-analysis excluding FinnGen
Effect sizes reported by FinnGen
Slope = 1.1
Effect sizes reported by LOO meta-analysis excluding HUNT
Effect sizes reported by HUNT
Slope = 0.87
Effect sizes reported by LOO meta-analysis excluding Lifelines
Effect sizes reported by Lifelines
Slope = 1.4
Effect sizes reported by LOO meta-analysis excluding MGB
Effect sizes reported by MGB
Slope = 1.1
Effect sizes reported by LOO meta-analysis excluding MGI
Effect sizes reported by MGI
Slope = 0.81
Effect sizes reported by LOO meta-analysis excluding QSKIN
Effect sizes reported by QSKIN



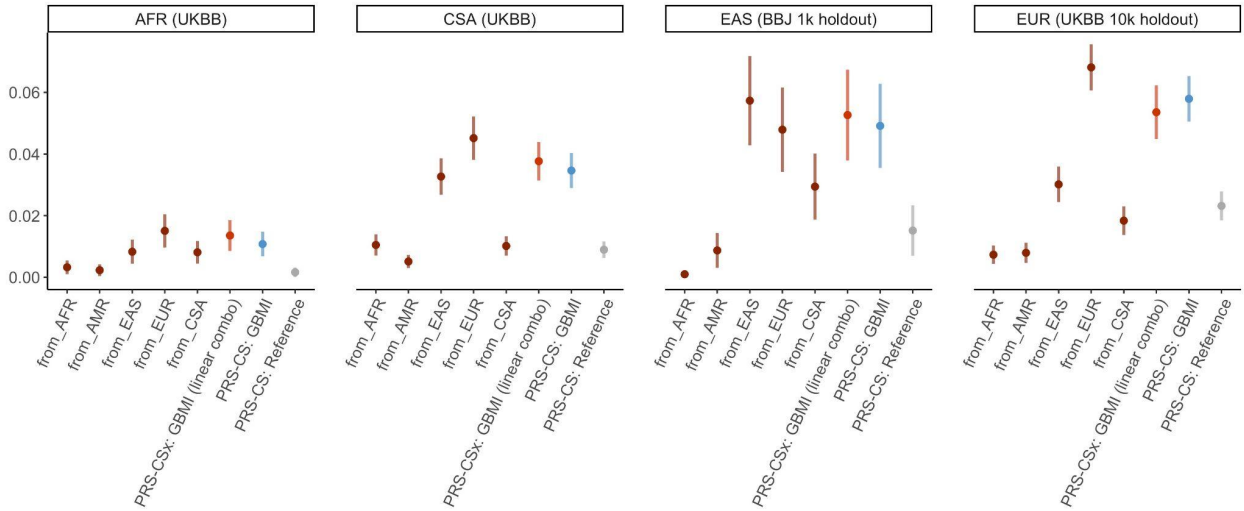
Supplementary Figure 3 | Consistency of asthma lead variants across biobanks using Deming regression. The effect sizes of the asthma lead variants as estimated in each biobank GWAS vs. in the corresponding leave-that-biobank-out meta-analysis were compared using the Deming regression method (Deming 1943). Intercepts were set to be 0; slopes from the regression analyses are reported here.



Supplementary Figure 4 | Genetic correlations between asthma and heritable diseases across UKBB and BBJ. Genetic correlations between asthma and diseases that were heritable in BBJ, UKBB EUR, or both. On x-axis: BBJ_BBJ = BBJ GWAS of asthma vs. BBJ GWAS of diseases on y-axis; GBMI_BBJ = GBMI-excluding-BBJ meta-analysis of asthma vs. BBJ GWAS of diseases on y-axis; GBMI_UKB = GBMI-excluding-UKB meta-analysis of asthma vs. UKB GWAS of diseases (EUR only) on y-axis; UKB_UKB = UKB GWAS of asthma vs. UKB GWAS of diseases (EUR only) on y-axis



Supplementary Figure 5 | Workflow for PRS-CSx analyses. The discovery data consisted of ancestry-specific meta-analyses, indicated by the squares on the left, that were inputs for PRS-CSx (Ruan et al. 2021). PRS-CSx returnsed separate sets of posterior effect size estimates for each input dataset, which were then used to construct PRS. The target cohorts were randomly evenly split; optimal weights for the linear combination of the PRS were selected in one subset and the linear combination of the PRS was evaluated in the other subset.



Supplementary Figure 6 | PRS performance of individual PRS vs. linear combination of PRS using PRS-CSx across ancestries. Each panel represents a target cohort. The performance of the individual PRS, computed from a single set of posterior effect size estimates corresponding to each input ancestry population from PRS-CSx, is plotted here. The prediction accuracy of the linear combination of the PRS from PRS-CSx, as well as the PRS from the PRS-CS analyses (shown in Fig. 5), are also plotted for comparison. PRS-CS results are from Wang et al. (Wang et al. 2021), which used the GBMI leave-BBJ-out meta-analysis and GBMI leave-UKBB-out meta-analysis as discovery data for the BBJ and

all UKBB target cohorts, respectively. The reference dataset used in the PRS-CS analyses was the TAGC meta-analysis (Demenais et al. 2018). Error bars represent standard deviation of R^2 on the liability scale across 100 replicates.

Supplementary Tables

STable 1: Description of 18 biobanks in GBMI that contributed summary statistics for asthma meta-analysis with sample size, ancestry, and recruitment strategy information.
STable 2: 180 lead variants discovered by GBMI asthma meta-analysis with annotations for nearby genes, missense variants, and fine-mapped cis-eQTLs.
STable 3: Description of asthma definition used by each biobank.
STable 4: 122 of the 180 lead variants present or with tagging variant in the TAGC study with effect sizes and p-values from GBMI and TAGC meta-analyses.
STable 5: Correlations of SNP effects of the 180 lead variants between each biobank and the corresponding leave-that-biobank-out meta-analysis.
STable 6: Heterogeneity p-values, computed from Cochran's Q statistic, for lead variants using effect sizes from the AFR, AMR, EAS, EUR, and CSA meta-analyses.
STable 7: 46 lead variants discovered by GBMI COPD meta-analysis with corresponding effect sizes, standard errors, and p-values from asthma meta-analysis.
STable 8: Genetic correlations estimated by LDSC between GBMI leave-UKBB-out meta-analysis and UKBB EUR GWAS of heritable phenotypes with phenotype descriptions.
STable 9: Genetic correlations estimated by LDSC between GBMI asthma meta-analyses and UKBB and BBJ GWAS of several diseases.
STable 10: Results from MAGMA gene analysis using GBMI EAS, CSA, and EUR asthma meta-analyses. Genes with p-values < Bonferroni-corrected p-value thresholds are reported.
STable 11: Results from MAGMA gene analysis using GBMI EAS and EUR COPD meta-analysis. Genes with p-value < Bonferroni-corrected p-value thresholds are reported.
STable 12: Results from MAGMA gene-set enrichment analyses for asthma. Gene sets with FDR < 0.05 are reported.
STable 13: Results from MAGMA gene-set enrichment analyses for COPD. Gene sets with FDR < 0.05 are reported.
STable 14: Description of discovery data used for PRS-CSx and PRS-CS analyses.

References

Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., Graubert, A., Hadley, K., Handsaker, R.E., Huang, K.H., Kashin, S., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science.

Akinbami, L.J., Moorman, J.E., Bailey, C., Zahran, H.S., King, M., Johnson, C.A., and Liu, X. (2012). Trends in asthma prevalence, health care use, and mortality in the United States, 2001-2010. NCHS Data Brief 1–8.

Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467.

Akmatov, M.K., Ermakova, T., Holstiege, J., Steffen, A., von Stillfried, D., and Bätzing, J. (2020). Comorbidity profile of patients with concurrent diagnoses of asthma and COPD in Germany. Sci. Rep. *10*, 17945.

Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354.

Asher, M.I., García-Marcos, L., Pearce, N.E., and Strachan, D.P. (2020). Trends in worldwide asthma prevalence. Eur. Respir. J. *56*.

Battle, N.C., Choudhry, S., Tsai, H.-J., Eng, C., Kumar, G., Beckman, K.B., Naqvi, M., Meade, K., Watson, H.G., LeNoir, M., et al. (2007). Ethnicity-specific Gene–Gene Interaction between IL-13 and IL-4Rα among African Americans with Asthma. Am. J. Respir. Crit. Care Med. *175*, 881–887.

Borish, L., and Culp, J.A. (2008). Asthma: a syndrome composed of heterogeneous diseases. Ann. Allergy Asthma Immunol. *101*, 1–8; quiz 8–11, 50.

Boughton, A.P., Welch, R.P., Flickinger, M., VandeHaar, P., Taliun, D., Abecasis, G.R., and Boehnke, M. (2021). LocusZoom.js: Interactive and embeddable visualization of genetic association study results. Bioinformatics.

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

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198–1213.e14.

Dalakas, M.C., and Spaeth, P.J. (2021). The importance of FcRn in neuro-immunotherapies: From IgG catabolism, FCGRT gene polymorphisms, IVIg dosing and efficiency to specific FcRn inhibitors. Ther. Adv. Neurol. Disord. *14*, 1756286421997381.

Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat. Genet. *50*, 42–53.

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

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

Dharmage, S.C., Perret, J.L., and Custovic, A. (2019). Epidemiology of Asthma in Children and Adults. Front Pediatr *7*, 246.

El-Husseini, Z.W., Gosens, R., Dekker, F., and Koppelman, G.H. (2020). The genetics of asthma and the promise of genomics-guided drug target discovery. Lancet Respir Med *8*, 1045–1056.

Ferreira, M.A.R., Mathur, R., Vonk, J.M., Szwajda, A., Brumpton, B., Granell, R., Brew, B.K., Ullemar, V., Lu, Y., Jiang, Y., et al. (2019). Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. Am. J. Hum. Genet. *104*, 665–684.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.

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

Han, Y., Jia, Q., Jahani, P.S., Hurrell, B.P., Pan, C., Huang, P., Gukasyan, J., Woodward, N.C., Eskin, E., Gilliland, F.D., et al. (2020). Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. Nat. Commun. *11*, 1776.

Hobbs, B.D., de Jong, K., Lamontagne, M., Bossé, Y., Shrine, N., Artigas, M.S., Wain, L.V., Hall, I.P., Jackson, V.E., Wyss, A.B., et al. (2017). Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. Nat. Genet. *49*, 426–432.

Hosseini, M., Almasi-Hashiani, A., Sepidarkish, M., and Maroufizadeh, S. (2019). Global prevalence of asthma-COPD overlap (ACO) in the general population: a systematic review and meta-analysis. Respir. Res. *20*, 229.

Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C.A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature *547*, 173–178.

International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature *478*, 103–109.

Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., Sakaue, S., Matoba, N., Low, S.-K., Okada, Y., et al. (2020). Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat. Genet. *52*, 669–679.

Jefferson, J.A., and Shankland, S.J. (2007). Familial nephrotic syndrome: PLCE1 enters the fray. Nephrol. Dial. Transplant *22*, 1849–1852.

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.

Kanai, M., Ulirsch, J.C., Karjalainen, J., Kurki, M., Karczewski, K.J., Fauman, E., Wang, Q.S., Jacobs, H., Aguet, F., Ardlie, K.G., et al. (2021). Insights from complex trait fine-mapping across diverse populations.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2021). Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *590*, E53.

Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M.P., Kuzmin, I., Trevanion, S.J., et al. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. *53*, 1290–1299.

Kousha, A., Mahdavi Gorabi, A., Forouzesh, M., Hosseini, M., Alexander, M., Imani, D., Razi, B., Mousavi, M.J., Aslani, S., and Mikaeili, H. (2020). Interleukin 4 gene polymorphism (-589C/T) and the risk of asthma: a meta-analysis and met-regression based on 55 studies. BMC Immunol. *21*, 55.

Kumbhare, S., Pleasants, R., Ohar, J.A., and Strange, C. (2016). Characteristics and Prevalence of Asthma/Chronic Obstructive Pulmonary Disease Overlap in the United States. Ann. Am. Thorac. Soc. *13*, 803–810.

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

Li, D., Achkar, J.-P., Haritunians, T., Jacobs, J.P., Hui, K.Y., D'Amato, M., Brand, S., Radford-Smith, G., Halfvarson, J., Niess, J.-H., et al. (2016). A Pleiotropic Missense Variant in SLC39A8 Is Associated With Crohn's Disease and Human Gut Microbiome Composition. Gastroenterology *151*, 724–732.

Loo, T.H., Ye, X., Chai, R.J., Ito, M., Bonne, G., Ferguson-Smith, A.C., and Stewart, C.L. (2019). The mammalian LINC complex component SUN1 regulates muscle regeneration by modulating drosha activity. Elife *8*.

Majara, L., Kalungi, A., Koen, N., Zar, H., Stein, D.J., Kinyanda, E., Atkinson, E.G., and Martin, A.R. (2021). Low generalizability of polygenic scores in African populations due to genetic and environmental diversity.

Maselli, D.J., and Hanania, N.A. (2018). Asthma COPD overlap: Impact of associated comorbidities. Pulm. Pharmacol. Ther. *52*, 27–31.

Massoud, A.H., Charbonnier, L.-M., Lopez, D., Pellegrini, M., Phipatanakul, W., and Chatila, T.A. (2016). An asthma-associated IL4R variant exacerbates airway inflammation by promoting conversion of regulatory T cells to TH17-like cells. Nat. Med. *22*, 1013–1022.

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

Nakata, T., Creasey, E.A., Kadoki, M., Lin, H., Selig, M.K., Yao, J., Lefkovith, A., Daly, M.J., Graham, D.B., and Xavier, R.J. (2020). A missense variant in SLC39A8 confers risk for Crohn's disease by disrupting manganese homeostasis and intestinal barrier integrity. Proc. Natl. Acad. Sci. U. S. A. *117*, 28930–28938.

Nebert, D.W., and Liu, Z. (2019). SLC39A8 gene encoding a metal ion transporter: discovery and bench to bedside. Hum. Genomics *13*, 51.

Nie, W., Zhu, Z., Pan, X., and Xiu, Q. (2013). The interleukin-4 −589C/T polymorphism and the risk of asthma: A meta-analysis including 7345 cases and 7819 controls. Gene *520*, 22–29.

Ober, C., and Yao, T.-C. (2011). The genetics of asthma and allergic disease: a 21st century perspective. Immunol. Rev. *242*, 10–30.

Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.

Postma, D.S., and Rabe, K.F. (2015). The Asthma–COPD Overlap Syndrome. N. Engl. J. Med. *373*, 1241–1249.

Purcell, S., and Chang, C. PLINK 1.9.

Qi, T., Wu, Y., Zeng, J., Zhang, F., Xue, A., Jiang, L., Zhu, Z., Kemper, K., Yengo, L., Zheng, Z., et al. (2018). Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nat. Commun. *9*, 2282.

Riar, S.S., Banh, T.H.M., Borges, K., Subbarao, P., Patel, V., Vasilevska-Ristovska, J., Chanchlani, R., Hussain-Shamsy, N., Noone, D., Hebert, D., et al. (2019). Prevalence of Asthma and Allergies and Risk of Relapse in Childhood Nephrotic Syndrome: Insight into Nephrotic Syndrome Cohort. J. Pediatr. *208*, 251–257.e1.

Ruan, Y., Lin, Y.-F., Feng, Y.-C.A., Chen, C.-Y., Lam, M., Guo, Z., He, L., Sawa, A., Martin, A.R., Qin, S., et al. (2021). Improving Polygenic Prediction in Ancestrally Diverse Populations.

Sakornsakolpat, P., Prokopenko, D., Lamontagne, M., Reeve, N.F., Guyatt, A.L., Jackson, V.E., Shrine, N., Qiao, D., Bartz, T.M., Kim, D.K., et al. (2019). Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations.

Nat. Genet. *51*, 494–505.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

Sembajwe, G., Cifuentes, M., Tak, S.W., Kriebel, D., Gore, R., and Punnett, L. (2010). National income, self-reported wheezing and asthma diagnosis from the World Health Survey. Eur. Respir. J. *35*, 279–286.

Song, S., Jiang, W., Hou, L., and Zhao, H. (2020). Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. PLoS Comput. Biol. *16*, e1007565.

Sordillo, J.E., Lutz, S.M., Jorgenson, E., Iribarren, C., McGeachie, M., Dahlin, A., Tantisira, K., Kelly, R., Lasky-Su, J., Sakornsakolpat, P., et al. (2021). A polygenic risk score for asthma in a large racially diverse population. Clin. Exp. Allergy *51*, 1410–1420.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J. 'an, Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. *42*, 937–948.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Therneau, T. (2018). deming: Deming, Theil-Sen, Passing-Bablock and Total Least Squares Regression.

Ulirsch, J.C., and Kanai, M. An annotated atlas of causal variants underlying complex traits and gene expression.

Van Wonderen, K.E., Van Der Mark, L.B., Mohrs, J., Bindels, P.J.E., Van Aalderen, W.M.C., and Ter Riet, G. (2010). Different definitions in childhood asthma: how dependable is the dependent variable? Eur. Respir. J. *36*, 48–56.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). 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.

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

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

Wei, T., and Simko, V. (2021). R package "corrplot": Visualization of a Correlation Matrix.

Zhou, W., Kanai, M., Wu, K.-H.H., Humaira, R., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2021). Global Biobank Meta-analysis Initiative: powering genetic

discovery across human diseases.

Zhu, Z., Lee, P.H., Chaffin, M.D., Chung, W., Loh, P.-R., Lu, Q., Christiani, D.C., and Liang, L. (2018). A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat. Genet. *50*, 857–864.

Zhu, Z., Zhu, X., Liu, C.-L., Shi, H., Shen, S., Yang, Y., Hasegawa, K., Camargo, C.A., Jr, and Liang, L. (2019). Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. Eur. Respir. J. *54*.

Zhu, Z., Guo, Y., Shi, H., Liu, C.-L., Panganiban, R.A., Chung, W., O'Connor, L.J., Himes, B.E., Gazal, S., Hasegawa, K., et al. (2020). Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. J. Allergy Clin. Immunol. *145*, 537–549.

Zhu, Z., Hasegawa, K., Camargo, C.A., and Liang, L. (2021). Investigating asthma heterogeneity through shared and distinct genetics: Insights from genome-wide cross-trait analysis. J. Allergy Clin. Immunol. *147*, 796–807.

(2015). Asthma. Nat Rev Dis Primers *1*, 15036.

UK Biobank — Neale lab. http://www.nealelab.is/uk-biobank/.

Pan UKBB. https://pan.ukbb.broadinstitute.org/.