Genome-wide association studies in a large Korean cohort identify novel quantitative trait loci for 36 traits and illuminates their genetic architectures

Yon Ho Jee[1], Ying Wang[2,3], Keum Ji Jung[4], Ji-Young Lee[4], Heejin Kimm[4], Rui Duan[5], Alkes L. Price[1,5,6], Alicia R. Martin[2,3,7], Peter Kraft[1,8]

[1]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.*
[2]*Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.*
[3]*Stanley Center for Psychiatric Research and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.*
[4]*Institute for Health Promotion, Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea.*
[5]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA.*
[6]*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.*
[7]*Department of Medicine, Harvard Medical School, Boston, MA, USA.*
[8]*Transdivisional Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, MD, USA.*

Contact: Peter Kraft, Transdivisional Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, MD, USA. (phillip.kraft@nih.gov)
Keywords: genome-wide association study, complex traits, Korean population, genetic architecture
Version: May 2024

Abstract

Genome-wide association studies (GWAS) have been predominantly conducted in populations of European ancestry, limiting opportunities for biological discovery in diverse populations. We report GWAS findings from 153,950 individuals across 36 quantitative traits in the Korean Cancer Prevention Study-II (KCPS2) Biobank. We discovered 616 novel genetic loci in KCPS2, including an association between thyroid-stimulating hormone and *CD36*. Meta-analysis with the Korean Genome and Epidemiology Study, Biobank Japan, Taiwan Biobank, and UK Biobank identified 3,524 loci that were not significant in any contributing GWAS. We describe differences in genetic architectures across these East Asian and European samples. We also highlight East Asian specific associations, including a known pleiotropic missense variant in *ALDH2*, which fine-mapping identified as a likely causal variant for a diverse set of traits. Our findings provide insights into the genetic architecture of complex traits in East Asian populations and highlight how broadening the population diversity of GWAS samples can aid discovery.

Introduction

Large-scale biobanks integrating genomic and electronic health record data enable genome-wide association studies (GWAS) to identify numerous genetic associations and provide insights into the biological mechanisms of human complex traits and diseases.[1,2] In turn, the combined effects of these genetic markers can be summarized as polygenic risk score (PRS) to estimate individuals' genetic predispositions for complex diseases, which have successfully identified individuals with a high risk of disease.[3,4] However, current genetic discovery efforts heavily underrepresent non-European populations globally and thus limit further discoveries of variants that are rare or absent in European (EUR) populations but common in other ancestry groups.[5] Furthermore, this genomic research imbalance could lead to health disparities if the genomic discoveries benefit only European ancestry individuals in clinical practice.[6]

Early efforts toward diversifying GWAS in East Asian (EAS) populations, including Biobank Japan (BBJ),[7–9] Korean Genome and Epidemiology Study (KoGES),[10–13] and China Kadoorie Biobank,[14] Taiwan Biobank (TWB)[15] have made significant contributions and facilitated various genetic studies in these populations. Despite these efforts, the representation of EAS groups in genetic research remains low, compared to European groups (e.g., UK Biobank [UKB][16], FinnGen[17], HUNT[18], and deCODE[19]). For example, one of the most extensively used resources is UKB, which includes approximately 500,000 British individuals with deep phenotyping and genomic data.[20] According to the GWAS Diversity Monitor,[21] over 90% of total GWAS participants are from European-ancestry samples, while only 4% of participants are of Asian origin despite making up 59% of the global population. The inclusion of additional EAS biobanks is warranted to empower genetic discovery and elucidate the genetic architecture of complex traits and diseases within East Asia.

Here we conducted GWAS for 36 quantitative traits from 153,950 individuals in the Korean Cancer Prevention Study-II (KCPS-II) Biobank[22], a prospective cohort study of the Korean population with genomic data and a wide range of measured phenotypes. Following the GWAS in KCPS2, we meta-analyzed 21 traits across KCPS2,

KoGES, BBJ, TWB, and UKB to identify significant loci across East Asian and European ancestry populations. We compared the genetic architectures of these traits across populations leveraging GWAS summary statistics from KCPS2, BBJ, TWB, and UKB. Lastly, we pinpointed putatively causal variants through fine-mapping and conducted colocalization to understand the biological mechanisms underlying these traits.

Results

A total of 153,950 participants were genotyped, including 64,812 participants on the GSA-chip array and 89,138 participants on the Korean-chip array in this study. We subsequently conducted genotype quality control (QC) and imputation. Figure 1 provides an overview of the KCPS2 samples, the traits examined, their abbreviations, and the analyses conducted in this study (Table S1). We analyzed 36 quantitative traits including 4 anthropometric traits, 7 metabolic biomarkers, 5 liver function enzymes, 1 thyroid hormone, 1 tumor marker, 3 kidney function traits, 10 hematological traits, 2 cardiovascular traits, and 3 lifestyle factors.

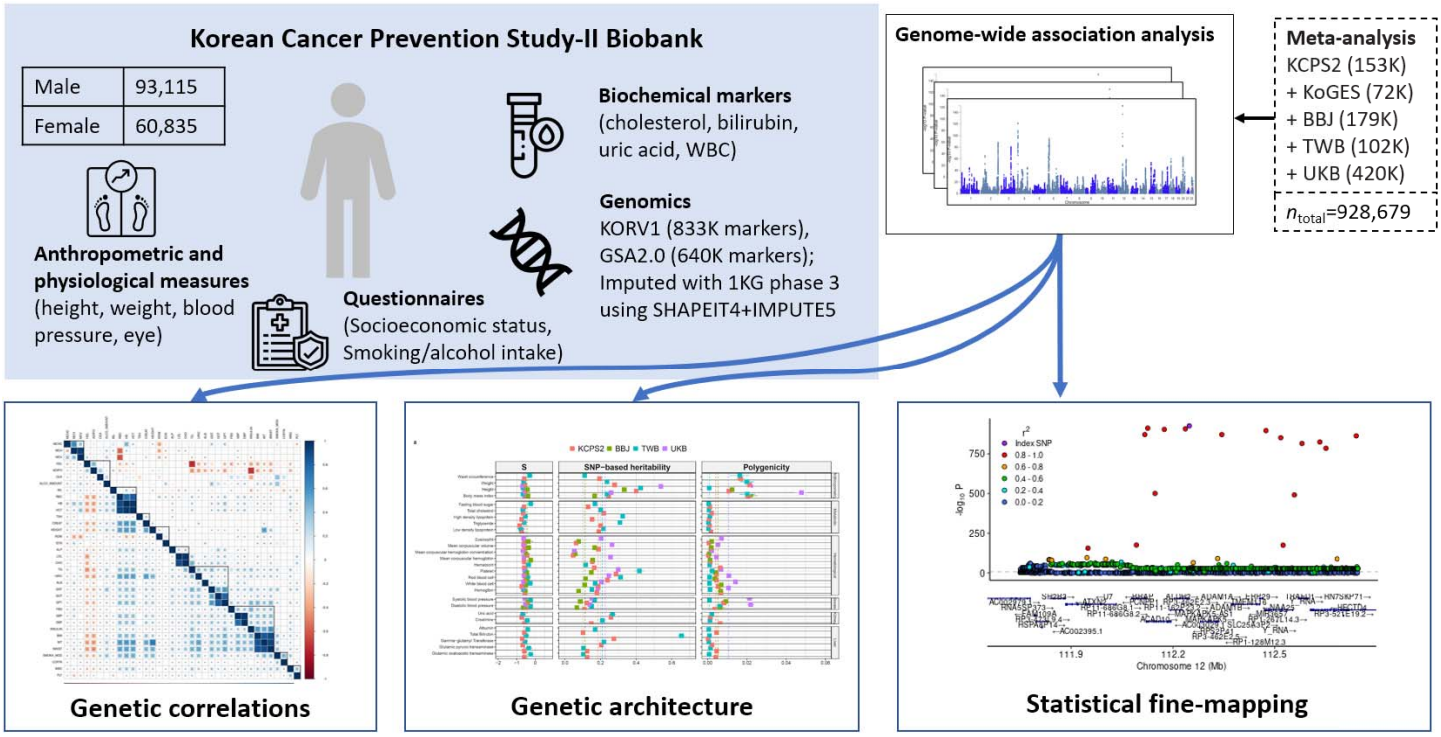


Figure 1 | Overview of the Korean Cancer Prevention Study-II Biobank and analysis. Detailed descriptions of the 36 quantitative traits examined in this study are shown in Table S1. After QC, the data were phased using SHAPEIT4[23] and imputed using IMPUTE5[24] with 1000 Genomes Project Phase 3 data.

GWAS of KCPS2 and pleiotropy analysis

We conducted GWAS of 36 human quantitative traits in the KCPS2 Biobank (n=153,950). We used a linear mixed model implemented in SAIGE[25] for association testing to maximize statistical power and included age, sex, 10 principal components (PCs), and SNP array as covariates. None of the GWAS exhibited striking systematic inflation in test statistics indicative of population stratification or other artifact (median λ_{GC} 1.25, median S-LDSC intercept 1.05) (Table S1). Using S-LDSC with the baseline-LD model, we estimated the SNP-based heritability for each trait (Table S1), which ranged from 0.033 (alcohol intake) to 0.345 (height).

Our analysis discovered 2,962 independent genome-wide significant loci (median 68, range 1-428 loci; 2,631 unique loci) across 36 traits using the 1000 Genome phase 3 EAS samples as the LD reference (Table S2). Among these, 616 loci (median 10, range 0-55 loci) were not reported in previous GWAS[26] related to the corresponding trait using Experimental Factor Ontology (EFO) term (Figure 2a, Table S3), with the greatest fraction of novel loci (novelty rate) for carcinoembryonic antigen [CEA, 10/16 (63%) novel loci], followed by thyroid-stimulating hormone [TSH, 55/100 (55%) novel loci]. The novel loci tend to be more common in KCPS2 than in 1000 Genome phase 3 EUR samples (median KCPS2 minor allele frequencies [MAF]: 0.207 vs. median EUR MAF: 0.118) (Figure S1). We also identified widespread pleiotropy: 4,960 gene regions contained variants associated with one or more traits (mean 2.3 traits, range 1-27). For example, out of 36 traits, variants near *ALDH2* were associated with 26 traits, including blood pressure and liver enzyme values (Figure 2b, Table S4).

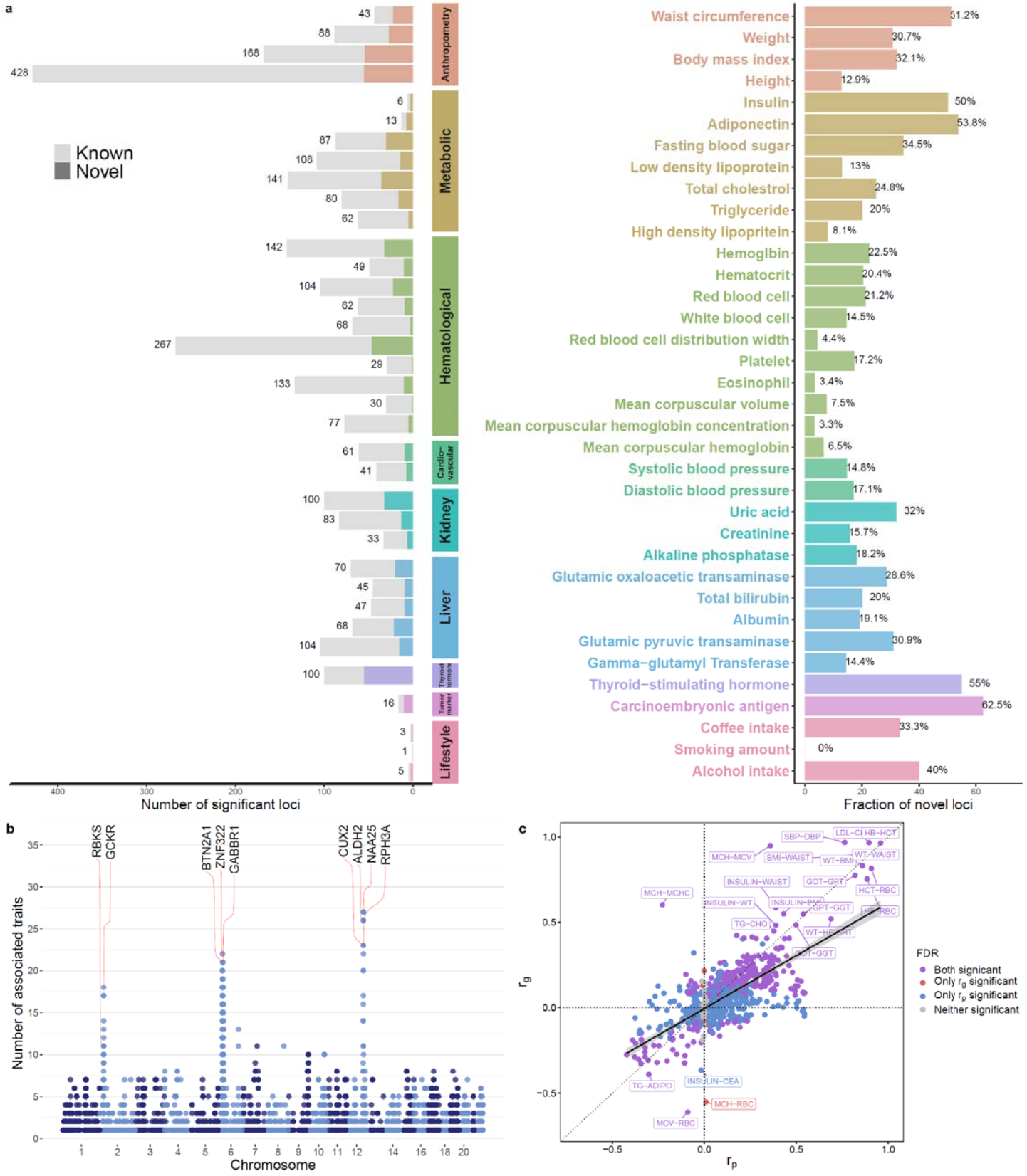


Figure 2 | GWAS results for 36 quantitative traits in the Korean Cancer Prevention Biobank-II (KCPS2). (a) Number of known and novel variants identified in KCPS2 compared to the Open Target Genetics[27] using EFO terms (Table S2-S3). **(b)** A summary of genome-wide significant loci associated with the 36 traits in KCPS2. Each locus was mapped to a gene using FUMA[28] with a 1000 Genome Phase 3 East Asian reference panel. We

then counted the number of associated traits (out of 36 traits) per gene (Table S4). **(c)** Comparisons of pairwise genetic correlations (r_g) between phenotypic correlations (r_p) for the 36 traits in KCPS2. r_g was estimated using bivariate LDSC based on association test statistics from linear regression. Significant r_g and r_p after false discovery rate (FDR) correction is indicated by purple if both r_g and r_p were significant, red if only r_g was significant, blue if only r_p was significant, and gray if neither was significant. The black solid line was estimated by spline smoothing from a linear regression model. The complete set of r_g and r_p, is available in Table S5.

Genetic and phenotypic correlations between the 36 traits in KCPS2

By estimating pairwise genetic correlations (r_g) between traits, we identified clusters of highly genetically correlated traits, including cardiometabolic risk factors (e.g., fasting blood sugar [FBS], systolic blood pressure [SBP], diastolic blood pressure [DBP], insulin, body mass index [BMI], weight, and waist circumference) and liver enzyme traits (e.g., albumin, glutamic oxaloacetic transaminase [GOT], glutamic pyruvic transaminase [GPT], and gamma-glutamyl transferase [GGT]) (Table S5, Figure 2c, Figure S2). The slope of the relationship between pairwise genetic correlations and phenotypic correlations (r_p) for 36 traits was 0.634 (standard error [s.e.]=0.026). We identified significantly negative genetic and phenotypic correlations of high-density lipoprotein [HDL] cholesterol and adiponectin with the majority of cardiometabolic risk factors including FBS, insulin, and BMI (mean cardiometabolic traits r_g=-0.24, r_p=-0.25 for HDL; r_g=-0.27, r_p=-0.27 for adiponectin). These findings are consistent with a known cardioprotective role of HDL[29] and beneficial effects of adiponectin on obesity-associated metabolic and vascular disorders.[30,31] In contrast, the genetic correlations between bilirubin and a number of cardiometabolic risk factors (mean cardiometabolic traits [FBS, insulin, BMI, waist circumference] r_g=-0.09), low-density lipoprotein cholesterol (r_g=-0.13 [FDR=0.012]), and WBC (r_g=-0.11 [FDR=0.001]) were significantly negative, although the phenotypic correlations were significantly positive (e.g., mean correlation between bilirubin and cardiometabolic traits r_p=0.04). Bilirubin levels have been shown to be inversely correlated with cardiovascular disease risk by inhibiting cholesterol synthesis and modulating the immune system,[32,33], which is supported by our genetic correlations results. Liver enzyme values such as GGT were positively associated with alcohol consumption both genetically and phenotypically (r_g=0.31

[FDR=0.0001], r_p=0.33 [FDR<0.0001]), consistent with a previous Mendelian randomization (MR) study.[34] Similarly, smoking, alcohol consumption, and hemoglobin showed significantly positive genetic and phenotypic correlations with a number of cardiometabolic risk factors (mean cardiometabolic traits r_g=0.18, r_p=0.22 for smoking; r_g=0.14, r_p=0.16 for alcohol consumption; r_g=0.14, r_p=0.28 for hemoglobin). For hemoglobin, previous MR showed evidence for lower hemoglobin levels being associated with lower BMI, better glucose tolerance and other metabolic profiles, lower inflammatory load, and blood pressure.[35]

Meta-analysis of 21 traits across KCPS2, KoGES, BBJ, and UKB

We meta-analyzed 21 traits across KCPS2 (153K), KoGES (72K), BBJ (179K), TWB (102K), and UKB (420K) and discovered a total of 11,861 loci associated with the 21 traits, among which 3,524 were not significant in any of the other four contributing GWAS (Figure 3a, Figure S3, Table S6-S7). The median MAF in KCPS2 for the lead variants at the loci which were only significant in the meta-analysis but not significant in the other individual GWAS, was lower than the MAF in KCPS2 for the lead variants at the loci that were only significant in the KCPS2 (median MAF 0.26 versus 0.29, respectively) (Figure 3b).

We compared effect sizes from KCPS2 to effect sizes by study for the lead variants at the 11,861 genome-wide significant loci from meta-analysis (Figure 3c). The correlation with KCPS2 effect sizes was greatest for the BBJ (regression slope=0.967, s.e.=0.017), followed by with KoGES (slope=0.772, s.e.=0.008), TWB (slope=0.725, s.e.=0.008), and UKB (slope=0.418, s.e.=0.007). To further explore novel associations, we compared effect sizes from the meta-analysis to MAF by study (Figure S4). There was an inverse relationship between MAF and effect size, due in part to the restriction to genome-wide significant variants. The lead variants from genome-wide significant loci identified in the meta-analysis had in general similar study-specific MAF in East Asian populations (KCPS2 [median MAF=0.26], KoGES [median MAF=0.26], and BBJ [median MAF=0.28]) compared to European ancestry populations in UKB (median MAF=0.26).

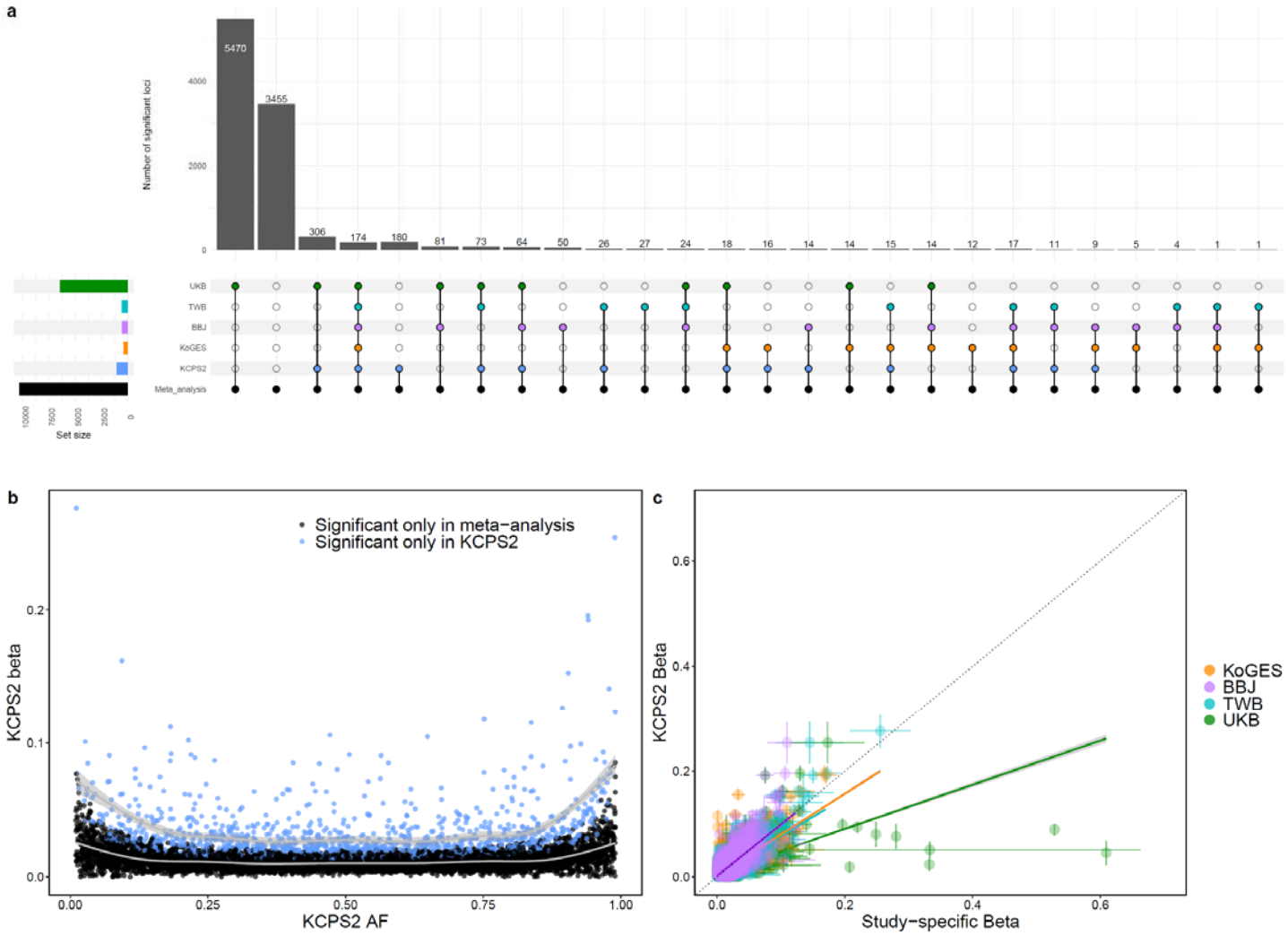


Figure 3 | Meta-analysis of 21 traits across KCPS2, KoGES, BBJ, and UKB. (a) Genome-wide significant loci identified in the meta-analysis, Color of dots indicate significance in meta-analysis (black), KCPS2 (blue), KoGES (orange), BBJ (purple), and UKB (green). Multiple dots in a bar represent simultaneous significance in multiple cohorts. **(b)** Comparisons of allele frequency and effect sizes in KCPS2 for the genome-wide significant variants discovered only in KCPS2 (blue) versus those identified only in the meta-analysis (black). **(c)** Comparisons of effect sizes in KCPS2 and study-specific effect sizes for the lead variants at the 11,861 meta-analysis genome-wide significant loci. The solid lines were estimated by spline smoothing from generalized additive model (b) or linear regression model (c). Full meta-analysis results are shown in Table S6-7.

Genetic architecture compared between KCPS2, BBJ, TWB, and UKB

We investigated the genetic architecture in KCPS2 (Figure S5) and compared it with BBJ, TWB, and UKB across six trait categories (Figure 4a, Table S8). The *S* parameters linking MAF and effect sizes were similar across the

biobanks (median S=-0.59, range -0.84, -0.22), suggesting a pervasive action of negative selection on the trait-associated variants.[36] The SNP-heritability estimates (h^2_g) varied widely across different biobanks and categories. For example, compared to BBJ, KCPS2 has higher heritability estimates for anthropometry (median h^2_g=0.26 vs. 0.25), cardiovascular (median h^2_g=0.12 vs. 0.08), and hematological traits (median h^2_g=0.17 vs. 0.11). For hematological traits, UKB has the largest heritability estimates (median h^2=0.23), with the exception of platelet (h^2_g=0.42) and RBC (h^2_g=0.31) being the largest in TWB. Compared to TWB, KCPS2 has lower heritability estimates for metabolic (median h^2_g=0.19 vs. 0.22), liver (median h^2_g=0.13 vs. 0.17), and kidney traits (median h^2_g=0.15 vs. 0.24). Overall, the heritability estimates of KCPS2 had lower correlations with TWB (Pearson correlation *r*=0.68, 95% confidence interval [CI]: 0.37-0.85), particularly for metabolic and hematological traits, whereas we observed higher correlations with BBJ (*r*=0.86, 95% CI: 0.57-0.96) and UKB (*r*=0.9, 95% CI: 0.67-0.97) (Figure 4b-d). However, we note that our TWB heritability estimates were higher than those reported by Chen et al., (2023)[15] using LDSC and in-sample LD, especially for metabolic and hematological traits (Figure S6A). When we used the heritability estimates reported by Chen and colleagues, the correlation in heritability between KCPS2 and TWB improved (*r*=0.80, 95% CI: 0.58-0.91) (Figure S6B,C). The most polygenic traits (weight, BMI, and waist circumference) had about 2% SNPs with nonzero effects, whereas the least polygenic traits (coffee intake, bilirubin, and MCHC) were affected by about 0.006-0.04% common SNPs in KCPS2. The median polygenicity estimates for the 8 traits available in all four studies were largest in UKB (median π=0.02), followed by BBJ (median π=0.007), KCPS2 (median π=0.006), and TWB (median π=0.001), which follows the same order as the sample sizes of the biobanks. Nevertheless, the genetic correlation (r_g) estimates within EAS were close to 1 (KCPS2-KoGES median r_g=0.997, KCPS2-BBJ median r_g=0.885, KCPS2-TWB median r_g=0.926) and were in general higher than the r_g between EAS and EUR (KCPS2-UKB median r_g=0.815) for these traits (Figure S7, Table S9).

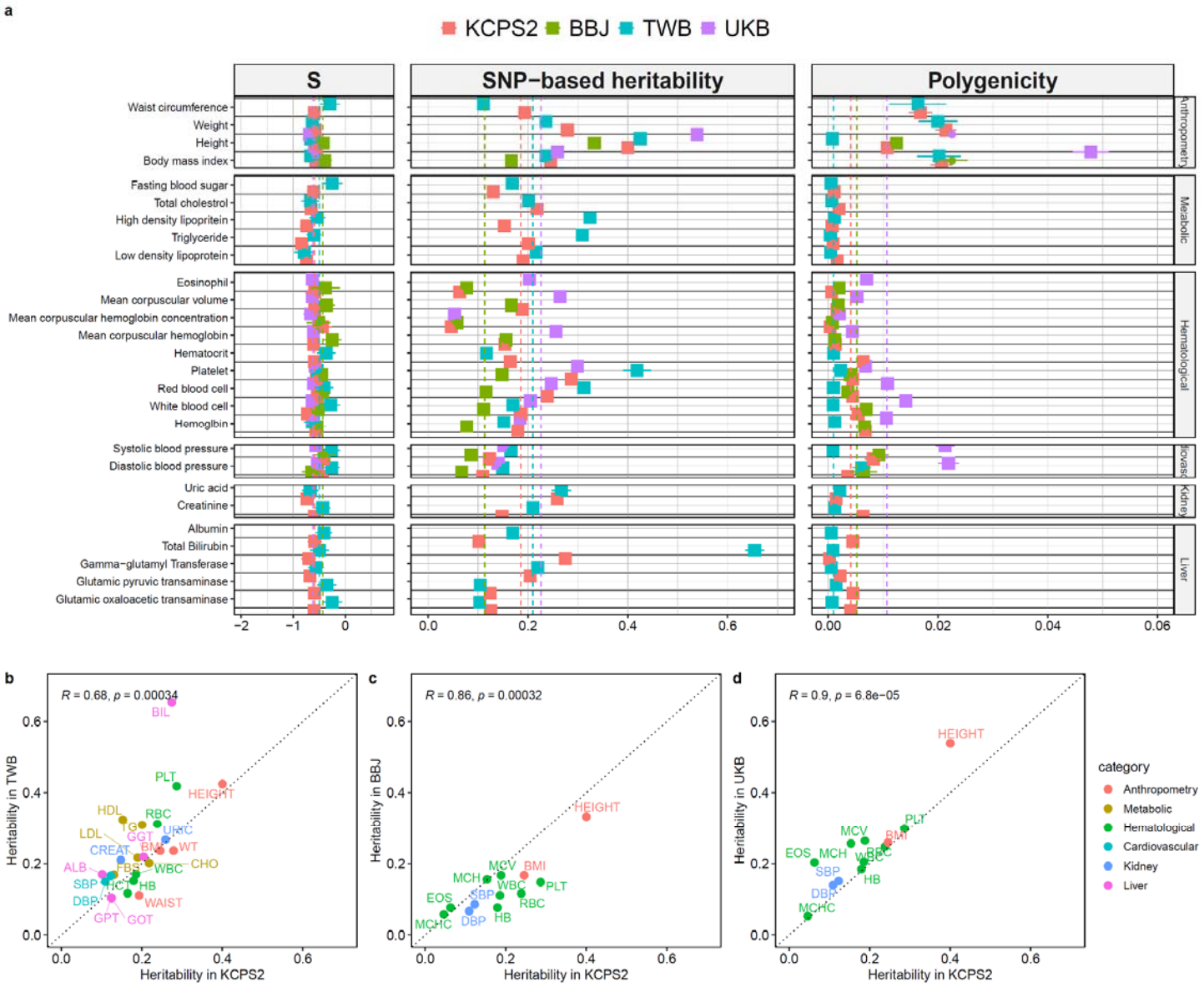


Figure 4 | Genetic architecture of complex traits across KCPS2, BBJ, TWB, and UKB. (a) The dots represent posterior means and horizontal bars represent standard errors of the parameters for each trait. The vertical dashed line shows the median of the estimates across traits. Full results are shown in Table S8. **(b-d)** Pearson correlations of SNP-heritability between KCPS2 and TWB (b), BBJ (c), and UKB (d) across the traits shown in a. Data are presented as posterior means of SNP-heritability. The trait categories are indicated by different colors labeled with their trait names.

Fine-mapping and colocalization analysis

To identify potential causal variants, we performed a single-population fine-mapping using SuSiE[37] in KCPS2.

Specifically, we fine-mapped 26 traits associated with the region spanning *ALDH2* on chromosome 12 (± 500kb

from rs671, which is known to be functionally related to alcohol metabolism). 1,476 variants in this region

were fine-mapped to a total of 56 credible sets, among which 17 contain exactly one variant (median 17.5, range 1-470 variants) (Table S10). rs671, a non-synonymous SNP associated with alcohol metabolism and alcohol intake, had a posterior inclusion probability (PIP) of greater than 90% for 8 traits including alcohol intake, GOT, GPT, GGT, SBP, DBP, coffee intake, and triglyceride (Figure 5a-b). For alcohol intake, we found seven credible sets with exactly one variant, all with PIP=1, including rs671, rs555501971, rs141043717, rs61055528, rs149178839, rs11066008, and rs550463060 ; in a sensitivity analysis setting the maximum number of credible sets to one (L=1), only rs671 remained in the credible set (Table S11).

To examine the mechanisms underlying these pleiotropic associations, we performed colocalization by pairing GWAS for alcohol intake with GWAS for each of traits where the PIP of rs671 was greater than 90%. These traits included liver enzymes (GOT, GPT, GGT), blood pressure (SBP, DBP), coffee intake, and triglyceride. rs671 was colocalized between alcohol intake and all of these traits with PP4=100%, supporting a shared causal variant for these two traits at the locus (Figure 5b).

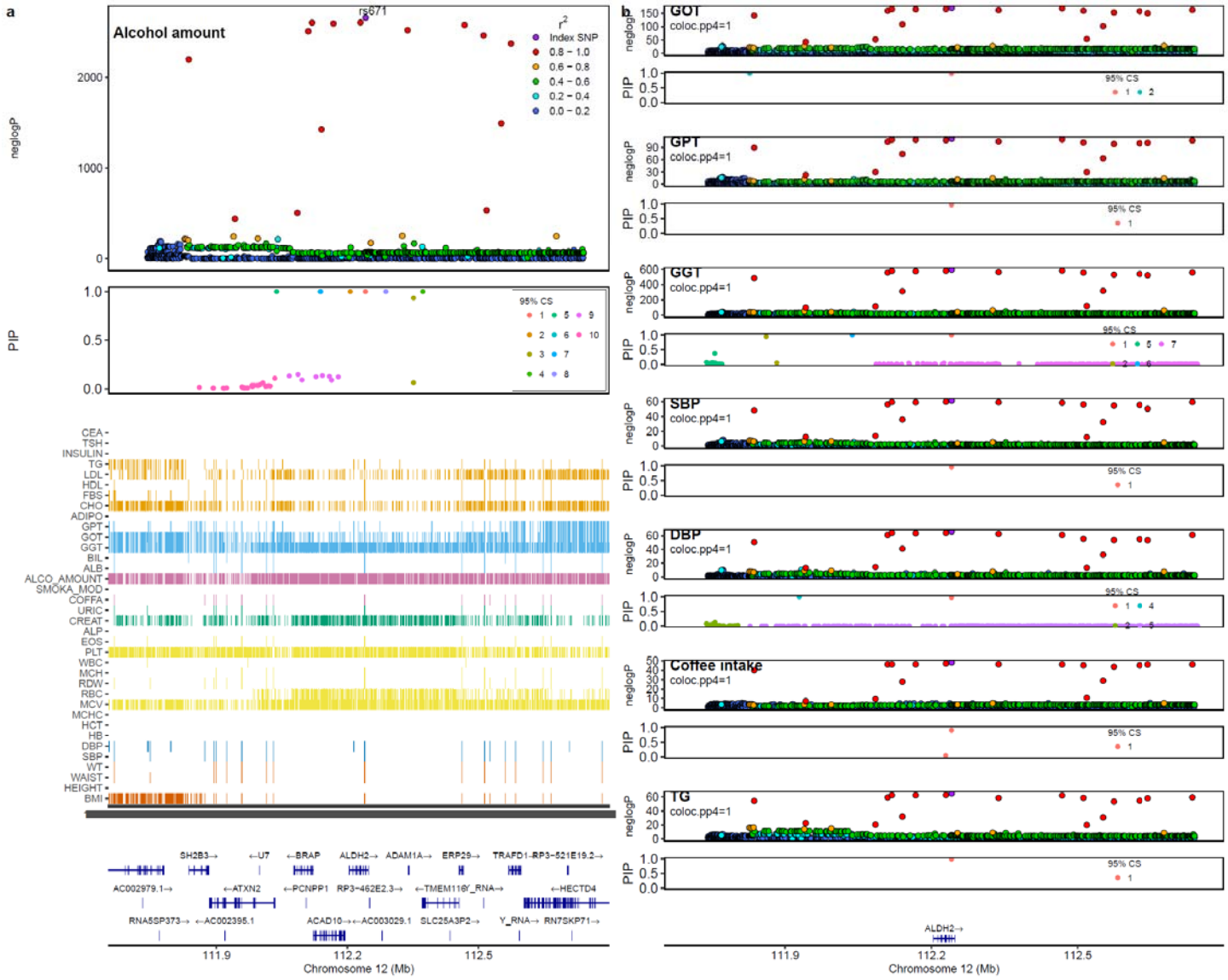


Figure 5 | Fine-mapping and colocalization analysis of *ALDH2* region in KCPS2. (a) Association between *ALDH2* (12q24.12) and alcohol intake in KCPS2. Colors in the Manhattan panels represent r^2 values to the lead variant rs671. In the posterior inclusion probability (PIP) panels, only fine-mapped variants in the 95% credible sets (CS) are colored. Heatmap represents significant variants ($P < 5.0 \times 10^{-8}$) for the other quantitative traits. **(b)** Colocalization analysis between alcohol intake and six traits that showed PIP>0.9 for rs671 was done in the same region. Coloc.PP4 represents the posterior probability of colocalization at the specified region. Each regional plot shows associations of each locus for the most significantly associated trait, which was all rs671 with PIP>0.9. Full fine-mapping results are shown in Table S10.

Discussion

In this study, we identified novel quantitative trait loci for 36 complex traits and investigated the genetic architecture of complex traits in 153,950 Korean individuals. Our analysis discovered 616 novel genetic loci that were not reported in previous GWAS related to the corresponding trait. We also demonstrated widespread pleiotropy and variants near *ALDH2* were associated with 26 traits. Meta-analysis of 21 traits across KCPS2, KoGES, BBJ, TWB, and UKB identified 3,524 loci that were not significant in any of the other four contributing GWAS. We compared the genetic architecture of these traits in KCPS2, BBJ, TWB, and UKB, and pinpointed one of the most pleiotropic regions (*ALDH2*) through fine-mapping, which colocalized with high probability with a diverse set of traits such as liver enzyme values.

Our study underscores the importance of enhancing the ancestral diversity and sample size of GWAS samples to facilitate genetic discovery and provide insights into the biological mechanisms of human quantitative complex traits. We discovered 616 novel loci that have lower median MAF in European ancestry individuals than in East Asian populations, which was enabled by leveraging samples from diverse ancestry groups. In particular, the novelty rate was high for TSH (55 out of 100 loci) in KCPS2. The most strongly associated lead variant with TSH, rs10799824 (Beta=-0.14, P=2.98x10^{-139}), was previously reported in GWAS of TSH[38] (Beta=-0.11, P=4.0x10^{-21}) and strict autoimmune hypothyroidism[17] (Odds ratio=0.87, P=4.1x10^{-18}) in European ancestry individuals. Several lead TSH loci were mapped to nearby genes previously linked to thyroid function, including rs13030651 (Beta=0.035, P=1.09x10^{-13}) in the thyroid adenoma associated gene (*THADA*)[39,40] and rs2160215 (Beta=0.065, P=1.75x10^{-51}) in thyrotropin receptor gene (*TSHR*).[41,42] Notably, a novel missense variant *CD36* p.Pro90Ser (rs75326924; Beta=-0.052, P=9.08x10^{-11}) found in our TSH GWAS (AF_{KCPS2}=0.068) has low frequency in the gnomAD v4.1.0 East Asian genetic ancestry group (MAF=0.03) but is exceedingly rare outside of that group (MAF<10^{-5}) and entirely absent from European genetic ancestry groups.[43] *CD36* (also known as fatty acid translocase [FAT]) facilitates the transport of fatty acids into cells and participates in

triglyceride storage.[44] A study in hypothyroid rats showed a reduced fatty acid absorption in the liver compared to euthyroid rats[45] and decreased hepatic FAT expression has been demonstrated in rats with postnatal hypothyroidism.[46] Moreover, recent studies have revealed that *CD36* contributes to the tumorigenesis and development of multiple cancer types by reprogramming the metabolism of glucose and fatty acid,[47–49] providing new insights for developing potential therapeutic target and prognostic biomarker in the clinical setting. We also found high novelty rate for GWAS of CEA in KCPS2, which recapitulates several known tumor biomarkers in various cancer types, including a non coding transcript exon variant (rs149037075; Beta=0.179, P=4.8x10^{-477}) in *ABO*[50,51] and a missense variant (rs28362459; Beta=0.068, P=4.85x10^{-128}) in *FUT3* (also known as Lewis gene),[52–55] consistent with previous CEA GWAS.[56] Previous studies show that determinants of the blood A and B antigens and Lewis antigens and of CEA share the same glycoprotein carrier molecules,[57,58] which might explain the association of CEA concentrations with the *ABO* and the *FUT3* locus. Several variants not previously reported in CEA GWAS were mapped to genes with potential role in cancer such as *C15orf39*[59] (rs143001709; P=1.31x10^{-8}), *ST6GAL1*[60] (rs73187787; P=3.36x10^{-11}), and *CCDC138*[61] (rs10179849; P=6.55x10^{-19}). Further studies are warranted to investigate the potential functional importance of these associations.

As a global effort to broaden the population diversity of genetic studies in East Asia, the KCPS2 GWAS enhanced our understanding of the genetic basis of complex traits in a Korean population. Our genetic correlations across 36 quantitative traits recapitulated known biology, including negative genetic correlations of HDL, adiponectin, and bilirubin with cardiometabolic risk factors [29,31–33] and positive genetic correlations of smoking, alcohol intake, and hemoglobin with cardiometabolic traits.[35,62] Notably, most of the significant genetic correlations were consistent with phenotypic correlations, which underscores the robustness and potential of the genetics-based approaches to understand biological architectures of complex traits. Many of the significant findings were consistent with BBJ,[8] KoGES,[13] TWB,[15] and UKB,[63] suggesting a similar genetic

architecture for these quantitative traits within EAS populations and across EAS and EUR populations as shown by our work and previous findings on within- and cross-ancestry genetic correlation analysis.[6,9,15]

Our findings provide opportunities to investigate the genetic architecture of complex traits within East Asian and across continental populations. While similar negative selection patterns were observed across traits and populations, the heritability estimates vary within EAS and across EAS and EUR populations, which may be attributed to several factors such as phenotype data collection, biobank design and environmental influences. For instance, compared to a hospital-based cohort such as BBJ, participants of a population-/community-based cohort such as KCPS2, KoGES, TWB, and UKB may have different distributions of disease-related traits due to healthy-volunteer effects.[64] Hence, the comparison of heritability estimates across biobanks requires careful consideration of technical differences, potential collider bias, and variability in baseline health status among studies.[6] Moreover, we demonstrated the correlation in heritability between KCPS2 and TWB improved when the heritability of TWB were replaced by the previously reported estimates[15] that used LDSC and in-sample LD, especially for metabolic and hematological traits. Thus, in addition to the phenotype heterogeneity, heritability may be affected by different heritability estimation methods and LD matrices. Further research is needed to explore the impact of these factors on genetic architecture comparisons. Nevertheless, our study highlights the importance of increasing genetic diversity to understand genetic architecture of diverse populations, which is crucial to achieve equitable delivery of genomic knowledge to global populations.[6,65,66]

The KCPS2 GWAS facilitated pinpointing causal variants through fine-mapping. For example, a missense variant rs671 (AF_{EAS}=0.2254 vs. $AF_{EUR}=2.4x10^{-5}$ in non-Finnish EUR populations in gnomAD v4.0.0[43]) was identified as the causal variant in the *ALDH2* gene for 8 traits including alcohol intake, GGT, GOT, GPT, SBP, DBP, coffee intake, and triglyceride through fine-mapping. *ALDH2* gene is the target of drug for alcoholism which irreversibly inactivate catalytic Cys302 in *ALDH2* by carbamylation in the substrate site of the

enzyme.[67,68] Furthermore, our colocalization results suggest that alcohol intake is a causal risk factor for liver enzymes (GGT, GOT, GPT), blood pressure, and triglyceride levels at the rs671 region, which is supported by evidence for causation from previous MR studies among East Asian populations.[34,69,70] Our findings demonstrated the potential of diversifying EAS GWAS to uncover genetic associations that are common in EAS populations but rare in EUR populations, which could not be discovered even with very large European sample sizes. Furthermore, discovery of such variants may help identify targets for prevention and treatment, thus offering equitable access to precision medicine to diverse populations.

We note several limitations to our study. First, we only conducted GWAS of continuous traits due to limited power for disease phenotypes. Further investigation into disease outcomes should be conducted. Second, we conducted fine-mapping in KCPS2 only for a particular locus, which might cause a concern about potential LD tagging effects for observed pleiotropy. Recent studies suggest that multi-ancestry fine-mapping can improve refinement of causal variants by leveraging different LD patterns across ancestries.[71,72] We will explore these potential extensions in the near future. Third, for the estimation of genetic architecture parameters, in-sample LD was used for KCPS2, BBJ, and UKB but not for TWB. Since we were unable to find publicly available data to estimate LD in a Taiwanese population, we estimated the genetic architecture parameters using the LD matrix based on the 50K individuals from KCPS2. Such disagreement between the genetic associations and the correlation matrix may induce spurious results due to different LD patterns between the Taiwanese population and Korean individuals from KCPS2, even though both are East Asian populations.

Our findings highlight how broadening the population diversity of GWAS samples can aid discovery and post-GWAS analyses. Our results also provide insights into the genetic architecture of complex traits in East Asian populations. By increasing the sample size and ancestral diversity of GWAS samples, our analysis may help

identify novel population-specific targets for prevention and treatment, thus offering equitable access to precision medicine to diverse populations.

References

1. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).
2. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am J Hum Genet* **110**, 179–194 (2023).
3. 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).
4. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* **28**, R133–R142 (2019).
5. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat Med* **28**, 243–250 (2022).
6. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).
7. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat Genet* **49**, 1458–1467 (2017).
8. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat Genet* **50**, 390–400 (2018).
9. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415–1424 (2021).
10. Kim, T., Park, A. Y., Baek, Y. & Cha, S. Genome-Wide Association Study Reveals Four Loci for Lipid Ratios in the Korean Population and the Constitutional Subgroup. *PLOS ONE* **12**, e0168137 (2017).
11. Cho, H.-W., Jin, H.-S. & Eom, Y.-B. A Genome-Wide Association Study of Novel Genetic Variants Associated With Anthropometric Traits in Koreans. *Frontiers in Genetics* **12**, (2021).
12. Park, J. sung, Kim, Y. & Kang, J. Genome-wide meta-analysis revealed several genetic loci associated with serum uric acid levels in Korean population: an analysis of Korea Biobank data. *J Hum Genet* **67**, 231–237 (2022).
13. Nam, K., Kim, J. & Lee, S. Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits. *Cell Genomics* **2**, 100189 (2022).

14. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom* **3**, 100361 (2023).
15. Chen, C.-Y. *et al.* Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genomics* **3**, 100436 (2023).
16. Karczewski, K. J. *et al.* Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. 2024.03.13.24303864 Preprint at https://doi.org/10.1101/2024.03.13.24303864 (2024).
17. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
18. Brumpton, B. M. *et al.* The HUNT study: A population-based cohort for genetic research. *Cell Genom* **2**, 100193 (2022).
19. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435–444 (2015).
20. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
21. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet* **52**, 242–243 (2020).
22. Jee, Y. H. *et al.* Cohort Profile: The Korean Cancer Prevention Study-II (KCPS-II) Biobank. *International Journal of Epidemiology* **47**, 385–386f (2018).
23. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10**, 5436 (2019).
24. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet* **16**, e1009049 (2020).
25. 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).
26. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977–D985 (2023).

27. 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).
28. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
29. Perswani, P. *et al.* Rethinking HDL-C: An In-Depth Narrative Review of Its Role in Cardiovascular Health. *Current Problems in Cardiology* **49**, 102152 (2024).
30. Kawano, J. & Arora, R. The role of adiponectin in obesity, diabetes, and cardiovascular disease. *J Cardiometab Syndr* **4**, 44–49 (2009).
31. Tang, Y.-H. *et al.* Genetic and Functional Effects of Adiponectin in Type 2 Diabetes Mellitus Development. *International Journal of Molecular Sciences* **23**, 13544 (2022).
32. Wen, G., Yao, L., Hao, Y., Wang, J. & Liu, J. Bilirubin ameliorates murine atherosclerosis through inhibiting cholesterol synthesis and reshaping the immune system. *Journal of Translational Medicine* **20**, 1 (2022).
33. Hinds, T. D. & Stec, D. E. Bilirubin, a Cardiometabolic Signaling Molecule. *Hypertension* **72**, 788–795 (2018).
34. Xu, L. *et al.* Alcohol Use and Gamma-Glutamyltransferase Using a Mendelian Randomization Design in the Guangzhou Biobank Cohort Study. *PLoS One* **10**, e0137790 (2015).
35. Auvinen, J. *et al.* Systematic evaluation of the association between hemoglobin levels and metabolic profile implicates beneficial effects of hypoxia. *Sci Adv* **7**, eabi4822 (2021).
36. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* **50**, 746–753 (2018).
37. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1273–1300 (2020).
38. Porcu, E. *et al.* A Meta-Analysis of Thyroid-Related Traits Reveals Novel Loci and Gender-Specific Differences in the Regulation of Thyroid Function. *PLOS Genetics* **9**, e1003266 (2013).
39. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).

40. Drieschner, N. *et al.* A domain of the thyroid adenoma associated gene (THADA) conserved in vertebrates becomes destroyed by chromosomal rearrangements observed in thyroid adenomas. *Gene* **403**, 110–117 (2007).
41. Calebiro, D. *et al.* Frequent TSH receptor genetic alterations with variable signaling impairment in a large series of children with nonautoimmune isolated hyperthyrotropinemia. *J Clin Endocrinol Metab* **97**, E156-160 (2012).
42. Abramowicz, M. J., Duprez, L., Parma, J., Vassart, G. & Heinrichs, C. Familial congenital hypothyroidism due to inactivating mutation of the thyrotropin receptor causing profound hypoplasia of the thyroid gland. *J Clin Invest* **99**, 3018–3024 (1997).
43. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
44. Febbraio, M., Hajjar, D. P. & Silverstein, R. L. CD36: a class B scavenger receptor involved in angiogenesis, atherosclerosis, inflammation, and lipid metabolism. *J Clin Invest* **108**, 785–791 (2001).
45. Klieverik, L. P. *et al.* Thyroid hormone effects on whole-body energy homeostasis and tissue-specific fatty acid uptake in vivo. *Endocrinology* **150**, 5639–5648 (2009).
46. Santana-Farré, R. *et al.* Influence of neonatal hypothyroidism on hepatic gene expression and lipid metabolism in adulthood. *PLoS One* **7**, e37386 (2012).
47. Pascual, G. *et al.* Dietary palmitic acid promotes a prometastatic memory via Schwann cells. *Nature* **599**, 485–490 (2021).
48. Oh, D. S. & Lee, H. K. Autophagy protein ATG5 regulates CD36 expression and anti-tumor MHC class II antigen presentation in dendritic cells. *Autophagy* **15**, 2091–2106 (2019).
49. Ruan, C., Meng, Y. & Song, H. CD36: an emerging therapeutic target for cancer and its molecular mechanisms. *J Cancer Res Clin Oncol* **148**, 1551–1558 (2022).
50. Rummel, S. K. & Ellsworth, R. E. The role of the histoblood ABO group in cancer. *Future Science OA* **2**, (2016).
51. Huang, J. Y., Wang, R., Gao, Y.-T. & Yuan, J.-M. ABO blood type and the risk of cancer – Findings from the Shanghai Cohort Study. *PLoS One* **12**, e0184295 (2017).

52. Zhan, L., Chen, L. & Chen, Z. Knockdown of FUT3 disrupts the proliferation, migration, tumorigenesis and TGF-β induced EMT in pancreatic cancer cells. *Oncol Lett* **16**, 924–930 (2018).

53. do Nascimento, J. C. F., Beltrão, E. I. C. & Rocha, C. R. C. High FUT3 expression is a marker of lower overall survival of breast cancer patients. *Glycoconj J* **37**, 263–275 (2020).

54. Lin, L. *et al.* FUT3 facilitates glucose metabolism of lung adenocarcinoma via activation of NF-κB pathway. *BMC Pulmonary Medicine* **23**, 436 (2023).

55. He, C. *et al.* The DDX39B/FUT3/TGFβR-I axis promotes tumor metastasis and EMT in colorectal cancer. *Cell Death Dis* **12**, 1–12 (2021).

56. He, M. *et al.* A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9, carcinoembryonic antigen and α fetoprotein and their associations with cancer risk. *Gut* **63**, 143–151 (2014).

57. Holburn, A. M., Mach, J. P., MacDonald, D. & Newlands, M. Studies of the association of the A, B and Lewis Blood group antigens with carcinoembryonic antigen (CEA). *Immunology* **26**, 831–843 (1974).

58. Yamashita, K. *et al.* Structural studies of the carbohydrate moieties of carcinoembryonic antigens. *Cancer Res* **47**, 3451–3459 (1987).

59. Zhang, T. *et al.* Interrogating Kinase–Substrate Relationships with Proximity Labeling and Phosphorylation Enrichment. *J. Proteome Res.* **21**, 494–506 (2022).

60. Garnham, R., Scott, E., Livermore, K. E. & Munkley, J. ST6GAL1: A key player in cancer. *Oncol Lett* **18**, 983–989 (2019).

61. Anurag, M. *et al.* Multiomics profiling of urothelial carcinoma in situ reveals CIS-specific gene signature and immune characteristics. *iScience* **27**, 109179 (2024).

62. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237–244 (2019).

63. Wu, Y. *et al.* Fast estimation of genetic correlation for biobank-scale data. *Am J Hum Genet* **109**, 24–32 (2022).

64. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* **186**, 1026–1034 (2017).

65. Wang, Y. *et al.* Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genomics* **3**, 100241 (2023).

66. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635–649 (2017).

67. Lipsky, J. J., Shen, M. L. & Naylor, S. In vivo inhibition of aldehyde dehydrogenase by disulfiram. *Chem Biol Interact* **130–132**, 93–102 (2001).

68. Shen, M. L., Johnson, K. L., Mays, D. C., Lipsky, J. J. & Naylor, S. Determination of in vivo adducts of disulfiram with mitochondrial aldehyde dehydrogenase. *Biochem Pharmacol* **61**, 537–545 (2001).

69. Cho, Y. *et al.* Alcohol intake and cardiovascular risk factors: A Mendelian randomisation study. *Sci Rep* **5**, 18422 (2015).

70. Jee, Y. H., Lee, S. J., Jung, K. J. & Jee, S. H. Alcohol Intake and Serum Glucose Levels from the Perspective of a Mendelian Randomization Design: The KCPS-II Biobank. *PLoS One* **11**, e0162930 (2016).

71. Yuan, K. *et al.* Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. 2023.01.07.23284293 Preprint at https://doi.org/10.1101/2023.01.07.23284293 (2023).

72. Cai, M. *et al.* XMAP: Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias. *Nat Commun* **14**, 6870 (2023).

73. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

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

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

76. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case-control status and family history of disease increases association power. *Nat Genet* **52**, 541–547 (2020).

77. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

78. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527–1533 (2021).

79. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* **50**, 693–698 (2018).

80. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

81. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

82. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat Commun* **12**, 1164 (2021).

83. Wen, X. & Stephens, M. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The Annals of Applied Statistics* **4**, 1158–1182 (2010).

84. 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).

85. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genetics* **17**, e1009440 (2021).

86. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

Methods

Study population

The Korean Cancer Prevention Study-II Biobank (KCPS2) is a prospective cohort study based in Korea with genotype data and measurements of a wide range of phenotypes collected from 153,950 subjects (Figure 1). Participants in KCPS2 undertook routine health assessments at nationwide health promotion centers between 2004 and 2013. The study design and recruitment have been described in detail previously.[22] KCPS2 collects extensive phenotypes including demographics, socioeconomic status, environmental exposures, lifestyle, dietary habits, family history and self-reported disease status through structured questionnaires. The aAnthropometric measures as well as, and blood and urine samples were collected at recruitment, and several biomarkers were assayed subsequently. All participants in the KCPS2 were genotyped using either the Illumina Global Screening Array (GSA) v2.0 (78,260 samples) or the Korean Chip array v1 (90,245 samples). All participants provided written informed consent before participation.

Quality control (QC) and imputation were conducted separately for each of the two SNP arrays. First, SNPs with low call rate (<95%) were filtered out, along with samples with low call rate (<98%), gender discrepancy, excessive heterozygosity, excessive singletons, and duplicates. Additionally, SNPs with Hardy-Weinberg equilibrium p-value $<10^{-4}$ or minor allele frequencies (MAF) <0.01 were excluded. Following QC, the data were phased using SHAPEIT4[23] and imputed using IMPUTE5[24] with 1000 Genomes Project Phase 3 data. Variants with imputation INFO <0.8 were excluded after imputation. The two imputed data of GSA chip and the Korean chip were then merged, resulting in a total of 6,809,738 overlapping variants.

Genome-wide association analysis in KCPS2

We performed GWAS on 36 quantitative traits including anthropometric measures and biomarkers spanning 8 categories (metabolic, liver, thyroid hormone, tumor marker, kidney, hematological, cardiovascular,

arthrometry, and lifestyle factors). For each trait, we excluded samples with measurements that were more than 6 standard deviations away from the sample average.

We used linear mixed models implemented in SAIGE (v.1.1.9)[25] for association testing, controlling for age, sex, 10 PCs, and SNP array. The SAIGE method contains two main steps: in step 1, we used a subset of linkage disequilibrium (LD)-pruned variants with R^2<0.2 (158,729 variants) to obtain the genetic relationship matrix. We included age, sex, 10 PCs, and SNP array as covariates in step 1. Single-variant association testing was performed in step 2 where the phenotypes were inverse rank-based normal transformed and leave-one-chromosome-out scheme to remove the proximal contamination. We used FUMA[28] with 1000 Genome Project Phase 3[73] EAS samples as LD reference to identify independent genome-wide significant loci ($p<5x10^{-8}$) for each trait, window size of 5 Mb, and LD threshold R^2 of 0.1.

Linkage disequilibrium score regression (LDSC)[74] was applied to estimate cross-trait genetic correlations (r_g) in KCPS2. We ran stratified-LDSC (S-LDSC)[75] with a full baseline-LD v1.2 model[75] to compute LDSC intercept. To correctly specify effective sample size in LDSC or S-LDSC analysis, we used GWAS summary statistics generated from simple linear regression models instead of linear mixed models, which have a different effective GWAS sample size than the study sample size.[76] We ran linear regression using PLINK2[77] for association testing, controlling for age, sex, 10 PCs, and SNP array. All phenotypes used in these GWAS were inverse rank-based normal transformed.

Novel association identification

We mapped each trait to a term in the Experimental Factor Ontology (EFO) each trait (Table S3). For each of the independent loci we identified to be associated with a given trait, wWe queried the Open Target Genetics database (release 22.09)[27,78] for each of the independent loci we identified to be associated with a given trait to identify any previously reported associations (with the same EFO term or category, see below) within ±500

kb of the lead variant at that locus. If none were identified for a locus, we considered that locus to be novel. Given the widespread pleiotropy and phenotypic heterogeneity,[79] we may overcount novel associations. We therefore also used EFO categories, which are more generic than EFO terms to evaluate novelty. For example, "body height" (EFO_0004339) is an EFO term that maps to the broader EFO category of "body measurement" (EFO_0004324) by GWAS catalog.[26] We further exhaustively searched for previous reports of genetic association in a given trait using the GWAS Catalog, which might not be included in the Open Target Genetics database. Since the recent GWAS results of height by Yengo et al. (2022)[80] were not listed in the GWAS Catalog at the time of the curation, we additionally excluded variants that were genome-wide significant in the GWAS.

Evaluation of gene pleiotropy

We investigated gene pleiotropy, where a gene affects multiple traits in KCPS2. We defined the degree of pleiotropy as the number of significant associations per gene ($p<5x10^{-8}$). The list of genes mapped to each SNP in KCPS2 GWAS results was taken from FUMA[28] to map SNPs in GWAS results to a gene with the 1000 Genome Phase 3[73] EAS reference panel. We then quantified the degree of pleiotropy per gene by aggregating and counting the number of genome-wide significant associations across 36 traits.

Meta-analysis of EAS and EUR GWAS

We conducted meta-analysis of 21 traits across KCPS2, KoGES, BBJ, TWB, and UKB (European ancestry samples) to further identify novel loci across East Asian and European ancestry populations. We implemented inverse-variance-weighted fixed-effect meta-analysis using METAL.[81] We then used FUMA[28] to identify genome-wide significant loci in the meta-analysis after clumping variants with p-values$<5x10^{-8}$, window size of 5 Mb, and LD threshold R^2 of 0.1. We identified the association as novel if none of the variants within the locus reached genome-wide significance ($P< 5x10^{-8}$) in KoGES, BBJ, TWB, or UKB GWAS.

Genetic architecture of complex traits in KCPS2, BBJ, TWB, and UKB

We used SbayesS[82] to estimate the SNP-based heritability (h^2_g), polygenicity (π; proportion of SNPs with nonzero effects), and the relationship between minor allele frequency (MAF) and SNP effects (*S* parameter) for 36 traits in KCPS2. We constructed a full LD correlation matrix based on 50K individuals from KCPS2 and shrunk the matrix to ignore small LD correlations due to sampling variance using the shrinkage method from Wen and Stephens (2010).[83] To calculate the LD matrix shrinkage estimate, we used a genetic map for East Asian populations, with the effective population sample size of 12,239,[73] while using the default shrinkage cutoff (10^{-5}). We then compared the genetic architecture of KCPS2 with BBJ, TWB, and UKB across six categories including anthropometry, cardiovascular, hematological, kidney, liver, and metabolic traits (among which 12 traits overlap between KCPS2 and BBJ/UKB, 23 traits overlap between KCPS2 and TWB, and 8 traits available in all biobanks: height, body mass index, platelet, red blood cell, white blood cell, hemoglobin, systolic blood pressure, and diastolic blood pressure). For BBJ and UKB, we used the previously reported genetic architecture parameter estimates,[65] which were constructed using GWAS summary statistics generated from linear regression models and in-sample LD for the corresponding population. For a fair comparison of these parameters between KCPS2, BBJ, TWB, and UKB, we applied SbayesS to GWAS summary statistics generated from linear regression models in KCPS2 and TWB instead of linear mixed models. For TWB, we estimated the genetic architecture parameters using the LD matrix based on the 50K individuals from KCPS2 because we were unable to find publicly available data to estimate LD in a Taiwanese population. We did not include KoGES for the genetic architecture comparisons because 1) summary statistics from linear regression models were not publicly available, and 2) their relatively small sample size might lead to estimates with a higher degree of uncertainty, since the genetic architecture parameters are sample size dependent.[82]

Cross-biobank genetic correlation

To estimate cross-biobank genetic effect correlations within EAS (KCPS2-KoGES, KCPS2-BBJ, and KCPS2-TWB), we used LDSC[74] to estimate r_g using the 1000 Genomes phase 3 EAS reference panel. We used Popcorn

(v.1.0)[84] to estimate cross-biobank genetic-effect correlation between KCPS2 and UKB GWAS with precomputed cross-population scores for EUR and EAS 1000 Genomes Project populations provided by the authors. For a fair comparison, we restricted to HapMap3 SNPs that were shared across all five biobanks. We applied the analysis to traits with heritability calculated by LDSC or Popcorn >0.01 and their GWAS summary statistics generated from linear mixed models from all biobanks which were publicly available.

Fine-mapping and colocalization analysis

We fine-mapped one of the most pleiotropic regions identified by GWAS of the 36 traits above, a 500kb region flanking *ALDH2* in KCPS2. We applied SuSiE[37] to GWAS summary statistics and in-sample LD on 1,476 SNPs in this region. We implemented colocalization analysis to further investigate whether two traits share a causal variant. We applied *coloc.susie*[85] which allows multiple signals to be distinguished using SuSiE, and then performed colocalization analysis on all possible pairs of signals between the traits. We performed colocalization analysis in a 500kb window centered on an identified causal variant between alcohol intake and the other traits with PIP of rs671 being greater than 0.9 from fine-mapping results. We reported posterior probability of colocalization (PP4) for each of these pairs at the specified region. We applied LocusZoom[86] to visualize the colocalization analysis.

Acknowledgements

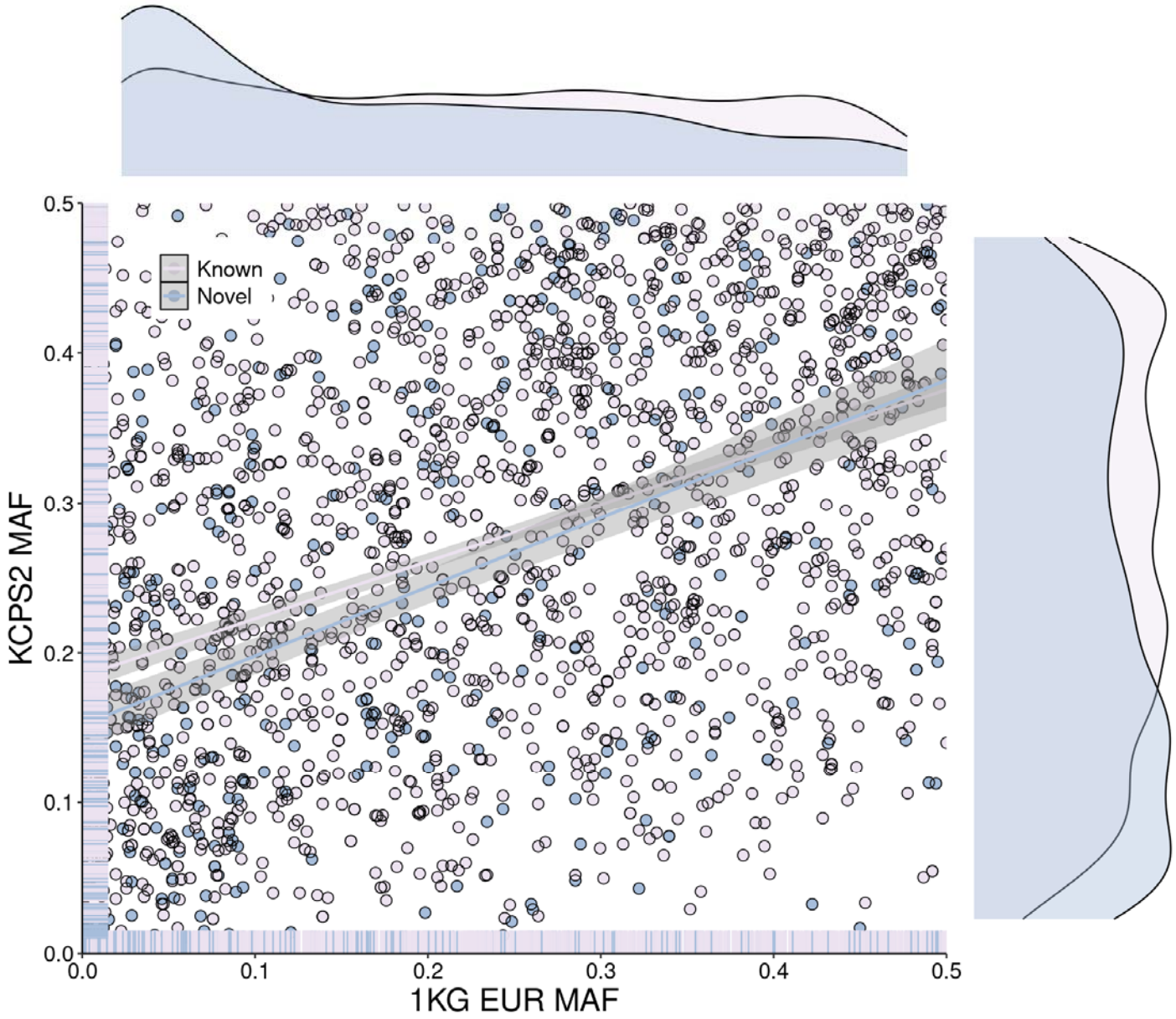
We acknowledge the participants of KCPS2 and the management team and leadership of KCPS2 for their outstanding support in collecting samples and clinical data. We thank KoGES, BBJ, TWB, and UKB for providing resources and releasing the GWAS summary statistics, which made this study possible. K.J.J, J,-Y.L, and H.K acknowledge support from the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (RS-2023-00239122). R.D. was supported by R01 GM148494. A.L.P was supported by R01 HG006399. A.R.M. was supported by funding from the National Institutes of Health (K99/R00MH117229) and funding from a Broad Next Generation Fund. P.K. was supported by U01CA261339.

Data availability

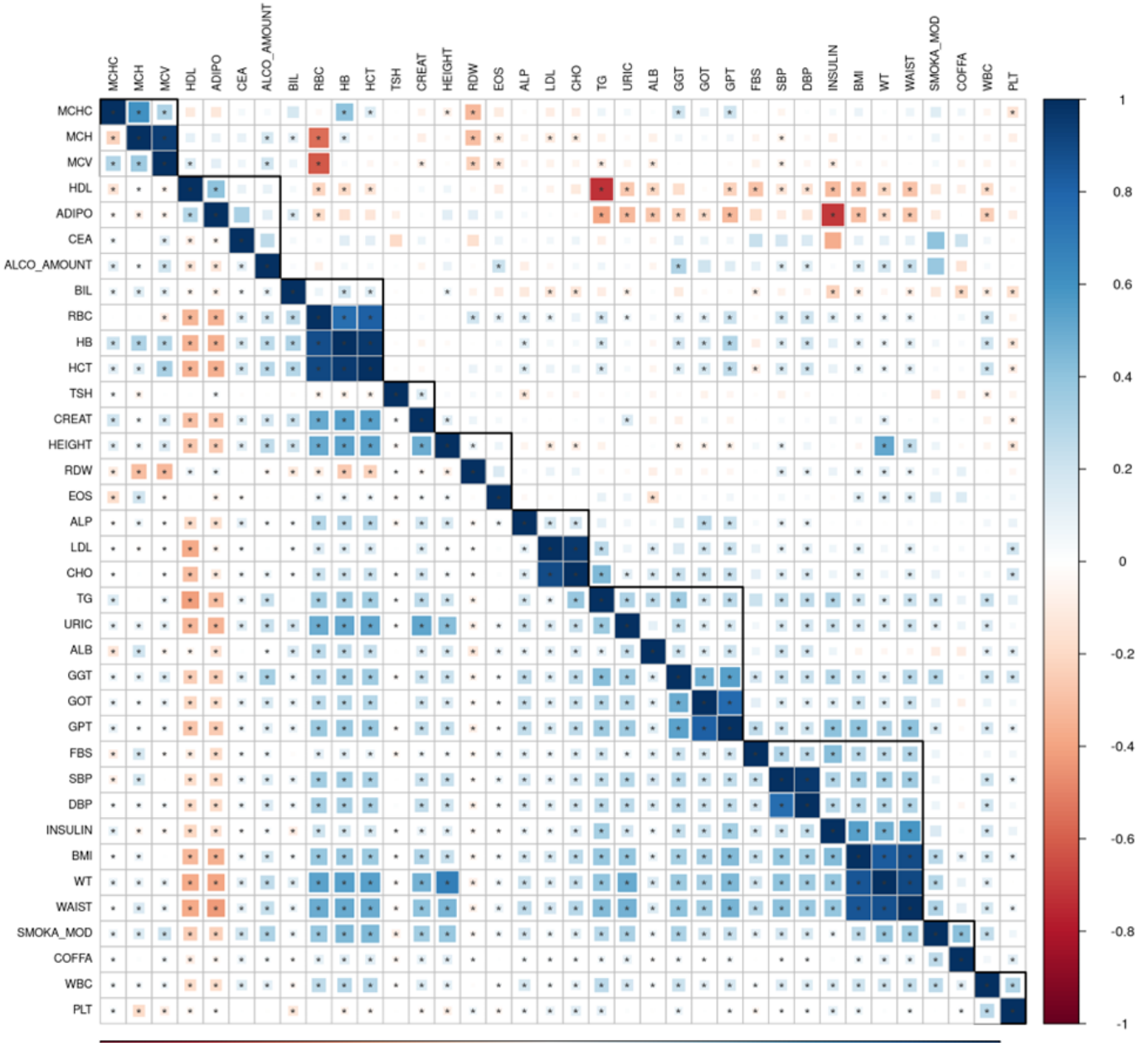
GWAS summary statistics in this study will be made publicly available prior to publication. The summary statistics for KoGES used in this study were downloaded from the KoGES Zendo (https://zenodo.org/record/7042518), BBJ summary statistics from the Biobank Japan PheWeb (https://pheweb.jp/), TWB summary statistics from GWAS Catalog (https://www.ebi.ac.uk/gwas/publications/38116116) , and summary statistics for Europeans in UKB were downloaded from Pan-UK Biobank (https://pan.ukbb.broadinstitute.org/).

Supplementary Figures

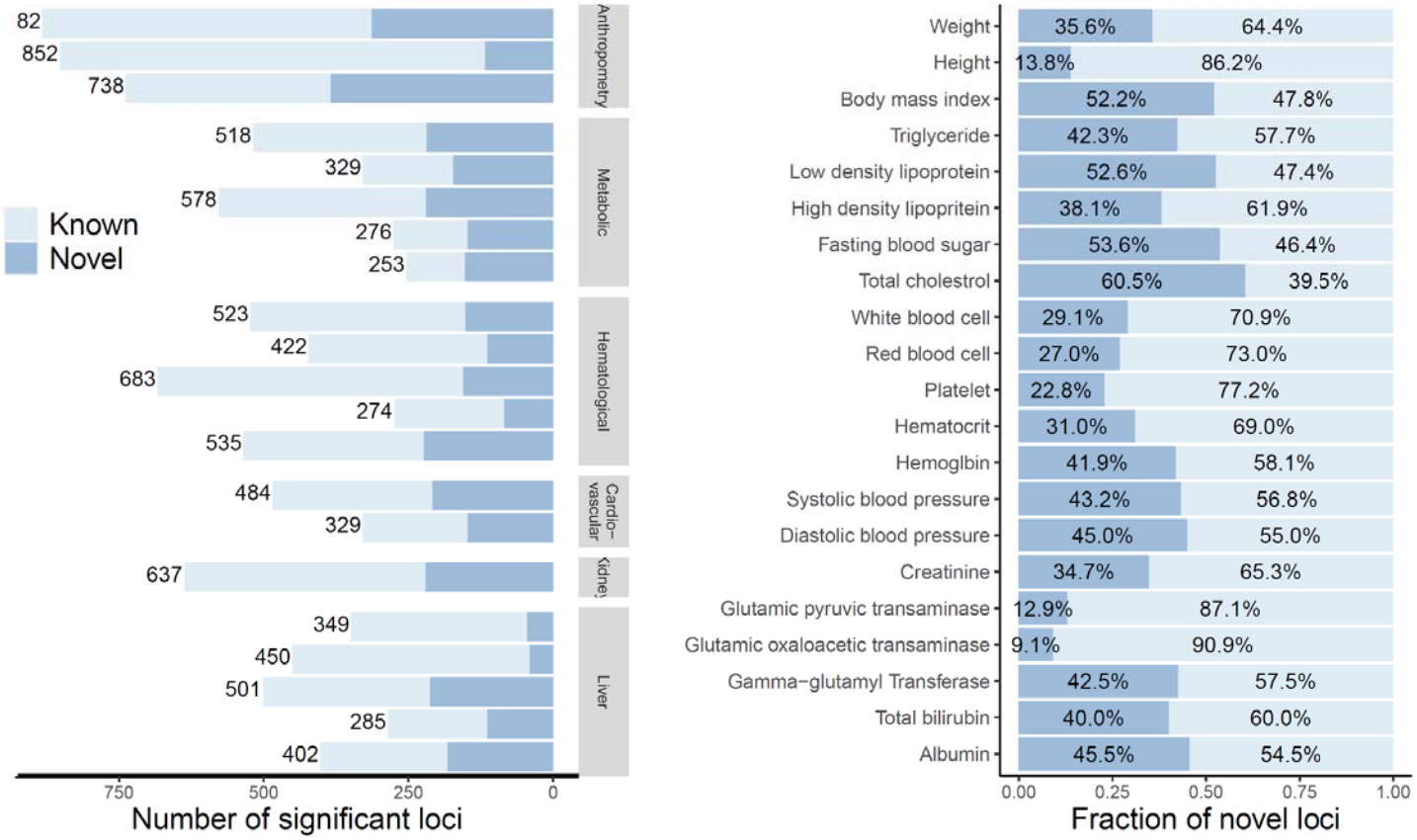
Supplementary Figure 1. MAF comparisons of 2,962 independent genome-wide significant loci in KCPS2 with MAF in 1000 Genomes Project (1KG) EUR population.



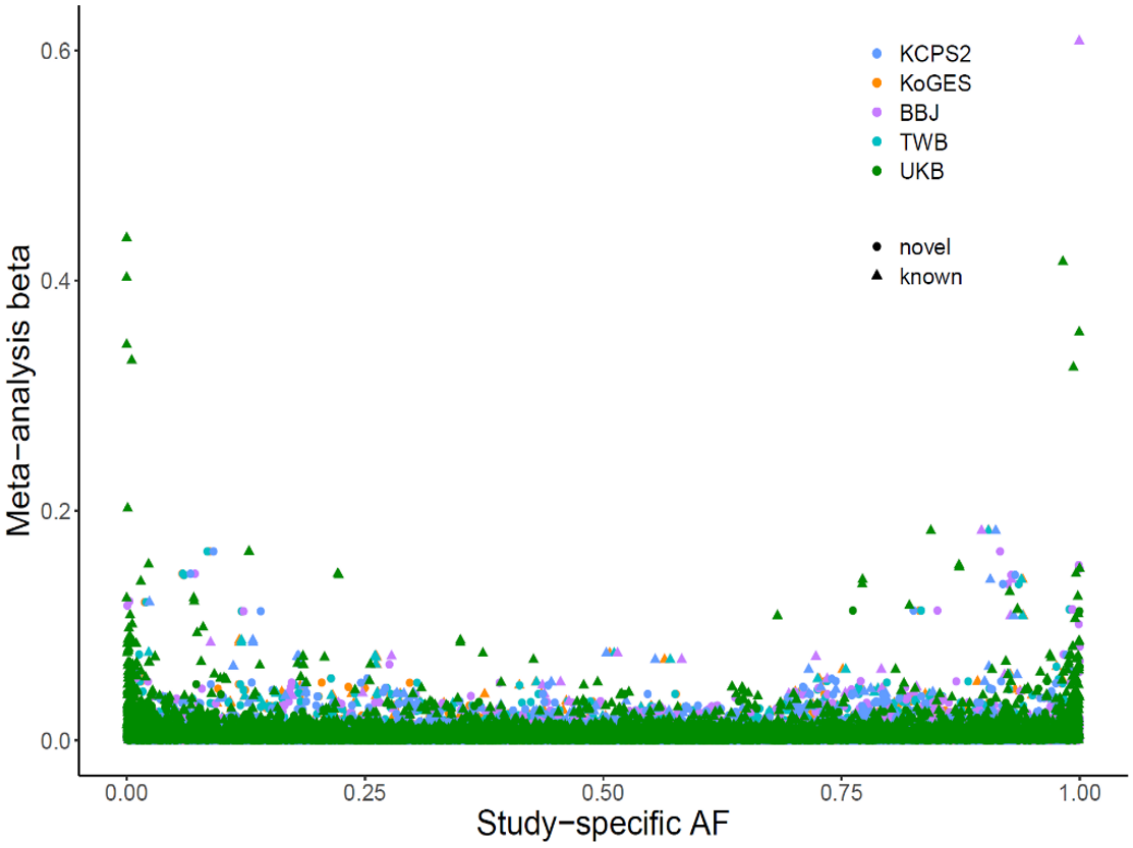
Supplementary Figure 2. Pairwise genetic correlations (r_g, upper diagonal) and phenotypic correlations (r_p, lower diagonal) between the 36 traits in KCPS2. r_g was estimated using bivariate LDSC based on association test statistics from linear regression. Significant r_g and r_p after false discovery rate correction is indicated by an asterisk sign (two-sided Wald test). The complete set of r_g and r_p, is available in Table S5.



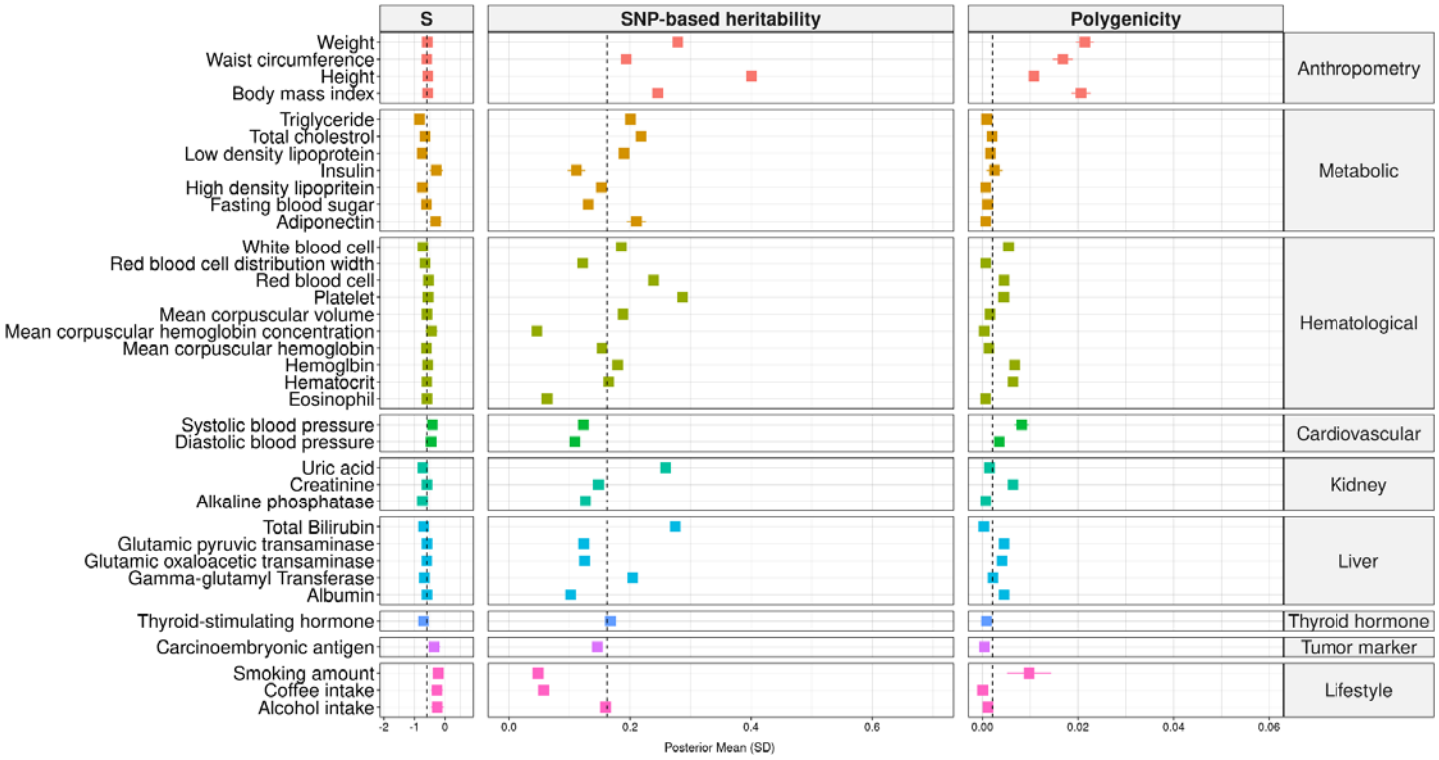
Supplementary Figure 3. Number of known and novel variants identified in the meta-analysis across Korean Cancer Prevention study-II (KCPS2), Biobank Japan (BBJ), Korean Genome and Epidemiology Study (KoGES), Taiwan Biobank (TWB), and UK Biobank (UKB). We identified the association as novel if none of the variants within the locus reached genome-wide significance ($P < 5x10^{-8}$) in KoGES, BBJ, TWB, or UKB GWAS.



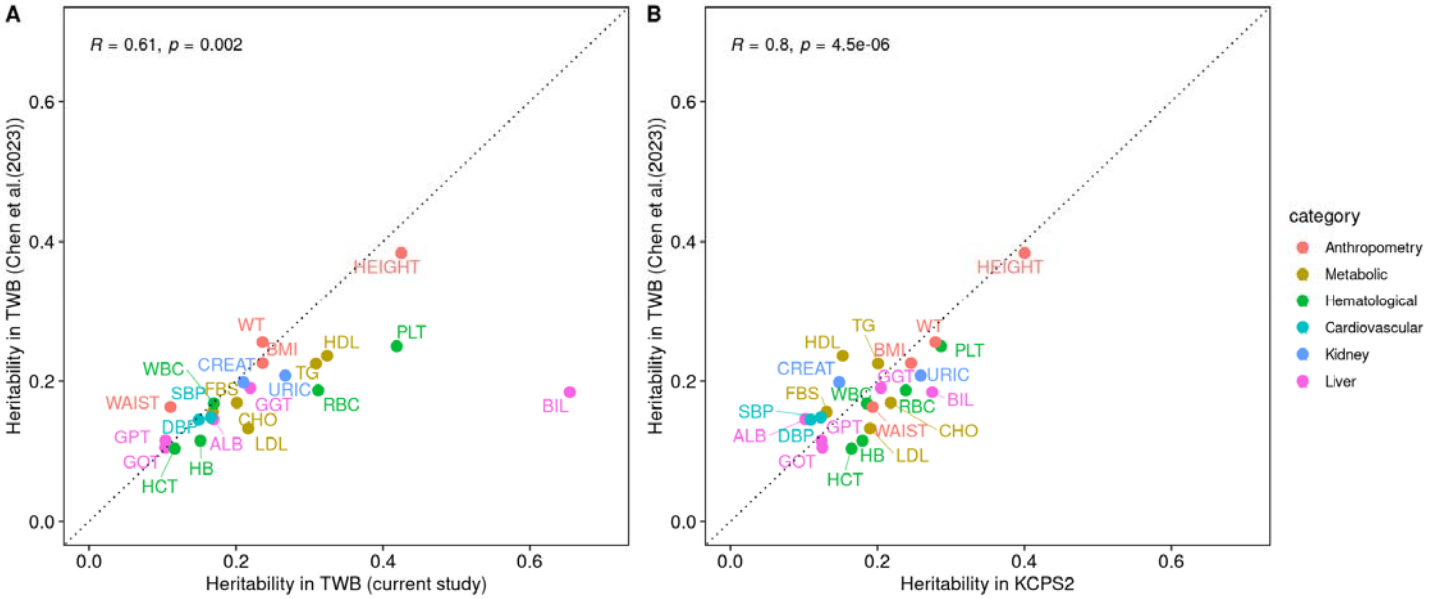
Supplementary Figure 4. Comparisons of study-specific allele frequencies and effect sizes estimated from the multi-ancestry meta-analysis. We identified the association as novel if none of the variants within the locus reached genome-wide significance ($P< 5x10^{-8}$) in KoGES, BBJ, TWB, or UKB GWAS.

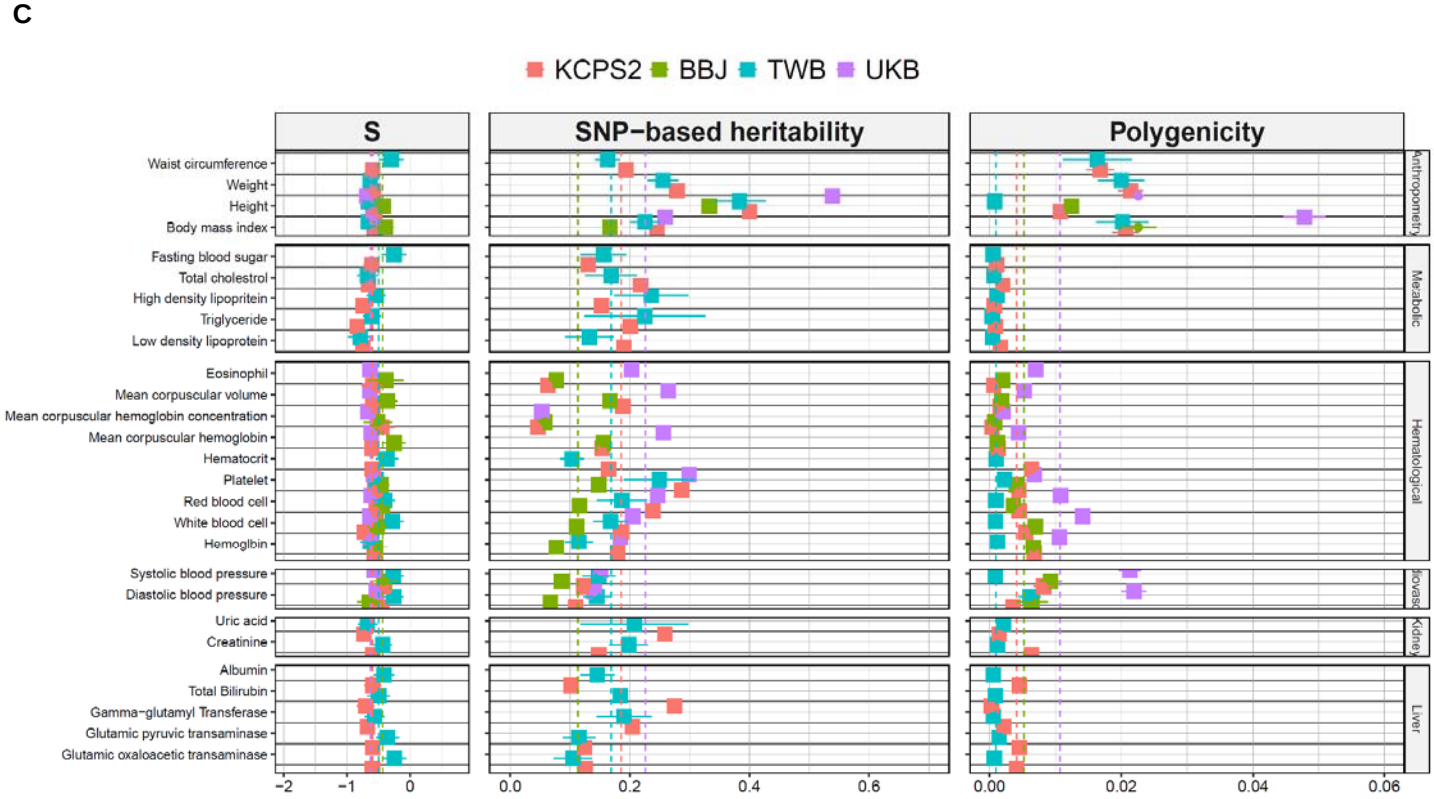


Supplementary Figure 5. Genetic architecture of 36 traits in Korean Cancer Prevention study-II (KCPS2). The dots represent posterior means and horizontal bars represent standard errors of the parameters for each trait. The vertical dashed line shows the median of the estimates across traits.



Supplementary Figure 6. Comparison of TWB heritability estimates using SbayesS and KCSP2 LD matrix vs. TWB heritability estimates reported by Chen et al., (2023)[15] using LDSC and in-sample LD from TWB. **A)** Comparisons between TWB heritability estimates using SbayesS and KCSP2 LD matrix (X-axis) and. TWB heritability estimates reported by Chen et al., (2023)[15] using LDSC and in-sample LD from TWB (Y-axis). **B)** Comparisons between heritability estimates in KCPS2 (X-axis) and heritability in TWB reported by Chen et al., (2023)[15] (Y-axis). **C)** Genetic architecture of these traits when replacing TWB heritability with the estimates reported by Chen et al., (2023)[15].





Supplementary Figure 7. Comparison of within- and cross-biobank genetic correlation estimates for 21 quantitative traits in KCPS2, KoGES, BBJ, TWB, and UKBB (r_g). (a-c) The r_g estimates within EAS were computed in LDSC[74] using 1000 Genomes Project EAS reference panel. (d-f) The cross-biobank population genetic effect correlations between KCPS2 and UKB were estimated in Popcorn (v.1.0)[84] using precomputed cross-population scores for EUR and EAS 1000 Genomes Project populations (Tables S9).