Genome-wide association study of 448,558 East Asians reveals ancestry-specific and shared genetic architecture of educational attainment

Ruilei Ma[1,2,3], Keiko Hikino[4,5,6,7], Takahisa Kawaguchi[8], Xiaoxi Liu[4], Taisei Mushiroda[4], Yuxuan Zhao[9], Qi Liu[9,10], Linxuan Li[11,12], Hebin Zhang[11,12], Han Xiao[13], Lizhong Wang[14,15,16], Senwei Tang[14,15,16], Teng Li[14,15,16], Hao Mi[17], Fengzhe Xu[18], Kouichi Ozaki[19], Yasuharu Tabara[8], Robin G. Walters[20], Zhengming Chen[20], Jun Lv[9,21,22,23], Liming Li[9,21,22], Aifen Zhou[13,24], Huanhuan Zhu[11,25], Xiong-Fei Pan[26], An Pan[27], Yu-Ming Chen[28], Yimin Zhu[29], Dan Zhou[30], Peter M. Visscher[31,32], Ting Qi[2,3,33], Ju-Sheng Zheng[*18], Kun Zhou[*17], Gang Chen[*14,15,16,34], Xin Jin[*11,25,35,36,37], Dianjianyi Sun[*9,21,22], Chikashi Terao[*4,7,38,39], Jian Yang[*2,3]

Educational attainment (EA) is a complex trait that has been extensively studied through genome-wide association studies (GWAS) in individuals of European (EUR) ancestry. However, the genetic basis of EA in non-EUR ancestries remains underexplored. To bridge this gap, we conducted the largest EA GWAS in East Asian (EAS) ancestry to date ($n = 448{,}558$), identifying 33 independent genome-wide significant associations and 11 EAS-specific signals independent of EUR-reported associations. Leveraging multiple gene prioritization methods, we identified 28 high-confidence genes whose functions implicate neurobiological mechanisms underlying EA. For instance, we identified an EAS-specific signal in a coding variant of *CBLN2*, a gene experimentally shown to maintain cognitive function in mice. By integrating human single-cell and mouse spatial transcriptomics data with EA GWAS across and within ancestries, we resolved the EA genetic associations to specific cell populations, including region-dependent neuronal cell types and non-neuronal subpopulations. These findings point to a spatially constrained cellular mechanism conserved across ancestries, through which the genetic component of EA manifests its biological function. Our study highlights the importance of incorporating GWAS data from diverse ancestries to identify novel genetic associations and provides a comprehensive understanding of the polygenic architecture of EA across ancestries.

https://doi.org/10.15302/vita.2026.11.0089

INTRODUCTION

Educational attainment (EA), defined as the highest level of education an individual completes, is influenced by both environmental and genetic factors[1]. It is closely related to social outcomes such as income, civic engagement, and overall health, making it a subject of significant interest in social and genetic research[2,3]. Genome-wide association study (GWAS) is a widely used study design to explore the genetic architecture of complex traits such as EA. By analyzing large-scale population data, GWAS can associate specific genetic polymorphisms with variation in educational level, which helps elucidate the biological underpinnings and relevant molecular pathways[4].

The genetic architecture of EA has been extensively studied in individuals of European (EUR) ancestry, with thousands of associated loci identified. These studies have revealed that EA heritability is enriched in functional genomic regions active in the central nervous system, particularly neuronal cell types, highlighting the neurobiological basis of the trait[5,6]. Beyond its social and behavioral implications, the polygenic risk score (PRS) for EA has significant predictive power and can improve the prediction of complex diseases, including cardiometabolic and psychiatric disorders[7]. Despite these advances, the genetic basis of EA in non-EUR ancestries remains underexplored. This is a substantial limitation because pronounced divergence in ancestry-specific allele frequencies and linkage disequilibrium (LD) patterns and potential heterogeneity in gene-by-environment interactions across global education systems mean that relying solely on EUR-based GWAS constrains a comprehensive understanding of the genetic components of EA. The consequences of this limitation are evident in the low cross-ancestry transferability of EA PRS[7,8], which can occur even in the absence of ancestry differences in causal genetic effects. Although a previous EA GWAS in East Asian (EAS) ancestry demonstrated improvements in fine-mapping and PRS power by incorporating multi-ancestry data[9], it failed to identify novel loci, and a comprehensive understanding of the ancestry differences in the genetic architecture of EA is still lacking. Considering the substantial disparity in sample sizes and the limited number of associa-

A list of affiliations appears at the end of the paper. *Correspondence: Ju-Sheng Zheng (zhengjusheng@westlake.edu.cn), Kun Zhou (kunbone@23mofang.com), Gang Chen (cg@wegene.com), Xin Jin (jinxin@genomics.cn), Dianjianyi Sun (dsun1@bjmu.edu.cn), Chikashi Terao (chikashi.terao@riken.jp), Jian Yang (jian.yang@westlake.edu.cn)
Received: April 2, 2026; Accepted: November 9, 2026; Published: November 17, 2026

tions discovered in EAS GWAS, significant gaps persist in understanding how the genetic architecture of EA diverges across ancestries.

To address this gap, we conducted a large-scale meta-analysis of EA GWAS comprising 448,558 individuals of EAS ancestry (Fig. 1). We identified a total of 33 independent loci, including novel loci not previously reported in EAS. We also defined EAS-specific association signals by conditioning on EUR-reported signals. Leveraging genetic correlation and Mendelian randomization (MR) analyses, we investigated genetic sharing between EA and other complex traits. Multiple gene prioritization methods and functional annotation datasets were used to further uncover the functional genes and potential mechanisms underlying these associations. Finally, we mapped region-specific cell type associations with EA and the spatial distribution of EA-associated cells in the brain by incorporating single-cell and spatial transcriptomics data.

RESULTS

Within- and cross-ancestry meta-analysis

We collected EA GWAS summary statistics from seven cohorts in mainland China (CKB, BGI-lcWGS, WGEG, WGEA, 23mofang, WeGut, and WGEC) and meta-analyzed GWAS data from six cohorts in Japan (denoted as JPN-meta), totaling 283,326 EAS individuals (Fig. 1; Supplementary Table S1). To assess and control for potential inflation caused by population stratification, we calculated the attenuation ratio using linkage disequilibrium score (LDSC) regression[10] for each cohort (Supplementary Table S2). For large cohorts with an LDSC ratio > 10% (BGI-lcWGS and JPN-meta), we corrected their test statistics using the LDSC intercept. We further incorporated the Chen *et al.*'s EA GWAS data from EAS ancestry, comprising 165,232 individuals from the Taiwan Biobank (TWB) and the Korean Genome and Epidemiology Study (KoGES)[9]. The genetic correlation between the meta-analysis of our eight EAS cohorts and that of Chen *et al.* is 0.994 (s.e. = 0.0426), confirms that the two datasets capture the same underlying genetic signals for EA. After quality control, we performed an inverse-variance-weighted meta-analysis[11] of all EAS datasets, representing the largest EA GWAS in EAS ancestry to date ($n = 448,558$). The LDSC ratio indicated no significant evidence of inflation from confounding bias in the test statistics (LDSC intercept = 1.019, s.e. = 0.014 ; attenuation ratio = 0.031, s.e. = 0.023; Supplementary Fig. S1). Of the 9,837,352 tested variants, we identified 732 common variants that passed the genome-wide significance threshold ($P < 5 \times 10^{-8}$; Fig. 2a). The standardized effect sizes of these variants ranged from 0.012 to 0.036 in absolute value, corresponding to approximately 2.6 to 7.9 weeks of schooling per allele. The SNP-based heritability was estimated at 6.1% by LDSC, which is lower than the previously reported estimate in the EAS cohort (9.0%), potentially reflecting increased cross-cohort heterogeneity from the more diverse datasets included in our meta-analysis, as well as cohort-specific characteristics and unmeasured gene-by-environment effects. LD-based clumping[12] of the GWAS summary statistics revealed 33 independent variants (Supplementary Table S3), including 25 variants located outside the seven genome-wide significant loci reported by Chen *et al.*[9] (Materials and Methods). Statistical fine-mapping at the 33 loci (1-Mb genomic regions centered on the lead variants) using CARMA[13] identified 29 credible sets (CSs) with a 99% probability of including a causal variant, and 50 genome-wide significant variants with a posterior inclusion probability (PIP) larger than 0.1 (Supplementary Tables S4, S5). Conditional and joint analysis (COJO)[14] revealed 31 independent lead variants, including two secondary signals that were masked in single-variant analysis (Supplementary Table S6). Conditioning on genome-wide significant variants reported in the previous EAS GWAS of EA, we identified 521 associated variants that remained genome-wide significant (23 independent loci after clumping; Supplementary Table S36), illustrating the increased statistical power and unique findings achieved by incorporating new datasets.

We calculated the genetic correlation of EA between EAS and EUR using the largest publicly available EUR summary statistics from Lee *et al.* ($n = 766,345$; EA3)[6]. The estimated genetic correlation was 0.768 (s.e. = 0.0245). Although the cross-ancestry genetic correlation reported by the previous EAS study using a different method[9] is higher than ours, the 95% confidence intervals overlap. We also performed age-, sex-, country-, and sampling-characteristic-stratified meta-analyses in our EAS cohorts, and the cross-ancestry genetic correlations varied across stratified groups (Supplementary Note 2 and Tables S44–S46). Overall, these cross-ancestry genetic correlation estimates are not significantly different from one another given the large estimation uncertainty, consistent with a generally shared genetic architecture of EA between ancestries. Because full summary statistics for the more recent EUR meta-analysis by Okbay *et al.* ($n = 3,037,499$; EA4)[7] were unavailable, we utilized EA3 for a cross-ancestry meta-analysis ($n_{meta} = 1,214,903$; Mean $\chi^2 = 2.815$), which identified 809 independent signals. We subsequently compared these signals with the 3,952 reported lead variants from EA4, revealing 23 potential novel signals located more than 500 kb away from, and showing LD $r^2 < 0.1$ in both EAS and EUR with, any previously reported lead variant in the largest EUR meta-analysis (Fig. 2b; Supplementary Table S7). Furthermore, we found that 73.6% of the genome-wide significant variants discovered only in the EAS GWAS exhibited a higher minor allele frequency (MAF) in EAS than in EUR (EA3). The mean ratio of expected heterozygosity between EAS and EUR across the EAS-identified variants was 1.45, indicating significantly higher genetic variation in EAS at these loci. Of these variants, 98.7% exhibited a consistent direction of effect with EUR, and 76.3% had larger absolute effect sizes in EAS; moreover, all variants discovered only in the EAS GWAS had larger absolute effect sizes in EAS than in EUR (Supplementary Fig. S2). Although we observed some EUR-only signals with larger effect sizes in EAS, none of these variants passed the genome-wide significance threshold in our EAS GWAS (Supplementary Table S40). This demonstrates that, due to the increase in the non-centrality parameter (NCP) of the test statistic driven by both MAF divergence and ancestry-specific effect sizes, our GWAS data can uncover novel associations that would otherwise remain undetectable in EUR studies despite their much larger sample sizes. These results underscore the significance of ancestry diversity in achieving a more comprehensive understanding of EA.

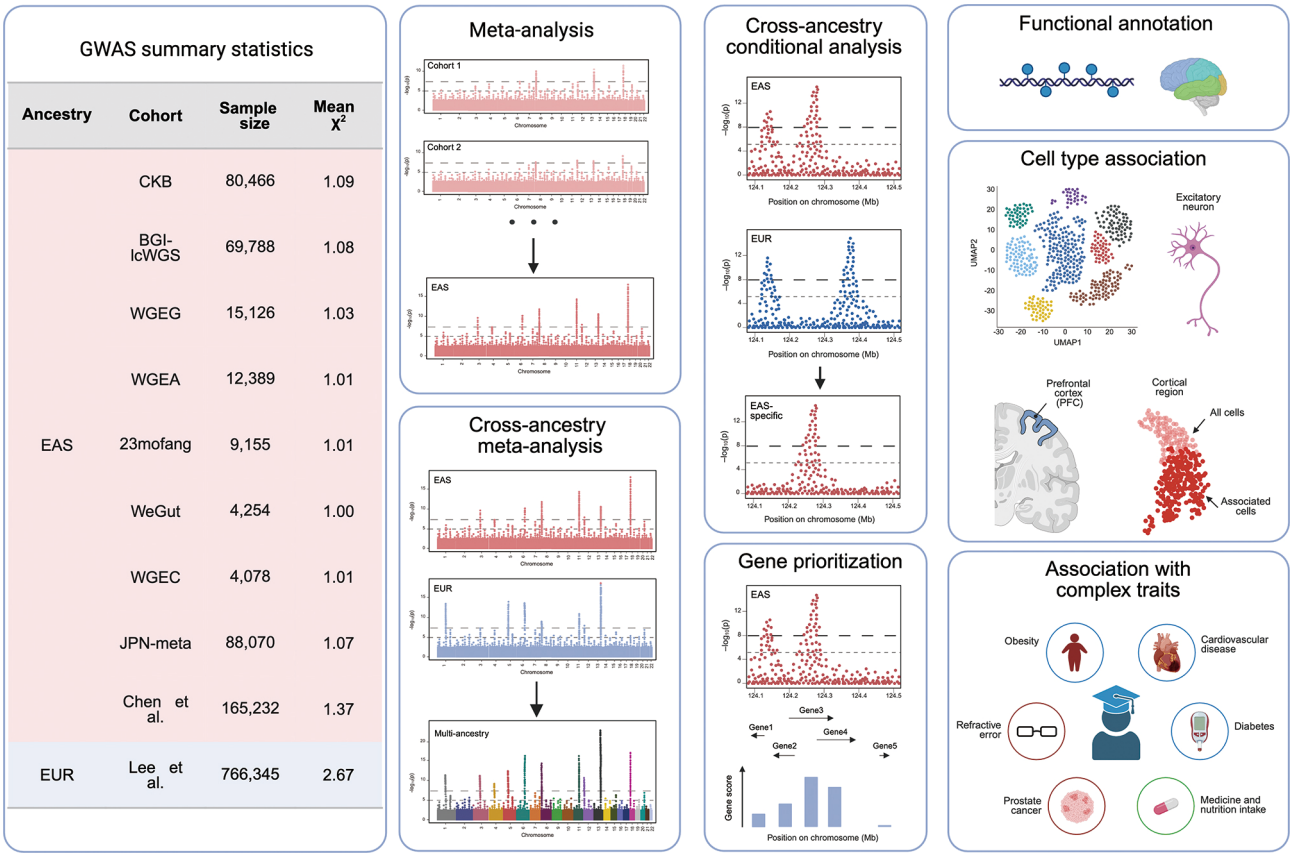


Fig. 1 Schematic overview of the study design. The analytical framework includes cohort-level GWAS, meta-analysis, conditional analysis, gene prioritization, genetic correlation and putative causal associations with other complex traits, functional annotation, and cell-type association analyses.

To evaluate the predictive power of our GWAS, we constructed PRS using our EAS GWAS data, the EUR GWAS data (EA3), and the tuned cross-ancestry ensemble of these two datasets (Materials and Methods). We quantified their performance via the incremental R^2 in a holdout sample of EAS individuals from the UK Biobank (UKB). The EAS-only and EUR-only PRS explained 1.58% and 0.70% of the phenotypic variance, respectively. The cross-ancestry PRS reached 1.59% (Supplementary Fig. S3 and Tables S8, S9). While the 95% confidence intervals of the prediction R^2 overlapped, reflecting the constraints of the test cohort's sample size, the nominal advantage of the EAS PRS over the EUR PRS, coupled with the marginal gain of the cross-ancestry PRS over the EAS PRS, is consistent with the influence of ancestry differences in LD and allele frequencies on PRS transferability, with possible additional contributions from ancestry-specific genetic architecture. Sensitivity analyses, including comparison with the previous EAS study, leave-one-cohort-out validation, and an alternative UKB-based LD reference, yielded consistent results, with the cross-ancestry PRS showing the clearest advantage in the leave-one-cohort-out analyses (Supplementary Note 4 and Table S37).

Identification of EAS-specific genetic associations

To identify EAS-specific genetic associations, we performed conditional analyses by adjusting the EAS GWAS summary statistics for genetic associations reported in EUR. For this analysis, we utilized the 3,952 independent lead variants reported by Okbay *et al.* (EA4), which represent the largest and most contemporary set of EA associations identified in EUR to date. Given the challenges posed by cross-ancestry LD differences, the potential absence of EUR lead variants in the EAS data, and the risk of collinearity among covariate variants when modeled using EAS LD, we implemented a two-step refinement process. First, we expanded the lead variants into a dense list using a EUR LD reference panel[15] to include all variants in LD with the lead variants (LD $r^2 > 0.1$), ensuring that proxy variants could be used even if certain EUR lead variants were absent in the EAS data. Second, to prevent collinearity in the conditional analysis model, we employed model selection implemented in GCTA-COJO to identify jointly significant variants from this expanded set based on EAS GWAS and LD data. This process was designed to maximize the inclusion of EUR signals as covariates in the conditional analysis model while accounting for two key factors: the unavailability of full summary statistics from EA4 and the inherent differences in LD structure between EAS and EUR (Materials and Methods).

We identified 88 genome-wide significant variants after conditioning on EUR signals (Fig. 2a). Notably, 18.8% of these variants were located in coding regions, and 28.3% were located in functional regions such as promoters and enhancers (Materials and Methods). Compared to all tested variants, these variants showed significant enrichment in coding regions ($OR = 2.66; P = 1.10 \times 10^{-3}$). LD-based clumping of the conditionally significant associations identified 11

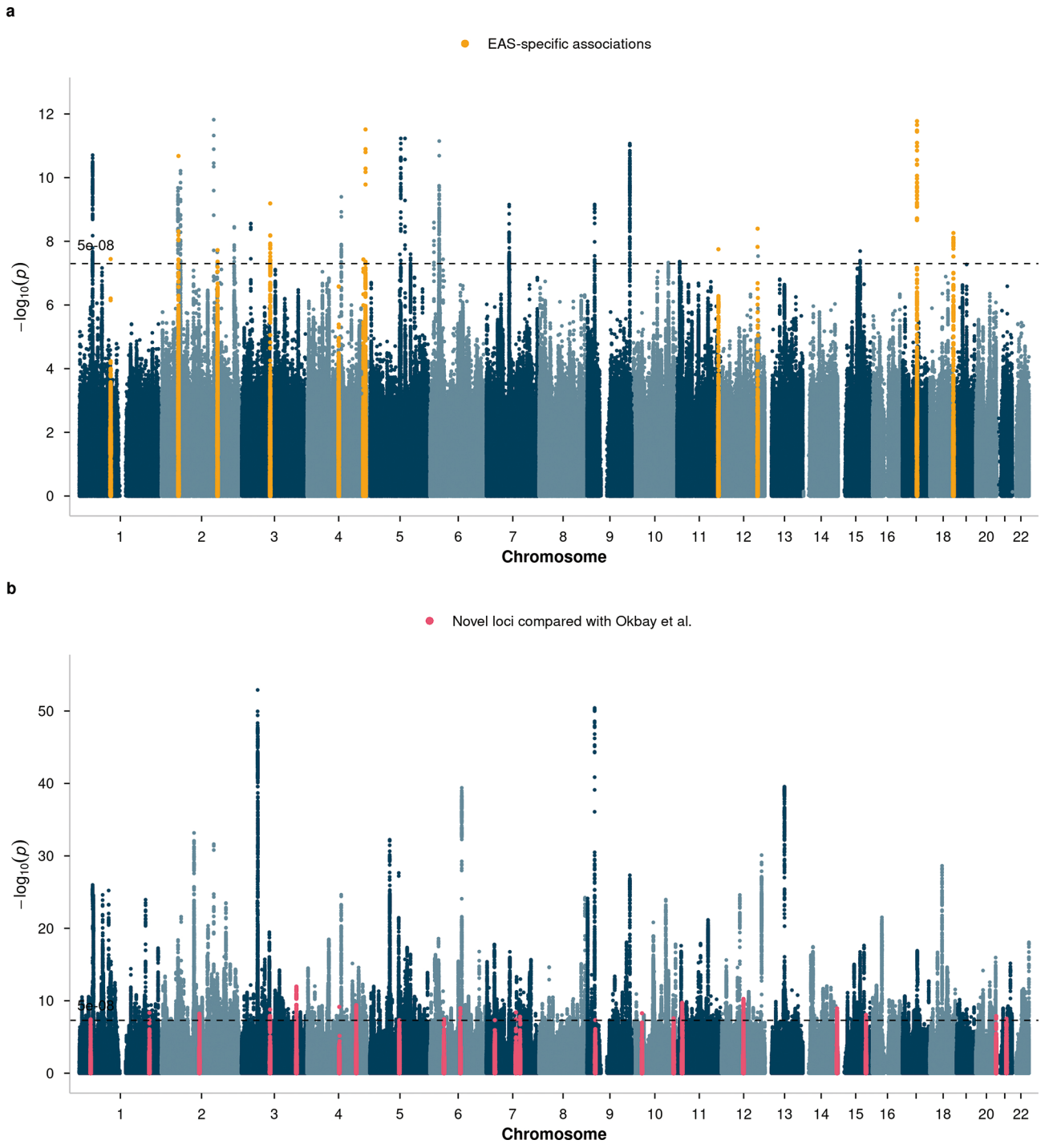


Fig. 2 Genome-wide association results for educational attainment in EAS and cross-ancestry meta-analyses. a Manhattan plot of GWAS in EAS meta-analysis. The dashed line represents the genome-wide significance threshold ($P = 5 \times 10^{-8}$). Yellow dots denote 11 EAS-specific association signals identified after conditioning on previously reported EUR signals. **b** Manhattan plot of the cross-ancestry meta-analysis of EA (EAS and EA3). Pink dots represent novel genomic regions identified in the meta-analysis when compared with the EUR GWAS from Okbay *et al.* (EA4).

independent EAS-specific associations (Supplementary Table S10). One of the associations, a secondary signal in the original EAS meta-analysis (Supplementary Table S6), also reached genome-wide significance after conditioning on the EUR associations. Four associated variants were located more than ± 500 kb from any EUR-reported lead variant (specifically rs10893782 on chr11, rs1844434 on chr3, rs4410534 on chr4, and rs3809297 on chr12). The remaining seven associated variants mapped within established EUR-reported loci, representing EAS-specific signals distinct from the EUR-reported signal at the same locus (Supplementary Fig. S4). These findings indicate potential ancestry-specific genetic architecture underlying EA.

Prioritization of high-confidence EA-associated genes

To dissect the function of EA-associated loci identified in our

EAS GWAS, we integrated multiple gene prioritization methods, functional annotation datasets, and molecular quantitative trait loci (xQTL) datasets to identify EA-associated genes and generated gene rankings based on the number of methods by which each gene was prioritized (Materials and Methods), thereby enhancing the robustness and interpretability of the prioritized genes. Our framework incorporates six gene discovery methods based on four rationales[16]: 1) gene-based association analysis (MAGMA[17] and mBAT-combo[18]); 2) variant-to-gene (V2G) mapping; 3) integrative analysis of GWAS and xQTL (SMR[19] and COLOC[20]); and 4) network-based prioritization (PoPS[21]). The V2G mapping incorporated four types of cis-regulatory element (cCRE)-gene connection annotations across various tissues and cell types, while the integrative analysis of GWAS and xQTL leveraged four layers of EAS xQTL datasets from multiple tissues and cell types.

Using this integrative framework, we identified 111 protein-coding genes underlying the genome-wide significant loci, nearly half of which (46.8%) were supported by multiple gene prioritization methods (Fig. 3a; Supplementary Tables S11–S17). While COLOC identified a high number of unique genes, the overall intersection across all methods was limited, highlighting the inherent challenges in achieving convergence among distinct prioritization methods. Consequently, we defined a 'high-confidence' set of 28 genes, supported by three or more independent methods (Fig. 3b), as such cross-method convergence provides robust evidence for their functional association with EA. For example, *MED27* was prioritized by all six methods and has not been reported by previous EA studies. The EA association signal at the *MED27* locus colocalized with the *MED27* eQTL signal (COLOC PP4 = 0.85), and the lead variant was linked to *MED27* by an active enhancer–promoter interaction detected by PCHiC in brain, suggesting a potential regulatory mechanism by which EA-associated variants impact its expression, implicating this gene in pathways relevant to EA (Fig. 3c). *MED27* is a known component of the mediator complex that activates transcription, while loss-of-function mutations in *MED27* have recently been reported to cause neurodevelopmental disorders[22-24], indicating its critical role in neurological function.

Moreover, we prioritized four high-confidence genes through EAS-specific associations, including *DCAKD*, *NMT1*, *PLCD3*, and *CBLN2*, which represent novel biological targets derived from the incorporation of EAS data. *DCAKD*, a gene predicted to be involved in the coenzyme A biosynthetic process, was also prioritized by all six methods. The lead variant (rs4525538 on chromosome 17) of its eQTL in a Chinese brain dataset suggested a potential mechanism in which increased *DCAKD* expression contributes to higher EA ($b_{SMR} = 0.156$; $P_{SMR} = 1.73 \times 10^{-6}$). The *DCAKD* locus also exhibits an EAS-specific association (the lead variant, rs7502512, was linked to *DCAKD* by a potential enhancer detected via the activity-by-contact interaction; Supplementary Fig. S14), with a similar MAF but a much larger effect size in EAS than in EUR ($MAF_{EAS} = 0.458$, $MAF_{EUR} = 0.347$; $\hat{b}_{EAS} = 0.0152$, $\hat{b}_{EUR} = 3.56 \times 10^{-3}$). Another one of these prioritized targets, *CBLN2* has been experimentally demonstrated to promote prefrontal spinogenesis, an essential process in neuron development[25], and has been shown to maintain cognitive function and task performance in mice[26], strengthening its functional association with EA. However, neither the *CBLN2* gene nor the GWAS locus has been reported in previous GWAS studies in EUR (Fig. 3d), likely due to the pronounced difference in allele frequencies of variants at this locus. For instance, a coding variant, rs8094389, in *CBLN2* showed a large divergence in its MAF between ancestries ($MAF_{EAS} = 0.487$; $MAF_{EUR} = 0.037$; expected-heterozygosity ratio = 7.01).

The remaining high-confidence genes can also be categorized into two major biological domains: neurobiology and transcriptional regulation. For example, *NRXN1*, *MCTP1*, and *LRFN2* are essential components of neurotransmission, with *NRXN1* and *MCTP1* involved in calcium signaling and *LRFN2* involved in synaptic adhesion. The critical nature of these genes is underscored by previous studies that linked *NRXN1* to several psychiatric disorders[27,28] and *LRFN2* to learning disabilities[29]. Additionally, we identified several genes involved in cellular regulation, including *CUL3* and *AUTS2*, both of which are established candidate genes for diverse neurodevelopmental disorders[30,31], as well as several zinc finger protein-coding genes, including *ZKSCAN4* and *ZSCAN9*.

These high-confidence genes demonstrate the power of multiple-method integration and ancestry diversity for gene discovery, suggesting that genetic variants may influence EA by regulating genes that mediate neurotransmission, neurodevelopment, and transcriptional pathways. Collectively, this list of high-confidence genes, prioritized through multiple independent lines of evidence, provides a strong foundation for downstream experimental validation of causal mechanisms.

Genetic correlation and causal inference across ancestries

We used LDSC to estimate the genetic correlation (r_g) between EA and 259 complex traits in EAS (Supplementary Table S18), identifying 48 traits with significant correlations (FDR < 0.05), spanning four domains: biomarkers, diseases, drug usage, and nutrition intake (Fig. 4). We observed a significant inverse genetic correlation between EA and the risk of metabolic and cardiovascular diseases, including type 2 diabetes and heart failure. These results align with the negative correlations between EA and its corresponding biomarkers such as body mass index (BMI), fasting glucose, and triglycerides, and the positive correlation with high-density lipoprotein-cholesterol (HDL-C). Moreover, EA showed negative correlations with corresponding drug usage, such as diuretics and calcium channel blockers, and positive correlations with nutrition intake. However, the risks of myopia (indicated by negative values of refractive error measures[32]), schizophrenia, and prostate cancer (PRCA) were observed to be positively correlated with EA. Most of these significant genetic correlations were consistent with those observed in EUR, with age at menarche, schizophrenia, pulmonary tuberculosis, and glucose in urine significantly genetically correlated with EA in EAS only. Notably, the risks of glaucoma and uterine fibroids showed significant but opposite signs of genetic correlations with EA in different ancestries (Supplementary Table S18).

To investigate putative causal relationship between EA and these traits, we performed GSMR[33] analyses using EA as the exposure. Because there is potential sample overlap between

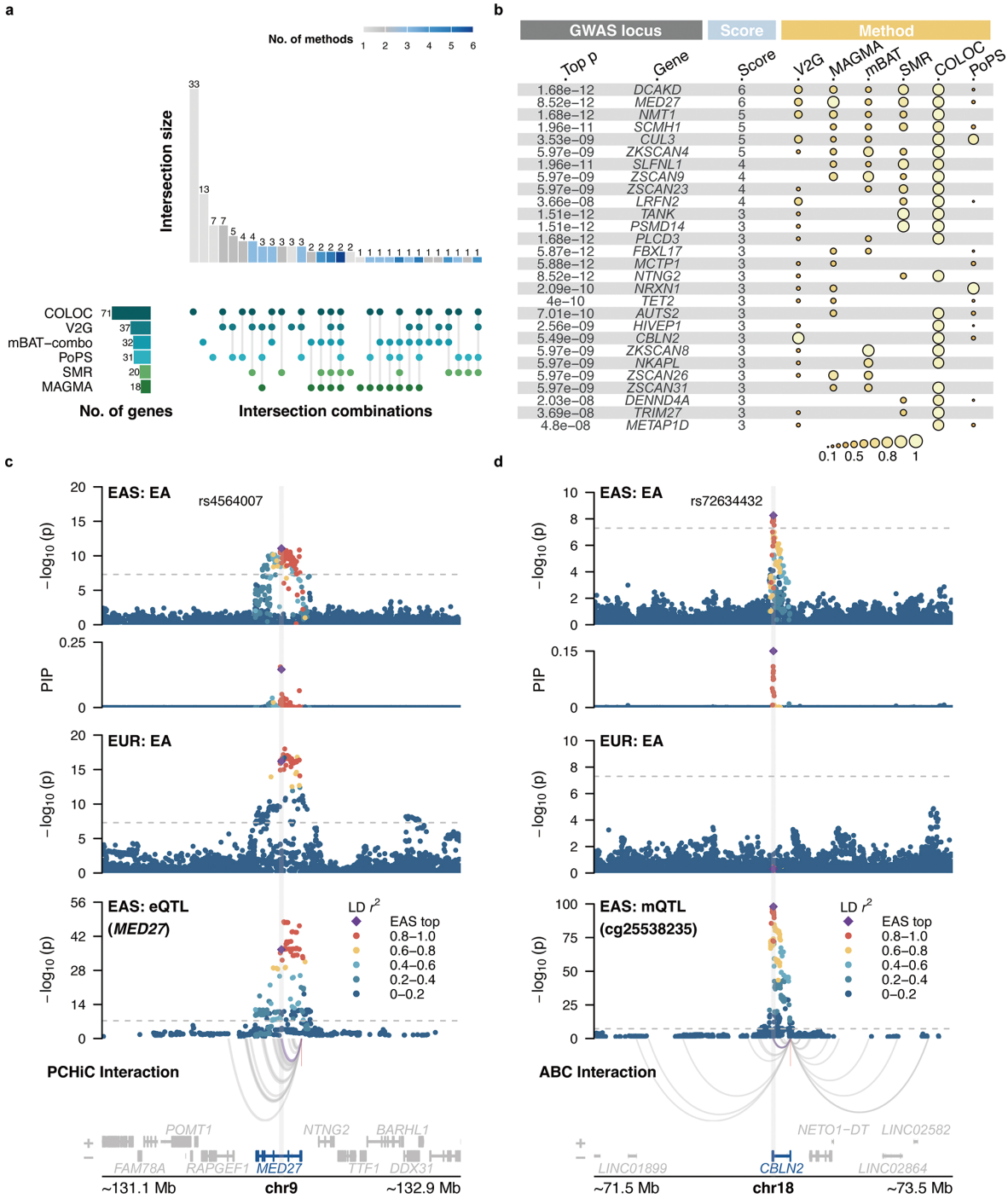


Fig. 3 Prioritization of EA-associated genes in EAS. a Intersection of prioritized gene sets across six methods. The horizontal bar plot (bottom-left) represents the number of genes prioritized by each method. The intersection matrix (bottom) and vertical bar plot (top) indicate the number of shared genes across specific methodological combinations, with bar colors corresponding to the count of supporting methods. **b** High-confidence EA-associated genes. Columns 1–2 list the gene name and the genome-wide association *P*-value for the corresponding lead SNP at the locus. Column 3 indicates the cumulative number of supporting methods (*i.e.*, the sum of binary indicators across the six methods), independent of dot size. The subsequent six columns illustrate specific methodological support for each gene, where dot size scales with the confidence metric of each approach: number of supporting functional annotation categories for V2G, posterior probability of colocalization (PP4) for COLOC, polygenic priority score for PoPS, and $-\log_{10}(P)$ for other methods. To accommodate different statistical outputs (probabilities and $-\log_{10}(P)$), results for each method were converted to method-specific percentiles across all observations. Circle size and color intensity scale with this relative percentile rank across the 28 high-confidence genes. For the corresponding absolute values for each method at key percentiles, see Supplementary Table S43. **c**, **d** Locus plots of the *MED27* and *CBLN2* genes. Each panel displays genetic associations of EA, fine-mapping PIP, and relevant molecular QTL signals in EAS, alongside genetic associations in EUR (EA3) at the locus. The GWAS lead SNP in EAS is denoted by a purple diamond, and the colors of other dots represent LD r^2 with the EAS lead SNP. Below the locus plots, chromatin interaction tracks illustrate epigenomic signatures, with red bars denoting promoter regions and purple arcs indicating interactions between the target gene promoter and the epigenomic marker harboring the lead SNP. The genomic position of the focal gene is shown at the bottom.

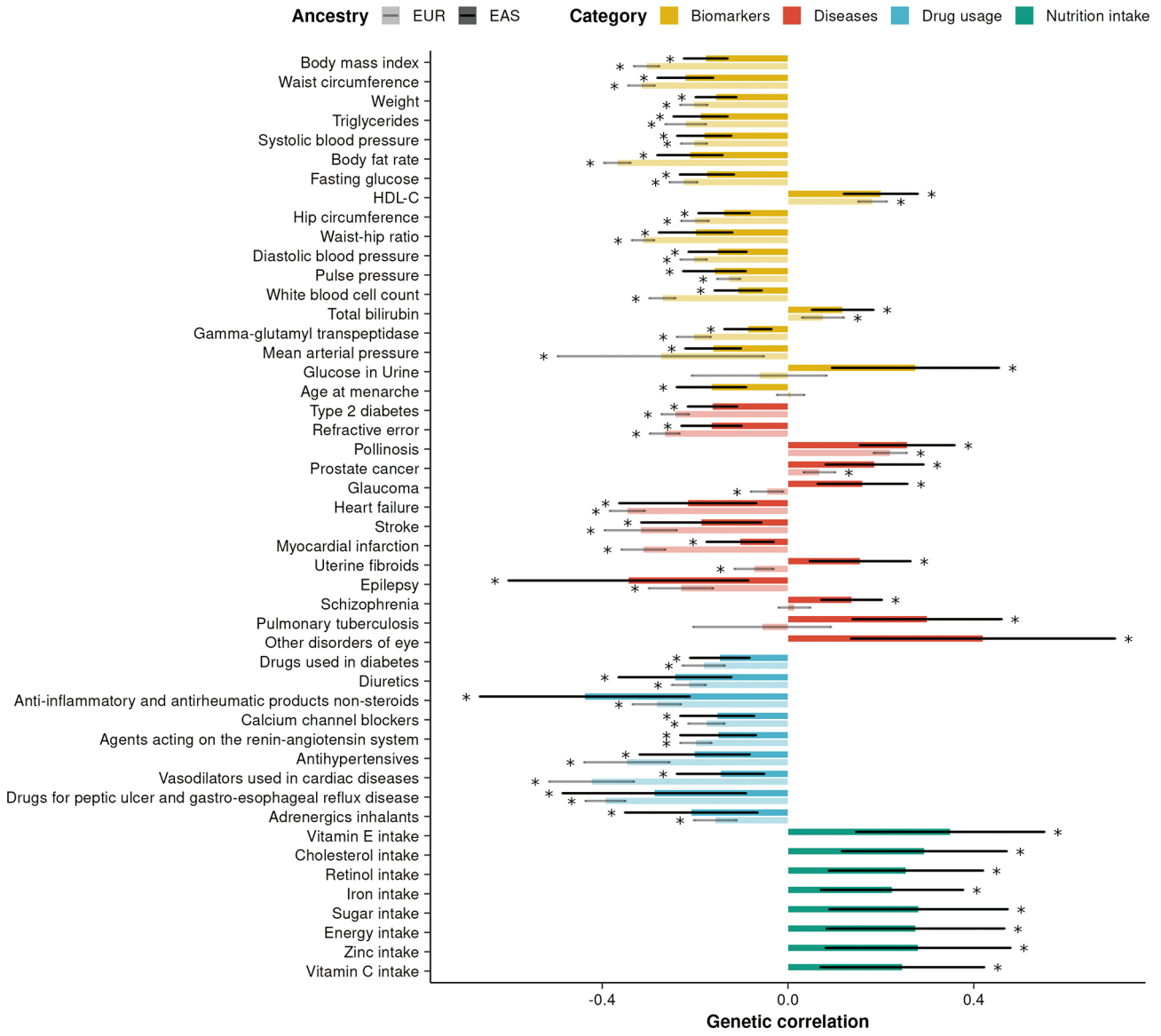


Fig. 4 Genetic correlation between EA and complex traits across ancestries. All 48 traits with significant genetic correlations with EA in EAS, together with their corresponding results in EUR, are shown, categorized into four groups: biomarkers, diseases, drug usage, and nutrition intake. EUR results for nutrition intake traits are not shown because the corresponding GWAS summary statistics were unavailable. Colors represent trait category, and opacity denotes ancestry (solid for EAS, transparent for EUR). Error bars denote 95% confidence intervals. Significant correlations after FDR correction (FDR < 0.05) are marked by asterisks.

some traits and EA, we incorporated two approaches for sensitivity analyses (Materials and Methods), one for traits with an independent GWAS dataset and another for traits without an independent dataset, and excluded traits that were significant in neither approach. We identified putative causal associations between EA and 28 traits in EAS (Supplementary Table S19), with the estimated effect sizes (β_{GSMR}) consistent in direction with the pairwise genetic correlations. BMI (FDR-adjusted $P_{r_g} = 9.31 \times 10^{-14}$; FDR-adjusted $P_{GSMR} = 1.03 \times 10^{-8}$ using an independent GWAS) exhibited the strongest evidence for both genetic correlation and causal relationship with EA. Triangulation across multiple MR methods yielded concordant effect directions; while not all individual causal estimates achieved statistical significance, this directional consistency across multiple models supports the robustness of the identified putative causal trends (Supplementary Figs. S5, S6). We also performed MR analyses using the same methods and the same traits in EUR, and identified 32 traits with putative causal relationships with EA (Supplementary Table S20). All traits significant in both ancestries exhibited the same direction of causal effects. Among the six traits with ancestry-dependent genetic correlations with EA, *i.e.*, opposite in sign between ancestries or significant only in EAS, age at menarche and uterine fibroids showed evidence of a causal relationship with EA only in EAS, with signs of effect sizes consistent with the genetic correlation results. We observed a potential reverse causal effect of BMI on EA across ancestries (Supplementary Tables S33, S34), indicating a bidirectional relationship between them.

Mapping EA-associated cell types and spatial distributions

As in previous studies, we conducted stratified-LDSC (S-

LDSC[34,35]) to characterize the biological contexts in which the genetic variants identified in EAS manifest their effects on EA. Partitioning heritability using S-LDSC with baseline annotations (Materials and Methods) revealed significant enrichment of heritability (FDR < 0.05) in non-synonymous variants and gene-related regulatory elements (Supplementary Fig. S7a and Table S21), such as chromatin regions marked by monomethylated lysine 4 on histone H3 (H3K4me1). Analysis leveraging annotations of tissue-specific activating histone marks further demonstrated significant heritability enrichment in brain-specific regulatory elements (Supplementary Fig. S7b and Table S22). Specifically, chromatin regions marked by H3K4me1 in the male fetal brain showed the strongest enrichment among all annotations (Enrichment $P = 4.41 \times 10^{-9}$), while those marked by acetylated lysine 27 on histone H3 (H3K27ac) in the dorsolateral prefrontal cortex showed the strongest enrichment compared to annotations of other brain regions (Enrichment $P = 1.95 \times 10^{-5}$). These heritability enrichment patterns in brain tissues were highly consistent with previous results in both EAS and EUR ancestries, reinforcing the central nervous system as the most relevant tissue domain for EA.

To better understand how brain function mediates genetic effects on EA, we mapped EA-associated cell types using high-resolution brain single-cell RNA sequencing data (scRNA-seq) and spatial transcriptomics (ST) data. While the EAS-specific genetic associations demonstrate the unique contribution of our data, the high genetic correlation observed between EAS and EUR ($r_g = 0.768$) and the concordant enrichment of biological annotations strongly support the hypothesis that the underlying molecular and cellular architecture of EA is conserved across ancestries. Moreover, the cross-ancestry genetic correlation between EA and IQ was 0.595 (s.e. $= 0.0298$), estimated by Popcorn, similar to the within-ancestry genetic correlation between EA and IQ in EUR ($r_g = 0.724$, s.e. $= 0.011$) by LDSC, suggesting that EA shares core biological mechanisms with IQ across ancestries. To maximize statistical power and deepen biological insights, we leveraged the cross-ancestry meta-analysis summary statistics (integrating our EAS GWAS with the EA3 GWAS) for cell mapping. By applying S-LDSC to the cross-ancestry meta-analysis, we confirmed the gain in power: it recovered the same enrichment profile as the EAS-only analysis (Pearson's r = 0.790 for the enrichment P-values of significantly associated functional genomic regions and 0.677 for tissue-specific open chromatin regions), while identifying more associated regions and tissues with smaller enrichment P-values (Supplementary Tables S38, S39). We then used the cross-ancestry meta-analysis of EA for cell mapping. Acknowledging that these cell-trait associations are driven largely by the larger EUR sample within the cross-ancestry meta-analysis, we validated them using ancestry-stratified summary statistics, which showed highly concordant cell-type association patterns across ancestries, with the cross-ancestry data offering increased power (Supplementary Note 5, Figs. S9–S11 and Tables S28–S32).

We first integrated scRNA-seq data from the human prefrontal cortex (PFC)[36] with GWAS summary statistics using scDRS[37]. Excitatory neurons, inhibitory neurons, and oligodendrocyte precursor cells (OPCs) were significantly associated with EA (FDR < 0.05; Fig. 5a–c). Excitatory neurons exhibited the highest trait score (referred to as 'disease score' in scDRS) and the most significant association among all cell types (Trait score $= 3.67$; $P = 9.99 \times 10^{-4}$; Supplementary Table S23), suggesting that EA-relevant genes have elevated expression in excitatory neurons. These results confirm findings from previous studies that EA heritability is more enriched in neurons than in glial cells, while also refining associations with specific neuronal cell types and identifying additional associations with OPCs, consistent with established evidence that OPCs play a role in neural circuits by forming synapses with neurons[38]. To validate this finding using an alternative method, we performed cell-type enrichment analysis via S-LDSC (Materials and Methods). Consistently, EA heritability was significantly enriched in excitatory and inhibitory neurons (FDR < 0.05; Supplementary Table S24). However, OPCs were not identified as a significantly associated cell type in the S-LDSC analysis. This discrepancy likely stems from differing mapping resolutions between the two frameworks: while scDRS preserves transcriptional heterogeneity within cell types and may capture specialized subpopulations or states of OPCs, the aggregation of cellular-level expression into broad cell-type annotations in S-LDSC may attenuate these signals.

To characterize the spatial distributions of cells associated with EA, we integrated ST data from the mouse brain[39] and GWAS summary statistics using gsMap[40]. We confirmed excitatory neurons as the most relevant cell type for EA ($P = 4.97 \times 10^{-25}$; Supplementary Table S25), and astrocytes exhibited the most significant association with EA among non-neuronal cell types in this dataset ($P = 2.12 \times 10^{-15}$). We also observed heterogeneous association strengths among excitatory neurons across brain regions (Fig. 5d), with the cerebral cortex harboring the most significantly associated excitatory-neuron subpopulations (Supplementary Table S26). These results were highly consistent with findings from additional human scRNA-seq data across 10 brain regions[41], confirming that the cerebral cortex is the primary associated brain region and demonstrating that neuronal subpopulations exhibit region-dependent association strengths (Supplementary Fig. S8 and Table S27). Furthermore, visualizing the spatial distribution of EA-associated cells based on ST data revealed that this distribution was not only heterogeneous across brain regions but also varied within brain regions (especially within the cortex), and that this spatial pattern was independent of predefined cell types (Fig. 5e, f; Supplementary Fig. S10b). Similar findings have been reported in a recent study in the mouse brain, where heterogeneous functions of neuronal subpopulations throughout the cortex were observed in the processes of decision-making[42]. These results strongly suggest that the neurobiological process underlying EA is influenced by spatially dependent cellular mechanisms and local microenvironmental factors.

DISCUSSION

This study represents the largest EA GWAS in individuals of EAS ancestry to date, providing a systematic exploration of the ancestry-specific and conserved genetic architecture of EA across EAS and EUR ancestries. By increasing the EAS sample size, we achieved a threefold increase in identified

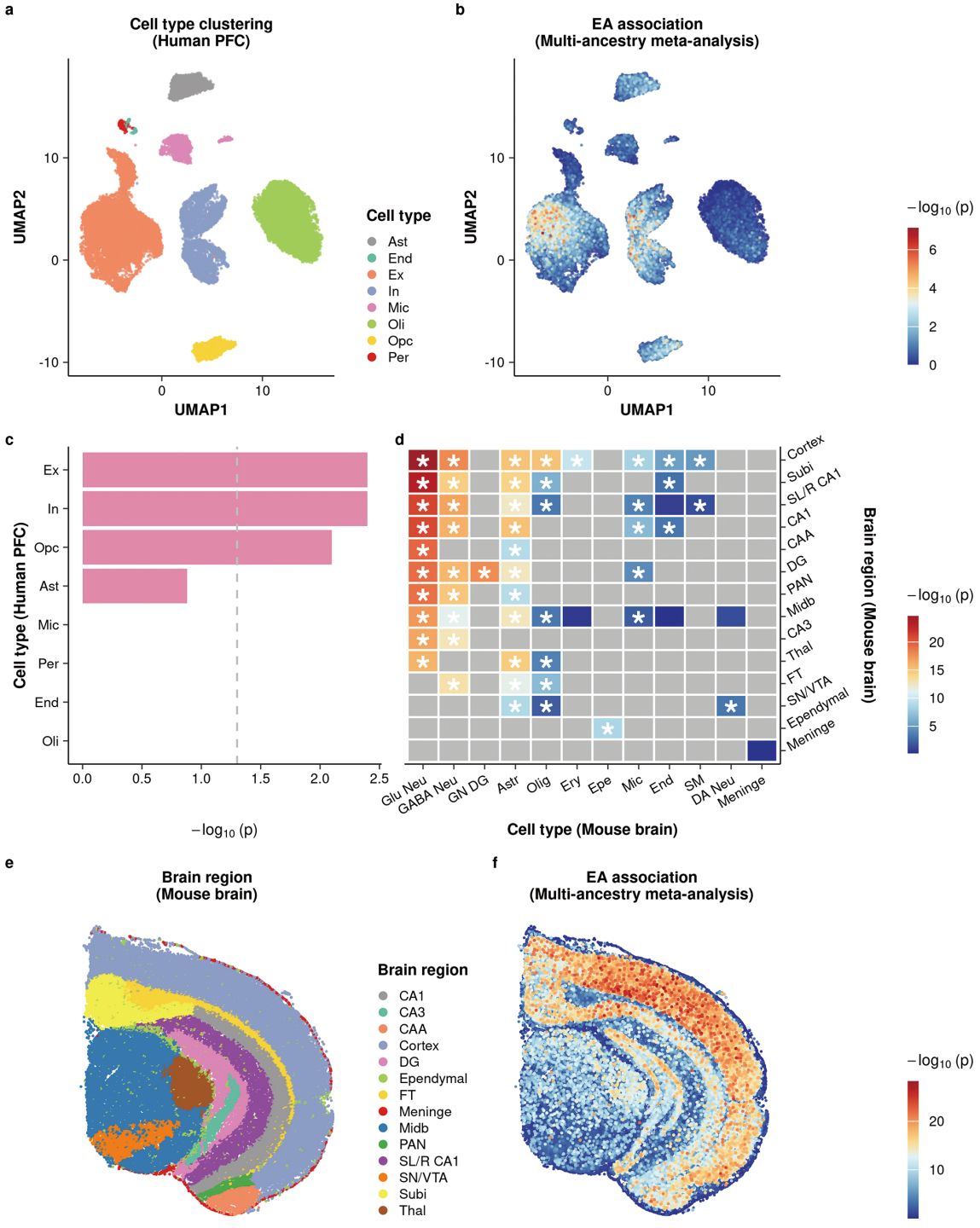


Fig. 5 Brain cell-type association based on cross-ancestry EA GWAS meta-analysis. a UMAP of human PFC single-cell transcriptomics. Ast, astrocyte; End, endothelial cell; Ex, excitatory neuron; In, inhibitory neuron; Mic, microglia; Oli, oligodendrocyte; Opc, oligodendrocyte precursor cell; Per, pericyte. **b** Cellular-level associations with EA in the human PFC. Cluster boundaries remain consistent with cell type clusters from panel **a**. The association significance of each cell is indicated by colors. **c** Cell-type associations in the human PFC. The dashed line represents the FDR threshold of 0.05. **d** Spatially resolved cell-type associations across mouse brain regions identified by gsMap. Rows denote anatomical brain regions, and columns represent established cell types. Colors indicate association significance, with significantly associated subpopulations (FDR < 0.05) marked by asterisks. Gray tiles represent excluded subpopulations with fewer than 100 cells. Subi, subiculum; CA1, cornu ammonis 1 region; SL/R CA1, stratum lacunosum-moleculare/radiatum of the hippocampal cornu ammonis area 1; CAA, cortical amygdalar area; DG, dentate gyrus; PAN, posterior amygdalar nucleus; Midb, midbrain; CA3, cornu ammonis area 3; Thal, thalamus; FT, fiber tract; SN/VTA, substantia nigra/ventral tegmental area; Glu Neu, excitatory glutamatergic neuron; GN DG, granule cell of dentate gyrus; Astr, astrocyte; GABA Neu, GABAergic neuron; Ery, erythrocyte; Olig, oligodendrocyte; Epe, ependymal cell; Mic, microglia; SM, smooth muscle cell; End, endothelial cell; DA Neu, dopaminergic neuron. **e** Mouse brain ST data, with spots colored by anatomical brain region. **f** Spatial mapping of EA-associated cellular signals in the mouse brain. Spatial distributions of cells remain consistent with those in panel **e**. The association significance of each cell is indicated by colors.

genetic associations compared to previous efforts, with 11 signals independent of EUR-reported signals. The heterogeneity in the genetic basis of EA across ancestries led to the identification of novel genes and ancestry-specific genetic correlations between EA and other complex traits. Leveraging the conserved biological mechanism of EA across ancestries, we mapped the specific subtypes and spatial locations of cells across brain regions that were significantly associated with EA. These results provide a more comprehensive landscape of the genetic component of EA.

One of the major findings of this study is the discovery of ancestry-specific genetic associations. By conditioning on the reported signals from the largest published EUR GWAS (EA4), we identified 11 lead variants that are statistically independent of the established EUR associations. We hypothesize that this independence arises from two primary scenarios: 1) Low MAF in EUR: This is supported by the fact that 9 of the 11 EAS-specific signals had a higher MAF in EAS than in EUR, with a mean ratio of expected heterozygosity of 2.27 between EAS and EUR, consistent with findings from recent cross-ancestry studies suggesting that most novel associations in non-EUR ancestries can be attributed to a higher ancestry-specific MAF[43]. With this increased discovery power, our data lay a foundation for detecting genetic associations that were underpowered in other ancestries due to a small MAF and facilitate novel discoveries in subsequent analyses. 2) Differences in tagging variants between ancestries: this is supported by several EAS-specific signals identified at EUR-reported loci, where the causal variant may be weakly tagged by the EUR lead variant but strongly tagged by a different variant in EAS owing to its distinct LD pattern. These loci benefit from cross-ancestry fine-mapping, as exploiting divergent LD patterns improves the resolution in separating tag variants from causal variants. For example, cross-ancestry fine-mapping results at a locus on chromosome 2 yielded five CSs (Supplementary Fig. S4a), suggesting the presence of five distinct putative causal signals in this region. CS 1, 4, and 5 each comprised a single variant with a PIP close to 1 in the center of the locus. These variants exhibited moderate LD ($r^2 = 0.424$) with the top SNP in EAS but weak LD ($r^2 = 0.030$) with the top SNP in EUR, indicating they represent shared causal variants that are tagged differently across ancestries. Compared to within-ancestry fine-mapping results at this locus, cross-ancestry fine-mapping identified more CSs and reduced the CS sizes in EUR. By applying cross-ancestry fine-mapping to the loci significantly associated with EA in both EAS and EUR, we identified more credible sets and more high-PIP variants than EUR-only fine-mapping (Supplementary Table S41), suggesting additional candidate causal variants and reduced uncertainty.

Through the integration of multiple gene prioritization methods, functional annotations, and xQTL data, we identified 28 high-confidence genes at the significant loci of our EAS GWAS. We defined confidence based on convergence across six methods, a stringent criterion given the distinct underlying assumptions and datasets of each approach. The functions of these high-confidence genes converge on three key biological domains: neurotransmission, neurodevelopment, and transcriptional regulation. Genes involved in neurotransmission (e.g., *NRXN1*, *CBLN2*, *MCTP1*) suggest that synaptic communication efficiency is a key biological factor underlying variation in EA, potentially by modulating complex brain functions such as working memory[44]. Similarly, neurodevelopmental genes (e.g., *MED27*, *CUL3*, *AUTS2*) likely influence early-life cognitive performance, thereby contributing to divergent educational outcomes. Our high-confidence gene set provides a strong foundation for future experimental validation of the biological mechanisms underlying EA. Furthermore, we identified novel target genes through EAS-specific associations. A notable example is *CBLN2*; despite its established role in synapse regulation, its association signal for EA was detected in EAS but not in EUR (Fig. 3d). In contrast, *NRXN1*, a conserved postsynaptic cell adhesion molecule in the same signaling pathway (*CBLN2-NRXN1-GLUD*[45]), was prioritized in both ancestries. This contrast reinforces the hypothesis that the absence of a *CBLN2* signal in EUR is primarily due to limited statistical power driven by allele frequency differences, suggesting that while the biological pathway is conserved, the specific variants are ancestry-dependent.

We further investigated the shared and divergent genetic architecture between EA and other complex traits across ancestries. Consistent with findings in EUR, we observed robust genetic correlations between EA and favorable metabolic profiles, biomarkers, and pharmacotherapeutic usage. This suggests that the genetic liability for higher EA is intrinsically linked to better health outcomes, highlighting the potential utility of integrating EA-related genetic factors into health equity frameworks. We also investigated the distinct genetic correlations between EA and diseases across ancestries (Supplementary Table S18). For instance, while we observed a significant positive genetic correlation between EA and the risk of uterine fibroids (UF) in EAS ($r_g = 0.155$, FDR-corrected $P = 0.028$), the corresponding estimate was negative in EUR ($r_g = -0.0732$, FDR-corrected $P = 7.31 \times 10^{-4}$), though we caution against over-interpreting this cross-ancestry difference given its modest magnitude. Although the high cross-ancestry genetic correlation of UF ($r_g = 0.847$, s.e. $= 0.155$) suggests that its core biology is shared across ancestries, its fluctuating SNP-based heritability[46] across ancestries indicates that it is prone to ancestry-specific gene-by-environment interactions, similar to EA. We performed local genetic correlation tests using LAVA[47] to further investigate the potential mechanism. Only one locus with a significant univariate signal was eligible for testing, and no significant local correlation was observed. Nevertheless, we observed a potential positive causal effect of EA on UF in EAS. The putative causal relationship between EA and UF was negative and significant in the GSMR result in EUR, but it did not pass the sensitivity analysis for sample overlap using MRlap[48]. In another case, the negative genetic correlation and putative causal inference between EA and age at menarche (AAM) were observed exclusively in EAS. The potentially overlapping biological mechanism underlying these two traits is supported by a recent study reporting that biological pathways, including head development, synaptic signaling, and behavior, are enriched among high-confidence AAM genes[49]. Overall, given the shared biology across ancestries for these traits (as illustrated by cross-ancestry genetic correlation results), we think the distinct genetic correlations between EA and complex traits such as UF and AAM across ancestries may

reflect unmeasured gene-by-environment effects. For the putative causal inference, EA could be an important exposure, but we suggest being cautious about this result regarding collider bias, and replication analyses using different datasets will be required in the future.

Moreover, we mapped the distinct subtypes and spatial locations of cells significantly associated with EA across brain regions by integrating scRNA-seq and ST data with the EA GWAS meta-analysis summary statistics. The convergence of multiple datasets and methods revealed that cortical neuronal cell types, particularly excitatory neurons, exhibited a stronger association with EA than other cell populations. The increased resolution provided by single-cell data enabled the identification of not only refined neuronal cell types, such as intratelencephalic excitatory neurons and distinct medium spiny neurons (MSNs), which express dopamine receptors and are implicated in reward-based behaviors (Supplementary Fig. S8 and Table S27), but also subpopulations of nonneuronal cells, including astrocytes and OPCs (Fig. 5b, d). By linking the EA-associated genetic variants to the specific expression patterns of these cell populations, we suggest that the genetic component of EA may involve the modulation of synaptic signaling and plasticity[50]. This process is known to be centered on excitatory neurons[51] and is precisely modulated by GABA released from inhibitory neurons[52], by the clearance of neurotransmitters by astrocytes[53], and by the direct interactions between OPCs and neurons[38]. The incorporation of multiple brain regions and spatial information indicated that the cellular mechanisms underlying EA depend jointly on cell type, brain region, and the spatial location of cells within regions. These results provide potential cellular mechanisms that influence EA, enhancing our understanding of its neurobiological basis.

The SNP-based heritability estimate of EA in our GWAS was lower than previously observed. Given the observed heterogeneity in EA genetic architecture between EAS and EUR, we expect differences in EA heritability across ancestries. Here, we mainly discuss potential factors underlying the difference in SNP-based heritability between our EAS GWAS and the previous EAS GWAS by Chen *et al*. We hypothesize that cross-cohort heterogeneity, cohort-specific characteristics, and unmeasured gene-by-environment effects could contribute to the difference in SNP-based heritability. Compared with the previous EAS GWAS (a meta-analysis of two prospective cohorts: TWB and KoGES), our new datasets included samples from prospective studies, clinical centers, and commercial genetic testing platforms. Multiple factors vary across these datasets and can contribute to cross-cohort heterogeneity and decreased heritability estimates in the meta-analysis, including genotyping processes (Supplementary Table S1), phenotype distributions, and cohort characteristics (Supplementary Table S35). For example, the BGI-lcWGS cohort leveraged an ultra-low-coverage whole-genome sequencing technique for genotyping and imputed genotypes to Chinese samples from the 1000 Genomes Project Phase 3 (1KGP3), while the JAPAN_2 to JAPAN_5 cohorts leveraged the Affymetrix Axiom Japonica arrays for genotyping and imputed genotypes to a reference panel built from 3,552 within-cohort samples. Phenotypic heterogeneity was also observed: mean years of schooling (*EduYears*) ranged from 8.8 (CKB) to 16.9 (WGEA), with distinct phenotype distributions across ISCED levels. Average birth year differed by up to 47 years across cohorts (1938 in JAPAN_1 vs 1985 in WGEA), and average *EduYears* and birth year were strongly correlated across the new cohorts (Pearson's $r = 0.830$, $P = 4.41 \times 10^{-4}$), indicating that era-dependent environmental factors, such as the expansion of educational opportunities across successive generations, contribute to the phenotypic differences among cohorts. Even among cohorts with similar birth years (1952–1956), average years of schooling differed between CKB (8.8 years) and the JAPAN_2 to JAPAN_5 cohorts (12.1–12.9 years), possibly reflecting differences in education systems between China and Japan. These factors contribute to cross-cohort heterogeneity and imperfect genetic correlations, thereby lowering the SNP-based heritability of the GWAS meta-analysis, consistent with a recent study showing that heritability estimates decrease as the number of cohorts in a meta-analysis increases when genetic correlations between cohorts are less than one or when cohort-specific heritability is heterogeneous[54]. Furthermore, cohort-specific characteristics and unmeasured gene-by-environment effects may also contribute to lower cohort-level heritability estimates. For the new datasets from the genetic testing platforms, we found their phenotypic standard deviations to be smaller than those of samples from TWB and KoGES (Supplementary Fig. S12a), which may indicate that these samples are more clustered. For the BGI-lcWGS cohort, samples were from pregnant women recruited at a hospital in Wuhan, which could lead to greater homogeneity among samples and an attenuated estimate of heritability. The WeGut cohort had characteristics similar to those of the KoGES in the previous EAS study. Its participants, who were born around 1957, were recruited from prospective studies and had *EduYears* distributions similar to those of KoGES participants. Consistent with these similarities, we observed a higher SNP-based heritability estimate in the WeGut data, though the relatively large standard error warrants conservative interpretation. For large cohorts such as CKB and JPN-meta, we found heterogeneous heritability estimates across age- and sex-stratified groups. For example, the SNP-based heritability estimate from rural samples in CKB was 0.0333 (s.e. $= 0.0113$), whereas the estimate from urban samples in CKB was 0.0826 (s.e. $= 0.0130$). The SNP-based heritability estimate from samples with an enrollment age (defined as birth year plus 6) before 1945 in JPN-meta was 0.0447 (s.e. $= 0.0207$), whereas the estimate from samples enrolled between 1946 and 1965 in JPN-meta was 0.0815 (s.e. $= 0.0134$). For samples enrolled in the same period (1946–1965), the SNP-based heritability estimate in CKB ($h^2_{SNP} = 0.0311$, s.e. $= 0.0128$) differed from that in JPN-meta (Supplementary Table S47). These observations may indicate interactions between environmental and genetic factors that were not captured in the current model. Although the SNP-based heritability estimated from our EAS GWAS was lower than in previous studies, the high genetic correlation between our newly collected cohorts and those of Chen *et al*., together with the increased power of the EAS meta-analysis, strongly supports the robustness of our data.

Several other limitations exist in our study. First, although we attributed the novel associations discovered only in EAS primarily to differences in MAF and tagging variants, we also

observed heterogeneity in variant effect sizes across ancestries. For example, the lead variant at the *DCAKD* locus on chromosome 17 showed a similar MAF but heterogeneous effect sizes across ancestries and was identified as an EAS-specific association in our conditional analysis (Supplementary Fig. S14). Furthermore, all the genome-wide significant variants discovered only in EAS showed consistent direction and larger absolute effect sizes than in EUR (Supplementary Table S40). These findings reflect heterogeneity in the genetic architecture of EA across ancestries, leading to differences in heritability estimates, imperfect genetic correlations, and a partial loss of PRS relative accuracy[7,8]. However, this factor is closely intertwined with MAF differences and correlated with unmeasured gene-by-environment interactions, which could be further investigated in future studies. Second, although our GWAS achieved a higher nominal predictive R^2 than the previous EAS study in predicting the phenotypes of EAS samples in the UKB (Supplementary Note 4), the overall limited SNP-based heritability and the large standard error of the prediction accuracy estimated from the small test cohort preclude the use of our EA PRS for polygenic prediction of other complex traits. Third, we primarily incorporated blood xQTL datasets in the integrative analyses for gene prioritization, given the scarcity of brain xQTL datasets in EAS. This limitation may have led to the omission of functional genes regulated by brain-specific genetic effects that colocalize with EA GWAS signals.

In summary, this study provides a deeper understanding of the genetic and biological components of EA, specifically within EAS ancestry. Our findings demonstrate that while the neurobiological architecture of EA is largely shared across ancestries, novel loci, gene targets, and genetic correlations with other complex traits emerged only through the inclusion of EAS data, underscoring the critical importance of ancestry diversity for a comprehensive understanding of the genetic landscape.

MATERIALS AND METHODS

Ethical approval

This study was approved by the Ethics Committee of Westlake University (approval no. 20200722YJ001).

Phenotype definition

We adopted a phenotype mapping approach similar to that of previous studies. Individual educational information collected from different types of questionnaires was first classified using the International Standard Classification of Education (ISCED) 2011 levels. These ISCED levels were subsequently converted to years of schooling (*EduYears*) using official country-specific mapping files from the ISCED website, with each cohort allowed the flexibility to adjust the exact level-to-year mapping within the range provided by those files. The resulting years of schooling were standardized ($mean = 0, sd = 1$) and used as phenotypes. Detailed mapping standards are provided in Supplementary Note 1 and Tables S49, S50, and cohort-specific ISCED mappings are provided in Supplementary Table S35.

Cohort descriptives

Our study incorporated seven Chinese cohorts representing diverse geographic regions: (1) China Kadoorie Biobank (CKB): A prospective cohort of 80,466 adults aged 30–79 from 10 geographic regions across China was genotyped using a customized Affymetrix Axiom array[55]. Variants were phased with SHAPEIT2 and imputed to the 1KGP3 EAS reference panel. CKB Data from baseline are available under the CKB Open Access Data Policy to *bona fide* researchers; full details of the CKB Data Sharing Policy are available at ckbiobank.org. The CKB study was approved by the Ethical Review Committee of the Chinese Center for Disease Control and Prevention (Beijing, China), the Institutional Review Board of Peking University (Beijing, China), and the Oxford Tropical Research Ethics Committee at the University of Oxford (UK). All participants provided written informed consent prior to data collection. (2) BGI low coverage WGS (BGI-lcWGS): A cohort of 69,788 pregnant women from Wuhan Children's Hospital was genotyped using the BGISEQ-500 WGS with ultra-low sequencing depth during their non-invasive prenatal testing[56,57]. Variants were imputed to 1KGP3 using STITCH. The BGI-lcWGS study received approval from the Bioethics and Biosafety of BGI (BGI-IRB 21088-T2). Each participant provided informed consent before enrollment. (3) WeGene cohorts (WGEG, WGEA, WGEC): Three cohorts comprising 15,126, 12,389, and 4,078 individuals recruited through the genetic testing service[58] from WeGene, Shenzhen Zaozhidao Technology Co., Ltd. were genotyped using Illumina Global Screening, customized Affymetrix WeGene V1, and Illumina Chinese Genotyping arrays, respectively. Variants were imputed via the China Metabolic Analytics Project (ChinaMAP) online server[59]. All WeGene cohorts were approved by the Ethical Committee of Shenzhen WeGene Clinical Laboratory. Informed consent for online research was obtained from all participants. (4) 23mofang: 9,155 samples recruited by the genetic testing service from 23mofang, Inc. were genotyped using customized Axiom Precision Medicine or Illumina Infinium arrays. Variants were phased with SHAPEIT4 and imputed to 1KGP3 using Minimac4. The 23mofang cohort was reviewed and approved by Ethical and Independent Review Services, a private institutional review board. All participants granted their informed consent and responded to online surveys according to the approved 23andMe human subjects protocol. (5) Westlake Gut (WeGut) consortium cohorts: including 4,254 participants from 4 cohorts, including Guangzhou Nutrition and Health study, Zhoushan Metabolic Syndrome Cohort, Westlake Precision Birth Cohort, and Tongji-Shuangliu Birth Cohort[60], which were genotyped using the Illumina Global Screening array. Variants were imputed on the ChinaMAP online server. The WeGut Project protocol was approved by the Ethics Committee of Westlake University (20220429ZJS001), and all participants provided written informed consent. We also incorporated the EA GWAS meta-analysis of six Japanese cohorts: (1) Cohort 1: 6,956 individuals recruited by the National Center for Geriatrics and Gerontology were genotyped using the Illumina Asian Screening array. Variants were phased with Eagle v2.4.1 and imputed to the JEWEL 3k and 1KGP3 reference panels with Minimac4. This study was approved by the Human Research Ethics Committee of the National Center for Geriatrics and Gerontology (approval No. 990-16). (2) Cohort 2–5: Four cohorts with 71,045 individuals recruited through the

Tohoku Medical Megabank Organization (ToMMo) community-based cohort study were genotyped using a customized Affymetrix Axiom array. Variants were phased with SHAPEIT2 and imputed to the cohort-specific reference panel using IMPUTE2, comprising 3,552 samples. These studies were approved by the Ethics Committee of RIKEN (RIKEN-Y-2022-002) and the ToMMo (2021-0032-1). (3) Cohort 6: 10,069 individuals recruited through the Nagahama Prospective Genome Cohort for Comprehensive Human Bioscience (the Nagahama Study) by Kyoto University were genotyped by Illumina Asian Screening or Japanese Screening arrays. Variants were phased with SHAPEIT4 and imputed to the TOPMed reference panel Version r2 on the online server[61]. This project is a collaboration between Kyoto University and Nagahama City[62], and this study was approved by the Ethics Committee of Kyoto University (Identifier: G1289). Written informed consent was obtained from all participants in each Japanese cohort. Further details on cohort-level genotyping, phasing, imputation, and association analyses can be found in Supplementary Table S1.

We incorporated the EA GWAS data from our study with the publicly available EAS EA GWAS summary statistics from Chen *et al.* for the within-ancestry meta-analysis, which included 165,232 individuals of EAS ancestry from the TWB and KoGES[9]. To ensure a compatible scale for phenotype and variant effect sizes across all cohorts, we performed *z*-score standardization of the summary statistics from Chen *et al.* Furthermore, a publicly available EUR GWAS summary statistics dataset comprising 766,345 individuals of EUR ancestry from the Lee *et al.* study (EA3)[6] was incorporated into the cross-ancestry meta-analysis. Since the available summary statistics from Lee *et al.* had already been standardized, the variant effect sizes were on the same scale as our data, and we performed the meta-analysis without further adjustment.

Meta-analysis

We conducted cohort-level quality control using the EasyQC R package[63], excluding variants with low imputation accuracy or invalid estimates. For each GWAS, we estimated SNP-based heritability and potential inflation using LDSC[10] (Supplementary Table S2). The BGI-lcWGS cohort ($n = 69{,}788$) and the Japanese meta-analysis (JPN-meta; $n = 88{,}070$) showed potential inflation from population stratification (LDSC ratio $> 10\%$). To address this, we adjusted the standard errors of variant effects for these cohorts by the square root of the LDSC intercept ($\sqrt{1.0423}$ for BGI-lcWGS and $\sqrt{1.1016}$ for JPN-meta).

We performed an inverse-variance-weighted (IVW) fixed-effects meta-analysis of nine quality-controlled GWAS results using OmicS-data-based Complex trait Analysis (OSCA[11]). We used OSCA because it performs allele checking before the meta-analysis, provides per-SNP sample sizes, and produces output in a format directly usable for our downstream analyses. We performed a sensitivity analysis using METAL[64] (the software used by previous studies) and validated the consistency of results between the two approaches using the Pearson's correlations of estimated effect sizes and association *P*-values for the 723 genome-wide significant variants ($r = 1, P < 2.2 \times 10^{-16}$; Supplementary Fig. S13). We also estimated the SNP-level heterogeneity using METAL. We found that 4.11% of variants had a *P*-value less than 0.05 in Cochran's Q test, and 6.46% of variants exhibited an I^2 larger than 50%, with a mean I^2 of around 10.5%, suggesting that variant-level heterogeneity across cohorts is low and that the fixed-effect model is compatible with our EAS GWAS meta-analysis of EA. The meta-analysis of EAS GWAS data showed no evident inflation. We calculated the approximate number of weeks of schooling per allele by multiplying the standardized effect sizes by the combined standard deviation of *EduYears* (4.19 years) across all EAS cohorts. We used consistent thresholds, as in previous EA studies, to identify independent genetic associations. We used an LD r^2 cutoff of 0.1 and a distance cutoff of 10 Mb to perform the LD-based clumping. The 10-Mb distance was set based on the same assumption used in GCTA-COJO, which assumes that variants more than 10 Mb away from each other are in complete linkage equilibrium. We also leveraged the GCTA-COJO model selection to identify conditionally independent signals and potential secondary signals. Using the ± 1 Mb distance from the clumping lead variants as fine-mapping windows, we performed statistical fine-mapping analyses using CARMA[13] on our EAS meta-analysis data. The genotypes from EAS samples in the 1KGP3[65] were used as an out-of-sample LD reference for all the analyses above. We compared the genetic associations identified in our EAS GWAS to those from the previous EAS GWAS by Chen *et al.* using two approaches. First, we reported the clumping-based lead variants located outside the previously identified genomic loci. For this analysis, we mapped the physical positions of our 33 clumping-based lead variants to the seven genomic regions defined in Table 1 of Chen *et al.*'s paper. Second, we used --cojo-slct in GCTA to select, from the variants reported in Supplementary Table S3 of Chen *et al.*, those that were jointly associated with EA in our EAS GWAS (using the 1KGP3 EAS samples as the LD reference), and then performed conditional analysis (--cojo-cond) with these variants as covariates. We reported the variants that remained genome-wide significant after conditioning, thereby identifying EA-associated variants statistically independent of the previously reported genome-wide significant associations in EAS. As a sensitivity analysis, we additionally performed Multi-Ancestry Meta-Analysis (MAMA[66]) using our EAS meta-analysis and the EUR meta-analysis from Lee *et al.* Compared with Chen *et al.*, we also identified more independent loci in the EAS-specific meta-analysis using our data (Supplementary Table S42).

We used Popcorn[67] to analyze the genetic correlation of EA between our EAS GWAS and the GWAS summary statistics from Lee *et al.* (EA3). An IVW cross-ancestry meta-analysis was then performed using OSCA to combine these two datasets. We used the genotypes of EUR samples from 1KGP3 as the LD reference for clumping the cross-ancestry meta-analysis. To ensure this did not introduce bias due to LD mismatch between the reference panel and the GWAS, we evaluated the LDSC intercept and attenuation ratio for the meta-analysis summary statistics using the EUR reference. The LDSC intercept was close to 1, and the attenuation ratio was less than 10% (LDSC intercept = 1.0721, s.e. = 0.0147; attenuation ratio = 0.0397, s.e. = 0.0081), indicating that using the EUR reference did not cause significant bias due to LD misspecification, consistent with previous studies[68,69].

Moreover, we performed sensitivity analyses leveraging a mixed EAS and EUR LD reference with a sample ratio consistent with our cross-ancestry meta-analysis. The resulting LDSC intercept and attenuation ratio values were biased upward (LDSC intercept = 1.1317, s.e. = 0.0164; attenuation ratio = 0.0723, s.e. = 0.0090), potentially due to both LD structure and differences in variant inclusion (Supplementary Note 3 and Table S48). We further evaluated LD-based clumping using the mixed LD reference. While all lead variants identified by the mixed reference were successfully captured by the EUR LD reference, a substantial number of additional lead variants were exclusively identified by the EUR LD reference. This discrepancy occurred because numerous variants present in our meta-analysis were absent from the mixed EAS and EUR LD reference. Given these practical advantages, we utilized the 1KGP3 EUR samples as the primary LD reference for cross-ancestry analyses. Potential novel signals were defined as the clumping lead variants of the meta-analysis that were located more than 500 kb away from, and exhibiting LD r^2 below 0.1 in both EAS and EUR with, any previously reported clumping lead variant from Okbay *et al.* (EA4)[7]. LD r^2 values were calculated using the EAS and EUR samples from 1KGP3.

Polygenic prediction

The EA PRS was evaluated in a holdout cohort of 1,615 individuals with self-reported EAS ancestry from the UKB[15] who passed quality control (excluding F-coefficient outliers, related samples, and samples without phenotypes). We used our EAS GWAS and the EUR GWAS from Lee *et al.* (EA3) as discovery GWAS, and the 1KGP3 EAS and EUR samples as LD references (formatted on the PRS-CS[70] website) to estimate the ancestry-specific posterior effect sizes of variants by PRS-CS and PRS-CSx[71]. We used the default values for the optional parameters in both software packages. PLINK2[72] was used to construct the PRS for each individual. We regressed the test cohort phenotypes on birth year, sex, the interaction between birth year and sex, and the first 10 PCs as the control model, and evaluated prediction accuracy through the difference in R^2 between the regressions with and without the EA PRS (incremental R^2). To determine the optimal ancestry-specific weights for building the cross-ancestry EA PRS via PRS-CSx, we employed a tuning approach based on 1,000 resampling iterations. For each iteration, we randomly partitioned the holdout cohort into a validation set ($n = 807$) for weight tuning and a test set ($n = 808$) for performance evaluation. We estimated the ancestry-specific posterior effect sizes and constructed the ancestry-specific EA PRS for each set. Then, we regressed the validation samples' phenotypes on their ancestry-specific PRS to obtain the regression coefficients for each ancestry. Finally, we used the linear combination of the ancestry-specific PRS from the test set as the cross-ancestry PRS, with the regression coefficients from the validation set as weights. The predictive accuracy was evaluated by the mean incremental R^2 across all 1,000 iterations to ensure robustness against sampling variance. Considering that the sample size of the hold-out UKB EAS cohort was small, we performed the following sensitivity analyses. First, we employed the same cross-ancestry EA PRS construction approach as in the previous EAS study, performing an inverse-variance-weighted meta-analysis of the posterior effect sizes across ancestries (using --meta in PRS-CSx). This approach bypasses the need for an independent tuning set and evaluates the prediction accuracy in the full UKB EAS cohort ($n = 1{,}615$). Second, we employed two EAS cohorts, 23mofang ($n = 9{,}155$) and WeGut ($n = 4{,}254$), in our study for leave-one-cohort-out PRS analyses. Third, we switched the LD reference for PRS construction to EAS and EUR samples from UKB (provided by the PRS-CSx authors), which have much larger sample sizes ($n_{EAS} = 2{,}181$; $n_{EUR} = 375{,}120$) than the 1KGP3 reference, but which may cause over-fitting since our test cohort is from the same population (Supplementary Note 4 and Table S37).

Cross-ancestry conditional analysis pipeline

To identify associations independent of the largest EUR EA GWAS from Okbay *et al.* (EA4), we developed a rigorous conditional analysis pipeline, addressing the missingness of lead variants and collinearity when modeled in EAS LD.

(1) Variant set expansion. Due to the unavailability of the full summary statistics of EA4, we manually expanded their 3,952 lead variants by including LD proxies ($r^2 > 0.1$) of these variants using a LD reference panel comprising 50,000 randomly sampled individuals of EUR ancestry from the UKB. For lead variants missing from the reference panel, we took a conservative approach and included all variants located less than ± 1 Mb of the lead variant. This step generated a dense variant list encompassing all genetic associations identified in EUR.

(2) Covariate variant selection. To avoid collinearity in conditional analysis, we selected independent variants within EAS LD patterns that were jointly significant in both EAS and EUR GWAS using the GCTA-COJO model selection with LD reference from 1KGP3 EAS samples.

(3) Conditional analysis. We finally performed conditional analysis on the EAS GWAS by adjusting for the covariate variants in step (2), ensuring that the remaining signals were independent of the EUR-reported associations.

Cross-ancestry fine-mapping

We leveraged SuSiEx to compare the cross-ancestry and within-ancestry fine-mapping results[73] at loci with genome-wide significant associations in both the EAS and EUR GWAS of EA. We incorporated our EAS GWAS, the EUR GWAS from Lee *et al.*, and the 1KGP3 LD references of EAS and EUR samples into this analysis. Loci used for analysis were first defined as 1-Mb genomic regions around the 33 LD-based clumping lead variants from our EAS GWAS. Then, only the 15 loci with genome-wide significant associations in the EUR GWAS from Lee *et al.* were included. We set the credible set coverage level to 95%, the maximum number of causal signals to 10, and the maximum number of iterations to 1,000 when applying SuSiEx. Other options were kept at their default values. We compared the number of credible sets, the median credible set size, and the maximum SNP PIP within each credible set across cross-ancestry, EUR-only, and EAS-only fine-mapping at the 15 common loci.

Gene prioritization

We incorporated six gene prioritization methods, four epige-

netic annotations, and four layers of EAS xQTL datasets to identify EA-associated genes. For all methods, we included only protein-coding genes with significant results after multiple testing correction at GWAS-significant loci. We used MAGMA[17] and mBAT-combo[18] for gene-based analyses, and genes with Bonferroni-corrected $P < 0.05$ were prioritized. For V2G mapping, we mapped genome-wide significant variants with CARMA PIP > 0.1 to their linked genes using the following annotations: the transcription start site (TSS) and exons of each protein-coding gene (using the annotation from GENCODE[74] v40), promoter capture Hi-C (PCHiC) annotations of two published studies[75,76], activity-by-contact (ABC) scores of two published studies[77,78], and epigenomics annotations from RoadMap[79] and EpiMap[80] consortia. For coding variants, we directly mapped them to their targeted genes. For TSS annotation, we mapped each variant to its closest TSS gene. For other annotations, we selected the gene exhibiting the strongest interaction with the enhancer in which each variant resided. We used two integrative analysis methods (SMR[19] and COLOC[20]) and four layers of omics data (mQTL, sQTL, eQTL, and pQTL) to integrate GWAS and xQTL data. We incorporated 54 xQTL datasets in this analysis, including eQTL datasets of 28 immune cell types and their meta-analysis[81], eQTL datasets of whole blood and brain tissue[82,83], sQTL datasets of 19 immune cell types and their meta-analysis[84], a serum pQTL dataset[85], a plasma pQTL dataset[82], and the meta-analysis of four blood mQTL datasets[86,87]. Most of the xQTL datasets were collected from pure EAS samples. Although the Asian Immune Diversity Atlas (AIDA) sQTL dataset includes East Asians, South Asians (SAS), and Southeast Asians (SEA), 74% of the samples are of EAS ancestry, and the association signals were mainly driven by EAS samples. The potential impact of using the AIDA sQTL dataset is a moderate loss of power due to AF differences. No splice sites or genes were identified through this dataset in combination with our EAS GWAS. Genes identified by any xQTL dataset with a Bonferroni corrected SMR P or SMR-multi $P < 0.05$, as well as a HEIDI test $P > 0.01$, were prioritized by SMR. Genes with a COLOC PP4 > 0.8 were prioritized by COLOC. We also used PoPS[21], a network-based method, for gene prioritization. The gene with the highest PoP score at each genome-wide significant locus was prioritized. Finally, we counted the number of methods by which each gene was prioritized, ranked them in descending order, and selected the genes prioritized by at least three independent methods as high-confidence genes associated with EA.

Genetic correlation and MR

We assessed genetic correlations between EA and 259 publicly available GWAS summary statistics of other complex traits from EAS using LDSC. We included traits from consortium GWAS papers using EAS samples, as well as traits collected from East Asian biobanks, including Biobank Japan (BBJ), the KoGES, the TWB, and the Taiwan Precision Medicine Initiative (TPMI). For traits shared among these biobanks, we first conducted a meta-analysis to increase statistical power and then performed the genetic correlation analysis. For traits shared between TWB and TPMI, we used only the TPMI GWAS due to its larger sample size. All data resources are listed in Supplementary Table S18. For traits showing significant genetic correlation with EA after FDR correction, we performed comparison analyses using the corresponding EUR data and EA3. The data sources were listed in Supplementary Table S18. We conducted bi-directional generalized summary-data-based Mendelian randomization (GSMR[33]) to evaluate causal inference between EA and the complex traits significantly correlated with EA, with FDR correction to control for false-positive discoveries. Furthermore, we leveraged two approaches to conduct sensitivity analyses, accounting for potential sample overlap between our EAS GWAS and the complex-trait GWAS from TWB, KoGES, or TPMI, and between the EUR GWAS from Lee *et al.* and the complex-trait GWAS datasets that included UKB samples. We first performed GSMR using complex-trait GWAS from independent datasets without overlapping samples with the EA GWAS. We used BBJ datasets for the analyses in EAS, and EUR datasets from the Million Veteran Program (MVP) or FinnGen for the analyses in EUR. Second, we used the MRlap R package, which performs IVW-MR and adjusted the results using the bivariate LD score regression intercept. All results were corrected for multiple testing using FDR. We also applied 10 commonly used MR methods to ensure the robustness of our findings using a previously developed pipeline[88], including IVW, Robust, weighted median, simple median, mode, MR-Egger, and contamination mixture method implemented in the MendelianRandomization package (v0.4.2, CRAN.R-project.org/package=MendelianRandomization), MR-PRESSO (v1.0, github.com/rondolab/MR-PRESSO), MRMix (v0.1.0, github.com/gqi/MRMix), and RAPS (v0.2, github.com/qingyuanzhao/mr.raps) package. Since the GSMR software specifically accounted for LD by leveraging a generalized least squares (GLS) regression model if the instrument variants are not fully independent, we retained its default LD threshold of 0.05. For MRlap and the other MR methods we used for triangulation, we used a stringent LD threshold of $r^2 < 0.001$ across instrument variants. We used the default values for all the other optional parameters.

Heritability enrichment across tissues

We used S-LDSC for EA heritability enrichment analyses in functional genomic regions incorporating the baseline annotation (v2.2). We also used the baseline annotation to annotate the function of the 88 genome-wide significant variants after adjusting for EUR signals. We used LDSC-SEG[35] and annotations of tissue-specific open chromatin regions to assess enrichment of heritability for EA across tissues. Significance in multiple testing was determined using FDR correction.

Mapping EA-associated cells through brain single-cell and spatial transcriptomics data

To identify cell populations associated with EA, we integrated human brain single-nucleus transcriptomics data from ROSMAP[36] ($n = 24$) with the multi-ancestry meta-analysis results using the single-cell disease relevance score (scDRS[37]). The cell-type-level association was calculated by comparing the top 5% quantile trait score of each cell type to the top 5% quantile control scores of all cell populations. We identified cell types with a significant Monte Carlo P-value after FDR correction as EA-associated. We also performed S-LDSC to

ensure the robustness of scDRS findings using the established pipeline from previous studies[89,90]. We computed gene expression specificity for each gene in each cell type as the expression level of a gene within a specific cell type divided by its total expression level across all cell types. Then, we ranked the proportional expression of all genes in each cell type and used the gene expression specificity of the top 10% genes as the annotation for that cell type. Finally, we used the baseline LDSC annotation as background and calculated the heritability enrichment for each cell type. We used the EUR samples in 1KGP3 as the reference panel.

For spatial mapping of EA-associated cells, we integrated mouse brain Stereo-seq data[39] and the cross-ancestry meta-analysis via gsMap[40], using the EUR samples in 1KGP3 as the reference panel and transforming mouse genes into human homologs. Cauchy's combination of the single-cell association *P*-values was employed to evaluate the cell-type association significance with EA across brain regions. We also used single-nucleus data from ten human brain regions ($n = 3$) [41] to perform cell-type association analyses across brain regions using scDRS for validation. FDR correction was applied to avoid false positive results.

DATA AVAILABILITY

Summary statistics of the EAS EA GWAS meta-analysis and the cross-ancestry (EAS-EUR) EA GWAS meta-analysis generated in this study are available at: yanglab.westlake.edu.cn/pubData/. Publicly available datasets used in this study include: Genotypes of 1KGP3 EAS samples (ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/); Genotypes of UKB EAS and EUR samples (ukbiobank.ac.uk/); eQTL data for 28 immune cell types (ddbj.nig.ac.jp/public/ddbj_database/gea/experiment/E-GEAD-000/E-GEAD-420/) ; sQTL data for 19 immune cell types (zenodo.org/records/8343365); Blood eQTL and pQTL from Japan COVID-19 Task Force (JCTF) cohort (humandbs-production.ddbj.nig.ac.jp/en/research/hum0343/v3); Brain tissue eQTL data (github.com/liusihan/population-compare-pipeline); Serum pQTL data (omics.lab.westlake.edu.cn/data.html); Blood mQTL data (biosino.org/node/project/detail/OEP002902, ngdc.cncb.ac.cn/omix/release/OMIX004116, https://yanglab.westlake.edu.cn/data/SMR/EAS.tar.gz); scRNA-seq data for the human PFC (synapse.org/Synapse:syn18485175); Mouse brain ST data (db.cngb.org/data_resources/project/CNP0001543); scRNA-seq data for 10 human brain regions (github.com/linnarsson-lab/adult-human-brain).

CODE AVAILABILITY

The following software and packages were used for data analyses: EasyQC (v23.8, uni-regensburg.de/medizin/institut-fuer-epidemiologie-und-praeventivmedizin/genetische-epidemiologie/software); PLINK 1.9 (cog-genomics.org/plink/1.9/); PLINK 2.0 (cog-genomics.org/plink/2.0/); OSCA (v0.46.1, yanglab.westlake.edu.cn/software/osca/#Download); METAL (github.com/statgen/METAL); LDSC (v1.0.1, github.com/bulik/ldsc); CARMA (github.com/ZikunY/CARMA/tree/master); GCTA (v1.94.1, yanglab.westlake.edu.cn/software/gcta/#Download); bedtools (v2.30.0, github.com/arq5x/bedtools2); mama (github.com/JonJala/mama); PRS-CS (github.com/getian107/PRScs); PRS-CSx (github.com/getian107/PRScsx); MAGMA (v1.10, cncr.nl/research/magma/); SMR (v1.03, yanglab.westlake.edu.cn/software/smr/#Download); COLOC (github.com/chr1swallace/coloc); PoPS (v0.2, github.com/FinucaneLab/pops/tree/master); MRlap (v0.0.3.3, github.com/n-mounier/MRlap); LAVA (v0.1.5, github.com/josefin-werme/LAVA); scDRS (v1.0.2, github.com/martinjzhang/scDRS); gsMap (github.com/JianYang-Lab/gsMap); and SuSiEx (github.com/getian107/SuSiEx). The custom code for the core analyses presented in this manuscript is publicly available at: github.com/RL-m/EAS_EA_GWAS_Ma_et_al.

ACKNOWLEDGEMENTS

We thank Feifei Cheng, Shuli Liu, Liyang Song, Weiyang Bai, Yazhou Guo, Wenjie Wei, Lounan Li, Junren Hou, and Kai Wang for their valuable comments and technical support. This work was supported by the National Key R&D Program of China (2024YFC3405800, 2024YFC3405802), the National Natural Science Foundation of China (U23A20165), the "Pioneer" and "Leading Goose" R&D Programs of Zhejiang (2024SSYS0032), and the New Cornerstone Science Foundation. The Japanese data acquisition was supported by a JSPS KAKENHI Grant (20H00462); the Japan Agency for Medical Research and Development (AMED) grants 21ek0109555, 21tm0424220, 21ck0106642, 23ek0410114, and 23tm0424225; the Medical Research Support Project of the Shizuoka Prefectural Hospital Organization and Takeda Hosho Grants for Research in Medicine. The CKB data of baseline education and genotyping were funded in whole or in part by the National Key R&D Program of China (2023YFC2509400, 2023ZD0510101, 2023ZD0510100, 2016YFC0900500, and 2016YFC1303904), and the National Natural Science Foundation of China (82103920, 82388102, 82192901, 82192904, 82192900, 81941018, 91843302, and 91846303). The CKB baseline survey and the first re-survey were supported by the Kadoorie Charitable Foundation in Hong Kong. DNA extraction and genotyping were funded by GlaxoSmithKline and the UK Medical Research Council (MC-PC-13049, MC-PC-14135). The Wellcome Trust (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z), and the UK Medical Research Council (MC_UU_00017/1, MC_UU_12026/2, and MC_U137686851), Cancer Research UK (C16077/A29186 and C500/A16896), and the British Heart Foundation (CH/1996001/9454) provide core funding to the Clinical Trial Service Unit and Epidemiological Studies Unit at Oxford University for the CKB project. We are grateful to all participants and researchers of the JPN-meta, CKB, BGI-lcWGS, WGEG, WGEA, WGEC, 23mofang, and WeGut cohorts for making this work possible. This research has been conducted using the UK Biobank Resource under Application Number 66982. We also acknowledge the research participants globally who shared their summary statistics and functional annotations. We thank the Westlake University High-Performance Computing Center for computational support.

AUTHOR CONTRIBUTIONS

J.Y. conceived and directed the study. R.M. and J.Y. designed the experiment. K.H., T.K., X.L., T.M., Y. Zhao, Q.L., Linxuan L., H. Zhang, H.X., L.W., S.T., T.L., H.M., F.X., K.O., Y.T., R.G.W., Z.C., J.L., Liming L., A.Z., H.Zhu, X.P., A.P., Y.C., Y. Zhu, D.Z.,

J.Z., K.Z., G.C., X.J., D.S., and C.T. contributed to data collection and the generation of GWAS summary statistics for the individual EAS cohorts (see Supplementary Note 6 for authorship contributions for each cohort). R.M. performed the meta-analysis and all subsequent analyses presented in the paper under the guidance of T.Q. and J.Y. P.M.V., C.T., and D.S. provided critical comments on data analysis and the interpretation of results. R.M. and J.Y. drafted the manuscript. All authors reviewed and approved the final manuscript.

COMPETING INTERESTS

L.W., S.T., T.L., and G.C. are employees of WeGene, Shenzhen Zaozhidao Technology Co., Ltd. H.M. and K.Z. are employees of Chengdu 23mofang Biotechnology Co., Ltd. All other authors declare no competing interest.

REFERENCES

1. Rietveld, C.A. et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).
2. Haider, Z.F. & von Stumm, S. Predicting educational and social-emotional outcomes in emerging adulthood from intelligence, personality, and socioeconomic status. *J. Pers. Soc. Psychol.* **123**, 1386–1406 (2022).
3. García-Mayor, J., Moreno-Llamas, A. & la Cruz-Sánchez, E.D. High educational attainment redresses the effect of occupational social class on health-related lifestyle: findings from four Spanish national health surveys. *Ann. Epidemiol.* **58**, 29–37 (2021).
4. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
5. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
6. Lee, J.J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
7. Okbay, A. et al. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**, 437–449 (2022).
8. Alemu, R., Young, A.S., Benjamin, D.J., Turley, P. & Okbay, A. Dissecting the predictive accuracy of polygenic indexes for behavioral phenotypes across genetic ancestries. *bioRxiv* https://doi.org/10.1101/2025.09.11.675704 (2026).
9. Chen, T.T. et al. Shared genetic architectures of educational attainment in East Asian and European populations. *Nat. Hum. Behav.* **8**, 562–575 (2024).
10. Bulik-Sullivan, B.K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
11. Zhang, F.T. et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* **20**, 107 (2019).
12. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
13. Yang, Z.K. et al. CARMA is a new Bayesian model for fine-mapping in genome-wide association meta-analyses. *Nat. Genet.* **55**, 1057–1065 (2023).
14. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
15. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
16. Qi, T., Song, L.Y., Guo, Y.Z., Chen, C. & Yang, J. From genetic associations to genes: methods, applications, and challenges. *Trends Genet.* **40**, 642–667 (2024).
17. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
18. Li, A. et al. mBAT-combo: a more powerful test to detect gene-trait associations from GWAS data. *Am. J. Hum. Genet.* **110**, 30–43 (2023).
19. Zhu, Z.H. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
20. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
21. Weeks, E.M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
22. Meng, L.Y. et al. *MED27* variants cause developmental delay, dystonia, and cerebellar hypoplasia. *Ann. Neurol.* **89**, 828–833 (2021).
23. Maroofian, R. et al. Biallelic *MED27* variants lead to variable ponto-cerebello-lental degeneration with movement disorders. *Brain* **146**, 5031–5043 (2023).
24. Li, X.C. et al. The indispensable role of Mediator complex subunit 27 during neurodevelopment. *Cell Biosci.* **15**, 83 (2025).
25. Shibata, M. et al. Hominini-specific regulation of *CBLN2* increases prefrontal spinogenesis. *Nature* **598**, 489–494 (2021).
26. Seigneur, E., Polepalli, J.S. & Südhof, T.C. Cbln2 and Cbln4 are expressed in distinct medial habenula-interpeduncular projections and contribute to different behavioral outputs. *Proc. Natl. Acad. Sci. USA* **115**, E10235–E10244 (2018).
27. Hu, Z.H., Xiao, X., Zhang, Z.H. & Li, M. Genetic insights and neurobiological implications from *NRXN1* in neuropsychiatric disorders. *Mol. Psychiatry* **24**, 1400–1414 (2019).
28. Sebastian, R. et al. Schizophrenia-associated *NRXN1* deletions induce developmental-timing- and cell-type-specific vulnerabilities in human brain organoids. *Nat. Commun.* **14**, 3770 (2023).
29. Thevenon, J. et al. Heterozygous deletion of the *LRFN2* gene is associated with working memory deficits. *Eur. J. Hum. Genet.* **24**, 911–918 (2016).
30. Morandell, J. et al. *Cul3* regulates cytoskeleton protein homeostasis and cell migration during a critical window of brain development. *Nat. Commun.* **12**, 3058 (2021).
31. Monderer-Rothkoff, G. et al. AUTS2 isoforms control neuronal differentiation. *Mol. Psychiatry* **26**, 666–681 (2021).
32. Cheng, F.F. et al. Multi-ancestry genome-wide association analyses of refractive error augment genetic discovery and polygenic prediction. *Nat. Genet.* **58**, 1030–1039 (2026).
33. Zhu, Z.H. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
34. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
35. Finucane, H.K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
36. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
37. Zhang, M.J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).
38. Buchanan, J.A., da Costa, N.M. & Cheadle, L. Emerging roles of oligodendrocyte precursor cells in neural circuit development and remodeling. *Trends Neurosci.* **46**, 628–639 (2023).
39. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792.e21 (2022).
40. Song, L.Y., Chen, W.H., Hou, J.R., Guo, M.M. & Yang, J. Spatially resolved mapping of cells associated with human complex traits. *Nature* **641**, 932–941 (2025).
41. Siletti, K. et al. Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).

42. Musall, S. et al. Pyramidal cell types drive functionally distinct cortical activity patterns during decision-making. *Nat. Neurosci.* **26**, 495–505 (2023).
43. Karczewski, K.J. et al. Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. *Nat. Genet.* **57**, 2408–2417 (2025).
44. Kandel, E.R., Dudai, Y. & Mayford, M.R. The molecular and systems biology of memory. *Cell* **157**, 163–186 (2014).
45. Dai, J.Y., Patzke, C., Liakath-Ali, K., Seigneur, E. & Südhof, T.C. GluD1 is a signal transduction device disguised as an ionotropic receptor. *Nature* **595**, 261–265 (2021).
46. Kim, J. et al. Genome-wide meta-analysis identifies novel risk loci for uterine fibroids within and across multiple ancestry groups. *Nat. Commun.* **16**, 2273 (2025).
47. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C.A. An integrated framework for local genetic correlation analysis. *Nat. Genet.* **54**, 274–282 (2022).
48. Mounier, N. & Kutalik, Z. Bias correction for inverse variance weighting Mendelian randomization. *Genet. Epidemiol.* **47**, 314–331 (2023).
49. Kentistou, K.A. et al. Understanding the genetic complexity of puberty timing across the allele frequency spectrum. *Nat. Genet.* **56**, 1397–1411 (2024).
50. Holtmaat, A. & Caroni, P. Functional and structural underpinnings of neuronal assembly formation in learning. *Nat. Neurosci.* **19**, 1553–1562 (2016).
51. Turrigiano, G.G. The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell* **135**, 422–435 (2008).
52. McFarlan, A.R. et al. The plasticitome of cortical interneurons. *Nat. Rev. Neurosci.* **24**, 80–97 (2023).
53. Lefton, K.B. et al. Norepinephrine signals through astrocytes to modulate synapses. *Science* **388**, 776–783 (2025).
54. Wang, X.T. et al. Polygenic risk prediction: why and when out-of-sample prediction R^2 can exceed SNP-based heritability. *Am. J. Hum. Genet.* **110**, 1207–1215 (2023).
55. Walters, R.G. et al. Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom.* **3**, 100361 (2023).
56. Xiao, H. et al. Genetic analyses of 104 phenotypes in 20,900 Chinese pregnant women reveal pregnancy-specific discoveries. *Cell Genom.* **4**, 100633 (2024).
57. Zhu, H.H. et al. cfGWAS reveal genetic basis of cell-free DNA end motifs. *Nat. Commun.* **17**, 1714 (2026).
58. Kang, K. et al. Direct-to-consumer genetic testing in China and its role in GWAS discovery and replication. *Quant. Biol.* **9**, 201–215 (2021).
59. Cao, Y.N. et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**, 717–731 (2020).
60. Gou, W.L. et al. Westlake Gut Project: a consortium of microbiome epidemiology for the gut microbiome and health research in China. *Med. Microecol.* **14**, 100064 (2022).
61. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
62. Setoh, K. & Matsuda, F. Cohort profile: the nagahama prospective genome cohort for comprehensive human bioscience (The Nagahama Study). In: Yano, M., Matsuda, F., Sakuntabhai, A. & Hirota, S. eds. Socio-life science and the COVID-19 outbreak: public health and public policy. Springer, 127–143 (2022).
63. Winkler, T.W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
64. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
65. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
66. Turley, P. et al. Multi-Ancestry Meta-Analysis yields novel genetic discoveries and ancestry-specific associations. *bioRXiv* https://doi.org/10.1101/2021.04.23.441003 (2021).
67. 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).
68. Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
69. Wang, Y. et al. Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genom.* **3**, 100241 (2023).
70. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A. & Smoller, J.W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
71. Ruan, Y.F. et al. Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
72. Chang, C.C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
73. Yuan, K. et al. Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. *Nature Genetics* **56**, 1841–1850 (2024).
74. Mudge, J.M. et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.* **53**, D966–D975 (2025).
75. Javierre, B.M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384.e19 (2016).
76. Jung, I. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442–1449 (2019).
77. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
78. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001.e19 (2021).
79. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
80. Boix, C.A., James, B.T., Park, Y.P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
81. Ota, M. et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006–3021.e17 (2021).
82. Wang, Q.S. et al. Statistically and functionally fine-mapped blood eQTLs and pQTLs from 1,405 humans reveal distinct regulation patterns and disease relevance. *Nat. Genet.* **56**, 2054–2067 (2024).
83. Chen, Y. et al. Cross-ancestry analysis of brain QTLs enhances interpretation of schizophrenia genome-wide association studies. *Am. J. Hum. Genet.* **111**, 2444–2457 (2024).
84. Tian, C. et al. Single-cell RNA sequencing of peripheral blood links cell-type-specific regulation of splicing to autoimmune and inflammatory diseases. *Nat. Genet.* **56**, 2739–2752 (2024).
85. Xu, F.Z. et al. Genome-wide genotype-serum proteome mapping provides insights into the cross-ancestry differences in cardiometabolic disease susceptibility. *Nat. Commun.* **14**, 896 (2023).
86. Hatton, A.A. et al. Genetic control of DNA methylation is largely shared across European and East Asian populations. *Nat. Commun.* **15**, 2713 (2024).
87. Peng, Q.Q. et al. Analysis of blood methylation quantitative trait loci in East Asians reveals ancestry-specific impacts on complex traits. *Nat. Genet.* **56**, 846–860 (2024).
88. Xue, A.L. et al. Genome-wide analyses of behavioural traits are subject to bias by misreports and longitudinal changes. *Nat. Commun.* **12**, 20211 (2021).
89. Bryois, J. et al. Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).
90. Adams, M.J. et al. Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. *Cell* **188**, 640–652.e9 (2025).

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.15302/vita.2026.11.0089.

Correspondence and requests for materials should be addressed to Ju-Sheng Zheng, Kun Zhou, Gang Chen, Xin Jin, Dianjianyi Sun, Chikashi Terao or Jian Yang.

Reprints and permission information is available at https://www.vita-journal.com/.

© The Author(s) 2026. Published by Higher Education Press. This is an Open Access article distributed under the terms of the CC BY license (https://creativecommons.org/licenses/by/4.0/).

1. College of Life Sciences, Fudan University, Shanghai, China. 2. New Cornerstone Science Laboratory, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China. 3. Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China. 4. RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan. 5. Department of Clinical Pharmacology and Therapeutics, The Jikei University School of Medicine, Tokyo, Japan. 6. Department of Statistical and Translational Genetics, Center for Genomic Medicine, Fujita Health University, Aichi, Japan. 7. Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. 8. Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan. 9. Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China. 10. Daxing Center for Disease Control and Prevention, Beijing, China. 11. Shenzhen Key Laboratory of Transomics Biotechnologies, BGI Research, Shenzhen, Guangdong, China. 12. College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. 13. Institute of Maternal and Child Health, Wuhan Children's Hospital (Wuhan Maternal and Child Health Care Hospital), Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China. 14. WeGene, Shenzhen Zaozhidao Technology Co., Ltd., Shenzhen, Guangdong, China. 15. Shenzhen WeGene Clinical Laboratory, Shenzhen, Guangdong, China. 16. Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, Hunan, China. 17. Chengdu 23mofang Biotechnology Co., Ltd., Chengdu, Sichuan, China. 18. Affiliated Hangzhou First People's Hospital, School of Medicine, Westlake University, Hangzhou, Zhejiang, China. 19. Medical Genome Center, Research Institute, National Center for Geriatrics and Gerontology, Aichi, Japan. 20. Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford, UK. 21. Peking University Center for Public Health and Epidemic Preparedness and Response, Beijing, China. 22. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China. 23. State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing, China. 24. Department of Obstetrics, Wuhan Children's Hospital (Wuhan Maternal and Child Health Care Hospital), Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China. 25. State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen, Guangdong, China. 26. Section of Epidemiology and Population Health & Department of Gynecology and Obstetrics, Ministry of Education Key Laboratory of Birth Defects and Related Diseases of Women and Children & Children's Medicine Key Laboratory of Sichuan Province, West China Second University Hospital, Sichuan University, Chengdu, Sichuan, China. 27. Department of Epidemiology and Biostatistics, Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China. 28. Guangdong Provincial Key Laboratory of Food, Nutrition and Health; Department of Epidemiology, School of Public Health, Sun Yat-sen University, Guangzhou, Guangdong, China. 29. Department of Epidemiology & Biostatistics, and Department of Respiratory Diseases of Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Huangzhou, Zhejiang, China. 30. The Second Affiliated Hospital, School of Public Health, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. 31. Nuffield Department of Population Health, University of Oxford, Oxford, UK. 32. Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, AU. 33. Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. 34. Graduate Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. 35. The Innovation Centre of Ministry of Education for Development and Diseases, School of Medicine, South China University of Technology, Guangzhou, Guangdong, China. 36. Shanxi Medical University-BGI Collaborative Center for Future Medicine, Shanxi Medical University, Taiyuan, Shanxi, China. 37. State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University Health Science Center, Beijing, China. 38. Department of Applied Genetics, School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. 39. Department of Statistical and Translational Genetics, Center for Genomic Medicine, Fujita Health University, Aichi, Japan.