Leveraging the local genetic structure for trans-ancestry association mapping

Authors

Jiashun Xiao, Mingxuan Cai, Xinyi Yu, Xianghong Hu, Gang Chen, Xiang Wan, Can Yang

Correspondence

wanxiang@sribd.cn (X.W.), macyang@ust.hk (C.Y.)

We introduce LOG-TRAM, a method to leverage the local genetic architecture for trans-ancestry association mapping. Not only can LOG-TRAM control type I errors at the nominal level, but it also can improve the statistical power of identifying risk variants in under-represented populations.

Xiao et al., 2022, The American Journal of Human Genetics *109*, 1317–1337
July 7, 2022
https://doi.org/10.1016/j.ajhg.2022.05.013

CellPress

Leveraging the local genetic structure for trans-ancestry association mapping

Jiashun Xiao,[1,2,6] Mingxuan Cai,[1,2,6] Xinyi Yu,[1,2] Xianghong Hu,[1,2] Gang Chen,[3] Xiang Wan,[4,5,*] and Can Yang[1,2,*]

Summary

Over the past two decades, genome-wide association studies (GWASs) have successfully advanced our understanding of the genetic basis of complex traits. Despite the fruitful discovery of GWASs, most GWAS samples are collected from European populations, and these GWASs are often criticized for their lack of ancestry diversity. Trans-ancestry association mapping (TRAM) offers an exciting opportunity to fill the gap of disparities in genetic studies between non-Europeans and Europeans. Here, we propose a statistical method, LOG-TRAM, to leverage the local genetic architecture for TRAM. By using biobank-scale datasets, we showed that LOG-TRAM can greatly improve the statistical power of identifying risk variants in under-represented populations while producing well-calibrated p values. We applied LOG-TRAM to the GWAS summary statistics of various complex traits/diseases from BioBank Japan, UK Biobank, and African populations. We obtained substantial gains in power and achieved effective correction of confounding biases in TRAM. Finally, we showed that LOG-TRAM can be successfully applied to identify ancestry-specific loci and the LOG-TRAM output can be further used for construction of more accurate polygenic risk scores in under-represented populations.

Introduction

Thousands of genome-wide association studies (GWASs) have been conducted to understand the genetic basis of complex human traits/diseases since the first GWAS publication in 2005.[1] As of April 2022, more than 370,000 genome-wide significant associations (p value $\leq 5 \times 10^{-8}$) have been identified between single-nucleotide polymorphisms (SNPs) and complex human traits.[2] The summary statistics from GWASs are accessible through public gateways, such as the GWAS Catalog.[2] These datasets contain rich information on genetic variations and phenotypes, providing an unprecedented research opportunity in biomedical and social science.[3]

Despite the great achievements, GWAS findings are limited by the lack of ancestral diversity. According to the GWAS Diversity Monitor (https://gwasdiversitymonitor.com), about 88.9% of GWAS participants have been of European ancestry (EUR) to date.[4] Non-European populations are severely under-represented in GWASs. For example, the proportions of participants of East Asian (EAS) and African ancestries are less than 6.9% and 0.4%, respectively.[4] Most of GWAS findings are based on individuals of European ancestry. Unfortunately, these findings may not be directly extrapolated to non-European ancestries.[5] As genetic studies with diversified ancestries are important in achieving global health equity, trans-ancestry association mapping (TRAM), which aims to identify risk genetic variants across ancestries (particularly in the under-represented populations), has become a critical step toward precision medicine.[5,6]

The challenges of TRAM arise from two major aspects. First, the genetic architectures of a phenotype are heterogeneous across ancestries.[5] Some trait-associated SNPs have vastly different allele frequencies between European and non-European ancestries.[5,7] SNP effect sizes and linkage disequilibrium (LD) patterns can also vary across ancestries.[8] Second, the publicly released GWAS summary statistics still suffer from confounding biases.[9] Although principal-component analysis (PCA)[10] and linear mixed models (LMMs)[11] are commonly used for association mapping in GWASs, the population stratification in biobank-scale data, such as socioeconomic status[12] or geographic structure,[13] may not be fully accounted for in these standard approaches.[14] Without correcting confounding biases hidden in GWAS summary statistics, TRAM will produce many false positive findings.

Much effort has been devoted to the development of statistical methods for TRAM. To handle heterogeneity of genetic architectures, several statistical methods have been developed for meta-analysis of GWAS data in various contexts.[15] To name a few, the random-effects (RE) model is a popular tool for meta-analysis of GWASs.[16,17] Despite its popularity, the RE model assumes that effect sizes vary greatly under the null hypothesis. This assumption can be fairly conservative.[18] To improve the power of RE models for TRAM, a new RE method named RE2[18] has been developed by assuming no heterogeneity under the null hypothesis. Later, these authors further improved the RE2

[1]Guangzhou HKUST Fok Ying Tung Research Institute, Guangzhou 511458, China; [2]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR, China; [3]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China; [4]Shenzhen Research Institute of Big Data, Shenzhen 518172, China; [5]Pazhou Lab, Guangzhou 510330, China
[6]These authors contributed equally
*Correspondence: wanxiang@sribd.cn (X.W.), macyang@ust.hk (C.Y.)
https://doi.org/10.1016/j.ajhg.2022.05.013.

method by allowing meta-analysis of correlated statistics, namely, RE2C.[19] Unlike the general hypothesis testing methods, MANTRA was specifically designed for meta-analysis of trans-ancestry GWAS data.[20] MANTRA assumes that the effect sizes should be closely matched for similar genetic ancestries; however, it allows for heterogeneity in effect sizes for more distal ancestries. MTAG[21] is a recently developed method to analyze multiple GWASs from a single ancestry. MTAG can greatly improve the statistical power of association mapping because it uses the global genetic correlation to borrow information across related traits. It assumes that the variance and covariance of marginal effect sizes are homogeneous across SNPs. This assumption may limit its usage in the trans-ancestry setting. Very recently, MAMA[22] has been developed to improve MTAG for TRAM by accounting for heterogeneity in LD across ancestries. However, both MTAG and MAMA are still not fully satisfactory when the local genetic architecture differs from the global architecture.

In this work, we develop a statistical method to leverage the local genetic architecture for trans-ancestry association mapping (LOG-TRAM). Not only can LOG-TRAM control type I errors at the nominal level, it can also improve the statistical power of identifying risk variants in under-represented populations. Below are the keys to the success of LOG-TRAM. First, it can greatly improve the statistical power of association mapping by making use of biobank-scale datasets of the auxiliary population (e.g., the UK Biobank [UKBB] dataset) while accounting for heterogeneity among multiple ancestries. Second, compared to existing meta-analysis methods that consider the global genetic architecture, LOG-TRAM focuses on the local genetic architectures to localize risk variants, including local heritability, local co-heritability, allele frequencies, SNP effect sizes, and LD patterns. Third, it is capable of correcting the confounding bias hidden in GWAS summary statistics to avoid inflated type I errors. Fourth, LOG-TRAM only takes summary statistics from multiple ancestries as inputs and outputs well-calibrated p values. With the innovations of our model design, LOG-TRAM is computationally efficient, as it has a closed-form solution at each step. Through comprehensive simulation studies, we demonstrated that LOG-TRAM largely outperformed existing meta-analysis approaches in terms of type I error rate and power. Then we applied LOG-TRAM to the GWAS summary statistics of 29 complex traits and diseases from BioBank Japan (BBJ) and UKBB. The analysis results show that LOG-TRAM can effectively account for confounding biases in the GWAS summary statistics and achieve a substantial gain in power for identification of risk variants. Furthermore, we showed the generality of our method in other under-represented populations by applying LOG-TRAM to integrate the GWASs of 17 traits from African and EUR/EAS. Finally, we successfully applied LOG-TRAM to identify ancestry-specific loci and showed that the LOG-TRAM output can be used for the construction of more accurate polygenic risk scores (PRSs) in under-represented populations.

Material and methods

Notation and problem setup

To introduce LOG-TRAM, we begin our formulation with the individual-level GWAS data. Without loss of generality, we consider two populations and the following model that relates genotypes with phenotypes in populations 1 and 2, respectively,

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon_1, \quad \mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon_2, \qquad \text{(Equation 1)}$$

where $\mathbf{y}_1 = [y_{1,1}, \ldots, y_{1,N_1}]^T \in \mathbb{R}^{N_1}$ and $\mathbf{y}_2 = [y_{2,1}, \ldots, y_{2,N_2}]^T \in \mathbb{R}^{N_2}$ are two phenotype vectors, $\mathbf{X}_1 = [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1M}] \in \mathbb{R}^{N_1 \times M}$ and $\mathbf{X}_2 = [\mathbf{x}_{21}, \ldots, \mathbf{x}_{2M}] \in \mathbb{R}^{N_2 \times M}$ are standardized genotype matrices whose columns have zero mean and unit variance, $\boldsymbol{\beta}_1 = [\beta_{11}, \beta_{12}, \ldots, \beta_{1M}]^T \in \mathbb{R}^M$ and $\boldsymbol{\beta}_2 = [\beta_{21}, \beta_{22}, \ldots, \beta_{2M}]^T \in \mathbb{R}^M$ are the SNP effect sizes, the residual vectors $\varepsilon_1 \in \mathbb{R}^{N_1}$ and $\varepsilon_2 \in \mathbb{R}^{N_2}$ are independent error terms, M is the number of SNPs, and N_1 and N_2 are the sample sizes of populations 1 and 2, respectively. We consider population 1 as the under-represented target population and population 2 as the auxiliary population with a biobank-scale sample size, i.e., $N_1 \ll N_2$. We expect that information from population 2 can be useful for association mapping in population 1. Here, we assume that the covariates (e.g., age, sex, and principal components) have been properly adjusted. More detailed treatment on covariate adjustment can be found in our previous work[23] and other related work.[24] We also implicitly assume that SNP effect sizes increase as the allele frequencies decrease when working with standardized genotype matrices.[23]

Suppose that individual-level data $\{\mathbf{y}_1, \mathbf{X}_1, \mathbf{y}_2, \mathbf{X}_2\}$ are not accessible but summary statistics $\{\widehat{\mathbf{b}}_1, \widehat{\mathbf{s}}_1\} = \{\widehat{b}_{1j}, \widehat{s}_{1j}\}_{j=1,\ldots,M}$ and $\{\widehat{\mathbf{b}}_2, \widehat{\mathbf{s}}_2\} = \{\widehat{b}_{2j}, \widehat{s}_{2j}\}_{j=1,\ldots,M}$ from the simple linear regressions are available:

$$\begin{aligned}
\widehat{b}_{1j} &= \mathbf{x}_{1j}^T\mathbf{y}_1 \big/ \mathbf{x}_{1j}^T\mathbf{x}_{1j} = \mathbf{x}_{1j}^T\mathbf{y}_1 \big/ N_1, \\
\widehat{s}_{1j} &= \sqrt{(\mathbf{y}_1 - \mathbf{x}_{1j}\widehat{b}_{1j})^T(\mathbf{y}_1 - \mathbf{x}_{1j}\widehat{b}_{1j}) \big/ \left(N_1\mathbf{x}_{1j}^T\mathbf{x}_{1j}\right)}, \\
\widehat{b}_{2j} &= \mathbf{x}_{2j}^T\mathbf{y}_2 \big/ \mathbf{x}_{2j}^T\mathbf{x}_{2j} = \mathbf{x}_{2j}^T\mathbf{y}_2 \big/ N_2, \\
\widehat{s}_{2j} &= \sqrt{(\mathbf{y}_2 - \mathbf{x}_{2j}\widehat{b}_{2j})^T(\mathbf{y}_2 - \mathbf{x}_{2j}\widehat{b}_{2j}) \big/ \left(N_2\mathbf{x}_{2j}^T\mathbf{x}_{2j}\right)}.
\end{aligned} \qquad \text{(Equation 2)}$$

Traditional association methods[25] often report genome-wide significant association status based on the Z scores $z_{1j} = \widehat{b}_{1j}/\widehat{s}_{1j}$ and $z_{2j} = \widehat{b}_{2j}/\widehat{s}_{2j}$. However, these association methods have several limitations. First, they do not account for the heterogeneity of LD patterns across ancestries. Second, they do not correct possible confounding factors hidden in summary statistics $\{\widehat{\mathbf{b}}, \widehat{\mathbf{s}}\} = \{\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{s}}_1, \widehat{\mathbf{s}}_2\}$, resulting in inflated type I errors. Third, they do not fully utilize the correlation between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, leading to a suboptimal statistical power.

Recent meta-analysis studies have revealed that statistical power of association mapping in under-represented populations can be improved by leveraging global cross-population genetic correlation.[20,22] However, current trans-ancestry GWAS meta-analysis relies on the assumption that variance and covariance of SNP effect sizes are homogeneous across the entire genome.[21,22] As a matter of fact, accumulating evidence suggests different genomic regions

contribute disproportionately to the heritability across the genome, inducing heterogeneous genetic similarity between populations in different local regions.[26–30] For example, a very recent study[28] has shown that the shared genetics between autism spectrum disorder (ASD) and cognitive performance (CP) and between ASD and intellectual disability are distinct at a local scale, which explains an empirical paradox: a positive genetic correlation between ASD and CP contradicts the comorbidity between ASD and intellectual disability. Therefore, it can be more appropriate to localize risk variants by leveraging the local genetic architecture rather than the global architecture. To see this more intuitively, let us consider the global genetic correlation $r_g = Corr(\beta_{1j}, \beta_{2j})$ averaged across the entire genome $j \in \{1, \ldots, M\}$ and the local genetic correlation $r_{g,\mathcal{R}} = Corr(\beta_{1j}, \beta_{2j})$ defined on a local genomic region $j \in \mathcal{R}$, where the local genetic correlation $r_{g,\mathcal{R}}$ can be very different from the global r_g.[27,28,31,32] On the one hand, $r_{g,\mathcal{R}}$ can be nearly zero even if the global correlation is substantial, e.g., $r_g = 0.7$. Clearly, no information should be borrowed from the auxiliary population for association mapping of the target population in the local region $\mathcal{R}$. In such a case, leveraging the global genetic correlation ($r_g = 0.7$) for TRAM leads to inflated type I errors. On the other hand, when using the global genetic correlation for TRAM, we lose statistical power substantially in the presence of strong local genetic correlation $r_{g,\mathcal{R}} = 0.7$ but weak global genetic correlation (e.g., $r_g \approx 0$).

To leverage the local genetic architecture for TRAM, we propose a statistical method named LOG-TRAM, which only requires GWAS summary statistics $\{\widehat{\mathbf{b}}, \widehat{\mathbf{s}}\}$ and a reference genome (e.g., The 1000 Genomes Project) as inputs (Figure 1). The output of LOG-TRAM $\{\widehat{\mathbf{b}}^{\text{TRAM}}, \widehat{\mathbf{s}}^{\text{TRAM}}\}$ has the following desired properties: (1) by borrowing information from biobank-scale European datasets through the local genetic architecture, the statistical power of non-EUR association studies can be largely improved; (2) the influence of confounding factors can be adjusted; (3) the LOG-TRAM estimate is unbiased and its standard error can be much smaller than that of standard GWASs (see supplemental methods, section 3.4); and (4) the LOG-TRAM output can be used for downstream analysis, such as construction of more accurate polygenic risk scores in the under-represented populations.

The LOG-TRAM model

To leverage the local genetic architecture, we first partition the genome into consecutive non-overlapping regions (e.g., 1 M base pair segments). Then we focus on identification of risk variants within a given region $\mathcal{R}$ at a time. Given the local region $\mathcal{R}$, we extend Equation 1 as:

$$\begin{aligned} \mathbf{y}_1 &= \sum_{j\in\mathcal{R}} \mathbf{x}_{1j}\beta_{1j} + \sum_{k\in\mathcal{B}} \mathbf{x}_{1k}\beta_{1k} + \varepsilon_1, \\ y_2 &= \sum_{j\in\mathcal{R}} \mathbf{x}_{2j}\beta_{2j} + \sum_{k\in\mathcal{B}} \mathbf{x}_{2k}\beta_{2k} + \varepsilon_2, \end{aligned} \quad \text{(Equation 3)}$$

where $\mathcal{B} = \{1, \ldots, M\}R$ is the set of all SNPs in the genome except the local region $\mathcal{R}$, ε_1 and ε_2 are vectors of independent noise with $\mathbb{E}[\varepsilon_1] = \mathbb{E}[\varepsilon_2] = 0$, $\text{Var}[\varepsilon_1] = \sigma^2_{\varepsilon_1}\mathbf{I}_{N_1}$, and $\text{Var}[\varepsilon_2] = \sigma^2_{\varepsilon_2}\mathbf{I}_{N_2}$. Partitioning the genome into the local region $\mathcal{R}$ and the background region $\mathcal{B}$ allows us to model the regional genetic architecture differently from the global pattern. To achieve this, we introduce the following probabilistic structure for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

$$\begin{aligned} \mathbb{E}\begin{pmatrix}\beta_{1j}\\ \beta_{2j}\end{pmatrix} = 0,\ \text{Var}\begin{pmatrix}\beta_{1j}\\ \beta_{2j}\end{pmatrix} = \boldsymbol{\Omega}_{\mathcal{R}} = \begin{pmatrix}\omega_{\mathcal{R},1} & \omega_{\mathcal{R},12}\\ \omega_{\mathcal{R},12} & \omega_{\mathcal{R},2}\end{pmatrix},\ \text{for } j\in\mathcal{R}, \\ \mathbb{E}\begin{pmatrix}\beta_{1k}\\ \beta_{2k}\end{pmatrix} = 0,\ \text{Var}\begin{pmatrix}\beta_{1k}\\ \beta_{2k}\end{pmatrix} = \boldsymbol{\Omega}_{\mathcal{B}} = \begin{pmatrix}\omega_{\mathcal{B},1} & \omega_{\mathcal{B},12}\\ \omega_{\mathcal{B},12} & \omega_{\mathcal{B},2}\end{pmatrix},\ \text{for } k\in\mathcal{B}, \end{aligned} \quad \text{(Equation 4)}$$

where $\boldsymbol{\Omega}_{\mathcal{R}}$ and $\boldsymbol{\Omega}_{\mathcal{B}}$ capture the genetic covariance of two populations in the local region $\mathcal{R}$ and background region $\mathcal{B}$, respectively. The diagonal elements represent the local/global per-SNP heritability for populations 1 and 2, and the off-diagonal elements represent the local/global per-SNP co-heritability between the two populations. Without loss of generality, we assume that the phenotypes have been standardized, i.e., $\text{Var}(y_{1,i}) = \text{Var}(y_{2,i}) = 1$. Based on Equations 3 and 4, we have $|\mathcal{R}|\cdot\omega_{\mathcal{R},1} + |\mathcal{B}|\cdot\omega_{\mathcal{B},1} + \sigma^2_{\varepsilon_1} = \text{Var}(y_{1,i}) = 1$ and $|\mathcal{R}|\cdot\omega_{\mathcal{R},2} + |\mathcal{B}|\cdot\omega_{\mathcal{B},2} + \sigma^2_{\varepsilon_2} = \text{Var}(y_{2,i}) = 1$, where $|\mathcal{R}|$ and $|\mathcal{B}|$ represent the number of SNPs in $\mathcal{R}$ and $\mathcal{B}$, respectively.

The flexible probabilistic structure given in Equation 4 has three salient properties. First, we allow the effect sizes $[\beta_{1j}, \beta_{2j}]^T$ of SNP j in the target region $\mathcal{R}$ to have different variances ($\omega_{\mathcal{R},1}$ and $\omega_{\mathcal{R},2}$) in different populations. Importantly, we achieve this through the moment conditions, which does not require any distributional assumptions on $[\beta_{1j}, \beta_{2j}]^T$. When only a subset of SNPs have non-zero effects, $\omega_{\mathcal{R},1}$ and $\omega_{\mathcal{R},2}$ measure the average variances in local region $\mathcal{R}$ induced by the non-zero effects in the corresponding populations. This formulation also offers the flexibility to capture the heterogeneous effect sizes across populations. In the extreme case, we allow $\omega_{\mathcal{R},1} = 0$ while $\omega_{\mathcal{R},2} \neq 0$ when SNPs in local region $\mathcal{R}$ have zero effects in population 1 but non-zero effects in population 2. Second, we characterize the relationship of cross-population effect sizes in region $\mathcal{R}$ by the genetic covariance term $\omega_{\mathcal{R},12}$. When a subset of SNPs have similar effect sizes in two populations, it will be reflected by a non-zero $\omega_{\mathcal{R},12}$. This property allows us to borrow local information from biobank-scale datasets. Third, rather than assuming that the genetic covariance of SNP effect sizes are homogeneous across the entire genome, i.e., $\boldsymbol{\Omega}_{\mathcal{R}} = \boldsymbol{\Omega}_{\mathcal{B}}$, we allow the per-SNP heritability of the local genomic region ($\omega_{\mathcal{R},1}$ and $\omega_{\mathcal{R},2}$) to be different from the genome background ($\omega_{\mathcal{B},1}$ and $\omega_{\mathcal{B},2}$) and the local per-SNP co-heritability ($\omega_{\mathcal{R},12}$) to be different from the global per-SNP co-heritability ($\omega_{\mathcal{B},12}$), effectively capturing the heterogeneous local structures. This assumption distinguishes LOG-TRAM from MTAG or MAMA, both of which assume a global covariance $\boldsymbol{\Omega}$ shared across all SNPs ($\boldsymbol{\Omega} = \boldsymbol{\Omega}_{\mathcal{R}} = \boldsymbol{\Omega}_{\mathcal{B}}$).

So far, we have described the method without covariates. In the presence of covariates such as gender, age, and principal-component scores, we extend Equation 3 as

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{Z}_1\mathbf{u}_1 + \sum_{j\in\mathcal{R}} \mathbf{x}_{1j}\beta_{1j} + \sum_{k\in\mathcal{B}} \mathbf{x}_{1k}\beta_{1k} + \varepsilon_1, \\ \mathbf{y}_2 &= \mathbf{Z}_2\mathbf{u}_2 + \sum_{j\in\mathcal{R}} \mathbf{x}_{2j}\beta_{2j} + \sum_{k\in\mathcal{B}} \mathbf{x}_{2k}\beta_{2k} + \varepsilon_2, \end{aligned} \quad \text{(Equation 5)}$$

where $\mathbf{Z}_1 \in \mathbb{R}^{N_1\times q_1}$ and $\mathbf{Z}_2 \in \mathbb{R}^{N_2\times q_2}$ are the covariate matrices of the target and auxiliary population, respectively, and $\mathbf{u}_1$ and $\mathbf{u}_2$ are the corresponding vectors of covariates effects. To get rid of the covariates, we define the projection matrices $\mathbf{P}_1 = \mathbf{I} - \mathbf{Z}_1(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\mathbf{Z}_1^T$ and $\mathbf{P}_2 = \mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2^T\mathbf{Z}_2)^{-1}\mathbf{Z}_2^T$ and obtain the new working model as

$$\begin{aligned} \mathbf{y}_1^{proj} &= \sum_{j\in\mathcal{R}} \mathbf{x}_{1j}^{proj}\beta_{1j} + \sum_{k\in\mathcal{B}} \mathbf{x}_{1k}^{proj}\beta_{1k} + \varepsilon_1^{proj}, \\ \mathbf{y}_2^{proj} &= \sum_{j\in\mathcal{R}} \mathbf{x}_{2j}^{proj}\beta_{2j} + \sum_{k\in\mathcal{B}} \mathbf{x}_{2k}^{proj}\beta_{2k} + \varepsilon_2^{proj}, \end{aligned} \quad \text{(Equation 6)}$$

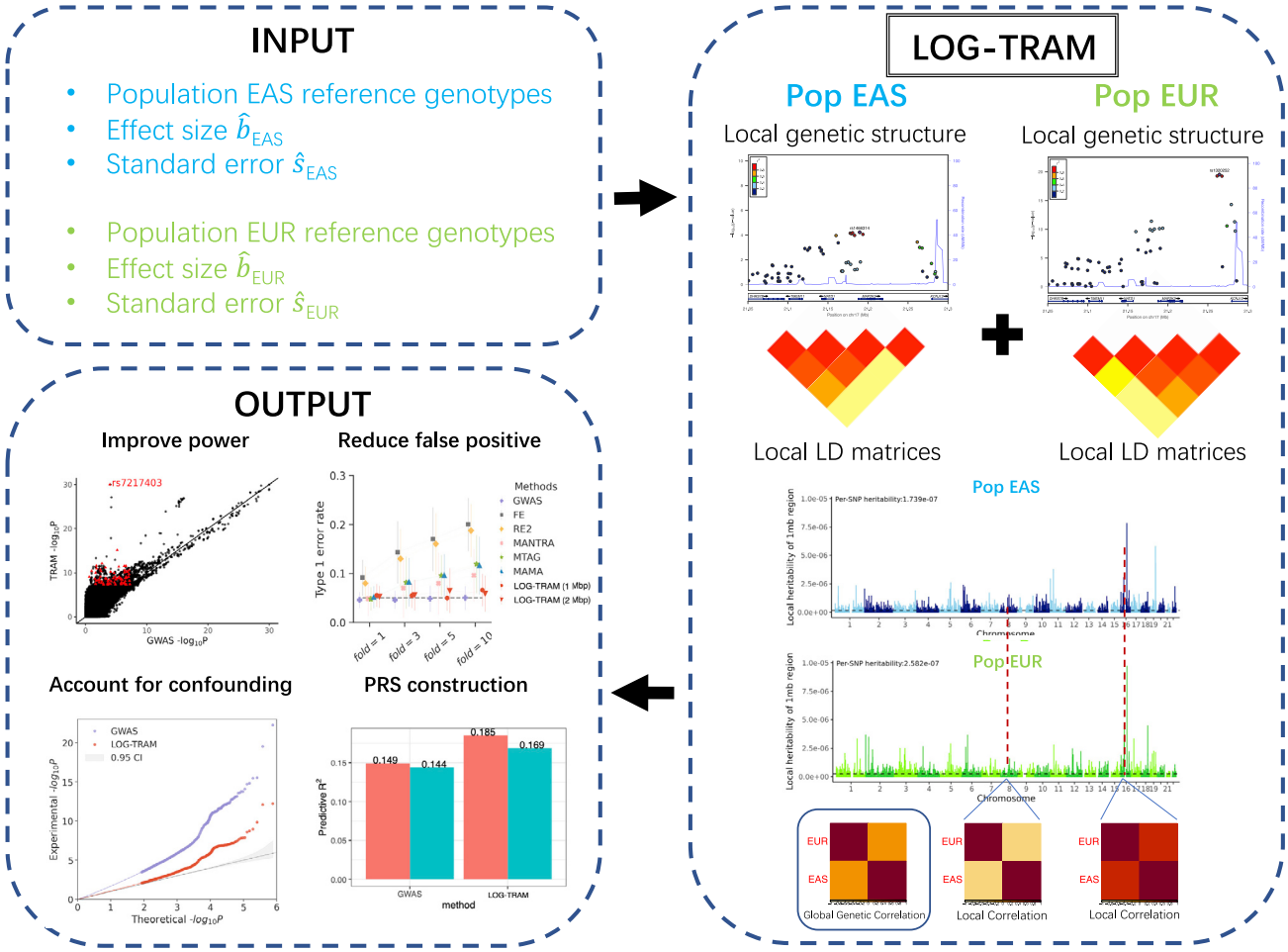

Figure 1. LOG-TRAM overview

LOG-TRAM only requires GWAS summary statistics $\{\widehat{\mathbf{b}}, \widehat{\mathbf{s}}\}$ and reference genomes of the considered populations as inputs. During the trans-ancestry meta-analysis, LOG-TRAM can handle the heterogeneity of genetic architectures (e.g., allele frequencies, LD patterns, and SNP effect sizes) across populations, and leverage information from biobank-scale European datasets through the local genetic architecture to increase statistical power and simultaneously reduce false positives for non-EUR association studies.

where $\mathbf{y}_1^{proj} = \mathbf{P}_1\mathbf{y}_1$, $\mathbf{y}_2^{proj} = \mathbf{P}_2\mathbf{y}_2$, $\mathbf{x}_1^{proj} = \mathbf{P}_1\mathbf{x}_1$, $\mathbf{x}_2^{proj} = \mathbf{P}_2\mathbf{x}_2$, $\varepsilon_1^{proj} = \mathbf{P}_1\varepsilon_1$, and $\varepsilon_2^{proj} = \mathbf{P}_2\varepsilon_2$. By the projection of covariates, Equation 5 reduces to Equation 3. In genomics, this projection approach is commonly used for association mapping[33] and heritability estimation.[34] In the main text, without loss of generality, we work on Equation 3 in the absence of covariates.

The LOG-TRAM model with summary-level data

The individual-level GWAS data are often not easily accessible because of privacy concerns. To overcome this difficulty, we consider the rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ as independent and identically distributed samples from the target population and auxiliary population, respectively. As such, we define the correlation between SNP j and SNP k as $r_{1jk} = \mathbb{E}[\mathbf{x}_{1j}^T\mathbf{x}_{1k}/N_1]$ and $r_{2jk} = \mathbb{E}[\mathbf{x}_{2j}^T\mathbf{x}_{2k}/N_2]$ for target and auxiliary populations, respectively. We can then define the underlying true marginal effect sizes as

$$b_{1j} = \mathbb{E}\left[\widehat{b}_{1j}|\boldsymbol{\beta}_1\right] = \mathbb{E}\left[\mathbf{x}_{1j}^T(\mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon_1)\Big/N_1|\boldsymbol{\beta}_1\right] = \sum_{k=1}^{M} r_{1jk}\beta_{1k},$$

$$b_{2j} = \mathbb{E}\left[\widehat{b}_{2j}|\boldsymbol{\beta}_2\right] = \mathbb{E}\left[\mathbf{x}_{2j}^T(\mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon_2)\Big/N_2|\boldsymbol{\beta}_2\right] = \sum_{k=1}^{M} r_{2jk}\beta_{2k}. \quad \text{(Equation 7)}$$

As we can observe, the true marginal effect size b_{1j} is the summation of underlying true effect sizes β_{1k} in LD with SNP j weighted by the pairwise SNP correlation between SNPs j and k. With the above model specification, we can show that the following relationship holds for the obtained summary statistics $\{\widehat{b}_{1j}, \widehat{b}_{2j}, \widehat{s}_{1j}, \widehat{s}_{2j}\}$ for SNP j in the local region $\mathcal{R}$,

$$\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix} = \begin{pmatrix}b_{1j}\\ b_{2j}\end{pmatrix} + \begin{pmatrix}e_{1j}\\ e_{2j}\end{pmatrix}, \quad j \in \mathcal{R}, \quad \text{(Equation 8)}$$

where e_{1j} and e_{2j} are the independent estimation errors with $\mathrm{Var}(e_{1j}) = \widehat{s}_{1j}^2$ and $\mathrm{Var}(e_{2j}) = \widehat{s}_{2j}^2$. Because the two datasets are from different populations, we have $\mathrm{Cov}(e_{1j}, e_{2j}) = 0$.

To establish the connection between observed summary statistics $\{\widehat{b}_{1j}, \widehat{b}_{2j}, \widehat{s}_{1j}, \widehat{s}_{2j}\}$ and model parameters $\{\omega_{\mathcal{R},1}, \omega_{\mathcal{R},2}, \omega_{\mathcal{R},12}, \omega_{\mathcal{B},1}, \omega_{\mathcal{B},2}, \omega_{\mathcal{B},12}\}$, we first introduce the LD scores in the context of the LOG-TRAM model. Given the SNP correlations within a population (e.g., r_{1jk} and r_{2jk}), we denote $l_{\mathcal{R},j1} = \sum_{k\in\mathcal{R}} r_{1jk}^2$ and $l_{\mathcal{B},j1} = \sum_{k\in\mathcal{B}} r_{1jk}^2$ as the single-population LD scores of SNP j in population 1 for regions $\mathcal{R}$ and $\mathcal{B}$, respectively, $l_{\mathcal{R},j2} = \sum_{k\in\mathcal{R}} r_{2jk}^2$ and $l_{\mathcal{B},j2} = \sum_{k\in\mathcal{B}} r_{2jk}^2$ as the single-population LD scores of SNP j in

population 2 for regions $\mathcal{R}$ and $\mathcal{B}$, respectively, and $l_{\mathcal{R},j12} = \sum_{k\in\mathcal{R}} r_{1jk}r_{2jk}$ and $l_{\mathcal{B},j12} = \sum_{k\in\mathcal{B}} r_{1jk}r_{2jk}$ as the cross-population LD scores of SNP j for regions $\mathcal{R}$ and $\mathcal{B}$, respectively. We leave the details of LD score calculation in supplemental methods, section 3.5.

On the basis of above, we can derive the following relationships (see supplemental methods, section 3.1):

$$\begin{aligned}
\mathbb{E}\left[\widehat{b}_{1j}\right] &= \mathbb{E}\left[\widehat{b}_{2j}\right] = 0,\\
\mathbb{E}\left[\widehat{b}_{1j}^2\right] &= \omega_{\mathcal{R},1}l_{\mathcal{R},j1} + \omega_{\mathcal{B},1}l_{\mathcal{B},j1} + \widehat{s}_{1j}^2,\\
\mathbb{E}\left[\widehat{b}_{2j}^2\right] &= \omega_{\mathcal{R},2}l_{\mathcal{R},j2} + \omega_{\mathcal{B},2}l_{\mathcal{B},j2} + \widehat{s}_{2j}^2,\\
\mathbb{E}\left[\widehat{b}_{1j}\widehat{b}_{2j}\right] &= \omega_{\mathcal{R},12}l_{\mathcal{R},j12} + \omega_{\mathcal{B},12}l_{\mathcal{B},j12}.
\end{aligned} \qquad \text{(Equation 9)}$$

As LD decays exponentially with distance,[35] Equation 9 implies that the obtained summary statistics largely depend on the local genetic architecture, including the local LD scores ($l_{\mathcal{R},j1}$, $l_{\mathcal{R},j2}$, and $l_{\mathcal{R},j12}$), local per-SNP heritability ($\omega_{\mathcal{R},1}$ and $\omega_{\mathcal{R},2}$), and local per-SNP co-heritability ($\omega_{\mathcal{R},12}$). Combining Equation 8 and Equation 9, we can further derive the relationship between the summary statistics and the model parameters $\mathbf{\Omega}_{\mathcal{R}}$ and $\mathbf{\Omega}_{\mathcal{B}}$,

$$\mathrm{Cov}\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix} = \mathrm{Cov}\begin{pmatrix}b_{1j}\\ b_{2j}\end{pmatrix} + \mathrm{Cov}\begin{pmatrix}e_{1j}\\ e_{2j}\end{pmatrix} = \underbrace{\begin{pmatrix}l_{\mathcal{R},j1} & l_{\mathcal{R},j12}\\ l_{\mathcal{R},j12} & l_{\mathcal{R},j2}\end{pmatrix}\circ\mathbf{\Omega}_{\mathcal{R}} + \begin{pmatrix}l_{\mathcal{B},j1} & l_{\mathcal{B},j12}\\ l_{\mathcal{B},j12} & l_{\mathcal{B},j2}\end{pmatrix}\circ\mathbf{\Omega}_{\mathcal{B}}}_{\mathbf{\Omega}_j} + \underbrace{\begin{pmatrix}\widehat{s}_{1j}^2 & 0\\ 0 & \widehat{s}_{2j}^2\end{pmatrix}}_{\widehat{\mathbf{S}}_j^2}, \qquad \text{(Equation 10)}$$

where $\circ$ is the element-wise product.

To account for the hidden confounding biases in GWAS summary statistics, we generalize linear Equation 3 based on the genetic drift model used in LDSC[9] and modify Equation 9 as (see supplemental methods, section 3.2)

$$\begin{aligned}
\mathbb{E}\left[\widehat{b}_{1j}\right] &= \mathbb{E}\left[\widehat{b}_{2j}\right] = 0,\\
\mathbb{E}\left[\widehat{b}_{1j}^2\right] &= \omega_{\mathcal{R},1}l_{\mathcal{R},j1} + \omega_{\mathcal{B},1}l_{\mathcal{B},j1} + \widehat{s}_{1j}^2c_1,\\
\mathbb{E}\left[\widehat{b}_{2j}^2\right] &= \omega_{\mathcal{R},2}l_{\mathcal{R},j2} + \omega_{\mathcal{B},2}l_{\mathcal{B},j2} + \widehat{s}_{2j}^2c_2,\\
\mathbb{E}\left[\widehat{b}_{1j}\widehat{b}_{2j}\right] &= \omega_{\mathcal{R},12}l_{\mathcal{R},j12} + \omega_{\mathcal{B},12}l_{\mathcal{B},j12} + \widehat{s}_{1j}\widehat{s}_{2j}c_{12},
\end{aligned} \qquad \text{(Equation 11)}$$

where c_1, c_2, and c_{12} are inflation constants that adjust for the confounding biases of GWAS standard errors. From Equation 11, we can see that the influence of population stratification on the variance of marginal estimates remains nearly constant across SNPs ($\widehat{s}_{1j}^2c_1 \approx c_1/N_1$, $\widehat{s}_{2j}^2c_2 \approx c_2/N_2$, and $\widehat{s}_{1j}\widehat{s}_{2j}c_{12} \approx c_{12}/\sqrt{N_1N_2}$), while the magnitudes of genetic effects are tagged by LD scores. Therefore, the Equation 10 can be updated as follows:

$$\mathrm{Cov}\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix} = \mathbf{\Omega}_j + \widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j, \quad \mathrm{Cov}\begin{pmatrix}e_{1j}\\ e_{2j}\end{pmatrix} = \widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j,$$
$$\mathbf{C} = \begin{pmatrix}c_1 & c_{12}\\ c_{12} & c_2\end{pmatrix}. \qquad \text{(Equation 12)}$$

In real data analysis, we find that inflation constants c_1 and c_2 are often larger than one, and c_{12} is very close to zero because of no overlapped samples in trans-ancestry association studies.

The LOG-TRAM estimator

We use the generalized method of moments (GMM)[22] to derive the LOG-TRAM estimator $\{\widehat{\mathbf{b}}^{\mathrm{TRAM}}, \widehat{\mathbf{s}}^{\mathrm{TRAM}}\} = \{\widehat{b}_{1j}^{\mathrm{TRAM}}, \widehat{b}_{2j}^{\mathrm{TRAM}}, \widehat{s}_{1j}^{\mathrm{TRAM}}, \widehat{s}_{2j}^{\mathrm{TRAM}}\}_{j=1,\ldots,M}$. Using Equation 12, we first obtain the conditional mean and conditional variance (see supplemental methods, section 3.3) as

$$\mathbb{E}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix}|b_{1j}\right] = \frac{\boldsymbol{\omega}_{j,\cdot1}}{\omega_{j,11}}b_{1j}, \qquad \text{(Equation 13)}$$

$$\mathrm{Var}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix}|b_{1j}\right] = \mathbf{\Omega}_j - \frac{\boldsymbol{\omega}_{j,\cdot1}\boldsymbol{\omega}_{j,\cdot1}^T}{\omega_{j,11}} + \widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j := \mathbf{\Lambda}_1^{-1}, \qquad \text{(Equation 14)}$$

where $\boldsymbol{\omega}_{j,\cdot1} = [\omega_{j,11}, \omega_{j,12}]^T$ is the first column of $\mathbf{\Omega}_j$. On the basis of Equation 13, we can define $\mathbf{m}(b) := \begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix} - \frac{\boldsymbol{\omega}_{j,\cdot1}}{\omega_{j,11}}b$ and give the first-order moment condition

$$\mathbb{E}\left[\mathbf{m}(b)|_{b=b_{1j}}\right] = \mathbb{E}\left[\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix} - \frac{\boldsymbol{\omega}_{j,\cdot1}}{\omega_{j,11}}b_{1j}\right] = 0. \qquad \text{(Equation 15)}$$

On the basis of GMM, we can obtain the LOG-TRAM estimator as

$$\begin{aligned}
\widehat{b}_{1j}^{\mathrm{TRAM}} &= \arg\min_b \mathbf{m}(b)^T\mathbf{\Lambda}_1\mathbf{m}(b)\\
&= \underbrace{\left(\frac{\boldsymbol{\omega}_{j,\cdot1}^T}{\omega_{j,11}}\mathbf{\Lambda}_1\frac{\boldsymbol{\omega}_{j,\cdot1}}{\omega_{j,11}}\right)^{-1}\frac{\boldsymbol{\omega}_{j,\cdot1}^T}{\omega_{j,11}}\mathbf{\Lambda}_1}_{\mathbf{w}_1^T}\underbrace{\begin{pmatrix}\widehat{b}_{1j}\\ \widehat{b}_{2j}\end{pmatrix}}_{\widehat{\mathbf{b}}_{\cdot j}} = \mathbf{w}_1^T\widehat{\mathbf{b}}_{\cdot j},
\end{aligned} \qquad \text{(Equation 16)}$$

and its variance as

$$\mathrm{Var}\left(\widehat{b}_{1j}^{\mathrm{TRAM}}\right) = \left(\frac{\partial\mathbf{m}^T}{\partial b}\mathbf{\Lambda}_1\frac{\partial\mathbf{m}}{\partial b}\right)^{-1} = \left(\frac{\boldsymbol{\omega}_{j,\cdot1}^T}{\omega_{j,11}}\mathbf{\Lambda}_1\frac{\boldsymbol{\omega}_{j,\cdot1}}{\omega_{j,11}}\right)^{-1}. \qquad \text{(Equation 17)}$$

Similarly, we can obtain $\widehat{b}_{2j}^{\mathrm{TRAM}}$ and its variance $\mathrm{Var}(\widehat{b}_{2j}^{\mathrm{TRAM}})$. Equation 16 implies that the LOG-TRAM estimator incorporates three parts of information. First, the LD heterogeneity between populations is properly adjusted by the cross-population LD scores $l_{\mathcal{R},12}$ and $l_{\mathcal{B},12}$ when constructing $\mathbf{\Omega}_j$ defined in Equation 10. When the LD structure is very different, the cross-population LD scores become smaller, which down-weighs the genetic covariance. Second, by introducing the local genetic covariance $\mathbf{\Omega}_{\mathcal{R}}$, LOG-TRAM is able to utilize the local genetic architecture to improve association mapping when $\omega_{\mathcal{R},12}$ is non-zero. When $\omega_{\mathcal{R},12} = 0$, indicating independence between effects sizes of the two populations in region $\mathcal{R}$, we show that LOG-TRAM statistics would be equivalent to the original GWAS statistics (see supplemental methods, section 3.6), which avoids incorrectly borrowing information from the auxiliary

population. Importantly, LOG-TRAM can estimate both the variances ($\omega_{\mathcal{R},1}$ and $\omega_{\mathcal{R},2}$) and the covariance $\omega_{\mathcal{R},12}$ from the data with the unbiased moment estimator (see the following section on parameter estimation and statistical inference), making it adaptive to the local genetic architecture in real data analysis. Third, LOG-TRAM corrects confounding biases from the GWAS data through the term $\widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j$ and thus produces well-calibrated statistics.

Parameter estimation and statistical inference

To obtain unknown parameters $\boldsymbol{\theta} = \{\omega_{\mathcal{R},1}, \omega_{\mathcal{R},2}, \omega_{\mathcal{R},12}, \omega_{\mathcal{B},1}, \omega_{\mathcal{B},2}, \omega_{\mathcal{B},12}, c_1, c_2, c_{12}\}$ in $\boldsymbol{\Omega}_j$ and $\mathbf{C}$ of Equation 16, we consider the following regressions derived from Equation 11:

$$\begin{aligned}
\widehat{b}_{1j}^2 &\sim \omega_{\mathcal{R},1} l_{\mathcal{R},j1} + \omega_{\mathcal{B},1} l_{\mathcal{B},j1} + c_1 \widehat{s}_{1j}^2, \\
\widehat{b}_{2j}^2 &\sim \omega_{\mathcal{R},2} l_{\mathcal{R},j2} + \omega_{\mathcal{B},2} l_{\mathcal{B},j2} + c_2 \widehat{s}_{2j}^2 \\
\widehat{b}_{1j}\widehat{b}_{2j} &\sim \omega_{\mathcal{R},12} l_{\mathcal{R},j12} + \omega_{\mathcal{B},12} l_{\mathcal{B},j12} + \widehat{s}_{1j}\widehat{s}_{2j} c_{12}.
\end{aligned} \qquad \text{(Equation 18)}$$

Practically, we regress the squared Z scores, computed as $z_{1j}^2 = \widehat{b}_{1j}^2/\widehat{s}_{1j}^2$, on $l_{\mathcal{R},j1}/\widehat{s}_{1j}^2$ and $l_{\mathcal{B},j1}/\widehat{s}_{1j}^2$ of population 1, and we regress the squared Z scores $z_{2j}^2 = \widehat{b}_{2j}^2/\widehat{s}_{2j}^2$ on the single-population LD scores $l_{\mathcal{R},j2}/\widehat{s}_{2j}^2$ and $l_{\mathcal{B},j2}/\widehat{s}_{2j}^2$ of population 2. The slopes of the two regressions are the estimates of per-SNP heritabilities ($\omega_{\mathcal{R},1}$, $\omega_{\mathcal{B},1}$) and ($\omega_{\mathcal{R},2}$, $\omega_{\mathcal{B},2}$), and the intercepts are the estimates of c_1 and c_2.

Similarly, we regress the products of Z scores from two populations $z_{1j}z_{2j} = \widehat{b}_{1j}\widehat{b}_{2j}/\sqrt{\widehat{s}_{1j}\widehat{s}_{2j}}$ on $l_{\mathcal{R},j12}/\sqrt{\widehat{s}_{1j}\widehat{s}_{2j}}$ and $l_{\mathcal{B},j12}/\sqrt{\widehat{s}_{1j}\widehat{s}_{2j}}$. The slopes of this regression are the estimates of the local per-SNP co-heritability $\omega_{\mathcal{R},12}$ and the background per-SNP co-heritability $\omega_{\mathcal{B},12}$, and the intercept is the estimate of c_{12}. We followed LDSC and estimated the regression coefficients by using weighted least-squares estimator.[9,36] To reduce the influence of outliers, we used a two-step estimator similar to LDSC. In the first step, we estimated the intercepts by excluding SNPs with $\chi^2 > 30$. In the second step, we estimated the heritabilities or co-heritatilities with all SNPs by fixing the intercepts to the value estimated in the first step.

$$\widehat{\boldsymbol{\Omega}}_j = \begin{pmatrix} \widehat{\omega}_{\mathcal{R},1} l_{\mathcal{R},j1} + \widehat{\omega}_{\mathcal{B},1} l_{\mathcal{B},j1} & \widehat{\omega}_{\mathcal{R},12} l_{\mathcal{R},j12} + \widehat{\omega}_{\mathcal{B},12} l_{\mathcal{B},j12} \\ \widehat{\omega}_{\mathcal{R},12} l_{\mathcal{R},j12} + \widehat{\omega}_{\mathcal{B},12} l_{\mathcal{B},j12} & \widehat{\omega}_{\mathcal{R},2} l_{\mathcal{R},j2} + \widehat{\omega}_{\mathcal{B},2} l_{\mathcal{B},j2} \end{pmatrix},$$
$$\widehat{\mathbf{C}} = \begin{pmatrix} \widehat{c}_1 & \widehat{c}_{12} \\ \widehat{c}_{12} & \widehat{c}_2 \end{pmatrix}. \qquad \text{(Equation 19)}$$

Because these estimates are derived with the method of moments (MoM) based on Equation 4, we do not require the normality of $[\beta_{1j}, \beta_{2j}]^T$. With moment conditions in Equation 4, the MoM estimators are unbiased and robust, which allow us to produce well-calibrated statistics (see supplemental methods, section 3.7). We then obtain the LOG-TRAM estimates $\{\widehat{b}_{1j}^{\text{TRAM}}, \widehat{b}_{2j}^{\text{TRAM}}\}$ from Equation 16 and the corresponding variance $\{\text{Var}(\widehat{b}_{1j}^{\text{TRAM}}), \text{Var}(\widehat{b}_{2j}^{\text{TRAM}})\}$ from Equation 17. The standard errors are computed by $\widehat{s}_{1j}^{\text{TRAM}} = \sqrt{\text{Var}(\widehat{b}_{1j}^{\text{TRAM}})}$ and $\widehat{s}_{2j}^{\text{TRAM}} = \sqrt{\text{Var}(\widehat{b}_{2j}^{\text{TRAM}})}$. With the above LOG-TRAM outputs, we can detect non-zero SNP effect sizes ($b_{1j} \neq 0$ and $b_{2j} \neq 0$) based on the Wald test.

Identification of variants with ancestry-specific effects by LOG-TRAM

Not only can LOG-TRAM provide inference on the SNP effect sizes from the two populations, but it also offers a statistical test to identify variants with ancestry-specific effects. Specifically, we denote the difference of the SNP effects size as $\delta_j = b_{1j} - b_{2j}$, where b_{1j} and b_{2j} are the underlying true marginal effect sizes defined in Equation 7. The estimate of δ_j can be obtained from the LOG-TRAM output:

$$\widehat{\delta}_j^{\text{TRAM}} = \widehat{b}_{1j}^{\text{TRAM}} - \widehat{b}_{2j}^{\text{TRAM}}. \qquad \text{(Equation 20)}$$

The variance of $\widehat{\delta}_j^{\text{TRAM}}$ is given as

$$\text{Var}\left(\widehat{\delta}_j^{\text{TRAM}}\right) = \text{Var}\left(\widehat{b}_{1j}^{\text{TRAM}}\right) + \text{Var}\left(\widehat{b}_{2j}^{\text{TRAM}}\right) - 2\text{Cov}\left(\widehat{b}_{1j}^{\text{TRAM}}, \widehat{b}_{2j}^{\text{TRAM}}\right). \qquad \text{(Equation 21)}$$

Here, the covariance term is derived as

$$\begin{aligned}
\text{Cov}\left(\widehat{b}_{1j}^{\text{TRAM}}, \widehat{b}_{2j}^{\text{TRAM}}\right) &= \frac{\left(\frac{\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\boldsymbol{\Lambda}_1\right)\text{Var}\left(\widehat{\mathbf{b}}_{\cdot,j}|\mathbf{b}_{\cdot,j}\right)\left(\boldsymbol{\Lambda}_2\frac{\boldsymbol{\omega}_{j,\cdot 2}}{\omega_{j,22}}\right)}{\left(\frac{\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\boldsymbol{\Lambda}_1\frac{\boldsymbol{\omega}_{j,\cdot 1}}{\omega_{j,11}}\right)\left(\frac{\boldsymbol{\omega}_{j,\cdot 2}^T}{\omega_{j,22}}\boldsymbol{\Lambda}_2\frac{\boldsymbol{\omega}_{j,\cdot 2}}{\omega_{j,22}}\right)} \\
&= \frac{\frac{\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\boldsymbol{\Lambda}_1\left(\boldsymbol{\Omega}_j - \left[\boldsymbol{\omega}_{j,\cdot 1}, \boldsymbol{\omega}_{j,\cdot 2}\right]\begin{bmatrix}\omega_{j,11} & \omega_{j,12} \\ \omega_{j,12} & \omega_{j,22}\end{bmatrix}^{-1}\left[\boldsymbol{\omega}_{j,\cdot 1}, \boldsymbol{\omega}_{j,\cdot 2}\right]^T + \widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j\right)\boldsymbol{\Lambda}_2\frac{\boldsymbol{\omega}_{j,\cdot 2}}{\omega_{j,22}}}{\left(\frac{\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\boldsymbol{\Lambda}_1\frac{\boldsymbol{\omega}_{j,\cdot 1}}{\omega_{j,11}}\right)\left(\frac{\boldsymbol{\omega}_{j,\cdot 2}^T}{\omega_{j,22}}\boldsymbol{\Lambda}_2\frac{\boldsymbol{\omega}_{j,\cdot 2}}{\omega_{j,22}}\right)} \\
&= \frac{\left(\frac{\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\boldsymbol{\Lambda}_1\right)\widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j\left(\boldsymbol{\Lambda}_2\frac{\boldsymbol{\omega}_{j,\cdot 2}}{\omega_{j,22}}\right)}{\left(\frac{\boldsymbol{\omega}_{j,\cdot 1}^T}{\omega_{j,11}}\boldsymbol{\Lambda}_1\frac{\boldsymbol{\omega}_{j,\cdot 1}}{\omega_{j,11}}\right)\left(\frac{\boldsymbol{\omega}_{j,\cdot 2}^T}{\omega_{j,22}}\boldsymbol{\Lambda}_2\frac{\boldsymbol{\omega}_{j,\cdot 2}}{\omega_{j,22}}\right)},
\end{aligned} \qquad \text{(Equation 22)}$$

Plugging in the estimated parameters $\widehat{\boldsymbol{\theta}}$, we can construct the estimates of $\boldsymbol{\Omega}_j$ and $\mathbf{C}$ as

where $\boldsymbol{\Lambda}_2$ is defined as $\boldsymbol{\Lambda}_2 := \boldsymbol{\Omega}_j - \frac{\boldsymbol{\omega}_{j,\cdot 2}\boldsymbol{\omega}_{j,\cdot 2}^T}{\omega_{j,22}} + \widehat{\mathbf{S}}_j\mathbf{C}\widehat{\mathbf{S}}_j$. With Equation 20 and Equation 21, we can obtain the LOG-TRAM-based

difference test statistic $z_{\delta j}^{\mathrm{TRAM}} = \widehat{\delta}_j^{\mathrm{TRAM}} / \sqrt{\mathrm{Var}(\widehat{\delta}_j^{\mathrm{TRAM}})}$ and apply the Wald test to test hypothesis $\mathrm{H}_0 : \delta_j = 0$ v.s. $\mathrm{H}_1 : \delta_j \neq 0$.

Recall that the traditional methods simply obtain the estimate of δ_j and its variance based on the original GWAS estimate: $\widehat{\delta}_j = \widehat{b}_{1j} - \widehat{b}_{1j}$, $\mathrm{Var}(\widehat{\delta}_j) = \widehat{s}_{1j}^2 + \widehat{s}_{2j}^2$, and then derive the GWAS-based test statistic $z_{\delta j}^{\mathrm{GWAS}} = \widehat{\delta}_j / \sqrt{\mathrm{Var}(\widehat{\delta}_j)}$. As LOG-TRAM can borrow information across populations and account for confounding bias, it can substantially improve the statistical power of identifying variants with ancestry-specific effects.

Compared methods

We compared the performance of LOG-TRAM with several commonly used GWAS meta-analysis methods, including fixed-effect IVW meta-analysis (e.g., FE[37]), random-effects model (e.g., RE2[18,19]), MANTRA,[20] MTAG,[21] and MAMA.[22] We are aware that some of these methods are not specifically designed for trans-ancestry association mapping (TRAM), such as FE and RE2. But it can be very helpful to evaluate statistical methods for TRAM, and their performance can serve as a reference. Specifically, the fixed-effect (FE) method[37] is developed to perform meta-analysis of homogeneous GWASs (e.g., meta-analysis of two different cohort studies of the same phenotype), where a common effect size is assumed across studies. The random-effect (RE) models[16] relax the fixed-effect model assumption and allow the heterogeneity of effect sizes. A popular RE method in GWASs is RE2,[18] which can improve the power of standard RE models by assuming no heterogeneity under the null hypothesis. RE2C[19] is further developed to improve RE2 by allowing correlated statistics. MANTRA,[20] a method specifically designed for TRAM, assumes that the effect sizes should be similar for similar genetic ancestries and be different for distal ancestries. MANTRA uses a Bayesian partition model to assign populations into ethnic clusters, where distance between a population and the cluster center is measured by the F-statistics (F_{ST}).[38] MTAG[21] is a recently developed method to analyze multiple GWASs from a single ancestry. Although MTAG is not designed for TRAM, it is extended as a new method named MAMA[22] for TRAM, where the global genetic correlation between the same traits is the key to borrow information across ancestries. Indeed, LOG-TRAM can also be viewed as an extension of MTAG. However, the way we extend MTAG is different from that of MAMA. As motivated by recently increasing evidence[26–30] that the local genetic architecture can be quite different from the global genetic architecture, we developed LOG-TRAM for TRAM by leveraging the local genetic architecture. Through comprehensive simulation studies and real data analysis, we show that LOG-TRAM has advantages when the local genetic architecture differs from the global genetic architecture, and it also has comparable performance when the local genetic architecture is consistent with the global genetic architecture.

Results

Simulation study

We performed simulations to compare LOG-TRAM with several existing methods, including FE, RE2, MANTRA, MTAG, and MAMA. To fairly compare different methods, we organized our simulation studies into two categories. (1) The local genetic architecture differs from the global genetic architecture. In this setting, LOG-TRAM can outperform existing statistical methods. (2) The local genetic architecture is consistent with the global genetic architecture. In this setting, LOG-TRAM is comparable to the methods that explore the global genetic correlation, such as MTAG and MAMA.

The local genetic architecture differs from the global genetic architecture

We first conducted the simulations to evaluate the type I error rate. We considered 18K samples from the EAS cohort[23] and 100K samples from UKBB as the target and auxiliary populations, respectively. We used their genotype matrices to mimic the different LD patterns and allele frequencies between populations. Precisely 17,248 HapMap3-matched SNPs from chromosome 20 were used in our simulation study. We partitioned chromosome 20 into two segments at base pair position 3,119,133 (GRCh37). One segment taking up about 95% of chromosome 20 was considered as the background region $\mathcal{B}$. We generated a polygenic scenario by setting 10% of SNPs with shared non-zero effects and varied the trans-ancestry genetic correlation (denoted as r_g) among $\{0, 0.2, 0.4, 0.6\}$. The shared effects were simulated from the bivariate normal distribution $\begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} h_1^2 & r_g h_1 h_2 \\ r_g h_1 h_2 & h_2^2 \end{bmatrix} \Big/ (0.1|\mathcal{B}|)\right)$ for $k \in \mathcal{B}$, where $h_1^2 = h_2^2 = 0.01 * (95\%)$ means that the total heritability contributed by chromosome 20 is around 0.01, and $|\mathcal{B}|$ is the number of SNPs in the background region $\mathcal{B}$. Another segment taking up 5% of chromosome 20 was treated as the local region $\mathcal{R}$. We then used the SNPs in $\mathcal{R}$ to assess the type I error rate of compared methods. Specifically, we set $\beta_{1j} = 0$ for $j \in \mathcal{R}$ for EAS while simulating the true effects of EUR (β_{2j}, $j \in \mathcal{R}$) by a normal distribution $\mathcal{N}\left(0, \frac{h_{\mathcal{R},2}^2 * fold}{0.1|\mathcal{R}|}\right)$, where $h_{\mathcal{R},2}^2 = 0.01 * 5\%$, $|\mathcal{R}|$ is number of SNPs in the local region $\mathcal{R}$, and 0.1 represents that 10% of SNPs jointly contribute to heritability, $h_{\mathcal{R},2}^2 * fold$, with per-SNP heritability $\frac{h_{\mathcal{R},2}^2 * fold}{0.1|\mathcal{R}|}$. Under this setting, 10% of SNPs in region $\mathcal{R}$ have non-zero effects in population 2, while all SNPs have zero effects in population 1. The parameter *fold* allows us to vary the enrichment fold of local heritability at 1, 3, 5, and 10, such that the local genetic architecture diversely differs from the global one. Given the effect sizes and genotype matrices, quantitative phenotypes in both populations were simulated with Equation 3. After that, we marginally regressed the simulated phenotypes on each SNP to obtain the Z scores of the two datasets. Finally, after performing meta-analysis, we reported the fraction of SNPs with a p value less than 0.05 in the local region $\mathcal{R}$ of EAS as the type I error rate.

As expected, Figure 2A shows that only the single-ancestry-based GWAS and LOG-TRAM have well-controlled type I error rates regardless of the enrichment

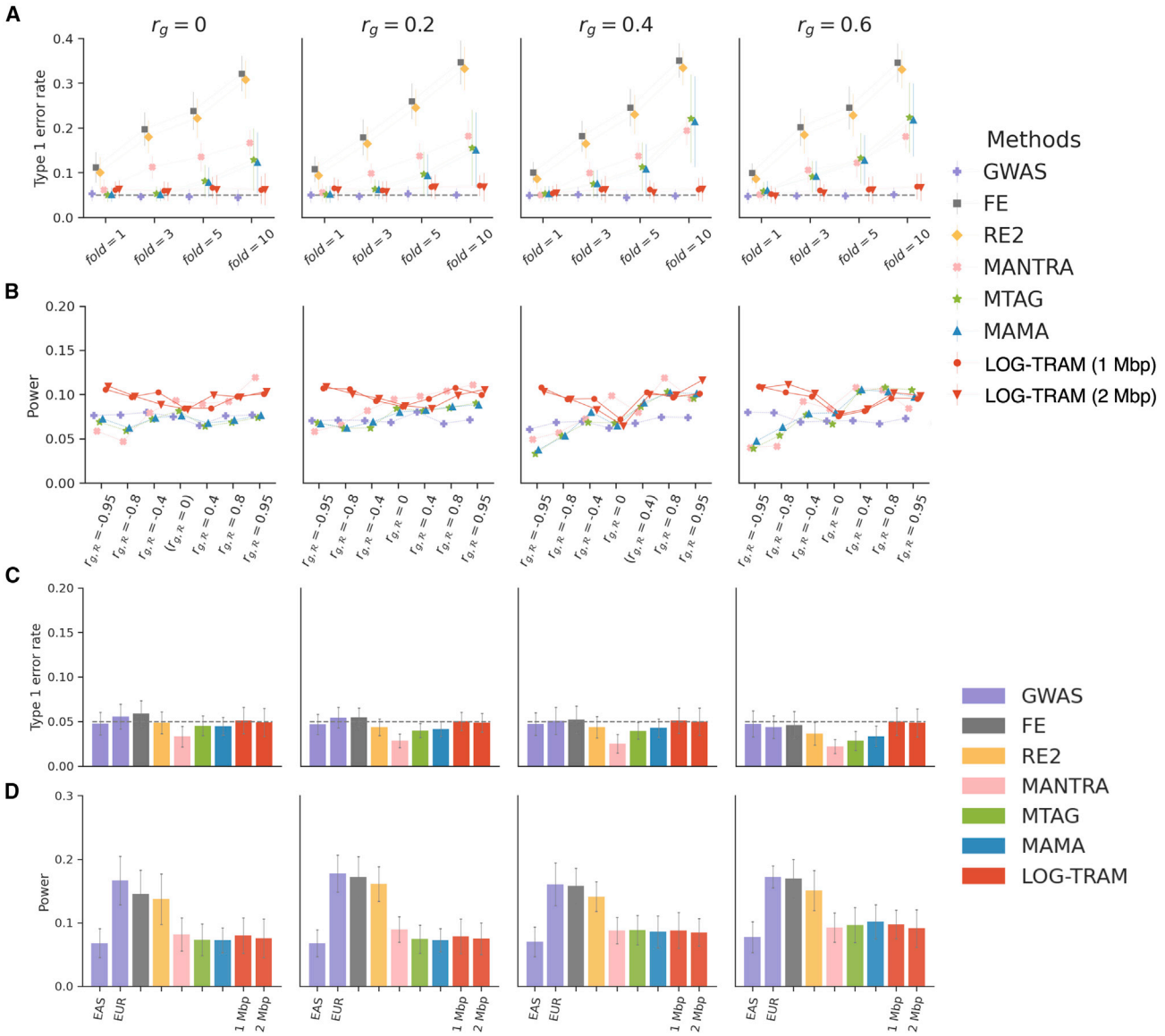

Figure 2. Comparisons among LOG-TRAM, MAMA, MTAG, MANTRA, RE2, FE, and single-ancestry GWASs in simulation studies across different genetic architectures

(A) Average type I error rates assessed under different settings of background trans-ancestry genetic correlations (r_g) and the fold enrichment (*fold*) of EUR local heritability. Error bars represent the standard errors of type I error rates evaluated on 50 replications.

(B) Average statistical power in multiple simulation scenarios with different combinations of background trans-ancestry genetic correlations (r_g) and local trans-ancestry genetic correlations ($r_{g,\mathcal{R}}$). Results are also summarized from 50 replications. More simulation results for different settings of total heritability (h^2) and proportion of causal SNPs are provided in Figures S1–S8.

(C) Average type I error rates assessed when SNPs in the same local region $\mathcal{R}$ of both populations were simulated as null SNPs.

(D) Average statistical power was evaluated in multiple simulation scenarios where background trans-ancestry genetic correlations (r_g) were the same with the local trans-ancestry genetic correlations ($r_{g,\mathcal{R}}$). Error bars represent the standard errors of power evaluated on 50 replications.

fold of local heritability for region $\mathcal{R}$ in EUR. In contrast, the type I error rates of other meta-analysis methods with the homogeneous assumption (e.g., FE, RE2, MANTRA, MTAG, MAMA) increase as the enrichment fold increases from 1 to 10. More specifically, the FE and RE2 approaches performed worst in most cases. When the background trans-ancestry genetic correlation (r_g) is non-zero, MANTRA, MTAG, and MAMA suffered from severely inflated type I error rates, as they misused or overused information from large-scale EUR GWAS summary statistics. We also examined whether the performance of LOG-TRAM could be sensitive to the size of window (or local genomic region). Specifically, we considered two window sizes, 1 M and 2 M base pairs, to partition the whole chromosome into multiple non-overlapping local regions. We observed that the type I error rates of LOG-TRAM were well controlled despite the different window sizes. Therefore, we set a window size of 1 M base pairs as the default

setting. In summary, LOG-TRAM has a satisfactory control of the type I error rate and its performance is insensitive to the size of a local region. Because FE and RE2 showed severe inflation of the type I error rate, we did not include them in the power comparison.

Next, we evaluated the power of compared methods in the cross-population setting. Different from the null SNP setting in the local region $\mathcal{R}$ of EAS for the evaluation of type I error rates, we generated the SNP effects of the two populations in local region $\mathcal{R}$ by a bivariate normal distribution $\begin{pmatrix}\beta_{1j}\\ \beta_{2j}\end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} h^2_{\mathcal{R},1} & r_{g,\mathcal{R}}h_{\mathcal{R},1}h_{\mathcal{R},2} \\ r_{g,\mathcal{R}}h_{\mathcal{R},1}h_{\mathcal{R},2} & h^2_{\mathcal{R},2}\end{bmatrix} \Big/ (0.1|\mathcal{R}|)\right)$ for $j \in \mathcal{R}$, where $h^2_{\mathcal{R},1} = h^2_{\mathcal{R},2} = 0.01 * 5\%$, $r_{g,\mathcal{R}}$ is the local trans-ancestry genetic correlation of region $\mathcal{R}$, and 0.1 represents that 10% of SNPs with non-zero effects jointly contribute to the local heritability. We considered a wide spectrum of local correlations to mimic the heterogeneous genetic similarity between populations in different gnomic regions: -0.95, -0.8, $-0.4, 0, 0.4, 0.8, 0.95$. We anticipated that LOG-TRAM would achieve substantial power gain in identifying SNPs with non-zero effects when the local genetic correlation was strong and the global genetic correlation was weak. As we described in the above, we simulated the quantitative phenotypes and obtained the Z scores of the two datasets. Finally, after performing meta-analysis, we reported the fraction of non-zero SNPs in the EAS local region $\mathcal{R}$ with a p value less than 0.05 as the power. As shown in Figure 2B, LOG-TRAM achieved the best performance in all settings. When the global/background trans-ancestry genetic correlation r_g is 0, LOG-TRAM was still able to increase the power by utilizing the information from EUR datasets through non-zero local genetic correlation $r_{g,\mathcal{R}}$. By contrast, MAMA and MTAG reduced to the standard single-ancestry GWAS because no information could be borrowed from the large-scale EUR datasets. When the global genetic correlation r_g became close to $r_{g,\mathcal{R}}$, MAMA and MTAG achieved comparable performance with LOG-TRAM.

The local genetic architecture is consistent with the global genetic architecture

Regarding the simulation analysis when local genetic architecture is consistent with the global genetic architecture, we randomly selected 10% of SNPs across the whole chromosome ($\mathcal{R} \cup \mathcal{B}$) as non-zero effect SNPs. Then we simulated their effect sizes from the bivariate normal distribution $\begin{pmatrix}\beta_{1k}\\ \beta_{2k}\end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} h^2_1 & r_g h_1 h_2 \\ r_g h_1 h_2 & h^2_2\end{bmatrix} \Big/ (0.1M)\right)$ for $k \in \mathcal{R} \cup \mathcal{B}$, where we set $h^2_1 = h^2_2 = 0.01$ meaning that the total heritability contributed by chromosome 20 is around 0.01, and r_g is the trans-ancestry genetic correlation varied at $\{0, 0.2, 0.4, 0.6\}$, $M = 17,248$ is the number of SNPs in the chromosome, and $0.1M$ means that 10% of SNPs had non-zero effects in both populations. Clearly, under this simulation setting, the local genetic architecture is consistent with the global genetic architecture.

We first used SNPs in $\mathcal{R}$ to evaluate the power of compared methods in the cross-population setting. As we described above, we reported the fraction of non-zero SNPs in the EAS local region $\mathcal{R}$ with a p value less than 0.05 as the power. In this setting, the stronger assumptions made by FE and RE2 were satisfied and thus they showed higher statistical power than other methods. As shown in Figure 2D, LOG-TRAM achieved comparable performance with meta-analysis methods with the homogeneous assumption (e.g., MTAG and MAMA).

Next, we still used SNPs in $\mathcal{R}$ to assess type I error rate of compared methods. To generate null SNPs, we set SNPs effect sizes $\beta_{1j} = \beta_{2j} = 0$ for $j \in \mathcal{R}$ in both EAS and EUR. Similarly, we reported the fraction of SNPs with a p value less than 0.05 in the local region $\mathcal{R}$ of EAS as the type I error rate. As shown in Figure 2C and Figure S9, all the compared methods have well-controlled type I error rates regardless of the trans-ancestry genetic correlation (r_g). Nevertheless, we observed that RE2, MANTRA, MTAG, and MAMA may produce slightly deflated p values when r_g was as high as 0.6, suggesting that they may be slightly conservative in this setting. In summary, LOG-TRAM has a comparable performance in power and satisfactory control of the type I error rate when the local genetic architecture is consistent with the global genetic architecture.

Under the setting where the local genetic architecture is consistent with the global genetic architecture, we further compared the performance of GWAS-based and LOG-TRAM-based difference test for identification of ancestry-specific loci in simulation studies. As shown in Figure S10, the LOG-TRAM-based difference test has significant higher power than the GWAS-based test while achieving well-controlled type I error rates. In summary, simulation studies suggest that LOG-TRAM has a great advantage over other methods by leveraging the local genetic architecture.

Real data analysis

Application of LOG-TRAM for trans-ancestry association mapping in East Asian population

To evaluate the performance of LOG-TRAM in real applications, we applied LOG-TRAM to publicly available GWAS summary statistics of 29 phenotypes from EAS and EUR. The details of datasets are given in Table S1. For each GWAS dataset, we used SNPs that overlapped the HapMap 3 list and removed the SNPs with ambiguous alleles. For a pair of phenotypes from different populations, we aligned the sign of their effect sizes to the same allele. The LD scores were estimated with a sliding window of 1 M base pair in the genome with 417 EUR and 377 EAS individuals from the 1000 Genomes Project. Figure 3A shows a summary of results for all analyzed phenotypes in the EAS population. Among all 29 traits, we observed that LOG-TRAM consistently identified more independent loci than the standard single-ancestry GWASs. In

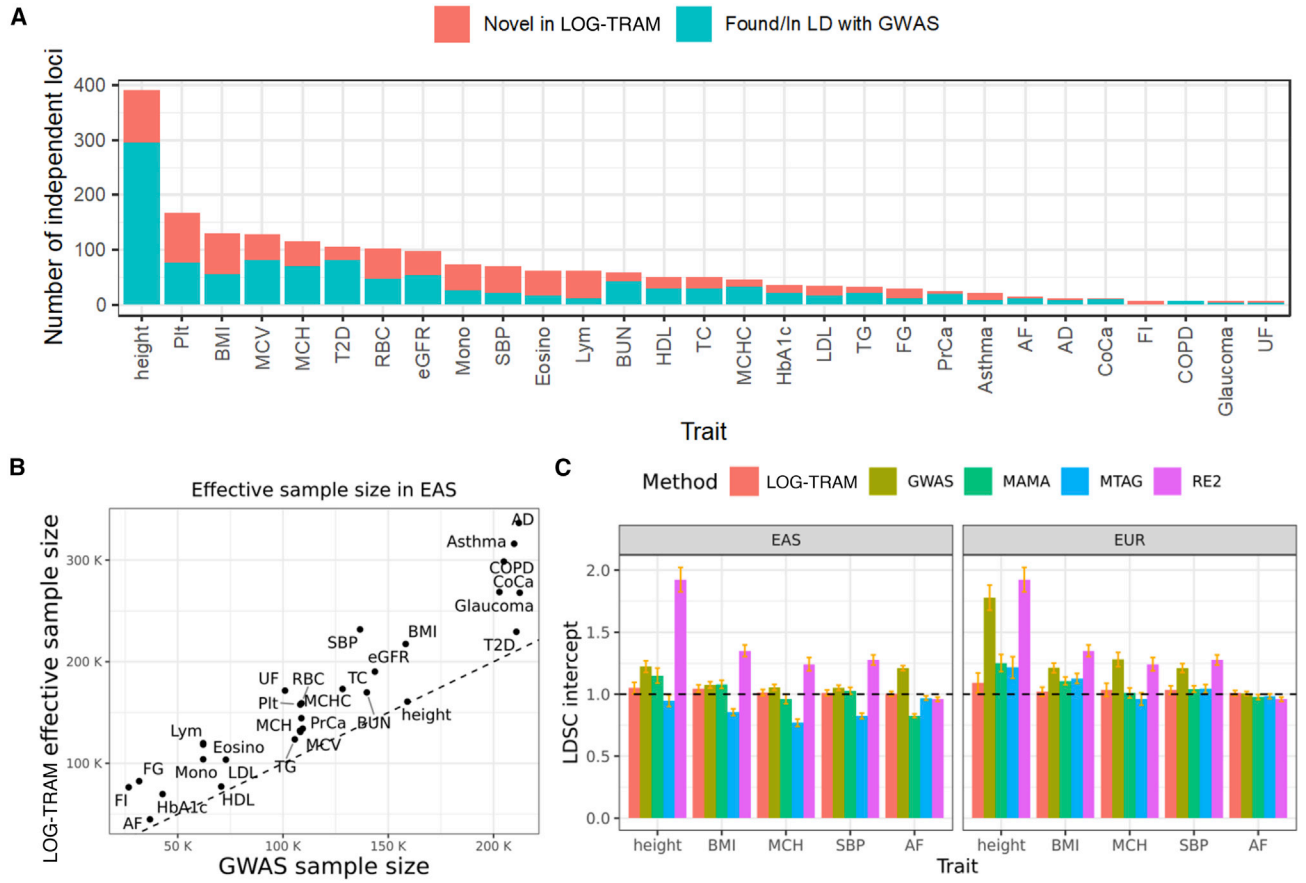

Figure 3. Summary of analysis results obtained by applying LOG-TRAM to combine the EAS and EUR GWAS summary statistics of 29 phenotypes

(A) LOG-TRAM identified independent novel lead SNPs for the 29 phenotypes in EAS compared to the original GWASs (Table S1).

(B) The effective sample size of association statistics output by LOG-TRAM was computed as $n_{\text{eff}} = M(\overline{\chi^2} - c)/(h^2\bar{l})$, where M is the number of SNPs, $\overline{\chi^2}$ is the mean χ^2 statistics of LOG-TRAM output, c is the LDSC intercept of LOG-TRAM summary statistics, h^2 is the heritability of the target trait, and $\bar{l}$ is the mean LD score.

(C) The estimated LDSC intercepts and standard errors (error bars) obtained by LOG-TRAM and alternative approaches, including the original GWAS, MAMA, MTAG, and RE2.

summary, LOG-TRAM identified 1,954 lead SNPs in EAS, among which 842 were not reported by the standard GWASs of BBJ and thus considered as novel loci. Besides, for each trait, we compared the original GWAS sample size with the LOG-TRAM effective sample size estimated by the mean χ^2 statistics. Specifically, we assessed how large of a sample size was needed to attain an equivalent increase in the mean χ^2 statistics of LOG-TRAM. Overall, we observed that the original GWAS sample size had to be increased in the range 1% (height, the small increase is mainly due to the unadjusted confounding bias of height GWASs in both populations) to 186% (fasting insulin, FI) to achieve an equivalent power gained by LOG-TRAM (Figure 3B and Table S2). We note that the significant power gain in EAS should be attributed to two indispensable key points: (1) the large sample size of EUR GWAS summary statistics (range from 89,858 to 560,658, see Table S2 for more details); (2) the non-zero local genetic covariance between the two populations (Figure 4E). As an example, compared to the 22 independent loci identified in the original GWAS of systolic blood pressure (SBP) from BBJ, LOG-TRAM achieved a substantially higher power for EAS associations by identifying $(69 - 22)/22 \approx 214\%$ more significant loci. Equivalently, LOG-TRAM increased the sample size of SBP from 136,597 in the original BBJ GWAS to 231,836, indicating that LOG-TRAM can borrow information from UKBB to perform association analysis in EAS. It is also worth noting that LOG-TRAM is computationally efficient. It only took 8 min on average to complete the analysis of the whole genome. The timing was evaluated by the Linux computing server with 20 CPU cores of Intel(R) Xeon(R) Gold 6230N CPU at 2.30 GHz processor, 1 TB of memory, and a 22 TB solid-state disk.

Next, we quantified the magnitude of confounding bias in the resulting meta-analysis association statistics by using the LDSC intercept.[9] Under the widely accepted LDSC assumption, the LDSC intercept should be one in the absence of confounding bias. As shown in Figure 3C, the LDSC intercept of the standard EAS height GWAS

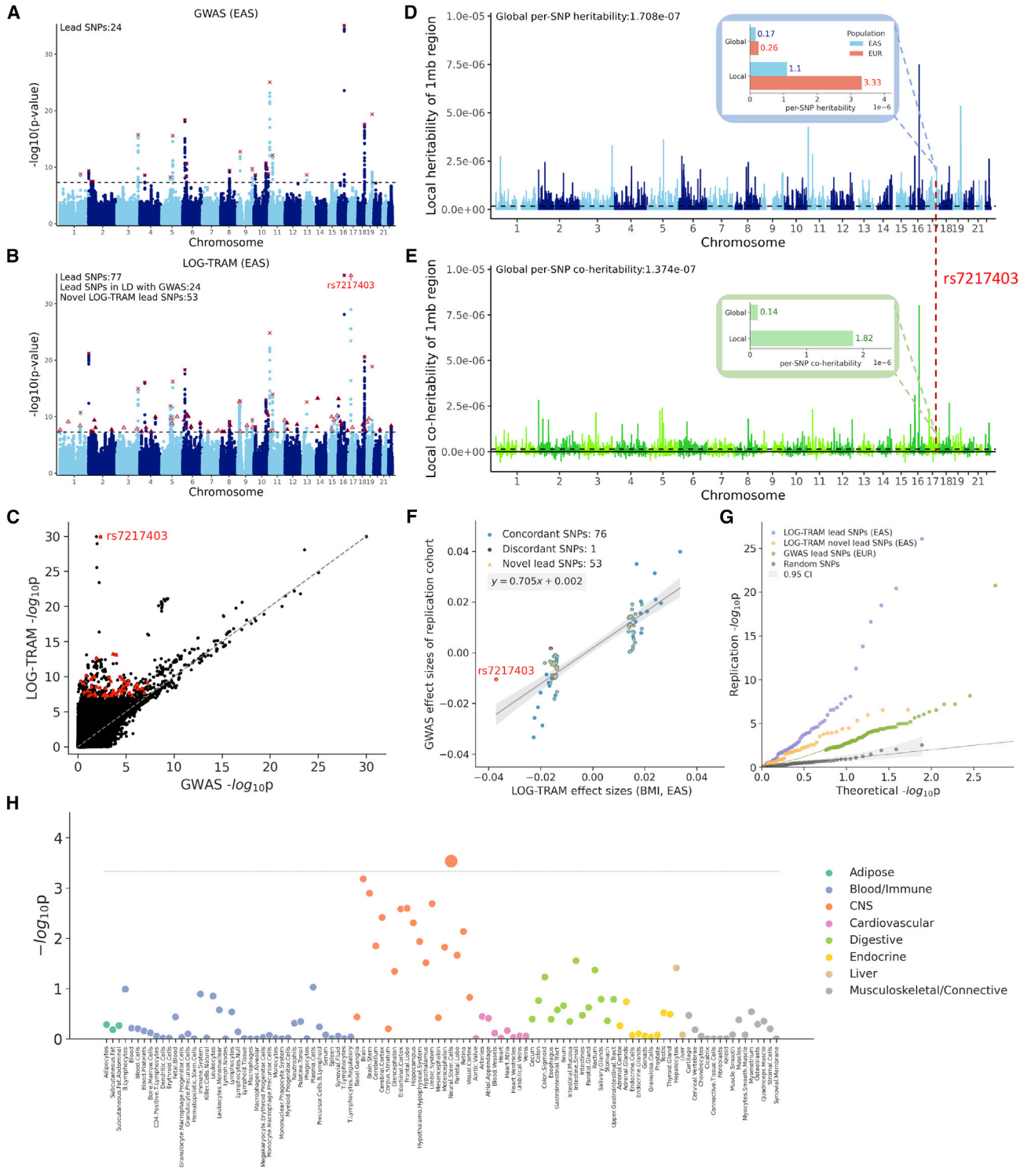

Figure 4. Trans-ancestry association mapping of BMI

(A) Manhattan plot of the BMI GWAS for the BBJ males, where 24 independent lead SNPs were identified.

(B) Manhattan plot of the LOG-TRAM results for the BBJ male. LOG-TRAM identified 77 independent lead SNPs, of which 53 were not identified in the original BBJ male GWAS. Lead SNPs identified by both LOG-TRAM and GWAS are marked by a cross. Novel lead SNPs are marked by a triangle.

(C) Comparison of the p value output by LOG-TRAM and original BMI GWAS p values of the BBJ male.

(D) Local per-SNP heritability estimated by LOG-TRAM on the basis of BBJ (male) and UKBB.

(E) Local per-SNP co-heritability between BBJ male and UKBB.

(F) Comparison of lead SNPs effect sizes output by LOG-TRAM and effect sizes of GWAS in the replication dataset (the BBJ female).

(legend continued on next page)

was estimated to be 1.22 (SE = 0.02), suggesting that the standard GWAS suffered from confounding bias. Through the special design of correction for confounding bias, the LDSC intercept of the LOG-TRAM result was reduced to 1.05 (SE = 0.02). By contrast, the LDSC intercept of the MTAG results was smaller than one, suggesting that MTAG over-corrected the confounding bias. The LDSC intercept of the MAMA results was 1.15 (SE = 0.03), indicating that MAMA was unable to fully correct the confounding biases. RE2 performed the worst with the LDSC intercept increased to 1.92 (SE = 0.05). Regarding the association result of EUR (right panel of Figure 3C), it has been well-known that EUR height GWAS from UKBB heavily suffers from uncorrected population structure with the LDSC intercept as high as 1.78 (SE = 0.05). After correction by LOG-TRAM, the LDSC intercept was decreased to 1.09 (SE = 0.04). By comparison, the LDSC intercepts of MTAG, MAMA, and RE2 were estimated to be 1.22 (SE = 0.04), 1.25 (SE = 0.03), and 1.92 (SE = 0.05), respectively, indicating that the results produced by these methods still suffered from confounding biases. Similarly for other traits, such as body mass index (BMI), mean corpuscular hemoglobin (MCH), SBP, and atrial fibrillation (AF), the LDSC intercepts of the LOG-TRAM results were all close to one. The above evidence clearly indicates that LOG-TRAM can effectively account for the confounding bias (see Figures S11, S12, and S13 and Tables S3 and S4 for more examples).

As a concrete example of TRAM, we applied LOG-TRAM to the GWASs of BMI, where the GWAS summary statistics of BBJ (male) and UKBB were used as the inputs. Then we obtained the LOG-TRAM results in EAS and EUR. We used the GWAS of BBJ females as a replication dataset to validate the LOG-TRAM output in EAS. Compared with the original GWAS of BBJ male (Figure 4A), LOG-TRAM identified 53 novel lead SNPs (Figure 4B). We first evaluated the credibility of signals discovered by LOG-TRAM. We compared the effect sizes and p values of the EAS lead SNPs identified by LOG-TRAM with those obtained from the independent replication GWAS (BBJ females). A higher consistency between the effect sizes of discovery and replication cohort would suggest higher credibility of the results. Similarly, a smaller p value in the replication GWAS indicates that the lead SNP is unlikely to be significant by chance. Specifically, we regressed the effect sizes of the lead SNPs obtained from LOG-TRAM output to those observed effect sizes in the replication GWAS. The regression slope can be viewed as a quantitative indicator of the replicability. Ideally, the slope should be close to 1.0 but it can deviate from 1.0 as a result of sex difference in BMI.[39,40] We observed a regression slope of 0.705 (SE = 0.04) for the LOG-TRAM results (Figure 4F). As a comparison, the regression slopes were much smaller for both MTAG (0.615, SE = 0.03, Figure S14C) and MAMA (0.630, SE = 0.03, Figure S14B). The significance of lead SNPs in the replication GWAS can also serve as an indicator of replicability. As shown in Figures 4G and S14D, the lead SNPs identified by LOG-TRAM showed more significant p values compared to those SNPs identified from the standard GWAS in European ancestry or the other meta-analysis methods. These results suggest a better replicability of LOG-TRAM.

Next, we focused on the interpretation of novel lead SNPs identified by LOG-TRAM. Among the 53 novel lead SNPs identified by LOG-TRAM, we observed the most significant novel lead SNP (Figure 4C), rs7217403. The local region harboring this variant has 1.1/0.17 = 6.47-fold and 3.33/0.26 = 12.8-fold enrichment of local heritability in EAS and EUR, respectively (Figure 4E). Furthermore, we also found that the local co-heritability of this region was higher than the global co-heritability (1.82 compared to 0.14, Figure 4F), suggesting that LOG-TRAM can effectively leverage the local genetic architecture to boost the power of association mapping. In replication analysis (Figure 4F), the highlighted SNP, rs7217403, showed a notable consistent effect with LOG-TRAM results. Although rs7217403 is located in an intergenic region, it maps near *MAP2K3*, which has been reported to be significantly associated with BMI across diverse populations, including American Indians,[41] Europeans,[42,43] and East Asians.[44] The expression level of *MAP2K3* was positively correlated with BMI in adipose tissue, and *in vitro* studies suggested that *MAP2K3* was activated during adipogenesis.[41] Given these statistical and biological evidences, *MAP2K3* appears to be a reproducible obesity locus, but knowledge for the molecular mechanisms underlying the association is still lacking. The intergenic variant, rs7217403, identified by LOG-TRAM in EAS may shed light on the genetic etiology of BMI and uncover biologically meaningful variation.

To cross-validate our LOG-TRAM result and reduce the influence of sex difference in our replication analysis, we used UKBB and the BBJ females as the discovery cohort and then replicated the LOG-TRAM output using the BBJ males. We observed that the LOG-TRAM results were highly consistent (Figures S15, S16, and S17). Compared with the original BMI GWAS of BBJ females, LOG-TRAM identified 63 novel lead SNPs (Figure S15B). Similarly, we observed the same most significant novel lead SNP (Figure S15C), rs7217403. The local region harboring this variant showed a consistent local genetic architecture with 1.81/0.19 = 9.53-fold enrichment for local

(G) The QQ plot compares the p values in the replication GWAS data: (1) LOG-TRAM lead SNPs, (2) LOG-TRAM novel lead SNPs, (3) lead SNPs identified from UKBB, and (4) randomly selected SNPs from the replication GWAS. Clearly, SNPs reported by LOG-TRAM are strongly supported in the replication study.
(H) Tissue/cell-type SEG annotation analysis for LOG-TRAM results of BMI in EAS. The p in the label of the y axis represents p values of one-sided test of stratified LDSC regression coefficients. Each circle represents a tissue or cell type from the Franke lab dataset; large circles pass the Bonferroni correction: p value $\leq 0.05/108$. Dashed line represents the significance threshold: $-\ log_{10}(0.05/108) \approx 3.33$.

heritability in EAS and 2.33/0.16 = 14.5-fold enrichment for local co-heritability (Figure S15E).

Beside the most significant novel lead SNP (rs7217403), other novel signals and their nearest genes discovered by LOG-TRAM also show potential biological functions related to BMI. As shown in Figure S18C, the local region harboring the novel lead SNP, rs987237, has 0.62/0.17 = 3.65-fold and 3.66/0.26 = 14.1-fold enrichment of local heritability in EAS and EUR, respectively. Meanwhile, the local co-heritability of this region was significantly higher than the global co-heritability (1.49 compared to 0.14). The local heritability/co-heritability estimations are consistent in the cross-validation analysis (Figure S18D). The annotated gene, *TFAP2B*, functions as both a transcriptional activator and repressor and regulates downstream genes involved in important biological functions, including face, body wall, and limb development.[45,46] Besides, studies have identified substantial associations between *TFAP2B* and BMI,[47–49] suggesting that *TFAP2B* plays a critical role in controlling the body formation. The other locus, rs879620, also shows strong (Figure S19C) and robust (Figure S19D) local heritability and co-heritability enrichments in EAS and EUR. SNP rs879620 is located in the UTR3 region of *ADCY9*, which is responsible for coding a membrane-bound enzyme that catalyzes the formation of ubiquitous second messenger cyclic adenosine monophosphate (cAMP) from adenosine triphosphate (ATP). Mutations in *ADCY9* have been reported to be associated with cardiovascular disease,[50] dyslipidemia,[51] and obesity in Europeans[8,52] and East Asians.[49,53] The connection between the molecular mechanism of these associated SNPs and the etiology of complex diseases needs to be further investigated.

Finally, we applied stratified LD score regression[54] to explore the tissues that are functionally relevant with BMI in EAS based on the BBJ GWAS summary statistics and the LOG-TRAM output. We used publicly available EAS samples of 1000 Genomes as LD reference and 108 Franke tissues/cell-type specifically expressed genes (SEGs) as annotations to construct LD scores and evaluated tissue/cell specificity. The Franke lab annotation[55] is the largest publicly available tissue/cell-type SEG dataset, comprising of 37,427 human samples. As a comprehensive SEG annotation dataset, it covers a wide spectrum of human tissues/cell types, including adipose, blood, immune, central nervous system (CNS), etc. Integrative analysis of the Franke lab annotations and GWAS summary statistics may offer novel biological insights to elucidate the etiologies of human complex traits/diseases. As shown in Figure S20, none of the tissues/cell types in the annotations were observed to be significantly associated with BMI when using the original BBJ GWAS. In contrast, a significant enrichment was detected in neural stem cells after applying LOG-TRAM (Figure 4H). Besides, consistent with previous tissue-specific enrichment analysis of BMI in European population,[54,56,57] LOG-TRAM results show robust and strong signals across a wide range of brain tissues/cell types, suggesting these brain tissues/cell types may also play a more important functional role for BMI in EAS.

Application of LOG-TRAM for trans-ancestry association mapping in African population

Besides the integrative analysis of GWASs from EAS and EUR, we further show that LOG-TRAM can be applied to integrate the GWASs from other populations. As a demonstration, we applied LOG-TRAM to improve association mapping in the African population (AFR). Specifically, we analyzed 17 phenotypes by combining AFR GWAS with EUR/EAS GWAS, including 11 hematological phenotypes (e.g., lymphocyte count [Lym] and eosinophil count [Eosino]),[58] three glycemic phenotypes (e.g., fasting glucose [FG] and fasting insulin [FI]),[59] and three lipid phenotypes (high-density lipoprotein [HDL], low-density lipoprotein [LDL], and total cholesterol [TC]).[60] The details of datasets are given in Table S5. For each GWAS dataset, we used SNPs that overlapped with the HapMap 3 list and removed the SNPs with ambiguous alleles. For a pair of phenotypes from different populations, we aligned the sign of their effect sizes to the same allele. We estimated the LD scores of the AFR population with 505 African samples from the 1000 Genomes Project and used the same definition of 1 M base pair widows as local regions. Similarly, we computed the effective sample sizes of the 17 phenotypes in AFR after applying LOG-TRAM to combine the GWAS from another population. As expected, the estimated effective sample sizes of LOG-TRAM outputs were larger than the sample sizes of original AFR GWASs when integrating with either EUR (Figure 5A) or EAS (Figure 5B). In particular, when the EUR GWASs were combined, LOG-TRAM identified more independent loci in all 17 phenotypes (Figure 5C). We also show the QQ plots of Lym (Figure 5D, left), Eosino (Figure 5D, middle), and FG (Figure 5D, right) as concrete examples. Compared to the integrating AFR with EAS, LOG-TRAM gained more power when combining AFR with EUR because of the larger sample size in EUR.

Identification of ancestry-specific loci

Taking the GWAS summary statistics from BBJ and UKBB as input, we applied the LOG-TRAM-based difference test to a number of complex traits, including mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), HDL, LDL, type 2 diabetes (T2D), and BMI. The test statistics $z_{\delta j}^{\mathrm{GWAS}}$ and $z_{\delta j}^{\mathrm{TRAM}}$ for overlapping SNPs across two populations were computed with Equation 20 and Equation 21. Here, we take MCHC as an example. In addition to the 23 ancestry-specific lead SNPs (p value $< 5 \times 10^{-8}$) identified by the standard GWAS-based difference test, the LOG-TRAM-based difference test identified six novel lead SNPs with significantly different effect sizes between EAS and EUR (Figures 6A and 6B). Although those novel SNPs (red dot in Figure 6C) show different effect sizes between EAS and EUR, they were not identified by the

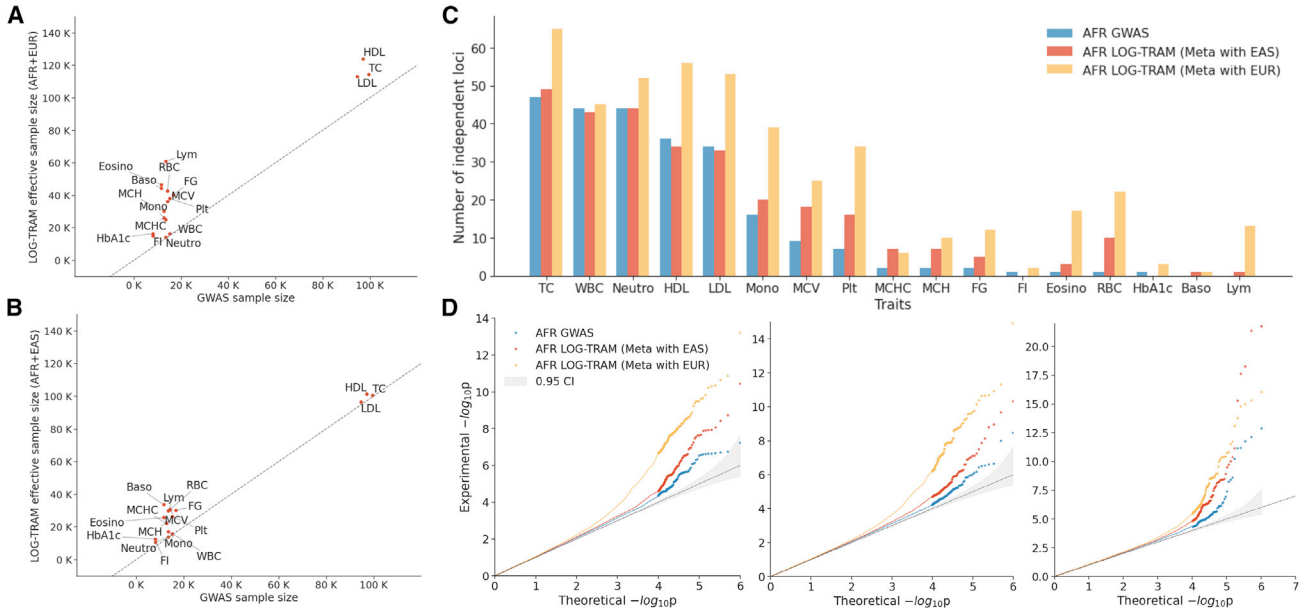

Figure 5. Summary of analysis results obtained by applying LOG-TRAM to combine the AFR and EUR/EAS GWAS summary statistics of 17 overlapped phenotypes

(A and B) The effective sample size of association statistics output by LOG-TRAM when combining AFR with EUR (A) or EAS (B) GWAS summary statistics of 17 phenotypes (Table S5).

(C) Comparison of the number of independent lead SNPs identified by LOG-TRAM and original GWASs for the 17 phenotypes in AFR.

(D) QQ plots of single-ancestry GWAS and LOG-TRAM results for lymphocyte count (left), eosinophil count (middle), and fasting glucose (right).

GWAS-based difference test because of the limited sample size. In contrast, the LOG-TRAM-based difference test successfully captured those signals, as LOG-TRAM can improve the power of association statistics with a larger effective sample size. From a biological perspective, these SNPs identified by the LOG-TRAM-based difference test were enriched in functional categories related to population differences, suggesting the functional importance of ancestry-specific SNPs. Notably, as shown in Figure 6D, SNPs in the top quantile of background selection statistic[61] have more significant p values, while SNPs in the top quantile of recombination rate[62] have less significant p values than the average. This phenomenon is consistent with the observation in Shi et al.,[30] where the background selection statistic is positively correlated to the depletion of trans-ancestry genetic correlation while recombination rate has a reverse pattern. The ancestry-specific loci identification results of five other traits are provided in Figures S21–S25.

The LOG-TRAM result for construction of polygenic risk scores

To demonstrate the utility of LOG-TRAM in the construction of polygenic risk scores (PRSs) in under-represented populations, we compared the predictive power of PRS constructed by using the standard GWAS summary statistics and the LOG-TRAM summary statistics. We considered construction of PRS for human height as an example. We first applied the most commonly used PRS method, LDpred,[63] to build an EAS height PRS model (denoted as GWAS-based PRS) by using the GWAS summary statistics of height from BBJ (n = 159,095). Using the GWAS summary statistics of height from BBJ and UKBB, we applied LOG-TRAM to generate association statistics for EAS population. Then we used the summary statistics (output of LOG-TRAM) to construct PRS, which is referred to as LOG-TRAM-based PRS. After that, we evaluated the prediction performance in an independent Chinese testing dataset.[23] In detail, we measured the prediction accuracy of each PRS model by using the predictive R^2, defined as the square of correlation of predicted PRSs and true residual phenotypes after regressing out covariates (e.g., gender, age, genomics PCs).[23,64]

As shown in Figure 6E, the predictive accuracy obtained by GWAS-based PRS is $R^2 = 0.144$. In contrast, PRS constructed from the LOG-TRAM association statistics achieved a (0.169–0.144)/0.144) ≈ 17% accuracy gain, indicating that the LOG-TRAM association statistics of EAS height successfully borrowed information from the large-scale UKBB datasets. The improvement of prediction accuracy was quite stable when different PRS approaches were applied. When we applied a recently developed PRS method named "dbslmm,"[65] we obtained predictive $R^2 = 0.149$ with the GWAS summary statistics. With the same PRS method, we achieved predictive $R^2 = 0.185$ with the LOG-TRAM association statistics, which was a (0.185–0.149)/0.149 ≈ 24% improvement. In summary, the above results suggest that the LOG-TRAM output can lead to a significant improvement in construction of accurate PRS for under-represented populations.

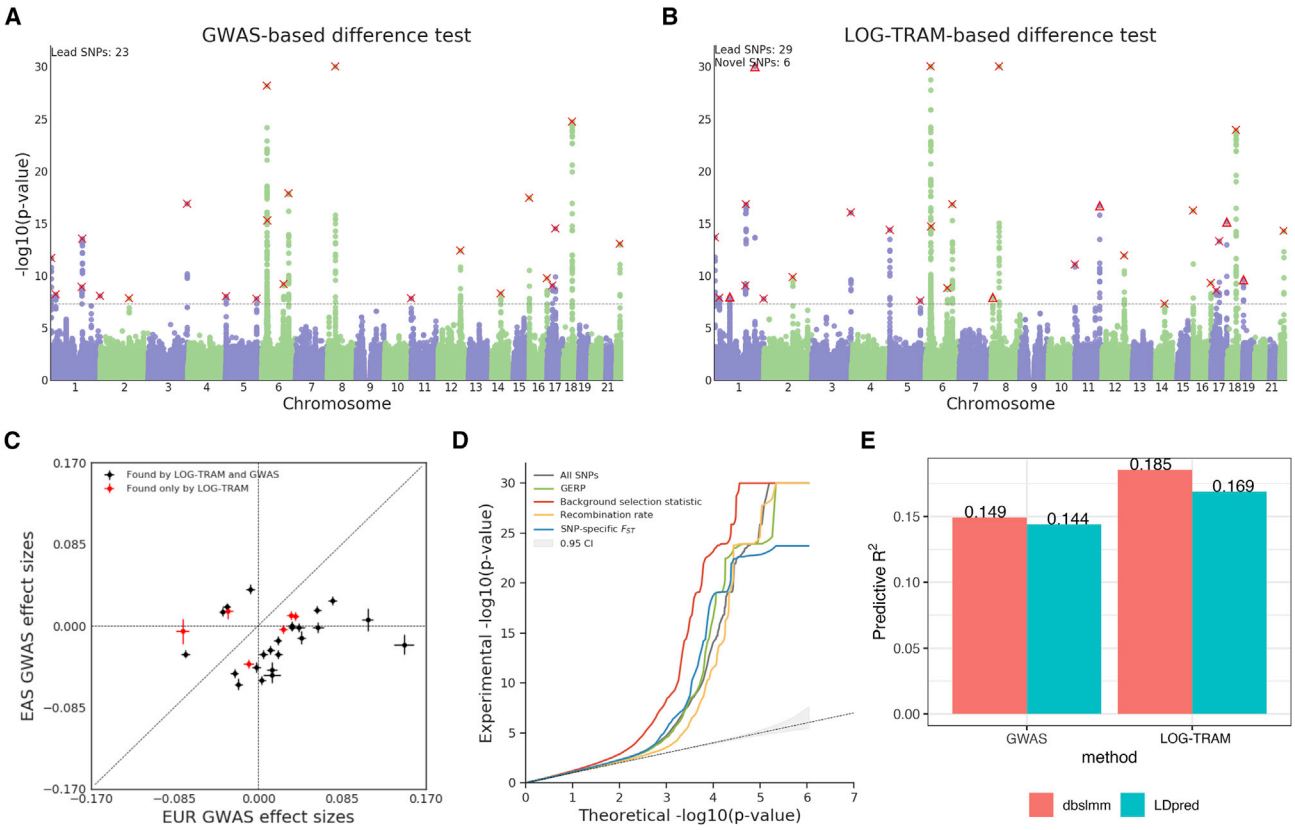

Figure 6. Applications of LOG-TRAM in ancestry-specific loci identification and PRS construction
(A) Manhattan plot of GWAS-based difference test results for MCHC. Lead SNPs are marked by a cross.
(B) Manhattan plot of LOG-TRAM-based difference test results for MCHC. The dashed line marks the threshold for genome-wide significance ($p = 5 \times 10^{-8}$). Lead SNP identified by both the LOG-TRAM and GWAS-based difference tests are marked by crosses. Novel lead SNPs are marked by triangles.
(C) Comparison of MCHC GWAS effect sizes for lead SNPs between EAS and EUR. Red points denote novel loci identified by the LOG-TRAM-based difference test; black points denote lead loci identified in both the LOG-TRAM and GWAS-based difference tests. Vertical and horizontal error bars represent standard errors of GWAS effect sizes in EAS and EUR, respectively.
(D) The QQ-plots of p values for all SNPs and SNPs in the top quantile of four continuous-valued annotations.
(E) Predictive R^2 for EAS height PRS models based on GWAS and LOG-TRAM association statistics.

Discussion

In this paper, we have introduced a trans-ancestry meta-analysis method, LOG-TRAM, aiming to improve the statistical power of GWASs in non-Europeans by leveraging locally shared genetic architectures with biobank-scale auxiliary datasets. Through comprehensive simulations, we showed that our method has greater statistical power while controlling the type I error rate compared to existing approaches, and our method is robust across various genetic architecture settings. We applied LOG-TRAM to GWASs of 29 complex traits and diseases from EAS and EUR, achieving substantial gains in power and effective correction of confounding biases. We found that LOG-TRAM results were reproducible in independent studies. We showed the ability of LOG-TRAM for integrating different populations by applying it to combine AFR GWAS with EAS/EUR GWASs. Finally, we demonstrated that the LOG-TRAM results can be further used for identification of ancestry-specific loci and construction of PRS in under-represented populations.

As we mentioned, accumulating evidence reveal that the shared genetic basis between phenotypes/populations varies across genomic regions.[26,27,29,30] For example, the effect sizes of SNPs can be phenotype-/population-specific, and the genetic architecture of multiple ancestries can be locally different. Therefore, assuming a constant variance-covariance matrix of effect sizes across the whole genome violates biological intuition in many circumstances. Stratified LDSC[56] is a representative method to explore the local genetic architecture. It assumes that SNPs in different genomic regions (e.g., functional categories) contribute disproportionately to the heritability and estimates the per-SNP heritability in each region by regressing the association statistics to the LD score corresponding to each region. Recently, several other statistical methods, including ρ-HESS,[27] SUPERGNOVA,[28] and LOGODetect,[31] have been developed to estimate local genetic correlation across traits in a single population. Their analysis results consistently suggest that the local genetic correlation can greatly differ from the global genetic correlation, offering new insights into complex human diseases and traits. However,

the methods that explore the local genetic architecture have not accounted for heterogeneity across ancestries. A direct application of these methods for association mapping in the trans-ancestry setting is still problematic. To develop the LOG-TRAM method, we considered a sliding window (e.g., 1 M base pair segment) to decompose the whole genome into local region $\mathcal{R}$ and background region $\mathcal{B}$, which allows us to model the regional genetic architectures differently from the global pattern. By successfully leveraging the local genetic architecture and accounting for confounding bias, not only can LOG-TRAM estimate local heritability but it also can be applied for association mapping in the trans-ancestry setting.

We have systematically compared the performance of LOG-TRAM with several commonly used methods for meta-analysis of GWAS data, including FE, RE2, MANTRA, MTAG, and MAMA, and demonstrated the benefit of leveraging local genetic architecture in TRAM. Besides these approaches, we are also aware of Tractor,[66] which is an individual-level method for TRAM in admixed populations. It first infers the local ancestry composition of admixed individuals in their genomes and then conducts association mapping. While both Tractor and LOG-TRAM aim to boost the power of association mapping in the multi-ancestry context and estimate ancestry-specific effects, they are different in the following aspects. First, they are designed to address different difficulties. Tractor is designed to account for the local ancestry of admixed individuals when conducting association mapping in admixed populations (e.g., African Americans and Latino/Hispanics). LOG-TRAM aims to improve the power of association mapping in an under-represented population by combining well-powered GWASs from European populations. Second, they take different levels of GWAS data as input. Tractor takes individual-level data with admixed individuals as input, while LOG-TRAM uses summary-level GWAS data from different populations as input. Third, they utilize the local genetic information in different ways. Tractor infers the local ancestry of admixed individuals while LOG-TRAM estimates the local genetic covariance across different populations. Therefore, Tractor and LOG-TRAM are designed for different purposes.

Due to the requirement of characterizing the local genetic architectures, LOG-TRAM may have its own limitations. First, the value of the local covariance matrix can depend on the definition of a "local" region. By expectation, the region defined by a smaller segment in the genome can capture the genetic architecture with higher resolution. However, it would be harder to accurately estimate the local variance and covariance of effect sizes with a small number of SNPs in a smaller region. In practice, it requires a trade-off between fine resolution and accurate parameter estimation. In genomic data analysis, 1 M base pair regions are often considered as local regions, e.g., Loh et al.[26] Therefore, we used a sliding window of 1 M base pair regions in our main analysis and demonstrated the robustness of this choice by comparing the results with those obtained with 2 M base pair regions. Second, the estimation errors of the local genetic covariance matrix can be larger than those of global ones. We conducted simulations to confirm that the variance of the local genetic covariance matrix becomes quite large when the sample size of discovery GWAS is less than 10,000, which leads to a reduction of power in association mapping (see supplemental methods, section 3.8). Therefore, for stable estimation of parameters, it is recommended to apply LOG-TRAM to GWASs with at least 10,000 samples. Nowadays, this sample size requirement is often satisfied in real data analysis (see Table S2 for more examples).

Although we have mainly focused on the local regions partitioned by the sliding window approach with a fixed window size in this study, it is worth mentioning that LOG-TRAM can also be applied to more generalized "local regions" defined by functional annotations or tissue/cell type SEG annotations. Indeed, SNPs with biological functions (e.g., gene regulatory elements,[67–70] epigenomic regulations,[56,71,72] and tissue-specific expressed genes[30,73]) have been well known for their enrichment in the heritability of complex traits, which re-emphasizes the widespread of heterogeneous genetic architectures across the genome. Consequently, these functional annotations are widely used in genetic studies to increase statistical power, including GWASs,[74–76] pleiotropy,[77] fine-mapping,[78,79] and polygenic risk scores.[80,81] Very recently, a study[82] suggests that functional annotations have great potential in improving the portability of trans-ancestry polygenic risk scores, indicating the substantial share of biological mechanisms across populations. Therefore, leveraging the functional/SEG annotations can also increase the power of trans-ancestry association mapping for under-represented populations. To see this, we considered SNPs annotated by 20 binary functional annotations from Baseline-LD-X model[30] as local regions and estimated their local genetic architectures for LOG-TRAM. Intuitively, each annotation can be analogized to one 1 M base pair segment defined in the previous section. We applied the functional-informed LOG-TRAM to the GWASs of BMI and T2D from BBJ and UKBB. Compared to the original EAS GWASs, we identified 85% and 16% more independently significant loci for BMI and T2D (Figures S28 and S31), respectively. In addition, we further used the 53 SEG annotations[30] as local regions and applied the SEG-informed LOG-TRAM to the same traits. As shown in Figures S28 and S31, SEG-informed LOG-TRAM achieved comparable power gains in EAS and identified 114% and 25% more independently significant loci for BMI and T2D, respectively.

Our LOG-TRAM approach needs more investigation in the following directions. First, it has been shown that pleiotropic effects are widespread in the genome.[83] In a systematic analysis of 4,155 publicly available GWASs, 90% trait-associated loci affect multiple phenotypes simultaneously.[2] Hence, joint modeling of multiple GWAS traits across populations may further boost the statistical power of trans-ancestry association mapping. Second,

LOG-TRAM assumes that, for a given population, SNPs with lower allele frequencies tend to have larger effect sizes. More specifically, we considered the relationship of per-allele effect size v and allele frequency f as $\mathrm{Var}(v) \propto [f(1-f)]^{\alpha}$, where $\alpha = -1$. This assumption has been shown to be a stable choice in simulation studies,[84] and it was also adopted in the previous trans-ancestry analysis.[85] Some recent studies on the estimation of α found that although estimated α are negative for most complex traits, they vary moderately across traits.[86,87] Therefore, it would be more appropriate to obtain an estimate of α for each trait rather than fixing $\alpha = -1$. Third, while LOG-TRAM can effectively integrate GWASs from different populations, its input GWAS data usually precludes admixed individuals from analysis. The exclusion of admixed individuals can reduce false positives at a cost of power in association mapping. Considering the increasingly mixed populations and enrichment of heritable diseases in admixed populations,[66] it would be interesting to integrate admixed populations (e.g., African Americans) with large-scale GWASs conducted with a more homogeneous ancestry background (e.g., UKBB) to improve power. The nature of LOG-TRAM in characterizing local genetic architectures could offer a convenient formulation to take the local ancestries in admixed populations into account. As inspired by Tractor, we may first infer the ancestries of local regions in the genome for the individuals from an admixed GWAS data. With the inferred local ancestry, the local cross-population genetic correlation between the target ancestry and the ancestry of an auxiliary GWAS data can be estimated and leveraged to improve the power of association mapping. We will explore these potential improvements in the near future.

Data and code availability

The publicly available GWAS summary statistics for meta-analysis were obtained from the links summarized in Tables S1 and S5. The UK Biobank data are from UK Biobank resource under application number 30186. The LOG-TRAM software and source codes in this study were publicly available in the GitHub repository of LOG-TRAM (https://github.com/YangLabHKUST/LOG-TRAM).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.05.013.

Acknowledgments

This work is supported in part by National Key R&D Program of China (2020YFA0713900), Hong Kong Research Grant Council [12301417, 16307818, 16301419, 16308120, 16307221], Hong Kong Innovation and Technology Fund [PRP/029/19FX], Hong Kong University of Science and Technology [startup grant R9405, Z0428 from the Big Data Institute], the Open Research Fund from Shenzhen Research Institute of Big Data [2019ORF01004], and Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen. The computational task for this work was partially performed with the X-GPU cluster supported by the RGC Collaborative Research Fund: C6021-19EF.

Declaration of interests

The authors declare no competing interests.

Received: January 30, 2022
Accepted: May 23, 2022
Published: June 16, 2022

Web resources

BBJ, http://jenger.riken.jp/en
BOLT-LMM, https://alkesgroup.broadinstitute.org/BOLT-LMM
Dbslmm, https://biostat0903.github.io/DBSLMM
LDpred, https://github.com/bvilhjal/ldpred
LDSC, https://github.com/bulik/ldsc
LOG-TRAM, https://github.com/YangLabHKUST/LOG-TRAM
MAMA, https://github.com/JonJala/mama
MTAG, https://github.com/JonJala/mtag
PLINK, https://www.cog-genomics.org/plink
S-LDXR, https://github.com/huwenboshi/s-ldxr
UKBB, https://www.ukbiobank.ac.uk
XPASS, https://github.com/YangLabHKUST/XPASS

References

1. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385–389. https://doi.org/10.1126/science.1109557.
2. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348. https://doi.org/10.1038/s41588-019-0481-0.
3. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484. https://doi.org/10.1038/s41576-019-0127-1.
4. Mills, M.C., and Rahal, C. (2020). The GWAS diversity monitor tracks diversity by disease in real time. Nat. Genet. *52*, 242–243. https://doi.org/10.1038/s41588-020-0580-y.
5. Gurdasani, D., Barroso, I., Zeggini, E., and Sandhu, M.S. (2019). Genomics of disease risk in globally diverse populations. Nat. Rev. Genet. *20*, 520–535. https://doi.org/10.1038/s41576-019-0144-0.
6. Hindorff, L.A., Bonham, V.L., Brody, L.C., Ginoza, M.E.C., Hutter, C.M., Manolio, T.A., and Green, E.D. (2018). Prioritizing diversity in human genomics research. Nat. Rev. Genet. *19*, 175–185. https://doi.org/10.1038/nrg.2017.89.
7. Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

8. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518. https://doi.org/10.1038/s41586-019-1310-4.
9. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295. https://doi.org/10.1038/ng.3211.
10. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.
11. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106. https://doi.org/10.1038/ng.2876.
12. Abdellaoui, A., Hugh-Jones, D., Yengo, L., Kemper, K.E., Nivard, M.G., Veul, L., Holtz, Y., Zietsch, B.P., Frayling, T.M., Wray, N.R., et al. (2019). Genetic correlates of social stratification in Great Britain. Nat. Human Behav. *3*, 1332–1342. https://doi.org/10.1038/s41562-019-0757-5.
13. Haworth, S., Mitchell, R., Corbin, L., Wade, K.H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani, G., Paternoster, L., Smith, G.D., et al. (2019). Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat. Commun. *10*, 333–339. https://doi.org/10.1038/s41467-018-08219-1.
14. Hu, X., Jia, Z., Lin, Z., Wang, Y., Peng, H., Zhao, H., Wan, X., and Yang, C. (2021). Mendelian randomization for causal inference accounting for pleiotropy and sample structure using genome-wide summary statistics. Preprint at bioRxiv. https://doi.org/10.1101/2021.03.11.434915.
15. Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. *6*. 91–14. https://doi.org/10.1186/s13073-014-0091-5.
16. DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. Contr. Clin. Trials *7*, 177–188. https://doi.org/10.1016/0197-2456(86)90046-2.
17. Evangelou, E., and Ioannidis, J.P.A. (2013). Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. *14*, 379–389. https://doi.org/10.1038/nrg3472.
18. Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am. J. Hum. Genet. *88*, 586–598. https://doi.org/10.1016/j.ajhg.2011.04.014.
19. Lee, C.H., Eskin, E., and Han, B. (2017). Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. Bioinformatics *33*, i379–i388. https://doi.org/10.1093/bioinformatics/btx242.
20. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822. https://doi.org/10.1002/gepi.20630.
21. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., James, J., Lee, M.A.F., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. *50*, 229–237. https://doi.org/10.1038/s41588-017-0009-4.
22. Turley, P., Martin, A.R., Goldman, G., Li, H., Kanai, M., Walters, R.K., Jala, J.B., Lin, K., Millwood, I.Y., Carey, C.E., et al. (2021). Multi-ancestry meta-analysis yields novel genetic discoveries and ancestry-specific associations. Preprint at bioRxiv. https://doi.org/10.1101/2021.04.23.441003.
23. Cai, M., Xiao, J., Zhang, S., Wan, X., Zhao, H., Chen, G., and Yang, C. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am. J. Hum. Genet. *108*, 632–655. https://doi.org/10.1016/j.ajhg.2021.03.002.
24. Luo, Y., Li, X., Wang, X., Gazal, S., Mercader, J.M., Neale, B.M., Florez, J.C., Auton, A., Price, A.L., Finucane, H.K., et al. (2021). Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. Hum. Mol. Genet. *30*, 1521–1534. https://doi.org/10.1093/hmg/ddab130.
25. Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. *10*, 681–690. https://doi.org/10.1038/nrg2615.
26. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392. https://doi.org/10.1038/ng.3431.
27. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. Am. J. Hum. Genet. *101*, 737–751. https://doi.org/10.1016/j.ajhg.2017.09.022.
28. Zhang, Y., Lu, Q., Ye, Y., Huang, K., Liu, W., Wu, Y., Zhong, X., Li, B., Yu, Z., Travers, B.G., et al. (2021). SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome Biol. *22*, 262–330. https://doi.org/10.1186/s13059-021-02478-w.
29. Werme, J., van der Sluis, S., Posthuma, D., and de Leeuw, C.A. (2022). An integrated framework for local genetic correlation analysis. Nat. Genet. *54*, 274–282. https://doi.org/10.1038/s41588-022-01017-y.
30. Shi, H., Gazal, S., Kanai, M., Koch, E.M., Schoech, A.P., Siewert, K.M., Kim, S.S., Luo, Y., Amariuta, T., Huang, H., et al. (2021). Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nat. Commun. *12*, 1098–1115. https://doi.org/10.1038/s41467-021-21286-1.
31. Guo, H., Li, J.J., Lu, Q., and Hou, L. (2021). Detecting local genetic correlations with scan statistics. Nat. Commun. *12*, 2033–2113. https://doi.org/10.1038/s41467-021-22334-6.
32. van Rheenen, W., Peyrot, W.J., Schork, A.J., Lee, S.H., and Wray, N.R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. Nat. Rev. Genet. *20*, 567–581. https://doi.org/10.1038/s41576-019-0137-z.
33. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., Patterson, N., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290. https://doi.org/10.1038/ng.3190.
34. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). Gcta: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.
35. Wen, X., and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype

data. Ann. Appl. Stat. *4*, 1158. https://doi.org/10.1214/10-aoas338.
36. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241. https://doi.org/10.1038/ng.3406.
37. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). Metal: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.
38. Wright, S. (1965). The interpretation of population structure by f-statistics with special regard to systems of mating. Evolution *19*, 395. https://doi.org/10.2307/2406450.
39. Xi, B., Shen, Y., Reilly, K.H., Zhao, X., Cheng, H., Hou, D., Wang, X., and Mi, J. (2013). Sex-dependent associations of genetic variants identified by GWAS with indices of adiposity and obesity risk in a Chinese children population. Clin. Endocrinol *79*, 523–528. https://doi.org/10.1111/cen.12091.
40. Link, J.C., and Reue, K. (2017). Genetic basis for sex differences in obesity and lipid metabolism. Annu. Rev. Nutr. *37*, 225–245. https://doi.org/10.1146/annurev-nutr-071816-064827.
41. Bian, L., Traurig, M., Hanson, R.L., Marinelarena, A., Kobes, S., Kobes, S., Muller, Y.L., Muller, Y.L., Malhotra, A., Huang, K., et al. (2013). Map2k3 is associated with body mass index in American Indians and Caucasians and may mediate hypothalamic inflammation. Hum. Mol. Genet. *22*, 4438–4449. https://doi.org/10.1093/hmg/ddt291.
42. Winkler, T.W., Justice, A.E., Graff, M., Barata, L., Feitosa, M.F., Chu, S., Czajkowski, J., Esko, T., Fall, T., Kilpeläinen, T.O., et al. (2015). The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. PLoS Genet. *11*, e1005378. https://doi.org/10.1371/journal.pgen.1005378.
43. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206. https://doi.org/10.1038/nature14177.
44. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467. https://doi.org/10.1038/ng.3951.
45. Satoda, M., Zhao, F., Diaz, G.A., Burn, J., Goodship, J., Davidson, H.R., Pierpont, M.E.M., and Gelb, B.D. (2000). Mutations in TFAP2B cause Char syndrome, a familial form of patent ductus arteriosus. Nat. Genet. *25*, 42–46. https://doi.org/10.1038/75578.
46. Zhao, F., Weismann, C.G., Satoda, M., Pierpont, M.E.M., Sweeney, E., Thompson, E.M., and Gelb, B.D. (2001). Novel TFAP2B mutations that cause Char syndrome provide a genotype-phenotype correlation. Am. J. Hum. Genet. *69*, 695–703. https://doi.org/10.1086/323410.
47. Gong, J., Nishimura, K.K., Fernandez-Rhodes, L., Haessler, J., Bien, S., Graff, M., Lim, U., Lu, Y., Gross, M., Fornage, M., et al. (2018). Trans-ethnic analysis of metabochip data identifies two new loci associated with BMI. Int. J. Obes. *42*, 384–390. https://doi.org/10.1038/ijo.2017.304.
48. Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K.E., Xue, A., Zhang, M., Joseph E Powell, J.E., Goddard, M.E., Wray, N.R., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. Sci. Adv. *5*, eaaw3538. https://doi.org/10.1126/sciadv.aaw3538.
49. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. *53*, 1415–1424. https://doi.org/10.1038/s41588-021-00931-x.
50. Rautureau, Y., Deschambault, V., Higgins, M.-È., Rivas, D., Mecteau, M., Geoffroy, P., Miquel, G., Uy, K., Sanchez, R., Lavoie, V., et al. (2018). ADCY9 (adenylate cyclase type 9) inactivation protects from atherosclerosis only in the absence of CETP (cholesteryl ester transfer protein). Circulation *138*, 1677–1692. https://doi.org/10.1161/circulationaha.117.031134.
51. Tardif, J.-C., Rhéaume, E., Perreault, L.-.P.L., Grégoire, J.C., Feroz Zada, Y., Asselin, G., Provost, S., Barhdadi, A., Rhainds, D., L'Allier, P.L., Ibrahim, R., et al. (2015). Pharmacogenomic determinants of the cardiovascular effects of dalcetrapib. Circ. Cardiovasc. Genet. *8*, 372–382. https://doi.org/10.1161/circgenetics.114.000663.
52. Zhu, Z., Guo, Y., Shi, H., Liu, C.-L., Panganiban, R.A., Chung, W., O'Connor, L.J., Himes, B.E., Gazal, S., Hasegawa, K., et al. (2020). Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. J. Allergy Clin. Immunol. *145*, 537–549. https://doi.org/10.1016/j.jaci.2019.09.035.
53. Wen, W., Zheng, W., Okada, Y., Takeuchi, F., Tabara, Y., Hwang, J.-Y., Dorajoo, R., Li, H., Tsai, F.-J., Yang, X., et al. (2014). Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. Hum. Mol. Genet. *23*, 5492–5504. https://doi.org/10.1093/hmg/ddu248.
54. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629. https://doi.org/10.1038/s41588-018-0081-4.
55. Fehrmann, R.S.N., Karjalainen, J.M., Krajewska, M., Harm-Jan, W., Maloney, D., Simeonov, A., Pers, T.H., Hirschhorn, J.N., Jansen, R.C., Schultes, E.A., et al. (2015). Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat. Genet. *47*, 115–125. https://doi.org/10.1038/ng.3173.
56. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235. https://doi.org/10.1038/ng.3404.
57. Lu, Q., Powles, R.L., Wang, Q., He, B.J., and Zhao, H. (2016). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. PLoS Genet. *12*, e1005947. https://doi.org/10.1371/journal.pgen.1005947.
58. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and ancestry-specific

blood-cell genetics in 746,667 individuals from 5 global populations. Cell *182*, 1198–1213. https://doi.org/10.1016/j.cell.2020.06.045.

59. Chen, J., Spracklen, C.N., Marenne, G., Varshney, A., Corbin, L.J., Luan, J., Willems, S.M., Wu, Y., Zhang, X., Horikoshi, M., et al. (2021). The trans-ancestral genomic architecture of glycemic traits. Nat. Genet. *53*, 840–860. https://doi.org/10.1038/s41588-021-00852-9.
60. Graham, S.E., Clarke, S.L., Wu, K.H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 675–679. https://doi.org/10.1038/s41586-021-04064-3.
61. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427. https://doi.org/10.1038/ng.3954.
62. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321–324. https://doi.org/10.1126/science.1117196.
63. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001.
64. Xiao, J., Cai, M., Hu, X., Wan, X., Chen, G., and Yang, C. (2022). Xpxp: improving polygenic prediction by cross-population and cross-phenotype analysis. Bioinformatics *38*, 1947–1955. https://doi.org/10.1093/bioinformatics/btac029.
65. Yang, S., and Zhou, X. (2020). Accurate and scalable construction of polygenic scores in large biobank data sets. Am. J. Hum. Genet. *106*, 679–693. https://doi.org/10.1016/j.ajhg.2020.03.013.
66. Atkinson, E.G., Maihofer, A.X., Kanai, M., Martin, A.R., Karczewski, K.J., Santoro, M.L., Ulirsch, J.C., Kamatani, Y., Okada, Y., Finucane, H.K., et al. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. Nat. Genet. *53*, 195–204. https://doi.org/10.1038/s41588-020-00766-y.
67. Joseph, K.P. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
68. Yang, C., Wan, X., Lin, X., Chen, M., Zhou, X., and Liu, J. (2019). Comm: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. Bioinformatics *35*, 1644–1652. https://doi.org/10.1093/bioinformatics/bty865.
69. Cai, M., Chen, L.S., Liu, J., and Yang, C. (2020). Igrex for quantifying the impact of genetically regulated expression on phenotypes. NAR Genom. Bioinformatics *2*, lqaa010. https://doi.org/10.1093/nargab/lqaa010.
70. Shi, X., Chai, X., Yang, Y., Cheng, Q., Jiao, Y., Chen, H., Huang, J., Yang, C., and Liu, J. (2020). A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. Nucleic Acids Res. *48*, e109. https://doi.org/10.1093/nar/gkaa767.
71. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.
72. Lu, Q., Li, B., Ou, D., Erlendsdottir, M., Powles, R.L., Jiang, T., Hu, Y., Chang, D., Jin, C., Dai, W., et al. (2017). A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. Am. J. Hum. Genet. *101*, 939–964. https://doi.org/10.1016/j.ajhg.2017.11.001.
73. Zhu, X., and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat. Commun. *9*, 4361–4414. https://doi.org/10.1038/s41467-018-06805-x.
74. Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). Gpa: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet. *10*, e1004787. https://doi.org/10.1371/journal.pgen.1004787.
75. Ming, J., Dai, M., Cai, M., Wan, X., Liu, J., and Yang, C. (2018). Lsmm: a statistical approach to integrating functional annotations with genome-wide association studies. Bioinformatics *34*, 2788–2796. https://doi.org/10.1093/bioinformatics/bty187.
76. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. Am. J. Hum. Genet. *104*, 65–75. https://doi.org/10.1016/j.ajhg.2018.11.008.
77. Ming, J., Wang, T., and Yang, C. (2020). Lpm: a latent probit model to characterize the relationship among complex traits using summary statistics from multiple GWASs and functional annotations. Bioinformatics *36*, 2506–2514. https://doi.org/10.1093/bioinformatics/btz947.
78. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722. https://doi.org/10.1371/journal.pgen.1004722.
79. Li, Y., and Kellis, M. (2016). Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Res. *44*, e144. https://doi.org/10.1093/nar/gkw627.
80. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol. *13*, e1005589. https://doi.org/10.1371/journal.pcbi.1005589.
81. Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S.S., Furlotte, N., Adam, A., and Price, A.L. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andme data sets. Nat. Commun. *12*, 1–11.
82. Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., Terao, C., and Raychaudhuri, S. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354. https://doi.org/10.1038/s41588-020-00740-8.
83. Yang, C., Li, C., Wang, Q., Chung, D., and Zhao, H. (2015). Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. Front. Genet. *6*, 229. https://doi.org/10.3389/fgene.2015.00229.
84. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide

SNPs. Am. J. Hum. Genet. *91*, 1011–1021. https://doi.org/10.1016/j.ajhg.2012.10.010.

85. Yang, L., Neale, B.M., Liu, L., Lee, S.H., Wray, N.R., Ji, N., Li, H., Qian, Q., Wang, D., Li, J., et al. (2013). Polygenic transmission and complex neuro developmental network for attention deficit hyperactivity disorder: genome-wide association study of both common and rare variants. Am. J. Med. Genet. Part B: Neuropsychiatric Genetics *162*, 419–430. https://doi.org/10.1002/ajmg.b.32169.
86. Zeng, J., De Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746–753. https://doi.org/10.1038/s41588-018-0101-4.
87. Speed, D., Holmes, J., and Balding, D.J. (2020). Evaluating and improving heritability models using summary statistics. Nat. Genet. *52*, 458–462. https://doi.org/10.1038/s41588-020-0600-y.