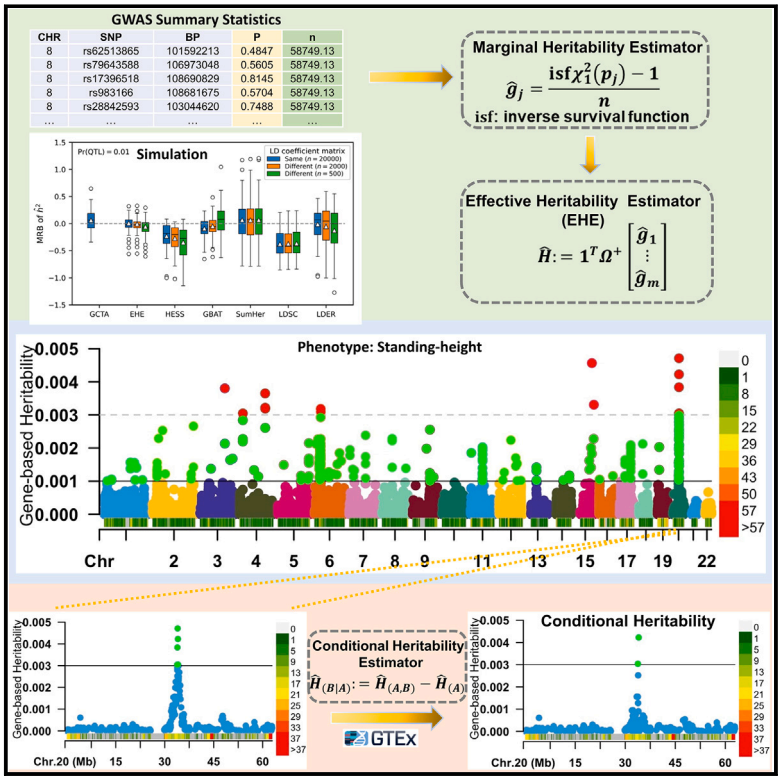
Dissecting the high-resolution genetic architecture of complex phenotypes by accurately estimating gene-based conditional heritability

Graphical abstract



Authors

Lin Miao, Lin Jiang, Bin Tang, Pak Chung Sham, Miaoxin Li

Correspondence

limiaoxin@mail.sysu.edu.cn

Miao et al. present EHE, a flexible method for estimating heritability at small genomic regions using p values of SNPs. EHE demonstrates higher accuracy and precision than alternative methods. Additionally, EHE allows for isolating non-redundant conditional heritability of nearby genes, providing valuable insights into the genetic architecture of complex phenotypes.

Miao et al., 2023, The American Journal of Human Genetics *110*, 1534–1548
September 7, 2023

https://doi.org/10.1016/j.ajhg.2023.08.006

CellPress

ARTICLE

Dissecting the high-resolution genetic architecture of complex phenotypes by accurately estimating gene-based conditional heritability

Lin Miao,[1,8] Lin Jiang,[1,8] Bin Tang,[2,3] Pak Chung Sham,[4,5] and Miaoxin Li[2,3,5,6,7,*]

Summary

Despite extensive research on global heritability estimation for complex traits, few methods accurately dissect local heritability. A precise local heritability estimate is crucial for high-resolution mapping in genetics. Here, we report the effective heritability estimator (EHE) that can use p values from genome-wide association studies (GWASs) for local heritability estimation by directly converting marginal heritability estimates of SNPs to a non-redundant heritability estimate of a gene or a small genomic region. EHE provides higher accuracy and precision for local heritability estimation among seven compared methods. Importantly, EHE can be applied to estimate the conditional heritability of nearby genes, where redundant heritability among the genes can also be removed further. The conditional estimation can be guided by tissue-specific expression profiles (or other functional scores) to prioritize and quantify more functionally important genes of complex phenotypes. Applying EHE to 42 complex phenotypes from the UK Biobank, we revealed the existence of two types of distinct genetic architectures for various complex phenotypes and found that highly pleiotropic genes are not enriched for more heritability compared to other candidate susceptibility genes. EHE provides an accurate and robust way to dissect the genetic architecture of complex phenotypes.

Introduction

Heritability gauges the proportion of genetic contribution to a phenotype.[1] Accurate quantification of heritability is fundamentally important in genetics and guides the strategies for genetic mapping, disease risk prediction, and breeding. Two types of heritability measurements are defined according to effect models: narrow-sense heritability for additive effects only and broad-sense heritability for additive effects, dominance effects, and epistatic effects.[2] Traditionally, heritability was estimated using related individuals, such as twins and family members.[3] However, such methods and strategies cannot accurately reveal the genetic contribution of specific genomic regions to understand the precise genetic architecture of phenotypes.

The advent of high-throughput genotyping technology provides unprecedented opportunities for estimating heritability using large samples of unrelated individuals[4] originally generated for genome-wide association studies (GWASs). The underlying idea is that heritability can be measured as the proportion of phenotype variance explained by many single-nucleotide polymorphisms (SNPs). Given genome-wide genotype and phenotype data of a large sample size, Yang et al. proposed using a linear mixed model to estimate heritability by all SNPs in a chromosome or a whole genome.[4,5] Then, the linear mixed model was extended, adding two random effects (genetic and environmental effects) to reduce potential estimation inflation.[6] A Bayesian mixed model was also proposed to speed up the heritability estimation in large cohort genomes.[7] However, the application of the methods is limited by the problem that individual-level data are not often accessible in a large-scale GWAS. So the SNP-based heritability estimation was extended to use GWAS summary statistics, which flexibly expanded its application to ultra-large GWASs. The linkage disequilibrium (LD) score regression (LDSC) is an early and widely used method that leverages LD information between SNPs to estimate whole-genome heritability under the polygenic model.[8] To fully use the LD information, high-definition likelihood (HDL)[9] and LD eigenvalue regression (LDER)[10] were applied to improve the estimation accuracy of SNP heritability recently.

In recent years, there has been growing interest in estimating local heritability in genomic regions using genetic association summary statistics to better understand subtle genetic contributions to complex phenotypes. The LDSC method has been extended to a stratified version, enabling partitioning of heritability in specific regions.[11] However, such methods may not be suitable for small regions, such as individual genes or exons, as they require many SNPs to estimate a meaningful standard error (SE).

[1]Research Center of Medical Sciences, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China; [2]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China; [3]Key Laboratory of Tropical Disease Control (Sun Yat-sen University), Ministry of Education, Guangzhou 510080, China; [4]State Key Laboratory of Brain and Cognitive Sciences, the University of Hong Kong, Hong Kong SAR 999077, China; [5]Department of Psychiatry, the University of Hong Kong, Hong Kong SAR 999077, China; [6]Center for Precision Medicine, Sun Yat-sen University, Guangzhou 510080, China; [7]Guangdong Provincial Key Laboratory of Biomedical Imaging and Guangdong Provincial Engineering Research Center of Molecular Imaging, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai 519000, China
[8]These authors contributed equally
*Correspondence: limiaoxin@mail.sysu.edu.cn
https://doi.org/10.1016/j.ajhg.2023.08.006.

To address this issue, Shi et al. proposed the Heritability Estimator from Summary Statistics (HESS), which is based on a quadratic form of SNP marginal effect estimates defined by the LD matrix.[12] However, this method has been found to have a large bias when the sample size is less than 10,000, as reported by Song et al.[10] Another challenge with estimating local heritability is the redundancy that occurs in physically close regions due to LD. Although methods have been developed for conditional gene-based association analyses,[13] no method is currently available to accurately remove redundant heritability from nearby genes for all we know. Therefore, the development of accurate methods is still needed to explore heritability in functional regions better.

This paper presents a heritability estimator for calculating conditional heritability at the regional level with GWAS summary statistics. The method removes the redundancy in marginal heritability of individual SNPs within an LD block by a linear combination of marginal heritability estimates with coefficients determined by the matrix of squared LD coefficients and is referred to as the effective heritability estimator (EHE) in this paper. By using ancestrally matched reference genotypes, such as those from the 1000 Genomes Project, EHE requires only GWAS p values and corresponding sample sizes to estimate heritability, making it highly scalable for large-scale GWASs. We evaluated its accuracy and precision through extensive simulations and compared it to six state-of-the-art or popular methods (GCTA,[5] HESS,[12] LDAK-GBAT,[14] LDAK-SumHer,[15] LDSC,[8] and LDER[10]). Furthermore, we applied EHE to explore the genetic contribution of protein-coding genes to 42 complex phenotypes.

Material and methods

The proposed method for local heritability estimation

Consider a linear regression model that incorporates random effects of m SNPs on a phenotype:

$$\mathcal{Y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \varepsilon, \varepsilon \sim N\left(0, \sigma_\varepsilon^2\right) \quad \text{(Equation 1)}$$

where $\mathcal{Y}$ is a standardized random variable denoting the phenotype, $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \ldots, \mathcal{X}_m)$ is a component-wise standardized random vector denoting the genotypes, $\boldsymbol{\beta}$ is a vector of the random effects, and ε is the residual. Suppose $\mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\beta}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. The heritability of the phenotype accounting for the m SNPs can be expressed as (refer to supplemental methods for the derivation)

$$h_g^2 := \frac{\mathrm{Var}(\boldsymbol{\mathcal{X}}\boldsymbol{\beta})}{\mathrm{Var}(\mathcal{Y})} = \sum_{j=1}^{m} \sigma_j^2 \quad \text{(Equation 2)}$$

Then, consider a sample of size n. Let z_j be the Z score of the marginal effect of the j$^{\text{th}}$ SNP and let $\boldsymbol{\nu} = \left(z_1^2, \ldots, z_m^2\right)^{\mathrm{T}}$. We propose the effective heritability estimator (EHE) for h_g^2 as

$$\widehat{H} := n^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{+}(\boldsymbol{\nu} - \mathbf{1}) \quad \text{(Equation 3)}$$

where $\boldsymbol{\Omega} = (\rho_{jk}^2)$ is the matrix of squared LD coefficients and $\boldsymbol{\Omega}^{+}$ is the Moore-Penrose inverse of $\boldsymbol{\Omega}$. We proved that $\widehat{H}$ is an unbiased estimator for h_g^2 before a realization of $\boldsymbol{\beta}$ (supplemental methods).

After m SNPs have been specified, their effects are realized and considered as unobservable fixed values. We define the marginal heritability for the j$^{\text{th}}$ SNP as $g_j := \mathrm{Var}(\mathcal{X}_j b_j)/\mathrm{Var}(\mathcal{Y}) = b_j^2$, where b_j is the marginal regression coefficient, and define the effective heritability as $H := \mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{+}(g_1, \ldots, g_m)^{\mathrm{T}}$, which is a linear combination of marginal heritability with $\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{+}$ as coefficients designed to eliminate the redundancy caused by LD. This is why it is called "effective." Then, EHE can be expressed as $\widehat{H} = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{+}(\widehat{g}_1, \ldots, \widehat{g}_m)^{\mathrm{T}}$, where $\widehat{g}_j = \left(z_j^2 - 1\right)/n$ is an estimator for g_j. We proved that a sufficient condition for $\mathrm{E}(\widehat{H}) = h_g^2$ after a realization of $\boldsymbol{\beta}$ is that all causal SNPs are uncorrelated (supplemental methods).

The variance of EHE after a realization of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\mathrm{Var}(\widehat{H}) &= \mathrm{Var}\left(n^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{+}[\boldsymbol{\nu} - \mathbf{1}]\right) = n^{-2}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{+}\mathrm{Var}(\boldsymbol{\nu})\boldsymbol{\Omega}^{+}\mathbf{1} \\ &= n^{-2}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{+}\mathbf{D}\mathbf{C}\mathbf{D}\boldsymbol{\Omega}^{+}\mathbf{1}\end{aligned} \quad \text{(Equation 4)}$$

where $\mathbf{C} = \mathrm{Corr}(\mathbf{v})$ and $\mathbf{D} = \mathrm{diag}\left(\mathrm{SD}\left(z_1^2\right), \ldots, \mathrm{SD}\left(z_m^2\right)\right)$. We estimate $\mathrm{SD}(z_j^2)$ by $\widehat{\mathrm{SD}}(z_j^2) = \sqrt{4z_j^2 - 2}$. For the estimation of $\mathbf{C}$, according to our previous study,[16] when either SNP has an effect in a large sample, there is $\mathrm{Corr}(z_j^2, z_k^2) \approx |\rho_{jk}|$ and it is $\mathrm{Corr}(z_j^2, z_k^2) \approx \rho_{jk}^2$ when neither SNP has an effect in the LD block.[13] Here, we employed our previous approach,[13,16] "effective chi-squared statistics" (ECS), to test whether variants in an LD block are associated with a phenotype. When the association p value (by ECS) of an LD block is smaller than 0.05, we set the correction as $|\rho_{jk}|$ and set as ρ_{jk}^2 otherwise (refer to supplemental methods for details).

Estimate heritability for dichotomous phenotypes

For a dichotomous phenotype of prevalence K in the population and a proportion P of affected individuals in the GWAS sample, under the liability threshold model, according to Lee et al.,[17] we can convert the EHE calculated by Equation 3 to the liability scale by

$$\widehat{H}_l = t\widehat{H}, t = \frac{(K[1 - K])^2}{(\varphi[\Phi^{-1}(K)])^2 P(1 - P)} \quad \text{(Equation 5)}$$

where $\widehat{H}_l$ denotes EHE of liability scale and $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and the cumulative distribution function of the standard normal distribution, respectively. Then, the variance of $\widehat{H}_l$ is given by

$$\mathrm{Var}(\widehat{H}_l) = t^2\mathrm{Var}(\widehat{H}) \quad \text{(Equation 6)}$$

Estimate conditional heritability

When multiple genomic regions are physically close, the heritability estimated by EHE at some genes may be because they had SNPs in LD with nearby causal genes. Therefore, we propose a conditional analysis among the proximal genes by EHE. First, the genome-wide gene-based association is carried out by ECS with summary statistics.[13,16] Suppose there are m significant genes. The significant genes in a large LD block are sorted by the association p values. According to the additive property of chi-squared statistics, the conditional heritability of multiple genes is estimated iteratively from the most to the least significant gene,

$$\begin{cases} \widehat{H}_1 \\ \widehat{H}_{2|1} = \widehat{H}_{1,2} - \widehat{H}_1 \\ \cdots \\ \widehat{H}_{m|1,2,\ldots,m-1} = \widehat{H}_{1,2,\ldots,m-1,m} - \widehat{H}_{1,2,\ldots,m-1} \end{cases} \quad \text{(Equation 7)}$$

where $\widehat{H}_1$ is the estimated heritability of the most significant gene, $\widehat{H}_{1,2}$ is the estimated heritability of the most and the second most significant genes when merged in a single region, $\widehat{H}_{2|1}$ is the heritability of the second most significant gene conditioning on the most significant region, and so on. The variance of the conditional EHE can be calculated by the formula

$$\mathrm{Var}(\widehat{H}_{2|1}) = \mathrm{Var}(\widehat{H}_{1,2}) + \mathrm{Var}(\widehat{H}_1) - 2\mathrm{Cov}(\widehat{H}_{1,2}, \widehat{H}_1) \quad \text{(Equation 8)}$$

where for each term, variance can be calculated by Equation 4. The same iteration procedure is carried out to approximate the SEs of the conditional heritability estimates for multiple genes in an LD block. The order of genes entering the iteration procedure can also be determined according to a third-party functional annotation of genes, e.g., the specific expression of genes in a phenotype-relevant tissue (DESE) or cell-type (PCGA).[18,19] Generally, the genes entering the iteration procedure earlier have more chance to keep their original heritability.

Software implementation of EHE

We have implemented EHE using Java on the KGGSEE software platform for estimating the local and conditional heritability of genes or regions. KGGSEE maps input SNPs onto genes defined in RefGene[20] and GENCODE[21] databases or customized genomic regions. The program takes genotypes of an ancestrally matched reference sample in variant call format (VCF)[22] to account for LD between SNPs. KGGSEE employs an algorithm to partition the input variants into multiple LD blocks. Starting from the SNPs with minimal position coordinates on a chromosome, the program explores the LD block boundaries by checking the LD of k (equals 100 by default) adjacent SNPs on the forward side. The boundary SNP must have no absolute LD value ($|r|$) with its preceding SNP larger than a cutoff c (equals 0.05 by default) to ensure that any couple of SNPs in different LD blocks must have $|r| \leq c$. This algorithm prevents the generation of a huge LD matrix of an entire chromosome. EHE is carried out in each LD block separately first, and the overall heritability is equal to the sum of the heritability estimates in all LD blocks. KGGSEE also implements ECS[13] to calculate association p values of genomic regions. For conditional heritability estimation, gene expression of multiple tissues or cell types can be input into KGGSEE, which implements the ECS and the cell type selective expression supervised iterations[18,19] for the conditional analyses.

Alternative heritability estimation methods for comparison

In this study, we compared the performance of EHE with six widely used or latest methods, namely GCTA,[5] HESS,[12] LDAK-GBAT,[14] LDAK-SumHer,[15] LDSC,[8] and LDER.[10] Among the seven methods, HESS is the only one that assumes a fixed-effects model, which can also be considered as a random effects model with no constraints on the variances of effect sizes. The random effects model assumed by HESS was named the generalized random effects (GREs) model,[23] and EHE also assumes the GRE model but is a different estimator. We used HESS v.0.5.3-beta in this study. Among the seven methods, GCTA is the only method that estimates heritability using genotype and phenotype data. GCTA performs REML estimation with the well-known GCTA model,[4,5] in which the variances of effect sizes of standardized genotypes are equal. When the effect sizes are generated by the GCTA model, we expect GCTA to be the most accurate. LDSC, LDER, and SumHer are all based on LD score regression, where LDSC and LDER assume the same genetic architecture as GCTA, and SumHer introduces an alpha parameter, thus allowing more general genetic architectures.[15,24] In this study, we used $\alpha = -0.25$ for all SumHer analyses. Since LDER makes full use of the LD information, it is expected to be more accurate than LDSC. GBAT is designed for gene-based association tests, which also calculates gene-based heritability. Similar to GCTA, GBAT also performs REML estimation; but unlike GCTA, GBAT uses summary statistics; and like SumHer, GBAT estimates with a user-specified alpha parameter. In this study, we run all GBAT analyses with $\alpha = -0.25$. Among the seven methods, LDSC, LDER, and SumHer are designed to estimate the heritability of a whole genome using summary statistics and estimate a SE using a jackknife estimator which is not suitable for a small amount of SNPs, so we didn't compare their SE estimates. In contrast, EHE, HESS, and GBAT are designed to estimate local heritability using summary statistics and are more comparable. Hou et al.[23] summarized the genetic architectures assumed by GCTA, LDSC, and SumHer by listing the additional assumptions in addition to the GRE model.

Computer simulation models and experiments

The LD patterns for the simulations

We performed extensive computer simulations to investigate the accuracy and precision of the EHE for gene-based heritability estimation. A total of 488 genes with different SNP sizes ranging from 6 to 2,094 were drawn from chromosome 1 of the human genome to perform the simulations. To make the sampling even, we first counted the number of SNPs with minor allele frequency larger than 0.01 in the EUR panel of the 1000 Genomes Project for each protein-coding gene (including 10 kb flanking regions on both sides) on chromosome 1 (1,944 genes in total). Then, we divided the genes into four groups by the three quantiles of the numbers of SNPs. Finally, for each group, we sorted the genes by coordinates, took the first gene out of every four, and got 122, 124, 121, and 121 genes in each of the four SNP-number ranges. We further took the first gene out of every eight in each of the four groups and got 61 genes to perform GCTA analyses.

Simulations for quantitative phenotypes

For each gene, we simulated a sample of 4.5×10^6 haplotypes by HapSim[25] with a reference being a phased VCF file of the EUR panel of the 1000 Genomes Project.[26] Each haplotype was coded by a vector of zeros and ones. Then, we generated a population of 2.25×10^6 diploid individuals by paring haplotypes and summing, and then standardizing for each SNP. We denote the standardized genotype matrix as $\boldsymbol{X}$. We realized $\boldsymbol{\beta}$ by sampling from $N(\mathbf{0}, \boldsymbol{\Sigma}_\beta), \boldsymbol{\Sigma}_\beta = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$, where different $\boldsymbol{\Sigma}_\beta$ were used to simulate different genetic architectures. According to Speed et al.,[24] we considered two alpha values (-1.0 and -0.25) in addition to three proportions of QTLs (0.01, 0.1, and 1) to mimic six different genetic architectures. For each genetic architecture of each gene, we realized $\boldsymbol{\beta}$ by (1) randomly choosing QTLs with the specified proportion; (2) for each selected SNP, setting $\sigma_j^2 = h_g^2 w_j / \sum_j w_j, w_j = [f_j(1-f_j)]^{1+\alpha}$, where f_j is the frequency of either allele of the j$^{\text{th}}$ SNP and h_g^2 is the target heritability, and for unselected SNPs, setting their variances to zero; and (3) sampling $\boldsymbol{\beta}$ from $N(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ and rescaling $\boldsymbol{\beta}$ to make the population variance of $\boldsymbol{X\beta}$ equals h_g^2.

Then, we realized the phenotype, took samples, and calculated Z scores and LD coefficient matrices.

We (4) realized the phenotype by $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, (1-h_g^2)\boldsymbol{I})$ and considered the realized $(\boldsymbol{y}, \boldsymbol{X})$ as the population; (5) took 100

samples of size n = 20,000 without replacement from $(\boldsymbol{y}, \boldsymbol{X})$ and denoted the ith sample as $(\boldsymbol{y}_i, \boldsymbol{X}_i)$, and denoted the LD matrix calculated by $\boldsymbol{X}_i$ as $\boldsymbol{R}_i$; (6) for each sample, e.g., the ith one, calculated the Z score vector as $\boldsymbol{z}_i = n^{-\frac{1}{2}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{y}_i$; (7) took 100 (and another 100) samples of sizes $\dot{n} = 2,000$ (and $\ddot{n} = 500$) without replacement from $\boldsymbol{X}$; calculated the LD matrix for each sample and denoted as $\dot{\boldsymbol{R}}_i$ ($\ddot{\boldsymbol{R}}_i$) for the ith sample.

For each gene, by the end of step (7), we have 100 samples, with the ith one denoted as $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{z}_i, \boldsymbol{R}_i, \dot{\boldsymbol{R}}_i, \ddot{\boldsymbol{R}}_i)$. We estimated heritability using the 100 samples of a gene by each of EHE, GCTA,[5] HESS,[12] LDAK-GBAT,[14] LDAK-SumHer,[15] LDSC[8], and LDER.[10] (8) With each of all the 100 samples, e.g., the ith one, we ran each of EHE, HESS, GBAT, SumHer, LDSC, and LDER with each of $(\boldsymbol{z}_i, \boldsymbol{R}_i)$, $(\boldsymbol{z}_i, \dot{\boldsymbol{R}}_i)$, and $(\boldsymbol{z}_i, \ddot{\boldsymbol{R}}_i)$; and ran GCTA with $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ of only the first 50 samples (refer to supplemental methods for the command line of each method). (9) For each of EHE, HESS, GBAT, SumHer, LDSC, and LDER, considering the 100 estimates as a sample; we calculated the sample mean (Mean($\widehat{h}^2$)) and the sample standard deviation (SD($\widehat{h}^2$)), and we calculated the mean relative bias (MRB) of heritability estimates as $\mathrm{MRB}(\widehat{h}^2) = \mathrm{Mean}(\widehat{h}^2)/h_s^2 - 1$. For each of EHE, HESS, and GBAT, we calculated the mean of the 100 SE estimates (Mean(SE($\widehat{h}^2$))) and calculated an MRB of the 100 SE estimates as $\mathrm{MRB}(\mathrm{SE}(\widehat{h}^2)) = \mathrm{Mean}(\mathrm{SE}(\widehat{h}^2))/\mathrm{SD}(\widehat{h}^2) - 1$.

For the analyses of comparing with GCTA, we calculated Mean($\widehat{h}^2$), SD($\widehat{h}^2$), MRB($\widehat{h}^2$), and MRB(SE($\widehat{h}^2$)) by only the first 50 samples of $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{z}_i, \boldsymbol{R}_i, \dot{\boldsymbol{R}}_i, \ddot{\boldsymbol{R}}_i)$. For EHE, HESS, GBAT, SumHer, LDSC, and LDER, we repeated steps (1)–(9) for all six genetic architectures. For the analyses comparing with GCTA, we only performed analyses of three genetic architectures ($\alpha = -1.0$ with proportions of QTLs being 0.01, 0.1, and 1) of 61 genes and 50 replicates for each gene. HESS computes eigenvalues and squared projections of the effect size vector onto the eigenvectors of the LD matrix in its first step and estimates local SNP heritability and the SE in its second step, which needs results of all 22 autosomes from the first step. We removed the loop of the 22 autosomes from the HESS code to go through the second step with the result of only one gene from the first step.

Simulations for dichotomous phenotypes

We simulated phenotypes using the liability threshold model.[17,27] For each gene, we simulated a population of 10^7 (or 10^8) diploid individuals for a prevalence of $K = 10\%$ (or $K = 1\%$) and denoted the standardized genotypes as $\boldsymbol{X}$ as described for quantitive phenotypes. We only simulated three genetic architectures with $\alpha = -0.25$ and proportions of QTLs being 0.01, 0.1, and 1. For each genetic architecture, we realized $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as described for quantitive phenotypes and converted $\boldsymbol{\gamma}$, the liability, into dichotomous phenotypic values with the threshold $\Phi^{-1}(K)$. Then, for each SNP, we built a marginal logistic regression model without intercept and performed an association test using the Logit class of Statsmodels v.0.13.5. Similar to the simulations of quantitative phenotypes, for the ith sample, we got $\boldsymbol{z}_i$ from the association tests and calculated $\dot{\boldsymbol{R}}_i$ and $\ddot{\boldsymbol{R}}_i$. But for dichotomous phenotypes, we only ran with $(\boldsymbol{z}_i, \dot{\boldsymbol{R}}_i)$ and $(\boldsymbol{z}_i, \ddot{\boldsymbol{R}}_i)$ for each of EHE, LDSC, LDER, HESS, SumHer, and GBAT. Finally, we calculated Mean($\widehat{h}^2$), SD($\widehat{h}^2$), MRB($\widehat{h}^2$), and MRB(SE($\widehat{h}^2$)) as described for quantitative phenotypes.

Validation with real datasets

GWAS summary statistics and heritability of protein-coding genes

We downloaded GWAS summary statistics of UK Biobank phenotypes from the Neale Lab. We selected 42 complex phenotypes with sample sizes greater than 100,000 and heritability calculated by LDER above 0.1 (Table S1). For a dichotomous phenotype, we calculated liability-scale heritability using the proportion of affected individuals in the GWAS sample as the prevalence, as no ascertainment was performed during sampling.[28] LDER was also used to calculate the heritability of all protein-coding genes, including transcribed regions (from the first exon to the last exon with introns included) and 10 kb flanking regions. We also ensured that the squared genetic correlation coefficient calculated by LDSC between each pair of phenotypes was less than 0.25 (Table S3). We used the 1000 Genomes Project European panel to calculate LD scores for LDER and LDSC.

Then, we compared the local heritability estimates of protein-coding genes among EHE, HESS, and GBAT for each of the 42 complex phenotypes. SNPs within the transcribed region and its 10 kb flanking regions are included for each gene. For all three methods, we used the European panel of the 1000 Genomes Project as the LD reference. For the second step of HESS, we decompose the total SNP heritability of protein-coding genes calculated by LDER into each gene. For GBAT, we removed genes warned in the running log and genes with a zero SE from the results. For dichotomous phenotypes, we calculated liability-scale heritability for EHE, GBAT, and LDER by Equation 5, and then we decomposed the liability-scale heritability of protein-coding genes calculated by LDER into each gene in the second step of HESS. We plotted Figure 4 by rMVP.[26]

Polygenicity and pleiotropy

Conditional EHE is a powerful tool for removing the heritability of a gene tagged by other genes due to LD. Leveraging EHE, we compared the conditional heritability among multiple groups of significant genes. First, we performed the unconditional ECS (uECS) test for each protein-coding gene for each of the 42 phenotypes and further performed the conditional ECS (cECS) test[13] for genes with a uECS-significant p value ($p < 2.5 \times 10^{-6}$). To determine the order of genes within an LD group to enter a series of cECS tests, we calculated their tissue selectivity scores[18] by integrating the GTEx v.8 tissue expression profile with genes' uECS p values. Second, for genes conditionally significantly associated with a phenotype (cECS $p < 2.5 \times 10^{-6}$; also called candidate susceptibility genes in the paper), we calculated their conditional heritability. For genes in an LD block, the order of entering the process of calculating conditional heritability was determined by their cECS tests. For instance, Figure S7 shows the heritability of protein-coding genes on chromosome 2 being uECS-significant for asthma. In group one, there are eight uECS-significant genes. Gene *IL1R1* (1-1), which had the highest tissue selectivity score, entered first into the series of cECS tests and the conditional heritability calculations, followed by gene *IL1RL1* (1-2), which had the second highest score. In this group, only the two genes are cECS significant, and thus, only their conditional heritability was further calculated. We calculated the heritability of candidate susceptibility genes by summing the conditional heritability over cECS-significant genes of a phenotype. We also estimated the heritability of the most significant half and the most significant quarter of the cECS-significant genes of a phenotype. Third, we separately collected genes conditionally significantly associated with ≥ 3, ≥ 5, and ≥ 7 phenotypes. For each phenotype, we calculated the heritability of genes affecting ≥ 3, ≥ 5, and ≥ 7 genes by summing their conditional heritability. Lastly, to investigate polygenicity, we calculated the enrichment of heritability of protein-coding genes in all of, in the most significant half of, and in the most significant quarter of the cECS-significant genes,

respectively. To investigate pleiotropy, we calculated the enrichment of heritability of cECS-significant genes in genes affecting ≥ 3, ≥ 5, and ≥ 7 phenotypes, respectively. Enrichment is defined as the ratio of the proportion of heritability to the proportion of genes.[11]

Results

Overview of EHE

EHE is a method proposed for accurately estimating local heritability in a genomic region or gene, i.e., the ratio of the genetic variance in a genomic region or gene to the phenotypic variance. Its rationale is that the heritability in a genomic region is the accumulation of non-redundant heritability from its SNPs. Technically, EHE converts the marginal heritability estimates at SNPs into a non-redundant heritability estimate of a set of SNPs by a linear combination of the marginal heritability estimates. The marginal heritability estimate of an individual SNP is calculated from its GWAS p value, equal to the corresponding chi-squared statistics of the p value with one degree of freedom minus one, and then divided by its sample size. This linear combination removes the redundancy in marginal heritability estimates due to LD and facilitates the calculation of the SE of an estimate. In a simulation study, EHE demonstrated higher accuracy for local heritability estimation than alternative methods originally designed for local or global heritability estimation (see Figure 1). Additionally, EHE can estimate conditional heritability with a correct SE by removing heritability tagged by nearby genes. The conditional analysis allows for quantifying genetic contribution among close genomic regions when integrated with functional annotations of genes. Theoretical analyses show that EHE can produce unbiased local heritability estimates (refer to material and methodsand supplemental methods for details).

We have created advanced algorithms, such as LD blocking and selective expression-supervised conditional iterations, to implement EHE in a user-friendly software named KGGSEE. This software can estimate local heritability and conditional heritability for genes or regions. EHE uses only p values as input and has high accuracy, making it an important tool for dissecting the genetic architecture of complex phenotypes using widely available large-scale GWAS summary data. In this paper, we demonstrate the use of EHE to investigate the local heritability of protein-coding genes and pleiotropic genes of 42 complex phenotypes.

Gene-based heritability estimation by EHE and the alternative methods in simulated datasets

We conducted simulations at 488 genes (a quarter of protein-coding genes in chromosome 1) with varying SNP sizes to investigate the accuracy of EHE in estimating local heritability. Additionally, we compared EHE's estimation accuracy and precision with six alternative methods (GCTA,[5] HESS,[12] LDAK-GBAT,[14] LDAK-SumHer,[15] LDSC,[8] and LDER[10]). Figure 1 presents the comparison results for 24 representative scenarios under six genetic architectures, assuming a true heritability of 0.1% in a gene for a quantitative phenotype. EHE exhibits three key characteristics for local heritability estimation using GWAS summary statistics of a gene. Firstly, EHE demonstrates the highest accuracy among the six methods when there are not many QTLs (quantitative trait loci) within a gene. For instance, when the proportion of QTLs in a gene is 0.01 (but at least one QTL), EHE consistently exhibits the smallest bias among all the methods, irrespective of the SNP sizes (ranging from 6 to 2,094) of the genes (Figures 1A, 1D, 1G, and 1J, etc.). In this condition, EHE is even more accurate than GCTA that needs individual-level genotypes and phenotypes as input (Figure 2). As the proportion of QTLs increases to 1, the estimation bias of EHE slightly increases but remains comparable to GBAT and LDER and is smaller than HESS, SumHer, and LDSC in most scenarios (Figures 1C, 1F, 1I, and 1L, etc.). Notably, EHE utilizes less information than all the other methods for estimation by disregarding the signs of LD coefficients and effect sizes. Secondly, EHE is proved robust against LD noises when LD coefficients are calculated from an independent reference sample. In most scenarios, EHE accurately and precisely estimates the true heritability when the LD matrix is calculated from an independent reference sample of 500 individuals, similar to a continental ancestry panel of the 1000 Genomes Project.[29] The accuracy of EHE was held for genes with a lower marker density, e.g., the HapMap3 SNP panel (Figure S1). Conversely, HESS, GBAT, and LDER exhibit increased estimation bias when using an independent sample of 500 individuals compared to using the original GWAS sample as the LD reference panel especially when the SNP size is large (Figures 1S, 1T, and 1U, etc.). Finally, EHE shows insensitivity to genetic architectures where the effect sizes are either negatively or not correlated with allele frequencies. This characteristic also holds for other alternative methods, as the power parameters -0.25 (for frequency-dependent effect) and -1.0 (for frequency-independent effect) result in similar MRBs of heritability estimates and their SEs (Figures 1 and 3). When the true heritability at a gene increased to 1%, EHE maintained its consistent relative advantages, as depicted in Figure S2. Similar to our observations for quantitative phenotypes, EHE demonstrated accurate, precise, and robust estimation of the true heritability for the majority of genes associated with dichotomous phenotypes, showcasing its superior performance compared to alternative methods (Figure S3).

Additionally, the simulation studies conducted in this study demonstrated that the estimator of SE of EHE is empirically unbiased and accurate, which is crucial for assessing the confidence level of a local heritability estimate. Figure 3 visually represents the consistency between EHE's SE estimates and the sample standard deviation calculated from 100 estimates (SD100), regardless of gene SNP sizes

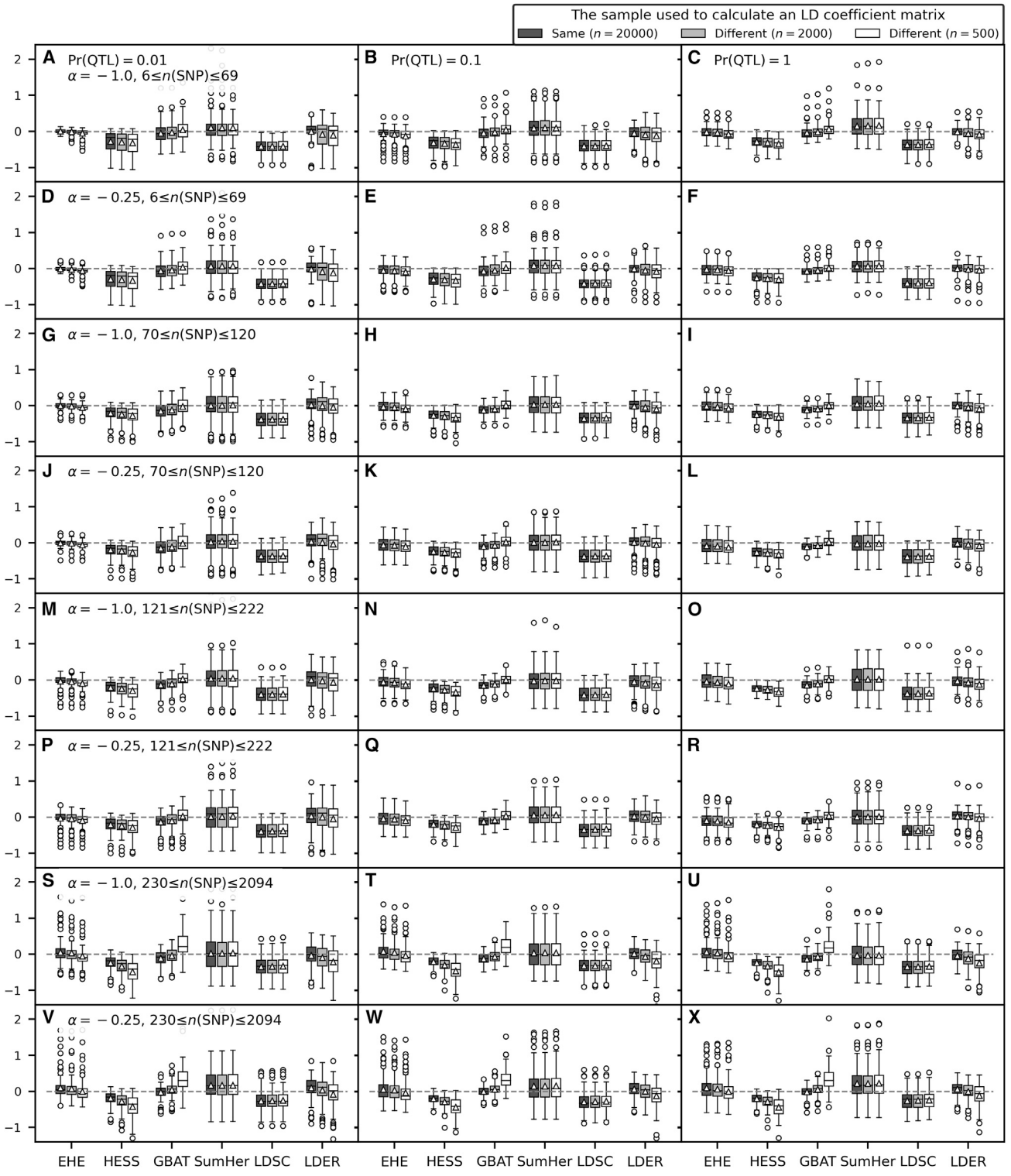


Figure 1. Mean relative bias of estimated heritability ($\mathrm{MRB}\left(\widehat{h}^2\right)$) by six methods for genes with varying SNP sizes for a quantitative phenotype

The scale of the y axes is $\mathrm{MRB}\left(\widehat{h}^2\right)$. The true heritability is 0.1% and the GWAS sample size is 20,000. The three panel columns represent the proportions of QTLs Pr(QTL) = 0.01, 0.1, or 1. However, each gene contains at least one QTL. The rows indicate quantile intervals of SNP sizes in a gene: first, second, third, and fourth. The odd-numbered rows correspond to a power parameter of $\alpha = -1.0$ for the effect size model, while the even-numbered rows correspond to $\alpha = -0.25$. For the boxplot of each method, each data point represents $\mathrm{MRB}\left(\widehat{h}^2\right)$ calculated by 100 repetitions for a gene; the whiskers extend to 1.5 times of the interquartile range; the white rectangle of each boxplot shows the mean.

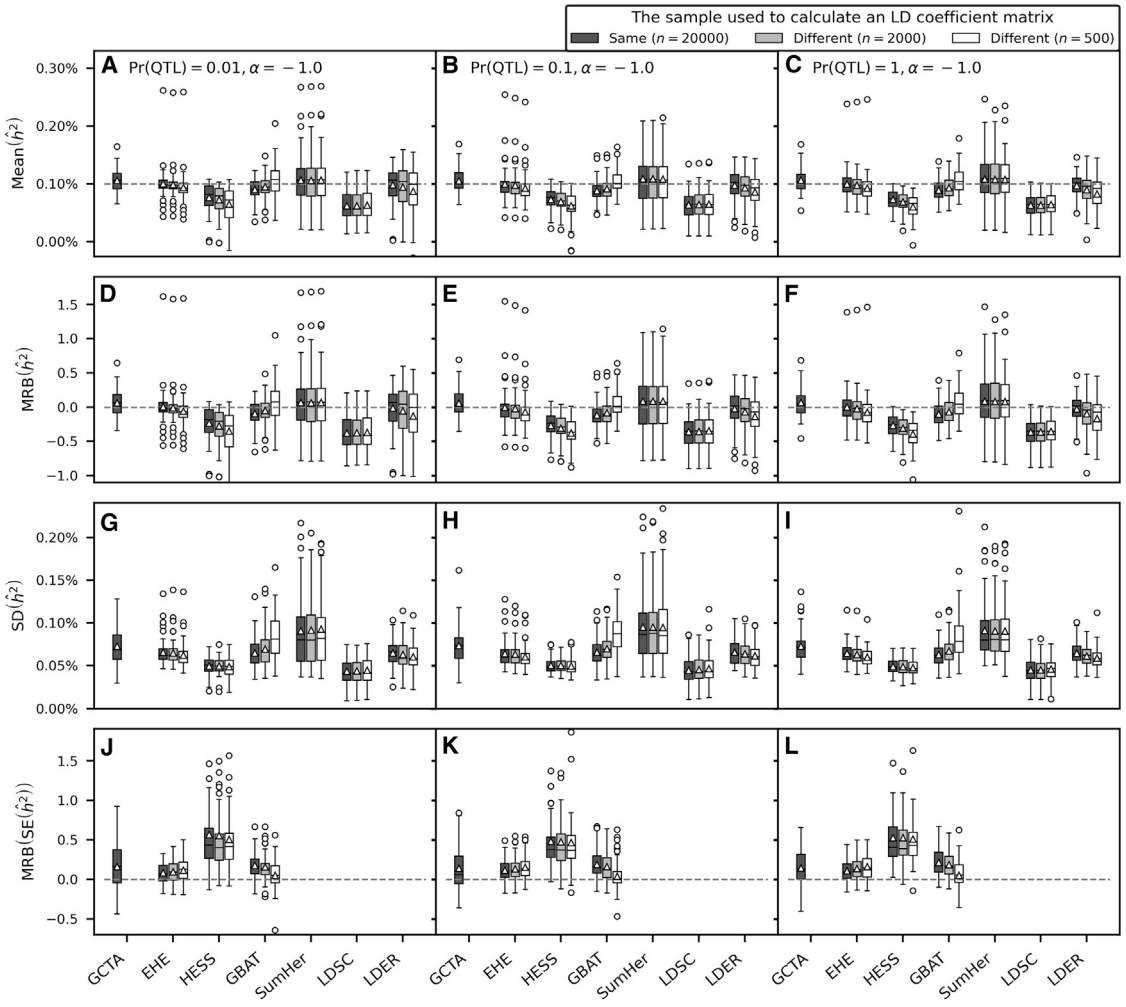


Figure 2. Comparison of gene-based heritability estimates among seven methods at 61 genes with 21 to 2,094 SNPs
The true heritability is 0.1%, the GWAS sample size is 20,000, and the power parameter α for the effect size model is −1.0. The three panel columns represent the proportions of QTLs, Pr(QTL) = 0.01, 0.1, and 1. However, each gene contains at least one QTL.
(A–C) Each data point represents the mean of 50 heritability estimates for a gene.
(D–F) Each data point represents the MRB of 50 heritability estimates for a gene.
(G–I) Each data point represents the standard deviation of 50 heritability estimates for a gene.
(J–L) Each data point represents the MRB of 50 SE estimates for a gene. The whiskers extend to 1.5 times of the interquartile range; the white rectangle of each boxplot shows the mean.

and genetic architectures. Notably, when the number of SNPs was less than 222, the SD100 of the majority of genes fell within the range between 0.05% and 0.075% for the true heritability of 0.1% (Figure S4). Interestingly, calculating LD coefficients using 500 independent individuals resulted in slightly smaller SD100. When comparing EHE to other methods, the SD100 of EHE was comparable to those of LDER and smaller than those of GBAT and SumHer (Figure S4). It is worth noting that the two methods (HESS and LDSC) that underestimated heritability exhibited the smallest SD100 among the compared methods. However, HESS showed the highest deviation between SD100 and SE estimates. GBAT displayed slightly larger deviations between SE estimates and corresponding SD100 in most scenarios (Figure 3). Note that the other three methods dedicated to global heritability (SumHer, LDSC, and LDER) utilized jackknife estimators for SE, which renders them unsuitable for estimating an SE of local heritability due to insufficient SNP numbers in most genes for the jackknife resampling.

Interestingly, LDER, originally designed for estimating global heritability, achieved the second most accurate estimation for local heritability in most scenarios. SumHer generally exhibited larger variance compared to LDER, while LDSC consistently produced estimates smaller than the true values for most genes. Among the two alternative tools specifically developed for local heritability estimation, HESS also yielded estimates smaller than the true values for most genes. GBAT appeared to be sensitive to LD noises, as it tended to produce estimates larger than the true values for genes with more than 230 SNPs (e.g., Figures 1S and 1V).

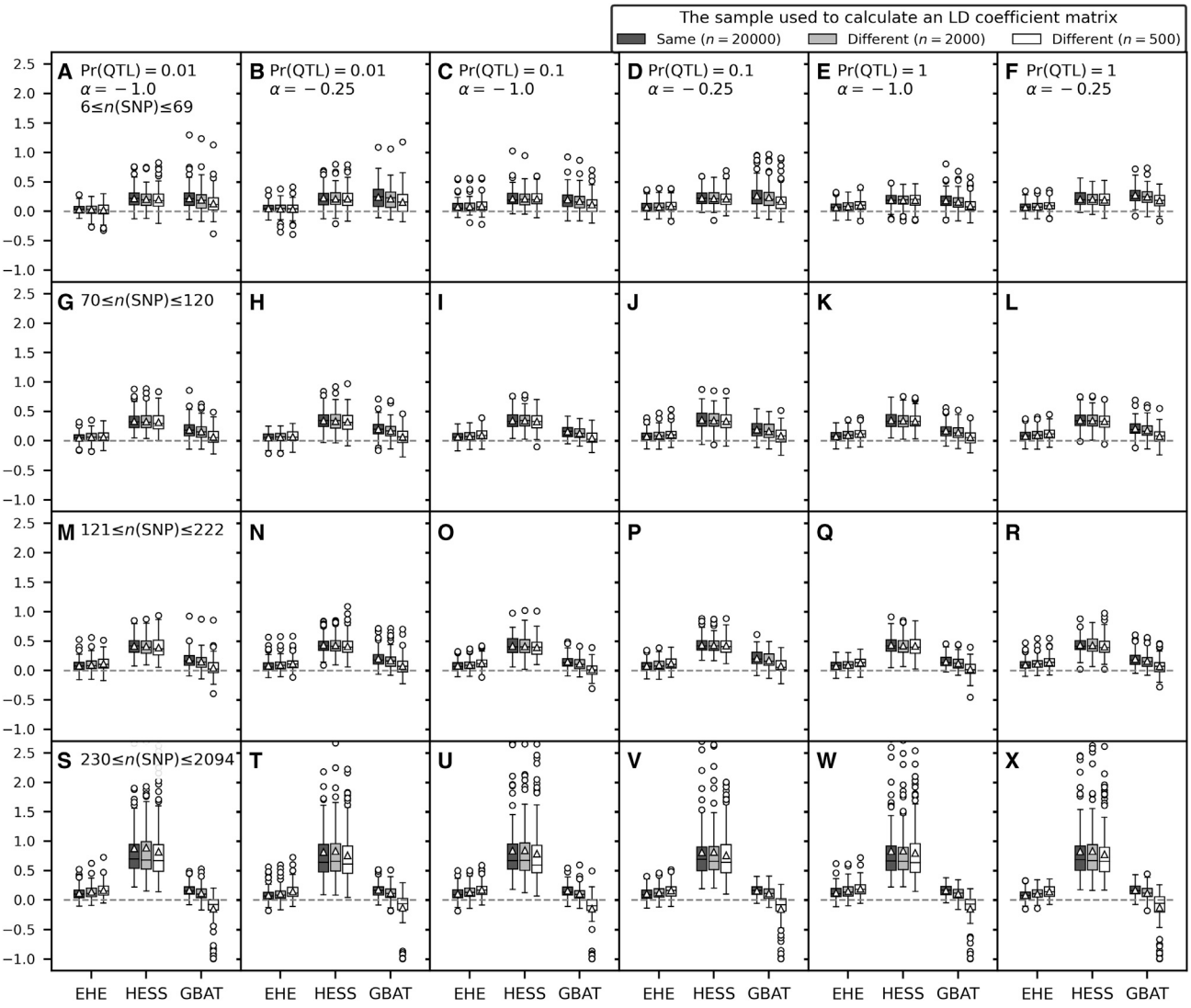


Figure 3. Mean relative bias of estimated standard errors of heritability estimates $\left(\mathrm{MRB}\left(\mathrm{SE}\left(\widehat{h}^2\right)\right)\right)$ by three methods for genes with varying SNP sizes

The scale of the y axes is $\mathrm{MRB}\left(\mathrm{SE}\left(\widehat{h}^2\right)\right)$. The true heritability is 0.1% and the GWAS sample size is 20,000. The panel columns represent different genetic architectures, including the proportion of QTLs (Pr(QTL) = 0.01, 0.1, or 1) in a gene and the power parameter ($\alpha = -1.0$ or -0.25) for the effect size model. The rows indicate quantile intervals of SNP sizes in a gene: first, second, third, and fourth. For the boxplot of each method, each data point represents $\mathrm{MRB}\left(\mathrm{SE}\left(\widehat{h}^2\right)\right)$ calculated by 100 repetitions for a gene; the whiskers extend to 1.5 times of the interquartile range; the white rectangle of each boxplot shows the mean.

EHE also accurately estimated conditional heritability

We conducted an additional simulation experiment to examine the potential of EHE for removing redundant heritability due to linkage disequilibrium (LD). When two genes are in LD (e.g., genes A and B), we define the conditional heritability of gene B conditioning on gene A as $h^2_{B|A} = h^2_{A,B} - h^2_A$, where $h^2_{A,B}$ denotes the overall heritability of both genes and h^2_A denotes the marginal heritability of gene A (and vice versa). We propose the conditional EHE as $\widehat{H}_{B|A} = \widehat{H}_{A,B} - \widehat{H}_A$, where $\widehat{H}_{A,B}$ and $\widehat{H}_A$ are EHEs of $h^2_{A,B}$ and h^2_A, respectively. We randomly selected an LD block of 100 SNPs, with only the first SNP being the causal SNP. The EHE calculated by the last 50 SNPs (from the 51st to the 100th) is non-zero due to LD (Figures S5A and S5C). Actually, there was no true heritability in the last 50 SNPs, as there was no causal SNP in this region. We applied EHE to estimate the heritability of the last 50 SNPs (no causal SNP) conditioning on the first 50 SNPs (with one SNP). As depicted in Figures S5B and S5D, the conditional EHE successfully removed the redundant heritability due to LD. The median of the estimated conditional heritability in the last 50 SNPs was close to zero, regardless of the GWAS sample size (10,000 or 50,000), true heritability (0.1% or 0.5%), and phenotype class (dichotomous or quantitative phenotype). The SEs of the conditional heritability were also similar to the empirical ones among 100 simulated datasets (Figure S6). Our findings suggest that EHE can remove redundant heritability in nearby regions or genes.

The estimated heritability of genes for 42 phenotypes

We further validated EHE for the heritability of protein-coding genes for each of the 42 phenotypes (Table S1) using real GWAS summary statistics provided by the Neale Lab. Figure 4A provides an overview of gene-level heritability estimates for the 42 phenotypes using EHE and two other methods (HESS and GBAT) dedicated for local heritability estimation (see Table S4 for the numeric results). Based on EHE's estimates ($\widehat{H}$), 16 phenotypes exhibited genes with $\widehat{H} > 1\%$, indicating the presence of genes with large effect sizes. Genes with large effect sizes (or heritability) usually have higher implications in precision medicine for phenotype prediction, treatment, and drug development. Notably, the quantitative phenotype with the largest number of genes (n = 24) showing $\widehat{H} > 1\%$ is sex hormone-binding globulin (SHBG). The gene with the highest heritability is *MPDU1* ($\widehat{H} = 2.85\% \pm 0.079\%$). Hypothyroidism shows the largest number of genes (n = 8) with $\widehat{H} > 1\%$ among all the seven dichotomous phenotypes, with *HLA-DQA1* exhibiting the highest heritability ($\widehat{H} = 1.51\% \pm 0.13\%$). However, it is important to note that there are also multiple genes with large heritability located within a proximity of less than 10 kb to these two genes (see Table S4 for details). The high estimated heritability of most of these genes is likely attributed to their LD with the true functional genes, emphasizing the need for conditional analysis to remove redundant gene heritability. Conversely, we also observed 17 phenotypes that did not exhibit genes with $\widehat{H} > 0.5\%$, suggesting that these phenotypes are highly polygenic and more challenging for precision medicine. Standing height and college degree attainment are typical quantitative and dichotomous polygenic phenotypes. For height, the gene with the highest heritability is *UQCC1* ($\widehat{H} = 0.47\% \pm 0.032\%$). However, it is in LD with the well-established gene growth differentiation factor 5 (*GDF5*, $\widehat{H} = 0.42\% \pm 0.028\%$), which is located at the same chromosomal locus, 20q11.22. Moreover, many other height-associated genes are distributed across different chromosomes with estimated heritability exceeding 0.1% (Figure 4B). Regarding college degree attainment, the region with the largest number of genes exhibiting heritability around 0.1% is 3p21.31, with the gene *RBM6* displaying the highest estimated heritability ($\widehat{H} = 0.13\% \pm 0.021\%$). Again, this locus also has multiple significant genes ($p < 2.5 \times 10^{-6}$) with similar heritability for the college degree.

We also observed differences between the three local heritability estimation methods in the real data analyses. HESS yielded more genes with very large heritability (>2%) than EHE. For instance, HESS estimated the largest gene-level heritability for SHBG as 4.90% ± 0.077%, whereas EHE yielded a value of 2.85% ± 0.079%. Similarly, for the phenotype platelet distribution width (PDW), HESS estimated the heritability of genes *TUBB1* and *ATP5F1E* on chromosome 20 as 2.75% ± 0.055% and 2.68% ± 0.055%, respectively, whereas EHE yielded values below 1% for both genes. On the other hand, HESS generally produced smaller estimated heritability for dichotomous phenotypes compared to EHE. For example, HESS did not identify genes with estimated heritability >1% for hypothyroidism, whereas EHE identified eight such genes. However, the overall relative number of genes with large heritability for every 42 phenotypes was consistent between HESS and EHE (Figure 4A). GBAT, on the other hand, generated numerous unrealistic estimates above 0.1, accompanied by warning messages. However, excluding genes with warning messages resulted in the omission of several genes with large estimated heritability identified by both EHE and HESS (Figure 4A). Notably, GBAT still produced large heritability estimates (>0.3%) for numerous genes associated with standing height (Figure 4C). It is worth noting that EHE demonstrated robustness in local heritability estimation compared to HESS and GBAT, as it does not require the consistency of reference allele definition for the LD and Z scores. This finding is consistent with our simulation study observations, further highlighting EHE's favorable performance for local heritability estimation.

Driver tissues of complex phenotypes inferred for the conditional heritability analyses

We proposed to use the selective expression of genes in potential driver tissues to guide the entry order of genes for the conditional heritability analysis by EHE. We assume that genes with higher selective expression in potential driver tissues are more likely to be the true susceptibility genes and are given higher priority to enter the iterative conditional heritability analysis procedure. However, as driver tissues are often unclear, we utilized DESE[18] to infer the driver tissues of the 42 UKBB phenotypes (Table S2) using expression profiles of autosomal protein-coding genes in GTEx.[30] Our results show that, for most phenotypes, the most significant driver tissues identified by DESE are consistent with established biological knowledge. For example, the pancreas was identified as the most significant driver tissue for diabetes, while the CNS was the most significant for a college degree, smoking status, chronotype, age of menarche, and neuroticism. Furthermore, the liver was identified as the most significant driver tissue for blood cholesterol concentration (CHOL). Based on these results, we assert that DESE is a reasonable approach for inferring driver tissues, and we assume that the method used by DESE to prioritize susceptibility genes among multiple genes in LD reflects the genetics in our empirical analyses. Therefore, we used the selective expression of genes in potential driver tissues identified by DESE to guide the conditional heritability analysis of genes within one LD block in the following analyses.

Heritability of pleiotropic genes for complex phenotypes and polygenicity of complex phenotypes

The heritability of pleiotropic genes for a complex phenotype and polygenicity of a complex phenotype are important characteristics of the genetic architecture of the

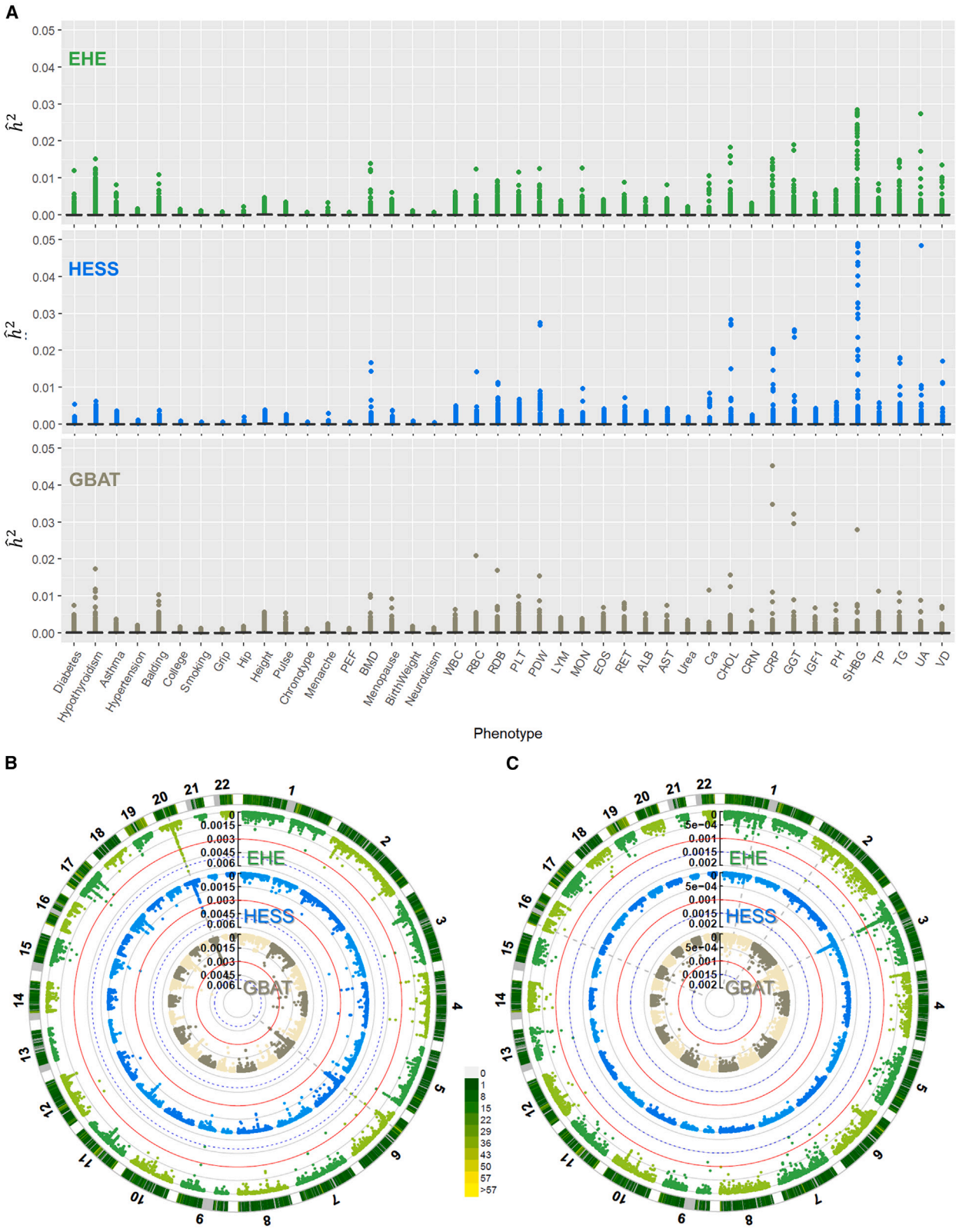


Figure 4. Comparison of estimated gene-based heritability using three local heritability estimators with real GWAS summary statistics

(A) Estimated gene-based heritability for 42 phenotypes using three methods.

(B) Manhattan plot illustrating the estimated gene-based heritability for standing height using three methods.

(C) Manhattan plot illustrating the estimated gene-based heritability for college degree awarding using three methods. Each dot represents a gene. The color bar indicates the density of gene counts per Mega base pairs in chromosomes, as shown in the outermost circle. Refer to Table S4 for the numeric results.

Figure 5. Heritability enrichment ($\Pr(\widehat{h^2})/\Pr(\text{genes})$) of candidate susceptibility genes and pleiotropic genes
(A) Heritability of protein-coding genes and heritability of cECS-significant protein-coding genes. The error bars show SEs.
(B) The enrichments of heritability of protein-coding genes in all, the most significant half, and the most significant quarter of the cECS-significant protein-coding genes.
(C) The enrichments of heritability of cECS-significant protein-coding genes in the genes affecting different numbers of phenotypes.

phenotype.[31] We found that LD can cause many genes to be indirectly associated with a phenotype, meaning they appear to be associated only because they correlate with a true susceptibility gene. The conditional analysis removes these indirectly associated genes and retains the directly associated ones, which are more likely to be the true susceptibility genes. Therefore, conditional analysis can help to identify the true genetic drivers of a phenotype. EHE is able to remove redundancy from the heritability of a gene due to the presence of LD with other genes. By taking advantage of EHE, we explored the polygenicity of the 42 phenotypes by comparing the heritability enrichments among conditionally significant genes and explored the heritability of genes affecting different numbers of phenotypes. We defined the heritability enrichment of a group of genes as the ratio of the proportion of heritability to the proportion of genes.[11]

First, we calculated the heritability of all cECS-significant protein-coding genes (also refered to as candidate susceptibility genes; Figure 5A). The proportion of heritability explained by candidate susceptibility genes highly depends on the number of cECS-significant genes of a phenotype (Pearson's $r = 0.71$, $p = 1.8 \times 10^{-7}$). For example, for height, there are 442 cECS-significant protein-coding genes, and the heritability of the 442 genes is 27% ± 0.23%, which accounts for 53.2% of the heritability of all protein-coding genes; for diabetes, there are 41 cECS-significant protein-coding genes, and the heritability of the 41 genes is 4.7% ± 0.096% which accounts for 22.4% of the heritability of all protein-coding genes.

Then, we compared the heritability enrichments of protein-coding genes in all significant, the most significant half, and the most significant quarter of the cECS-significant genes among the 42 phenotypes (Figure 5B). The heritability enrichments among the three groups of genes are highly correlated (Pearson's $r = 0.98$, $p = 3.1 \times 10^{-30}$ for all vs. the most significant half; Pearson's $r = 0.92$, $p = 3.1 \times 10^{-18}$ for all vs. the most significant quarter). The highest five heritability enrichments in all cECS-significant genes are observed for balding (enrich 78.6 times), hypothyroidism (enrich 83.9 times), age of menopause (enrich 96.3 times), vitamin D (VD) (enrich 105 times), and diabetes (enrich 111 times). The lowest five heritability enrichments in all cECS-significant genes are observed for hip circumference (enrich 18.3 times), white blood cell count (WBC) (enrich 21.1 times), height (enrich 21.7 times), college or university degree (enrich 23.7 times), and right-hand grip strength (enrich 25.5 times). A higher heritability enrichment in cECS-significant genes means that a small fraction of genes have a greater effect on a

phenotype, which indicates that the phenotype is less polygenic. In contrast, a lower heritability enrichment indicates that the phenotype is more polygenic.

Finally, we asked whether the heritability of a complex phenotype is more enriched in pleiotropic genes. We collected cECS-significant protein-coding genes for each of the 42 phenotypes. There are 1,164 genes affecting three or more phenotypes, 433 genes affecting five or more phenotypes, and 199 genes affecting seven or more phenotypes. The heritability enrichments of cECS-significant genes in pleiotropic genes affecting different numbers of phenotypes are highly correlated (Pearson's $r = 0.80$, $p = 3.0 \times 10^{-10}$ for genes affecting ≥ 3 vs. ≥ 5 phenotypes; Pearson's $r = 0.87$, $p = 9.6 \times 10^{-14}$ for genes affecting ≥ 5 vs. ≥ 7 phenotypes; Figure 5C). The average heritability enrichments of cECS-significant genes in genes affecting ≥ 7 phenotypes (enrich 1.17 ± 0.29 times) is slightly larger than that in genes affecting ≥ 5 phenotypes (enrich 1.14 ± 0.23 times), which in turn is also only slightly larger than that in genes affecting ≥ 3 phenotypes (enrich 1.11 ± 0.14 times). However, none of the differences is statistically significant ($p > 0.05$ for all pairs of Mann-Whitney U-tests).

Discussion

The present study introduces two important concepts for heritability estimation. First, we established a clear relationship between the marginal heritability[32,33] at individual SNPs and the overall heritability[8,10,11] of a genomic region. While there are multiple methods for estimating these two types of heritability, they have not been explicitly linked. Our finding that the overall heritability can be approximated by a linear combination of the marginal heritability is significant. The linear combination cleans redundant components in the marginal heritability due to LD, enabling the overall heritability to be interpreted as the non-redundant component of the summation of the marginal heritability values of all SNPs (named effective heritability). Furthermore, our explicit linear combination-based estimator of heritability is more accurate than widely used LD score regression-based methods,[8,10] leading to a straightforward and reliable calculation of SEs. Second, we proposed to estimate conditional heritability for regions or genes. This is important for local heritability estimation, as many regions are in LD, and some estimated heritability may only be "hitching a ride" on true heritability in nearby regions. Conditional heritability can remove indirect heritability. We also derived the theoretical SE of conditional heritability estimates and proposed a framework using the selective expression of tissues or cell types to supervise the conditional procedure, similar to our previous methods for estimating driver tissues or cell types of complex phenotypes.[18,19]

In addition to the two concepts introduced in this study, EHE is technically distinct from existing methods for local heritability estimation, such as HESS.[12] EHE is a linear combination of marginal heritability estimates of SNPs, while HESS is a quadratic form of marginal effect estimates of SNPs. However, a potential issue with using effect sizes is that the sign of an effect size depends on the genotype encoding in the GWAS sample, and inconsistencies between the encoding in GWASs and the encoding used for the LD reference panel can lead to inaccurate estimation. Since heritability is independent of genotype encoding, EHE does not need to consider the sign of LD coefficients. The marginal heritability used by EHE can be easily obtained from the GWAS p values of individual SNPs. Although both EHE and HESS calculate the Moore-Penrose inverse once, the simpler algebraic operation makes EHE more robust against random noises in the LD matrix and test statistics than HESS (Figure 1).

The EHE method proposed in this study is specifically designed to estimate the heritability of local genomic regions, rather than the entire genome or chromosomes. This is because the number of SNPs used for local heritability estimation is much smaller than that used for whole-genome heritability estimation, which makes the widely used LD score regression-based model[8,23] unsuitable. In contrast, EHE relies on the accurate estimation of marginal heritability for a limited number of SNPs and can reliably calculate local heritability and SE using a simple algebraic operation. Furthermore, we demonstrated that EHE can also be used to estimate conditional heritability to prioritize adjacent genomic regions, which is crucial for genetic fine-mapping analyses. However, the EHE method may not be appropriate for estimating the heritability of the entire genome because the aggregation of the heritability of genomic regions may lead to incorrect estimates due to accumulated estimation errors. Hence, we suggest that EHE may be better suited for gene-based or regional heritability estimation, especially for conditional heritability.

By applying EHE to various complex phenotypes, we uncovered intriguing patterns in the genetic spectrum of human phenotypes. Our analysis revealed two distinct genetic architectures among the 42 phenotypes examined. While some phenotypes, like hypertension and height, displayed high polygenicity without genes exceeding certain heritability thresholds (e.g., 0.16% or 0.5%), others, such as sex hormone-binding globulin and hypothyroidism, exhibited substantial heritability (>1%) attributed to specific genes. These findings suggest that for certain complex phenotypes, it may be easier to identify candidate genes for precise diagnosis, treatment, and potential druggable targets, while for others, a limited number of genes may be insufficient to fully understand the phenotype, requiring a different approach to precision medicine. Additionally, the presence of LD adds complexity to the observed genetic architectures. In most genomic regions, many genes exhibit significant heritability estimates for a given phenotype. However, it is crucial to consider that many of these genes likely reside in LD with the true causal genes. So, it is imperative to identify and exclude genes with redundant heritability estimates,

ensuring a more accurate delineation of the genes truly contributing to the phenotype. Fortunately, EHE's conditional heritability estimation offers an efficient solution to this challenge. Furthermore, our analysis demonstrated that highly pleiotropic genes, typically assumed to have higher heritability due to their connectivity and criticality in biological networks,[34,35] do not necessarily exhibit greater heritability than other genes associated with the same phenotypes. This unexpected finding suggests a sophisticated mechanism for developing complex diseases, where diverse pleiotropic genes contribute to phenotypes through intricate interactions and regulatory mechanisms. Overall, the application of EHE offers valuable insights into the genetic architecture of human phenotypes, including the distinction between genetic architectures, the impact of LD, and the complex relationship between gene pleiotropy and heritability, thus enhancing our understanding of complex diseases and informing precision medicine strategies.

It is important to note that addressing population stratification in a local region is more challenging than in the whole genome. EHE does not have parameters specifically designed to address population stratification as in LD-score-based regression models.[8,23] The presence of population stratification can lead to inflated chi-squared statistics and overestimated local heritability. However, in practice, when the phenotype is not too polygenic,[36,37] the marginal heritability of SNPs can be obtained by generating chi-squared statistics corrected by genomic control. Additionally, the gene-based heritability estimated by EHE can be adjusted by the inflation factor estimated by the global heritability estimator (e.g., LDER[10] and LDSC[8]). Therefore, an optional parameter on KGGSEE is provided to correct the input chi-squared statistics by an inflation factor before calculating the marginal heritability. It should be acknowledged that adjusting for population stratification is also routinely performed during GWASs by including principle components as covariates[38,39] to produce more reliable p values and chi-squared statistics in each GWAS cohort. Therefore, the limitation of EHE not having a specific parameter for addressing population stratification would not discourage its application to GWAS data.

Data and code availability

All algorithms in the paper have been implemented in a Java package, KGGSEE. KGGSEE is publicly available at http://pmglab.top/kggsee. Materials capable of repeating the simulation analyses in this study are available at https://github.com/PMGLab/EHEsim. Scripts capable of repeating the empirical analysis of Figure 4 are available at https://github.com/PMGLab/EHEemp.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.08.006.

Acknowledgments

This work was funded by National Natural Science Foundation of China (32170637 and 31970650) and Guangdong project (2017GC010644), and Guangdong Provincial People's Hospital High-level Full-time Talent Program of Peak Plan Special Fund (KJ012021008).

Author contributions

Conceptualization: L.J., P.C.S., and M.L.; Methodology: L.M. and B.T.; Software: M.L.; Formal Analysis: L.M. and L.J.; Writing – original draft: L.M., L.J., and M.L.; Writing – review & editing: L.M., L.J., B.T., P.C.S., and M.L.; Funding acquisition: L.J. and M.L.; Supervision: M.L.

Declaration of interests

The authors declare no competing interests.

Received: May 11, 2023
Accepted: August 10, 2023
Published: August 25, 2023

Web resources

GCTA, https://yanglab.westlake.edu.cn/software/gcta/#Overview
GWAS results of UK Biobank by the Neale Lab, http://www.nealelab.is/uk-biobank
Haplotypes of 1000 Genomes Project Phase3 of version 5, https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/
HESS, https://huwenboshi.github.io/hess/
KGGSEE, http://pmglab.top/kggsee/#/
LDAK, https://dougspeed.com/
LDER, https://github.com/shuangsong0110/LDER
LDSC, https://github.com/bulik/ldsc
PLINK, https://www.cog-genomics.org/plink/
Statsmodels, https://www.statsmodels.org/stable/index.html

References

1. Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era–concepts and misconceptions. Nat. Rev. Genet. *9*, 255–266. https://doi.org/10.1038/nrg2322.
2. Falconer, D., and Mackay, T. (1996). Introduction to Quantitative Genetics (Longman Group Ltd.).
3. Rice, T.K., and Borecki, I.B. (2001). Familial resemblance and heritability. Adv. Genet. *42*, 35–44. https://doi.org/10.1016/s0065-2660(01)42013-x.
4. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569. https://doi.org/10.1038/ng.608.
5. 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.
6. Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., Ekoru, K., Nsubuga, R.N., Ssenyomo, G., Kamali, A., et al. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation.

Proc. Natl. Acad. Sci. USA *113*, 7377–7382. https://doi.org/10.1073/pnas.1510497113.

7. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., 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.
8. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., 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.
9. Ning, Z., Pawitan, Y., and Shen, X. (2020). High-definition likelihood inference of genetic correlations across human complex traits. Nat. Genet. *52*, 859–864. https://doi.org/10.1038/s41588-020-0653-y.
10. Song, S., Jiang, W., Zhang, Y., Hou, L., and Zhao, H. (2022). Leveraging LD eigenvalue regression to improve the estimation of SNP heritability and confounding inflation. Am. J. Hum. Genet. *109*, 802–811. https://doi.org/10.1016/j.ajhg.2022.03.013.
11. 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.
12. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. *99*, 139–153. https://doi.org/10.1016/j.ajhg.2016.05.013.
13. Li, M., Jiang, L., Mak, T.S.H., Kwan, J.S.H., Xue, C., Chen, P., Leung, H.C.M., Cui, L., Li, T., and Sham, P.C. (2019). A powerful conditional gene-based association approach implicated functionally important genes for schizophrenia. Bioinformatics *35*, 628–635. https://doi.org/10.1093/bioinformatics/bty682.
14. Berrandou, T.-E., Balding, D., and Speed, D. (2023). LDAK-GBAT: Fast and powerful gene-based association testing using summary statistics. Am. J. Hum. Genet. *110*, 23–29. https://doi.org/10.1016/j.ajhg.2022.11.010.
15. Speed, D., and Balding, D.J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat. Genet. *51*, 277–284. https://doi.org/10.1038/s41588-018-0279-5.
16. Li, X., Jiang, L., Xue, C., Li, M.J., and Li, M. (2022). A conditional gene-based association framework integrating isoform-level eQTL data reveals new susceptibility genes for schizophrenia. Elife *11*, e70779. https://doi.org/10.7554/eLife.70779.
17. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305. https://doi.org/10.1016/j.ajhg.2011.02.002.
18. Jiang, L., Xue, C., Dai, S., Chen, S., Chen, P., Sham, P.C., Wang, H., and Li, M. (2019). DESE: estimating driver tissues by selective expression of genes associated with complex diseases or traits. Genome Biol. *20*, 233. https://doi.org/10.1186/s13059-019-1801-5.
19. Xue, C., Jiang, L., Zhou, M., Long, Q., Chen, Y., Li, X., Peng, W., Yang, Q., and Li, M. (2022). PCGA: a comprehensive web server for phenotype-cell-gene association analysis. Nucleic Acids Res. *50*, W568–W576. https://doi.org/10.1093/nar/gkac425.
20. Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35*. D61-D65. https://doi.org/10.1093/nar/gkl842.
21. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). Gencode 2021. Nucleic Acids Res. *49*. D916-D923. https://doi.org/10.1093/nar/gkaa1087.
22. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
23. Hou, K., Burch, K.S., Majumdar, A., Shi, H., Mancuso, N., Wu, Y., Sankararaman, S., and Pasaniuc, B. (2019). Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. Nat. Genet. *51*, 1244–1251. https://doi.org/10.1038/s41588-019-0465-0.
24. Speed, D., Cai, N., UCLEB Consortium, Johnson, M.R., Nejentsev, S., and Balding, D.J. (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. *49*, 986–992. https://doi.org/10.1038/ng.3865.
25. Montana, G. (2005). HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. Bioinformatics *21*, 4309–4311. https://doi.org/10.1093/bioinformatics/bti689.
26. Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., and Liu, X. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. Dev. Reprod. Biol. *19*, 619–628. https://doi.org/10.1016/j.gpb.2020.10.007.
27. Dempster, E.R., and Lerner, I.M. (1950). Heritability of Threshold Characters. Genetics *35*, 212–236. https://doi.org/10.1093/genetics/35.2.212.
28. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.
29. 1000 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. https://doi.org/10.1038/nature15393.
30. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center LDACC—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx eGTEx groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.
31. 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.
32. Yang, J., Wray, N.R., and Visscher, P.M. (2010). Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. Genet. Epidemiol. *34*, 254–257. https://doi.org/10.1002/gepi.20456.

33. Lynch, M., and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits (Sinauer).
34. Hämälä, T., Gorton, A.J., Moeller, D.A., and Tiffin, P. (2020). Pleiotropy facilitates local adaptation to distant optima in common ragweed (Ambrosia artemisiifolia). PLoS Genet. *16*, e1008707. https://doi.org/10.1371/journal.pgen.1008707.
35. Promislow, D.E.L. (2004). Protein networks, pleiotropy and the evolution of senescence. Proc. Biol. Sci. *271*, 1225–1234. https://doi.org/10.1098/rspb.2004.2732.
36. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004. https://doi.org/10.1111/j.0006-341x.1999.00997.x.
37. Zheng, G., Freidlin, B., and Gastwirth, J.L. (2006). Robust genomic control for association studies. Am. J. Hum. Genet. *78*, 350–356. https://doi.org/10.1086/500054.
38. 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.
39. Li, Q., and Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet. Epidemiol. *32*, 215–226. https://doi.org/10.1002/gepi.20296.