nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-020-0653-y

High-definition likelihood inference of genetic correlations across human complex traits

Zheng Ning[1,2], Yudi Pawitan[2] and Xia Shen[1,2,3] ✉

Genetic correlation is a central parameter for understanding shared genetic architecture between complex traits. By using summary statistics from genome-wide association studies (GWAS), linkage disequilibrium score regression (LDSC) was developed for unbiased estimation of genetic correlations. Although easy to use, LDSC only partially utilizes LD information. By fully accounting for LD across the genome, we develop a high-definition likelihood (HDL) method to improve precision in genetic correlation estimation. Compared to LDSC, HDL reduces the variance of genetic correlation estimates by about 60%, equivalent to a 2.5-fold increase in sample size. We apply HDL and LDSC to estimate 435 genetic correlations among 30 behavioral and disease-related phenotypes measured in the UK Biobank (UKBB). In addition to 154 significant genetic correlations observed for both methods, HDL identified another 57 significant genetic correlations, compared to only another 2 significant genetic correlations identified by LDSC. HDL brings more power to genomic analyses and better reveals the underlying connections across human complex traits.

Estimating genetic correlation is a key step toward understanding the shared genetic architecture between complex traits and diseases. The genetic correlation parameter describes how genome-wide genetic effects align between two complex phenotypes. To estimate genetic correlations using GWAS data, there are two widely used approaches—when individual-level data are available, genetic correlation is commonly estimated by restricted maximum likelihood (REML) for linear mixed models (LMMs)[1,2]; when only GWAS summary-level data are available, LDSC[3,4] can be used. A major appeal of summary statistics is their wide availability for many traits without the need to access individual-level data. As using GWAS summary statistics is more straightforward and computationally light, LDSC has been widely applied since its inception[5].

Although easy to use, the standard error (s.e.) values of genetic correlation estimates by LDSC are substantially larger than those of REML[4,6], affecting the power and precision in the detection and estimation of genetic correlations. This accuracy gap is often attributed to the mismatch between the GWAS sample and the reference sample from which the LD scores are estimated[7]. This mismatch introduces measurement errors into the LD scores and, consequently, decreases the accuracy of estimation. However, even when the GWAS sample and the reference sample are matched, the accuracy of LDSC is still evidently lower than that of REML[6].

In this report, we introduce an essential source that reveals the 'missing accuracy' of LDSC: LDSC uses only a small part of the LD information in the modeling of summary association statistics. To thoroughly exploit the information from GWAS summary-level data, we develop HDL, a full likelihood-based method for estimating genetic correlation using GWAS summary statistics. The full likelihood naturally extends the regression formula of LDSC. We compare the accuracy of HDL and LDSC based on simulated and real data from the UKBB[8]. We find that HDL is more accurate than LDSC, with a relative efficiency (ratio of estimator variance, which is equivalent to the ratio of sample size) of more than 2.5 in simulations. This leads to higher statistical power to detect genetic correlations between phenotypes and also more precise estimates. For the real data, of the 435 tests for genetic correlations across 30 behavioral and disease-related phenotypes, 57 significant genetic correlations were identified by HDL only, compared with 2 significant genetic correlations by LDSC only.

Results

Overview of methods. HDL is a natural extension of LDSC. LDSC is based on the fact that, for a polygenic trait, if a SNP is in higher LD with other SNPs, it will have a higher χ^2 test statistic on average due to more causal variants being tagged. Mathematically, under a polygenic model[9] where true genetic effects are normally distributed and population stratification is absent (Supplementary Note), for a single SNP j, the variance of its GWAS test statistic z_j is related to its LD with other SNPs as

$$\mathrm{Var}\left[z_j\right] = \mathrm{E}\left[z_j^2\right] = \frac{Nh^2}{M} l_{jj} + 1 \quad (1)$$

where N is the sample size; h^2 is the narrow-sense heritability; M is the number of SNPs; and $l_{jj} = \sum_{k=1}^{M} r_{jk} r_{kj} = \sum_{k=1}^{M} r_{jk}^2$ is defined as the LD score of j. LDSC is then developed using this relationship between the LD score of a single SNP and the variance of its test statistic.

In fact, not just the variance of the single-SNP test statistic but the whole variance–covariance matrix is determined from the LD matrix. For any two SNPs j and j', the covariance or expected product of z_j and $z_{j'}$ is given by

$$\mathrm{Cov}\left[z_j, z_{j'}\right] = \mathrm{E}\left[z_j z_{j'}\right] = \frac{Nh^2}{M} l_{jj'} + r_{jj'} \quad (2)$$

where $r_{jj'}$ is the LD between SNP j and SNP j' and $l_{jj'} = \sum_{k=1}^{M} r_{jk} r_{kj'}$. When $j=j'$, equation (2) becomes equation (1); the derivation is

[1]Biostatistics Group, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [3]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK. ✉e-mail: xia.shen@ed.ac.uk

shown in the Supplementary Note. To rewrite equation (2) in general matrix form, denoting the $M \times M$ full LD matrix as $\mathbf{R}$ with entries $\{r_{jj'}\}$, we defined the LD score matrix $\mathbf{L} := \mathbf{R}'\mathbf{R}$ with entries $\{l_{jj'}\}$. Then, for the vector of test statistics $\mathbf{z}$, its covariance matrix is given as

$$\mathrm{Cov}[\mathbf{z}] = \frac{Nh^2}{M}\mathbf{L} + \mathbf{R} \quad (3)$$

Note that the M diagonal elements of $\mathbf{L}$ are exactly the same as the LD scores of the M SNPs, and the M diagonal elements of $\mathrm{Cov}[\mathbf{z}]$ are the expected values of χ^2 statistics. Therefore, LDSC is actually a method of moments that only uses the diagonal information in equation (3).

For two traits, assuming the true genetic effects follow a joint normal distribution (Supplementary Note), LDSC can estimate their genetic covariance h_{12} based on

$$\mathrm{Cov}\left[z_{1j}, z_{2j}\right] = \mathrm{E}\left[z_{1j}z_{2j}\right] = \frac{\sqrt{N_1N_2}h_{12}}{M}l_{jj} + \frac{N_0(h_{12}+\rho_{12})}{\sqrt{N_1N_2}} \quad (4)$$

where z_{1j} and z_{2j} are the z-scores for a single SNP j from two studies of trait 1 and trait 2, respectively; N_i is the sample size of study i; N_0 is the overlapping sample size; and ρ_{12} is the residual covariance. Similarly to the extension in the one-trait scenario, equation (4) can be extended to

$$\mathrm{Cov}[\mathbf{z}_1, \mathbf{z}_2] = \frac{\sqrt{N_1N_2}h_{12}}{M}\mathbf{L} + \frac{N_0(h_{12}+\rho_{12})}{\sqrt{N_1N_2}}\mathbf{R} \quad (5)$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are z-score vectors of the M SNPs from two studies of trait 1 and trait 2, respectively. Under the same assumption of normality as for LDSC, from the likelihood based on equations (3) and (5), HDL exploits the information within the whole $\mathbf{L}$ matrix and the covariance matrix of z-scores, not only the information in their diagonal elements as used by LDSC.

Normalizing genetic covariance by heritabilities gives genetic correlation. The literature has suggested that, for LDSC, the estimates of genetic correlations are less susceptible to bias than the estimates of heritabilities[4,6,7,10]. Although HDL improves accuracy in estimating both heritability and genetic correlation, we also focus on the estimation of genetic correlation in this report. Similarly to LDSC, HDL can be applied to quantitative traits and binary traits, regardless of whether the samples overlap.

Simulations. We performed a series of simulations to compare the performance of HDL and LDSC and to evaluate the robustness of HDL with respect to the choice of reference samples and model assumptions. The simulations were mainly based on the UKBB Axiom Array data from 336,000 British individuals in the UKBB. For consistency with the literature[4,11], we took SNPs with minor allele frequency (MAF) above 5%. Further quality-control steps resulted in 307,519 SNPs (Methods). For both HDL and LDSC, the LD matrix was computed using these 307,519 SNPs from 336,000 individuals. Of these, a proportion of SNPs were randomly selected as causal variants. In each simulation replicate, to generate two phenotypes for genetic correlation estimation, we first drew true effect sizes of each causal variant from a bivariate normal distribution. Thereafter, the phenotypic values were generated by adding errors from another bivariate normal distribution. The summary statistics were then computed by genome-wide association analysis of the simulated phenotypes against the genotypes.

Figure 1 shows the genetic correlation estimates from 100 simulations where 30,752 (10% of 307,519) SNPs were causal. The true genetic correlation was set to 0.5. For both high- and low-heritability pairs of traits, HDL produced unbiased and more accurate estimates

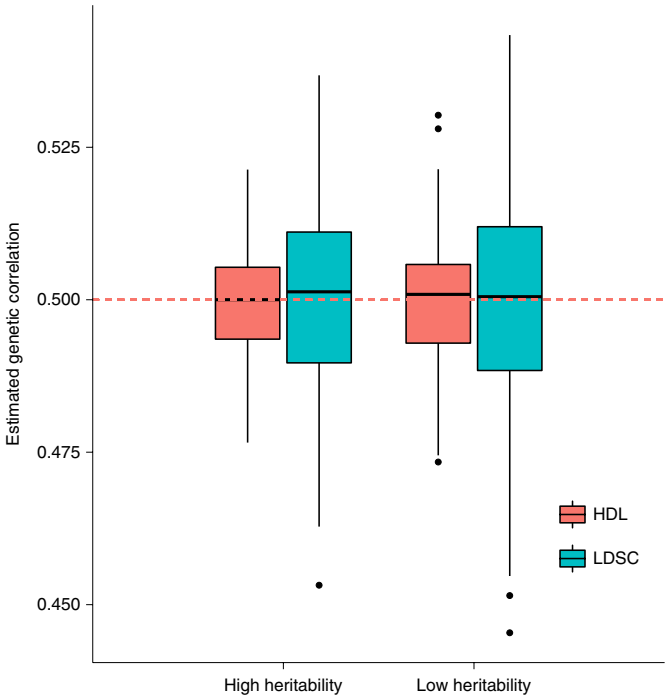


Fig. 1 | Relative efficiency of HDL against LDSC when 10% of SNPs are causal. Overall, 30,752 of 307,519 SNPs were randomly selected as causal variants. In each group, 100 replicates were simulated, where the true genetic and phenotypic correlations were both set to 0.5 for each pair of traits. In the high-heritability group, the heritability of the two traits was set to 0.6 and 0.8; in the low-heritability group, the heritability of the two traits was set to 0.2 and 0.4. Both HDL and LDSC were based on the LD matrix computed from 307,519 array SNPs from 336,000 individuals in the UKBB. Inside each box, the horizontal line represents the median, the central box indicates the interquartile range (IQR) and whiskers extend up to 1.5 times the IQR.

than those of LDSC. The relative efficiency was 2.58 (Levene's test, P value $= 7.1 \times 10^{-5}$) for high-heritability traits (with heritabilities of 0.6 and 0.8) and 2.93 (Levene's test, P value $= 1 \times 10^{-5}$) for low-heritability traits (with heritabilities of 0.2 and 0.4). The s.e. values from block jackknifing were consistent with the observed s.d. values (Supplementary Table 1). To further compare HDL and LDSC, we performed simulations when (1) all of the SNPs were simulated to be causal (Extended Data Fig. 1) and (2) model assumptions were violated (Extended Data Figs. 2 and 3). To compare HDL and LDSC when a large set of imputed SNPs were used as the reference panel, we first built an imputed reference panel based on 1,029,876 quality-controlled HapMap3 SNPs (Methods); we next simulated true phenotypes using these SNPs and then implemented HDL and LDSC, both using the imputed reference panel (Extended Data Fig. 4). Under all scenarios, the relative efficiency was around 2 or above.

Application to summary statistics from the UKBB. With higher efficiency, we can estimate genetic correlations more accurately and obtain higher statistical power to detect genetic correlations between phenotypes. To illustrate this using real data, we applied HDL and LDSC to estimate genetic correlations across 30 phenotypes in the UKBB. Most of the 30 phenotypes were behavioral traits, together with some disease-related and anthropometric traits. Based on our imputed reference panel including 1,029,876 quality-controlled HapMap3 SNPs, we obtained the genetic correlation estimates from HDL for the 435 pairwise combinations of

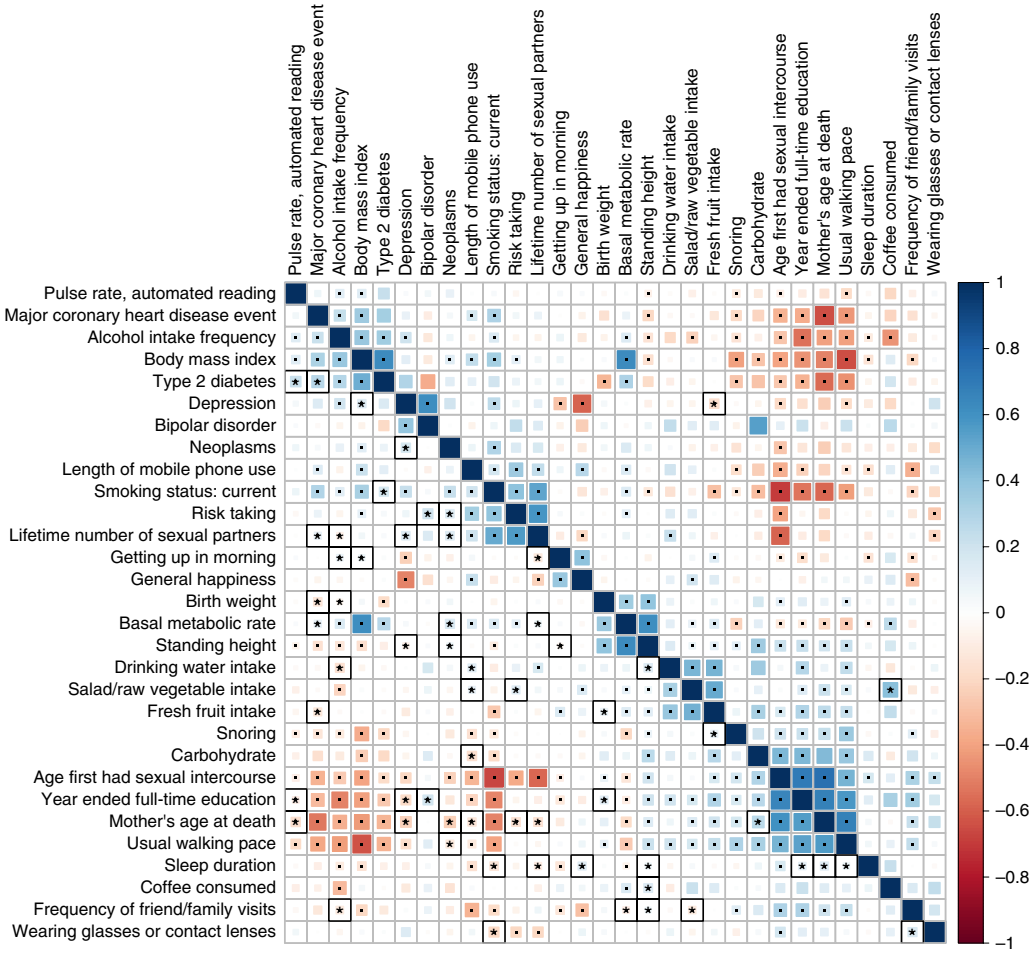


Fig. 2 | Genetic correlation estimates from HDL and LDSC among 30 phenotypes in the UKBB. Lower triangle: HDL estimates; upper triangle: LDSC estimates. The areas of the squares represent the absolute value of corresponding genetic correlations. After Bonferroni correction for 435 tests at a 5% significance level, genetic correlation estimates that were significantly different from zero in both methods (dot) and in only one method (asterisk and black square) are shown.

the 30 phenotypes and compared the results to the LDSC estimates (Fig. 2). For each pair of traits, the point estimates from the two methods were close. The s.e. values from HDL were generally (422 of 435) smaller than those from LDSC, with median relative efficiency = 2.35. The relative efficiency was positively correlated with the s.e. given by LDSC (Extended Data Fig. 5). The efficiency gains were larger among binary traits. Of the 435 tests for the genetic correlations (Supplementary Table 2), following Bonferroni correction ($P < 1.15 \times 10^{-4}$), 154 genetic correlations were significant for both methods, 57 correlations were significant for only HDL (Table 1) and 2 correlations were significant for only LDSC. Similar power gain was found when both HDL and LDSC used UKBB array SNPs as the reference panel (Extended Data Fig. 6).

Comparison with LMM results. An LMM fitted using individual-level data is known to be more accurate than LDSC in the estimation of heritability and genetic correlation[4,6]. If HDL has higher efficiency than LDSC, the gap of the genetic correlation estimates between HDL and LMM would be smaller than the gap between LDSC and LMM. To validate this, we extracted the results by Canela-Xandri et al.[11], where an LMM was fitted on UKBB individual-level data to estimate genetic correlations between hundreds of traits. Among the 30 traits analyzed, LMM-based results for 11 traits were available for comparison (Fig. 3 and Supplementary Table 3). For most pairs of traits, HDL estimates were close to LMM estimates ($R^2 = 0.80$), while LDSC estimates deviated more from LMM estimates ($R^2 = 0.67$).

Discussion

We have presented HDL, a full likelihood-based method for estimating genetic correlation using GWAS summary statistics. In contrast, LDSC uses only partial information based on the diagonal of the covariance matrix of *z*-scores. In both simulation and empirical applications, we have shown that HDL produces more accurate estimates than LDSC. As a result, HDL can detect more significant genetic correlations that might be missed by LDSC. Theoretically, the efficiency gain by HDL can be attributed to two reasons: (1) HDL uses more information on the relationship between test statistics and the LD structure than LDSC; and (2) likelihood-based methods such as HDL are more efficient than a method of moments such as LDSC when the underlying distributional assumption holds, which is typically the case for polygenic traits.

As an extension of LDSC, given that the underlying model is correct, HDL can also be used to quantify various properties. In single-trait HDL, the slope can be transformed to be an estimate of heritability (Extended Data Figs. 7 and 8), and the intercept evaluates population stratification; in double-trait HDL, the intercept implies phenotypic correlation and sample overlap. However, some concerns have been raised about estimating these quantities using LDSC[10,12–14]; therefore, we are cautious about interpreting the

Table 1 | Genetic correlation estimates significant in HDL but not in LDSC

Phenotype 1	Phenotype 2	r_g^{HDL} (s.e.)	r_g^{LDSC} (s.e.)	P_{HDL}	P_{LDSC}
Carbohydrate	Length of mobile phone use	−0.17 (0.03)	−0.24 (0.07)	1.2×10^{-6}	7.7×10^{-4}
Carbohydrate	Mother's age at death	0.26 (0.07)	0.43 (0.14)	1.0×10^{-4}	1.9×10^{-3}
Drinking water intake	Length of mobile phone use	0.12 (0.03)	0.20 (0.06)	4.6×10^{-5}	6.6×10^{-4}
Drinking water intake	Alcohol intake frequency	−0.15 (0.04)	−0.19 (0.06)	2.5×10^{-5}	2.6×10^{-3}
Drinking water intake	Standing height	0.13 (0.03)	0.14 (0.04)	3.6×10^{-7}	6.6×10^{-4}
Coffee consumed	Standing height	0.15 (0.03)	0.18 (0.06)	5.7×10^{-7}	2.9×10^{-3}
Pulse rate, automated reading	Year ended full-time education	−0.08 (0.02)	−0.10 (0.03)	1.8×10^{-5}	3.5×10^{-4}
Pulse rate, automated reading	Mother's age at death	−0.17 (0.03)	−0.15 (0.04)	4.5×10^{-8}	4.1×10^{-4}
Pulse rate, automated reading	Type 2 diabetes	0.21 (0.04)	0.23 (0.06)	1.8×10^{-8}	2.9×10^{-4}
Frequency of friend/family visits	Salad/raw vegetable intake	−0.11 (0.03)	−0.12 (0.04)	5.6×10^{-5}	1.6×10^{-3}
Frequency of friend/family visits	Alcohol intake frequency	−0.11 (0.02)	−0.11 (0.03)	1.2×10^{-8}	4.2×10^{-4}
Frequency of friend/family visits	Wearing glasses or contact lenses	0.16 (0.03)	0.18 (0.05)	3.4×10^{-6}	2.6×10^{-4}
Frequency of friend/family visits	Basal metabolic rate	−0.08 (0.02)	−0.09 (0.02)	3.5×10^{-7}	1.4×10^{-4}
Frequency of friend/family visits	Standing height	0.06 (0.01)	0.07 (0.02)	6.9×10^{-6}	2.0×10^{-3}
Length of mobile phone use	Salad/raw vegetable intake	0.09 (0.02)	0.10 (0.03)	3.4×10^{-5}	8.9×10^{-4}
Length of mobile phone use	Mother's age at death	−0.13 (0.03)	−0.21 (0.06)	2.3×10^{-6}	7.9×10^{-4}
Sleep duration	Smoking status: current	−0.14 (0.02)	−0.12 (0.03)	7.7×10^{-11}	6.8×10^{-4}
Sleep duration	General happiness	0.13 (0.03)	0.10 (0.04)	2.8×10^{-6}	1.5×10^{-2}
Sleep duration	Lifetime number of sexual partners	−0.10 (0.02)	−0.09 (0.03)	2.3×10^{-8}	5.2×10^{-3}
Sleep duration	Year ended full-time education	0.11 (0.02)	0.12 (0.03)	1.9×10^{-6}	1.2×10^{-4}
Sleep duration	Mother's age at death	0.13 (0.03)	0.05 (0.06)	7.7×10^{-5}	4.3×10^{-1}
Sleep duration	Standing height	0.07 (0.01)	0.05 (0.02)	2.4×10^{-8}	3.0×10^{-3}
Sleep duration	Usual walking pace	0.08 (0.01)	0.05 (0.02)	2.4×10^{-7}	2.8×10^{-2}
Getting up in morning	Alcohol intake frequency	0.08 (0.02)	0.08 (0.02)	4.9×10^{-6}	4.8×10^{-4}
Getting up in morning	Body mass index	0.07 (0.02)	0.07 (0.02)	8.9×10^{-6}	9.0×10^{-4}
Getting up in morning	Lifetime number of sexual partners	−0.12 (0.02)	−0.09 (0.03)	8.4×10^{-11}	7.1×10^{-4}
Getting up in morning	Standing height	−0.05 (0.01)	−0.06 (0.02)	5.8×10^{-5}	3.8×10^{-4}
Snoring	Fresh fruit intake	0.10 (0.02)	0.08 (0.03)	3.8×10^{-7}	2.8×10^{-3}
Salad/raw vegetable intake	Risk taking	0.12 (0.02)	0.13 (0.03)	2.7×10^{-7}	1.3×10^{-4}
Fresh fruit intake	Birth weight	0.09 (0.02)	0.06 (0.03)	6.7×10^{-6}	2.0×10^{-2}
Fresh fruit intake	Major coronary heart disease event	−0.12 (0.02)	−0.12 (0.04)	8.5×10^{-9}	2.0×10^{-3}
Alcohol intake frequency	Birth weight	−0.06 (0.01)	−0.06 (0.02)	3.9×10^{-6}	7.5×10^{-3}
Alcohol intake frequency	Lifetime number of sexual partners	−0.08 (0.02)	−0.06 (0.02)	3.9×10^{-6}	1.3×10^{-2}
Birth weight	Year ended full-time education	0.11 (0.02)	0.12 (0.03)	1.4×10^{-8}	1.5×10^{-4}
Birth weight	Major coronary heart disease event	−0.14 (0.03)	−0.15 (0.04)	7.4×10^{-8}	1.8×10^{-4}
Smoking status: current	Wearing glasses or contact lenses	−0.19 (0.03)	−0.18 (0.05)	5.1×10^{-10}	3.1×10^{-4}
Smoking status: current	Type 2 diabetes	0.16 (0.04)	0.19 (0.08)	8.4×10^{-5}	1.4×10^{-2}
Risk taking	Mother's age at death	−0.15 (0.04)	−0.19 (0.07)	4.4×10^{-5}	5.1×10^{-3}
Risk taking	Neoplasms	0.13 (0.03)	0.16 (0.05)	2.5×10^{-5}	2.6×10^{-3}
Risk taking	Bipolar disorder	0.19 (0.04)	0.25 (0.08)	3.5×10^{-6}	3.5×10^{-3}
Body mass index	Depression	0.13 (0.02)	0.11 (0.03)	8.7×10^{-9}	3.2×10^{-4}
Lifetime number of sexual partners	Basal metabolic rate	0.07 (0.01)	0.08 (0.02)	2.6×10^{-6}	1.8×10^{-4}
Lifetime number of sexual partners	Mother's age at death	−0.15 (0.03)	−0.20 (0.06)	3.5×10^{-6}	1.4×10^{-3}
Lifetime number of sexual partners	Major coronary heart disease event	0.10 (0.02)	0.08 (0.04)	4.1×10^{-6}	2.2×10^{-2}
Lifetime number of sexual partners	Neoplasms	0.14 (0.03)	0.16 (0.04)	2.8×10^{-7}	1.3×10^{-4}
Lifetime number of sexual partners	Depression	0.14 (0.03)	0.10 (0.04)	5.3×10^{-7}	1.5×10^{-2}
Year ended full-time education	Depression	−0.19 (0.04)	−0.17 (0.05)	4.4×10^{-7}	9.3×10^{-4}
Year ended full-time education	Bipolar disorder	0.19 (0.04)	0.22 (0.09)	7.6×10^{-6}	1.2×10^{-2}

Continued

Table 1 | Genetic correlation estimates significant in HDL but not in LDSC

Phenotype 1	Phenotype 2	r_g^{HDL} (s.e.)	r_g^{LDSC} (s.e.)	P_{HDL}	P_{LDSC}
Basal metabolic rate	Major coronary heart disease event	0.10 (0.02)	0.09 (0.03)	4.5×10^{-5}	2.6×10^{-3}
Basal metabolic rate	Neoplasms	0.16 (0.02)	0.16 (0.04)	4.7×10^{-16}	1.3×10^{-4}
Mother's age at death	Neoplasms	−0.24 (0.05)	−0.25 (0.09)	2.0×10^{-6}	4.1×10^{-3}
Mother's age at death	Depression	−0.22 (0.05)	−0.24 (0.09)	6.6×10^{-6}	7.6×10^{-3}
Standing height	Neoplasms	0.07 (0.02)	0.07 (0.04)	8.2×10^{-5}	6.0×10^{-2}
Standing height	Depression	−0.07 (0.02)	−0.08 (0.02)	8.8×10^{-5}	1.5×10^{-3}
Usual walking pace	Neoplasms	−0.12 (0.03)	−0.13 (0.04)	2.6×10^{-6}	9.9×10^{-4}
Major coronary heart disease event	Type 2 diabetes	0.28 (0.06)	0.33 (0.10)	9.2×10^{-6}	7.5×10^{-4}
Neoplasms	Depression	0.16 (0.04)	0.20 (0.07)	3.9×10^{-5}	3.1×10^{-3}

Results that passed Bonferroni correction (calculated by dividing the significance level by the number of tests, that is, 0.05/435) were reported as significant. r_g^{HDL} (s.e.), genetic correlation estimate and s.e. given by HDL; r_g^{LDSC} (s.e.), genetic correlation estimate and s.e given by LDSC; P_{HDL}, *P* value given by HDL; P_{LDSC}, *P* value given by LDSC.

intercept term and the single-trait HDL results, although HDL does improve heritability estimation (Extended Data Fig. 7). On the other hand, the LDSC estimates of genetic correlations have been shown to be unbiased under different circumstances[4,6,7,10]. This robustness has mainly been attributed to the ratio form of genetic correlation: the biases on the numerator and the denominator are in the same direction and therefore cancel out[4]. Given these considerations, we focused the application of HDL on estimating genetic correlations.

In application, the efficiency gain by HDL was more substantial when LDSC generated large s.e. values (Extended Data Fig. 5). This phenomenon was consistent with the simulation results—when the traits' heritabilities were low, LDSC s.e. values were larger and the relative efficiency was higher—indicating that it is more important to use the full LD information when the amount of genetic variance is limited. For example, as the observed heritabilities of binary traits are usually low, when they are involved in the genetic correlation estimation, the gain of HDL is higher (Extended Data Fig. 5). As diseases are mostly recorded as binary traits and are of interest in many GWAS projects and consortia, HDL would be more beneficial in such applications.

In some cases[15], the estimates of genetic correlations from LDSC are above 1. This is because the genetic covariance estimate is not constrained in the cross-trait LD score regression. Consequently, the randomness of genetic covariance estimates may result in a genetic correlation estimate above 1. HDL makes this less problematic by estimating heritability and genetic covariance parameters more precisely.

Although both the estimates from HDL and LDSC were compared to LMM estimates, it should be noted that, for binary phenotypes, LMM estimates were not used as the gold standard. The use of individual-level data allows LMMs to incorporate the full LD information; however, for binary outcomes, fitting a normal linear mixed model mis-specifies the likelihood function and thus is not optimal for statistical inference, while the HDL method models the GWAS test statistics whose distribution does not violate the normal assumption, even for binary outcomes. This is another theoretical advantage of applying HDL on summary association statistics for binary phenotypes.

Handling a large LD matrix requires numerical regularization. To regularize the LD matrix, instead of directly using the original LD matrix, we performed eigen-decomposition on the LD matrix and passed its top eigenvalues and eigenvectors to HDL. The selected eigenvalues and eigenvectors captured most of the information in the LD matrix (Extended Data Fig. 9). There are three benefits of this decomposition step: (1) improving the efficiency of HDL (Extended Data Fig. 10 and Supplementary Fig. 1), (2) saving computation time by avoiding matrix multiplication (Supplementary Note) and

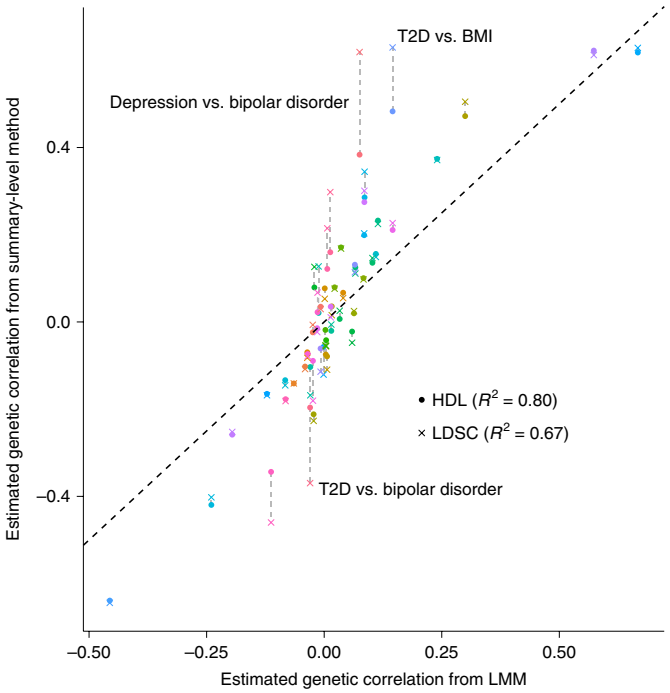


Fig. 3 | Comparing genetic correlation estimates from HDL and LDSC with those from LMMs across 11 phenotypes in the UKBB. HDL and LDSC estimates are shown as dots and crosses, respectively. For each pair of traits, genetic correlation estimates are in the same color and connected by a gray vertical dashed line. The black diagonal dashed line represents identity. BMI, body mass index; T2D, type 2 diabetes.

(3) saving storage space by only storing leading eigenvalues and eigenvectors for the reference panel that can be used across many GWAS summary-level data. Simulations suggest that taking the leading eigenvalues explaining 90% of the variance of the LD matrix has the highest estimation efficiency for the array SNP reference panel (Extended Data Fig. 10), and the top eigenvalues explaining 99% of the variance of the LD matrix have the highest estimation efficiency for the imputed SNP reference panel (Supplementary Fig. 1). Hence, in this report, when the array SNP reference panel was used, we implemented HDL based on the leading eigenvalues explaining 90% of the variance and their corresponding eigenvectors; when the imputed SNP reference panel was used, we implemented HDL based on the leading eigenvalues explaining 99% of

the variance and their corresponding eigenvectors. Note that, for heritability estimation, as mentioned above, consistent estimates are difficult to achieve for summary-statistics-based methods. For HDL, too little regularization of the LD matrix would lead to downward bias, whereas too much regularization would lead to lower estimation efficiency due to loss of information (Supplementary Fig. 2). Nevertheless, bias is not a concern for genetic correlation estimation (Supplementary Fig. 1).

In LDSC, 378 Europeans from the 1000 Genomes Project are often used as a reference sample to compute LD scores. However, because HDL uses more information from the LD matrix, a larger reference sample is preferred. Therefore, in the HDL software package, we took 336,000 genomic British individuals from the UKBB as a reference sample to compute the LD matrices and perform eigen-decomposition. These are stored in the software package so that the computation on user-input GWAS summary statistics is fast. In this report, the LD reference panel and GWAS summary statistics are both from UKBB, but in other applications this might not be the case. Hence, we performed a series of simulations to test the performance of HDL when the GWAS and reference samples were independent. In these simulations, we also evaluated the robustness of HDL under different scenarios where the LD matrix was (1) computed from different reference sample sizes (Supplementary Figs. 3 and 4) and (2) approximated by different numbers of its top eigenvalues and corresponding eigenvectors (Extended Data Fig. 10 and Supplementary Figs. 1 and 2). The results suggest that (1) when a large independent reference sample is used, HDL provides unbiased estimates of genetic correlation, the efficiency is almost equal to the efficiency when the GWAS sample and reference sample are identical and HDL is robust against the choice of top eigenvalues and corresponding eigenvectors; (2) HDL based on the leading eigenvalues explaining 90% of the variance still gives the optimal efficiency for the array SNP panel; and (3) when a small independent reference sample is used, HDL can still be unbiased but is less efficient and less robust against the choice of top eigenvalues and corresponding eigenvectors.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0653-y.

Received: 3 October 2019; Accepted: 26 May 2020;
Published online: 29 June 2020

References

1. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
2. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
3. Bulik-Sullivan, B. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
4. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
5. Zheng, J. et al. LD hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
6. Ni, G. et al. Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *Am. J. Hum. Genet.* **102**, 1185–1194 (2018).
7. Yang, J. et al. Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Hum. Mol. Genet.* **24**, 7445–7449 (2015).
8. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
9. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
10. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
11. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
12. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
13. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
14. Yengo, L., Yang, J. & Visscher, P. M. Expectation of the intercept from bivariate LD score regression in the presence of population stratification. Preprint at *bioRxiv* https://doi.org/10.1101/310565 (2018).
15. Ganna, A. et al. Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science* **365**, eaat7693 (2019).

Methods

Modeling and estimation of genetic correlation. Suppose there are two cohorts for two traits with sample sizes N_1 and N_2, where N_0 individuals are included in both cohorts. The number of SNPs is M in both cohorts. The z-score vector of the M SNPs from study i of trait i is denoted as $\mathbf{z}_i$. Then, under a polygenic model without population stratification[9], the covariance matrices are given as

$$\mathrm{Cov}[\mathbf{z}_i] = \frac{N_i h_i^2}{M}\mathbf{L} + \mathbf{R} \tag{6}$$

$$\mathrm{Cov}[\mathbf{z}_1, \mathbf{z}_2] = \frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{L} + \frac{N_0(h_{12}+\rho_{12})}{\sqrt{N_1 N_2}}\mathbf{R} \tag{7}$$

where $\mathbf{R}$ is the LD matrix of the M SNPs, $\mathbf{L} := \mathbf{R}'\mathbf{R}$ is the LD score matrix, h_i^2 is the narrow-sense heritability of trait i, h_{12} is the genetic covariance of the two traits and ρ_{12} is the environmental covariance. Denoting

$$\mathbf{\Sigma}_{ii} = \tfrac{N_i h_i^2}{M}\mathbf{L} + \mathbf{R}$$

$$\mathbf{\Sigma}_{12} = \tfrac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{L} + \tfrac{N_0(h_{12}+\rho_{12})}{\sqrt{N_1 N_2}}\mathbf{R}$$

based on equations (6) and (7), we have

$$\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{\Sigma}_{ii}) \tag{8}$$

$$\mathbf{z}_2|\mathbf{z}_1 \sim \mathcal{N}\left(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1, \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right) \tag{9}$$

Following equations (8) and (9), we used maximum likelihood to estimate h_1^2, h_2^2 and $r_g := h_{12}/\sqrt{h_1^2 h_2^2}$ (see the Supplementary Note for complete derivations).

The literature has shown that LDSC with a co nstrained intercept may produce substantially biased estimates[6,10], but LDSC with an unconstrained intercept is much more robust; therefore, in equations (6) and (7), we introduced parameters $\{c_{11}, c_{22}, c_{12}\}$ that were analogous to the unconstrained intercept in LDSC:

$$\mathrm{Cov}[\mathbf{z}_i] = \frac{N_i h_i^2}{M}\mathbf{L} + c_{ii}\mathbf{R} \tag{10}$$

$$\mathrm{Cov}[\mathbf{z}_1, \mathbf{z}_2] = \frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{L} + c_{12}\frac{N_0}{\sqrt{N_1 N_2}}\mathbf{R} \tag{11}$$

The diagonal elements in equations (10) and (11) are coincident with unconstrained-intercept LDSC. If the two traits are measured in the same study, and given that the underlying model is correct, $c_{12} = h_{12} + \rho_{12}$ will be the phenotypic correlation between the two traits. However, as mentioned in the Discussion, we should be cautious of interpreting the estimate of c_{12} in practice. Nevertheless, residual correlation does not have any obvious impact on the performance of HDL (Supplementary Fig. 5).

Quality control of UKBB genotype array data. In the UKBB, about 500,000 individuals aged 40–69 years were recruited between 2006 and 2010 from across the country. By March 2018, most of the participants had been genotyped on an Affymetrix chip including about 800,000 variants. Among the genotyped individuals, approximately 336,000 were identified as genetically unrelated white British individuals by the UKBB. These participants and their genotypes were taken forward. Because we used GWAS summary statistics by Neale's group ('Data availability') and compared HDL with LDSC, we took the SNPs overlapping between (1) UKBB array SNPs, (2) the list of SNPs for LDSC and (3) the SNPs in the GWAS from Neale's group to make a fair comparison when array SNPs were used as the reference panel. Following ref. [10] and LDSC, we excluded the major histocompatibility complex (MHC) region and SNPs with sample MAF below 5%. We further performed LD pruning and filtering on missing call rates with PLINK[16] using flags --geno 0.1 and --indep-pairwise 1000 5 0.95. We ended up with 307,519 autosomal SNPs for the analysis related to array SNPs in this report. For both simulation and application in which the reference panel consisted of array SNPs, the LD matrices used in HDL and LDSC were computed with 307,519 SNPs from 336,000 genetically unrelated white British individuals. This dataset was also used to simulate phenotypes in the simulation section whenever the comparison was based on array SNPs.

Quality control of imputed genotype data from the UKBB. When imputed SNPs were used as the reference panel, we took the SNPs overlapping between (1) the list of SNPs for LDSC and (2) the SNPs in the GWAS from Neale's group. We excluded SNPs that (1) were in the MHC region, (2) had sample MAF below 5%, (3) were multiallelic and (4) had an imputation quality < 0.9 and (5) had a call rate < 0.95. We converted the remaining genotype probabilities to hard calls for the construction of the LD reference. We ended up with 1,029,876 autosomal SNPs for the analysis related to imputed markers in this report. This panel was applied in HDL for analyses related to real UKBB GWAS summary statistics in the Results.

UKBB GWAS summary statistics. The UKBB GWAS summary statistics used in this report were from the second wave of results released in July 2018 by Neale's group. They performed association tests on the unrelated individuals of British ancestry for over 2,000 of the available phenotypes. For continuous traits, we took the GWAS version where phenotypes had been inverse rank normalized. We adjusted for the following covariates: age, age squared, inferred sex, age × inferred sex, age squared × inferred sex and principal components 1–20.

LDSC settings. When the reference panel consisted of array SNPs, the LD scores based on the 307,519 SNPs were computed using flags --l2 and --ld-wind-snps 500. We used 500-SNP windows to compute LD scores, because the LD matrix was computed by 500-SNP windows in HDL. Nevertheless, the LD scores computed by 500-SNP windows were highly consistent with those computed using a window size of 1 cM (Supplementary Fig. 6). When the reference panel consisted of imputed SNPs, the default 1000 Genomes panel was used. The estimation of genetic correlation was under the default setting with an unconstrained intercept. The same LD scores for both --w-ld-chr and --ref-ld-chr flags were used as recommended. For analyses related to real UKBB GWAS summary statistics in the Results, the default 1000 Genomes panel was applied.

Computational details of HDL. To speed up computation, we split the whole genome into pieces. When the reference panel consisted of array SNPs, each chromosome was on average cut into pieces with fewer than 10,000 SNPs, which resulted in 43 pieces for the whole genome. For each piece, we first banded its LD matrix into a band matrix with bandwidth 500. Then, we performed eigen-decomposition on the LD matrix and chose the leading eigenvalues explaining 90% of the variance and their correspondent eigenvectors (Extended Data Fig. 10). When the reference panel consisted of imputed SNPs, each chromosome was on average divided into pieces with fewer than 20,000 SNPs, resulting in 61 pieces for the whole genome. In eigen-decomposition, the leading eigenvalues explaining 99% of the variance and their correspondent eigenvectors were selected. After estimating heritabilities and genetic covariance for each piece, the piecewise results were integrated into one estimate for the whole genome. The s.e. of the genetic correlation estimate was computed via block jackknife with one piece left out (Supplementary Note).

Run times. When the leading eigenvalues and their corresponding eigenvectors of the LD matrices were available for loading, HDL took around 1.5 min to obtain the point estimate using 307,519 array SNPs as the reference panel on a single 2.8 GHz Intel Core i7 and another 4 min to obtain the s.e. values via jackknifing. When using 1,029,876 imputed markers as the reference panel on a single core, it took around 7 min to obtain the point estimate and another 8 min to get the s.e. values via jackknifing. The overall computation required about 1 GB of memory. When running in parallel with four threads, it took 5 min in total to acquire both the estimate and s.e. values.

Statistical testing. In simulations, Levene's test was used to assess the equality of variances between HDL estimates and LDSC estimates. As 100 replicates were simulated in each setting, the test statistic approximately followed an F distribution with 1 and 198 degrees of freedom. To test whether a genetic correlation was significantly different from 0, we used a two-sided Wald test with 1 degree of freedom.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The individual-level genotype and phenotype data are available by application from the UKBB (http://www.ukbiobank.ac.uk/). The UKBB GWAS summary statistics by the Neale laboratory can be obtained from http://www.nealelab.is/uk-biobank/. Source data are provided with this paper.

Code availability

HDL software is available at https://github.com/zhenin/HDL/. LDSC software is available at https://github.com/bulik/ldsc/. PLINK 2.0 (https://www.cog-genomics.org/plink/2.0/) was used to extract individual-level data of imputed SNPs from the UKBB. PLINK 1.9 (https://www.cog-genomics.org/plink/) and LDAK (http://dougspeed.com/ldak/) were used in LD correlation calculation and simulations.

References

16. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

Acknowledgements

We thank the UKBB resource, approved under application no. 14302 and 19655, for the individual-level genotype data used in LD correlation calculation and simulations. X.S. was in receipt of a Swedish Research Council starting grant (no. 2017-02543). Y.P. received a Swedish Research Council grant (no. 2016-04194). We thank the Edinburgh Compute and Data Facility (ECDF) for providing high-performance computing resources.

Author contributions

X.S. and Y.P. initiated and coordinated the study. Z.N. performed data analysis. All authors contributed to method development and manuscript writing.

Competing interests

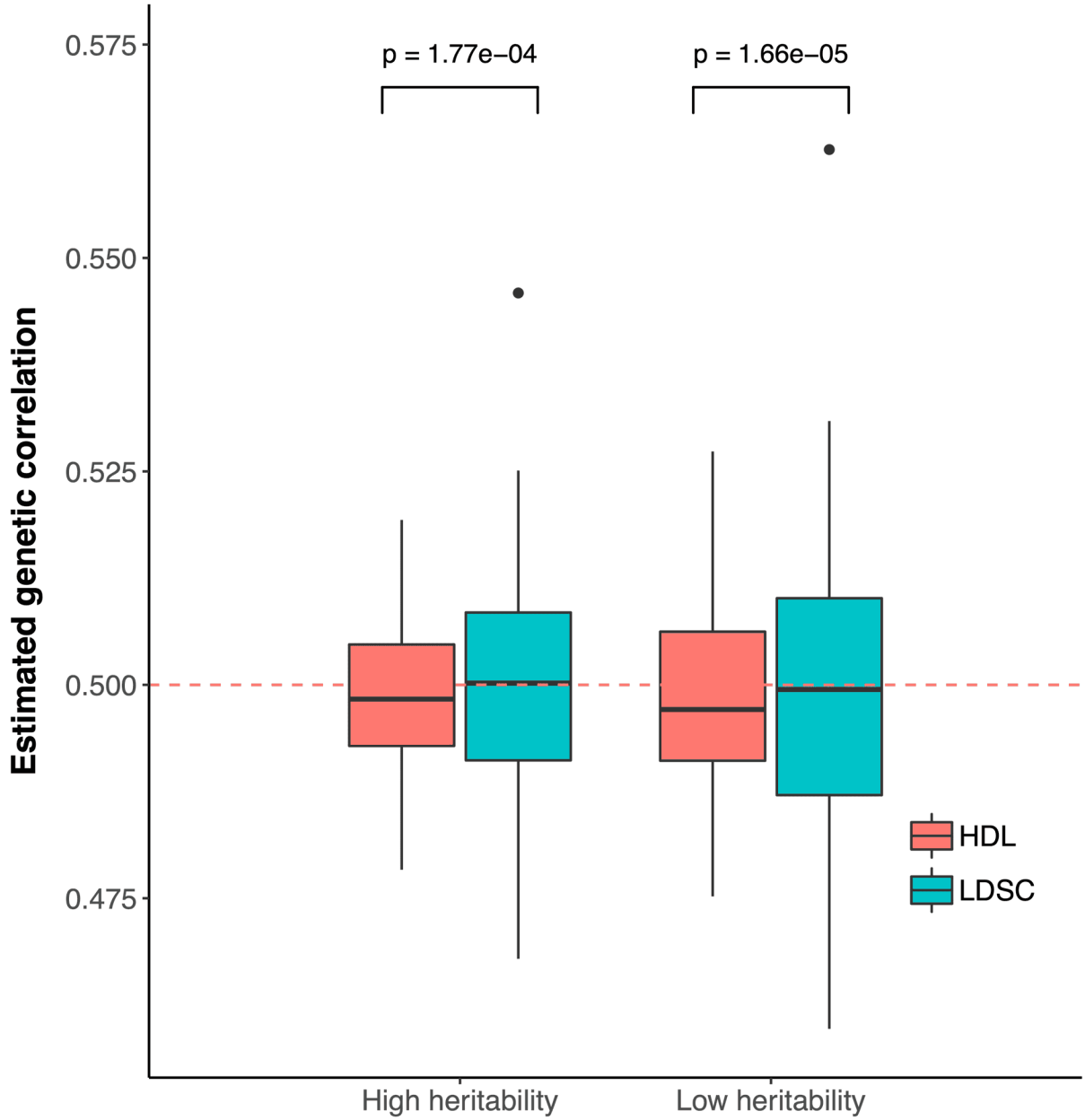
The authors declare no competing interests.

Additional information

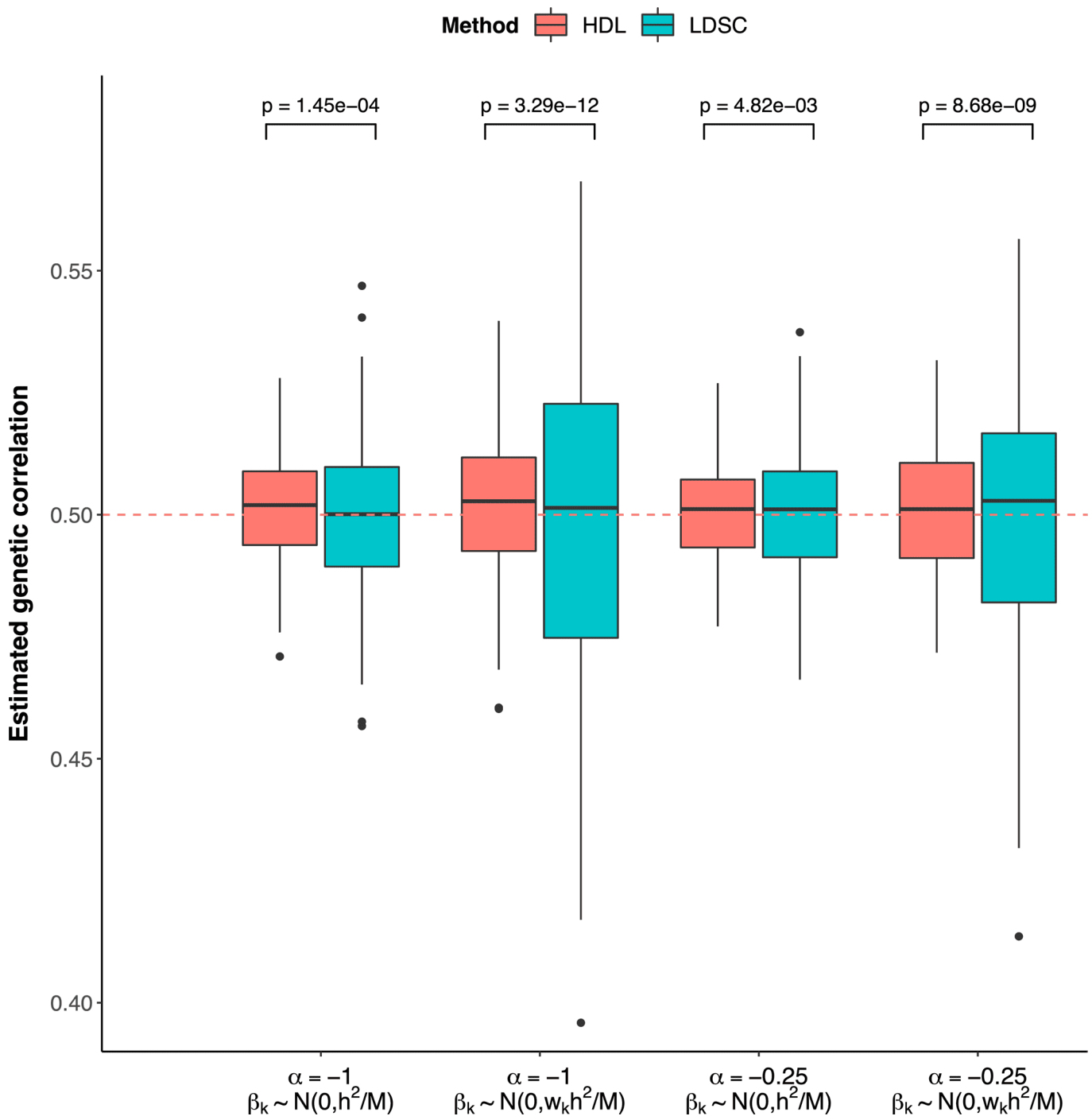
Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0653-y.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0653-y.

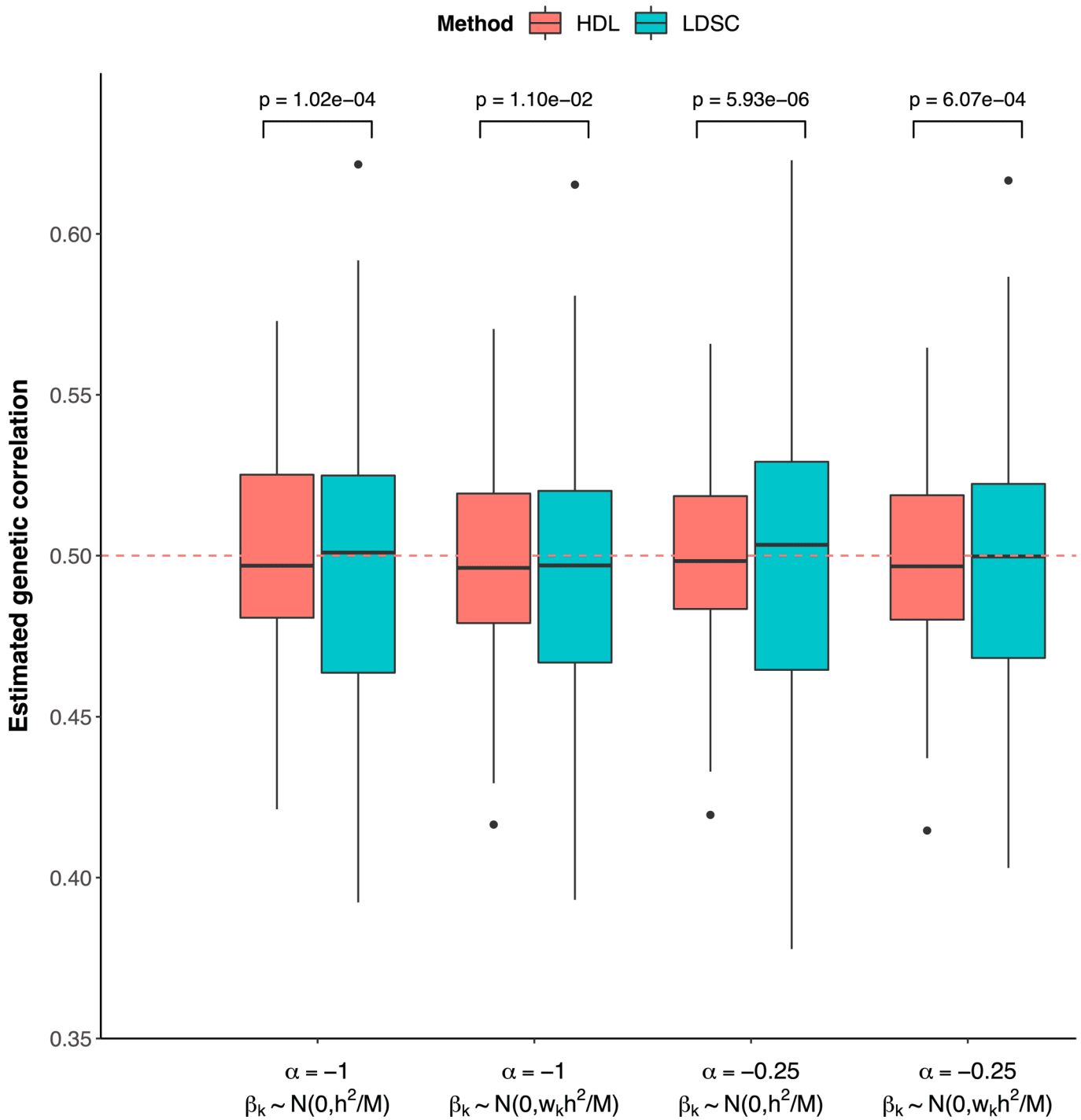
Correspondence and requests for materials should be addressed to X.S.



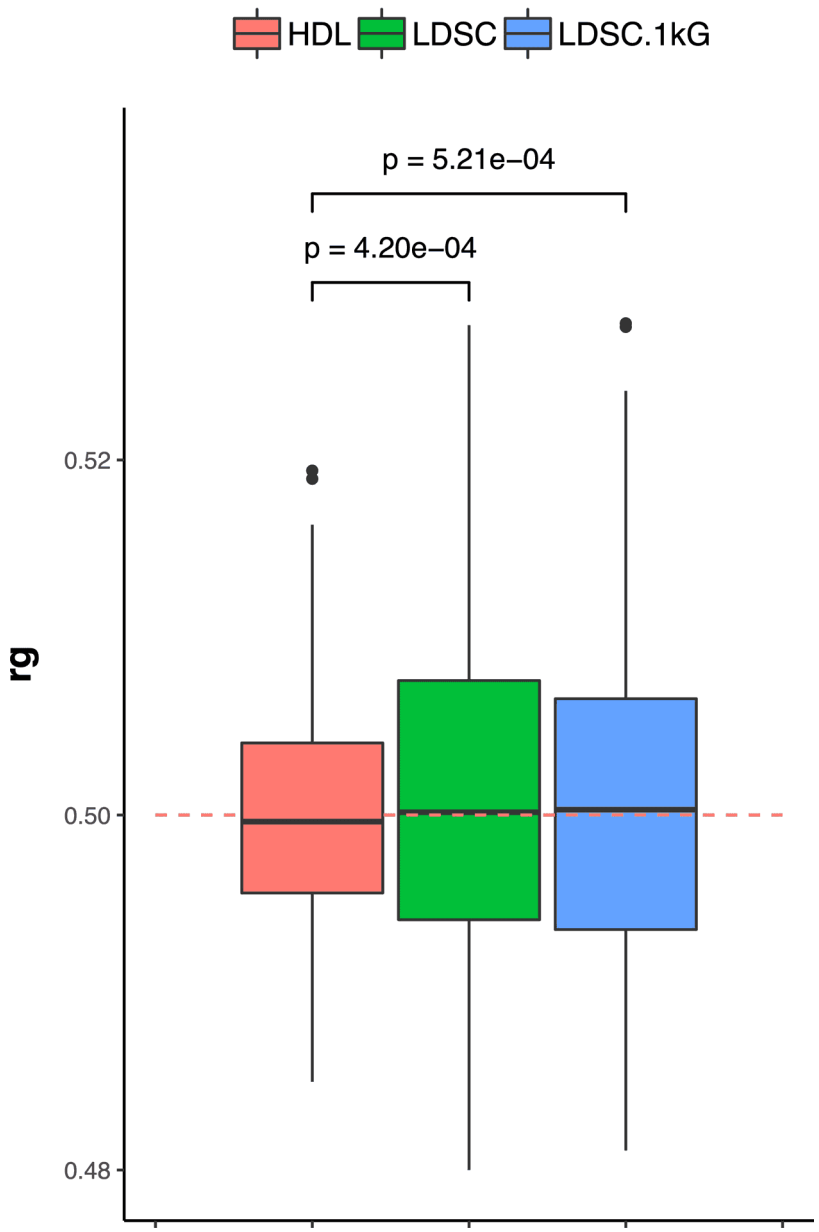
Extended Data Fig. 1 | Relative efficiency of HDL against LDSC when 100% SNPs are causal. In each heritability group, we generated 100 pairs of traits, where true genetic correlation and phenotypic correlation are 0.5. In the high heritability group, the heritability of the pair of traits is 0.6 and 0.8 separately; in the low heritability group, the heritability of the pair of traits is 0.2 and 0.4 separately. The 307,519 array SNPs of ~336,000 UKBB genomic British individuals were used to simulate true phenotypes and to compute the LD matrix for both HDL and LDSC. The P-values are from Levene's test for variance heterogeneity. Inside each box, the line indicates the median value, the central box indicates the interquartile range (IQR), and whiskers extend up to 1.5 times the IQR.



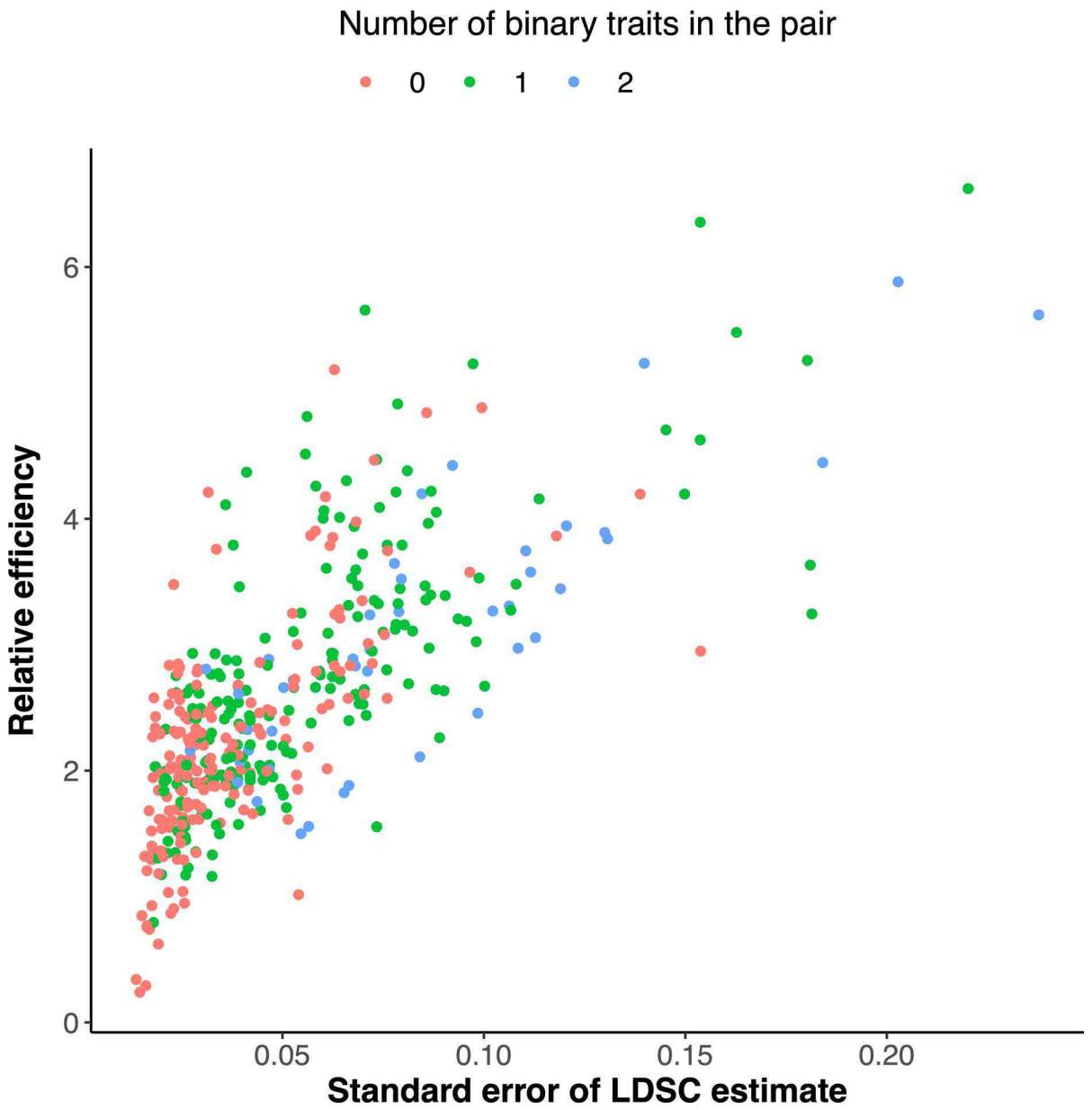
Extended Data Fig. 2 | Relative efficiency of HDL against LDSC under different model setups when 10% SNPs with MAF > 1% are causal. 52,914 out of 529,139 array SNPs with MAF > 1% were randomly selected as causal variants. 100 pairs of traits were generated, where true genetic correlation and phenotypic correlation are 0.5. The true phenotypes of trait *i* is generated from model $\mathbf{y}_i = \sum_{k=1}^{M} \mathbf{X}_{ik}\beta_{ik} + \epsilon_i$, where $\mathbf{X}_{ik} = (\mathbf{Z}_{ik} - 2p_k 1)[2p_k(1-p_k)]^{\alpha/2}$; $\mathbf{Z}_{ik}$ are the original genotypes of SNP *k* for trait *i*; p_k is the MAF of SNP *k*; M is the number of causal variants. Four scenarios were simulated: (1) $\alpha = -1$, and the marginal distribution of β_{ik} is $N(0, h_i^2/M)$; (2) $\alpha = -1$, and the marginal distribution of β_{ik} is $N(0, w_k h_i^2/M)$, where w_k is the LDAK weight of SNP *k* which is inversely proportional to its LD score; (3) $\alpha = -0.25$, and the marginal distribution of β_{ik} is $N(0, h_i^2/M)$ and (4) $\alpha = -0.25$, and the marginal distribution of β_{ik} is $N(0, w_k h_i^2/M)$. After $\boldsymbol{\beta}_i$ were generated, they were rescaled by multiplying the same constant so that the true heritabilities were 0.5 for both traits. The 307,519 array SNPs of ~336,000 UKBB genomic British individuals were used to simulate true phenotypes and to compute LD matrix for both HDL and LDSC. The P-values are from Levene's test for variance heterogeneity. Inside each box, the line indicates the median value, the central box indicates the interquartile range (IQR), and whiskers extend up to 1.5 times the IQR.



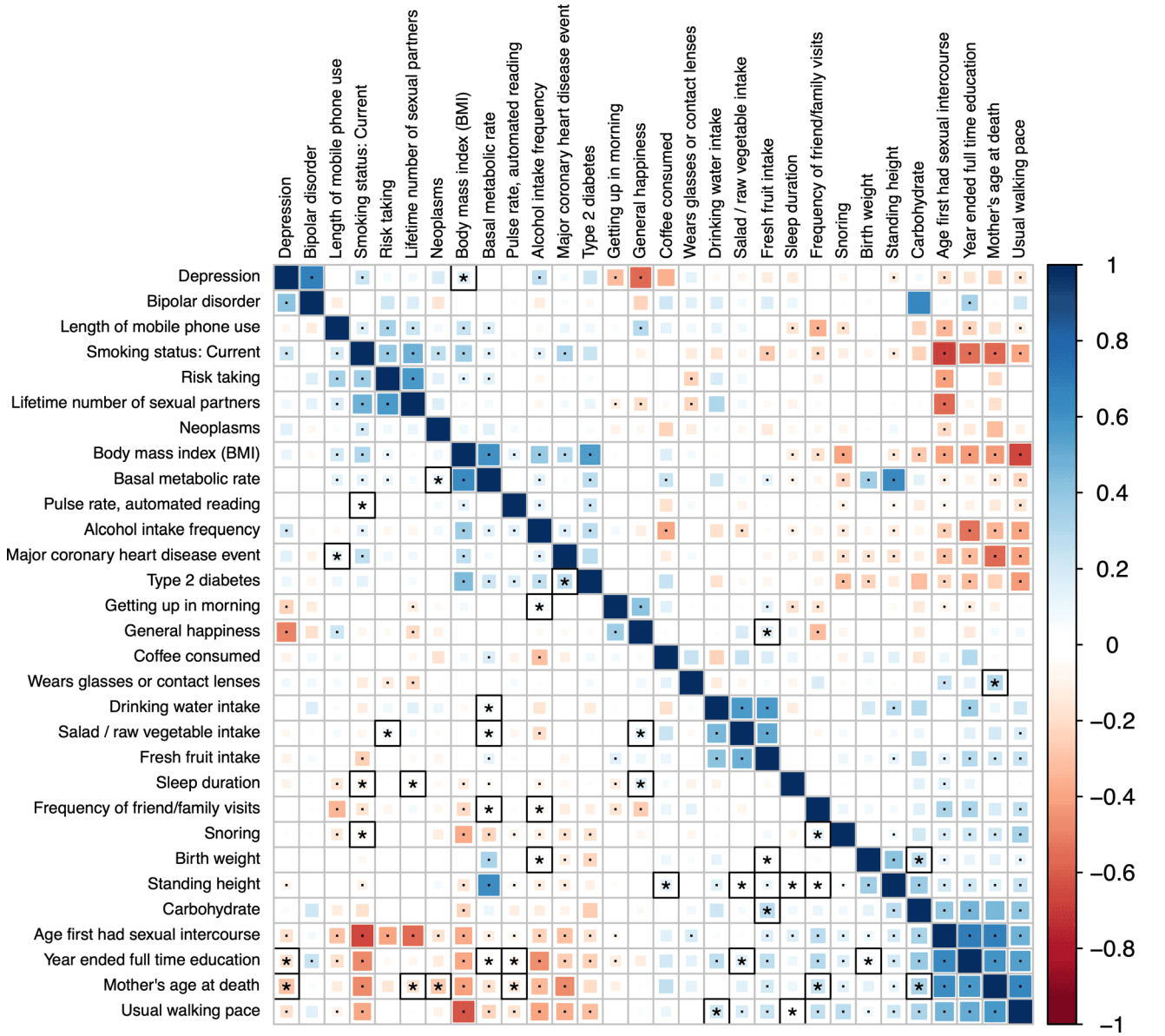
Extended Data Fig. 3 | Relative efficiency of HDL against LDSC under different model setups when 10% SNPs with 5% > MAF > 1% are causal. 52,914 out of 221,620 array SNPs with 5% > MAF > 1% were randomly selected as causal variants. 100 pairs of traits were generated, where true genetic correlation and phenotypic correlation are 0.5. The true phenotypes of trait i is generated from model $\mathbf{y}_i = \sum_{k=1}^{M} \mathbf{X}_{ik}\beta_{ik} + \epsilon_i$, where $\mathbf{X}_{ik} = (\mathbf{Z}_{ik} - 2p_k 1)[2p_k(1-p_k)]^{\alpha/2}$; $\mathbf{Z}_{ik}$ are the original genotypes of SNP k for trait i; p_k is the MAF of SNP k; M is the number of causal variants. Four scenarios were simulated: (1) $\alpha = -1$, and the marginal distribution of β_{ik} is $N(0, h_i^2/M)$; (2) $\alpha = -1$, and the marginal distribution of β_{ik} is $N(0, w_k h_i^2/M)$, where w_k is the LDAK weight of SNP k which is inversely proportional to its LD score; (3) $\alpha = -0.25$, and the marginal distribution of β_{ik} is $N(0, h_i^2/M)$ and (4) $\alpha = -0.25$, and the marginal distribution of β_{ik} is $N(0, w_k h_i^2/M)$. After $\boldsymbol{\beta}_i$ were generated, they were rescaled by multiplying the same constant so that the true heritabilities were 0.5 for both traits. The 307,519 array SNPs of ~336,000 UKBB genomic British individuals were used to simulate true phenotypes and to compute LD matrix for both HDL and LDSC. The P-values are from Levene's test for variance heterogeneity. Inside each box, the line indicates the median value, the central box indicates the interquartile range (IQR), and whiskers extend up to 1.5 times the IQR.



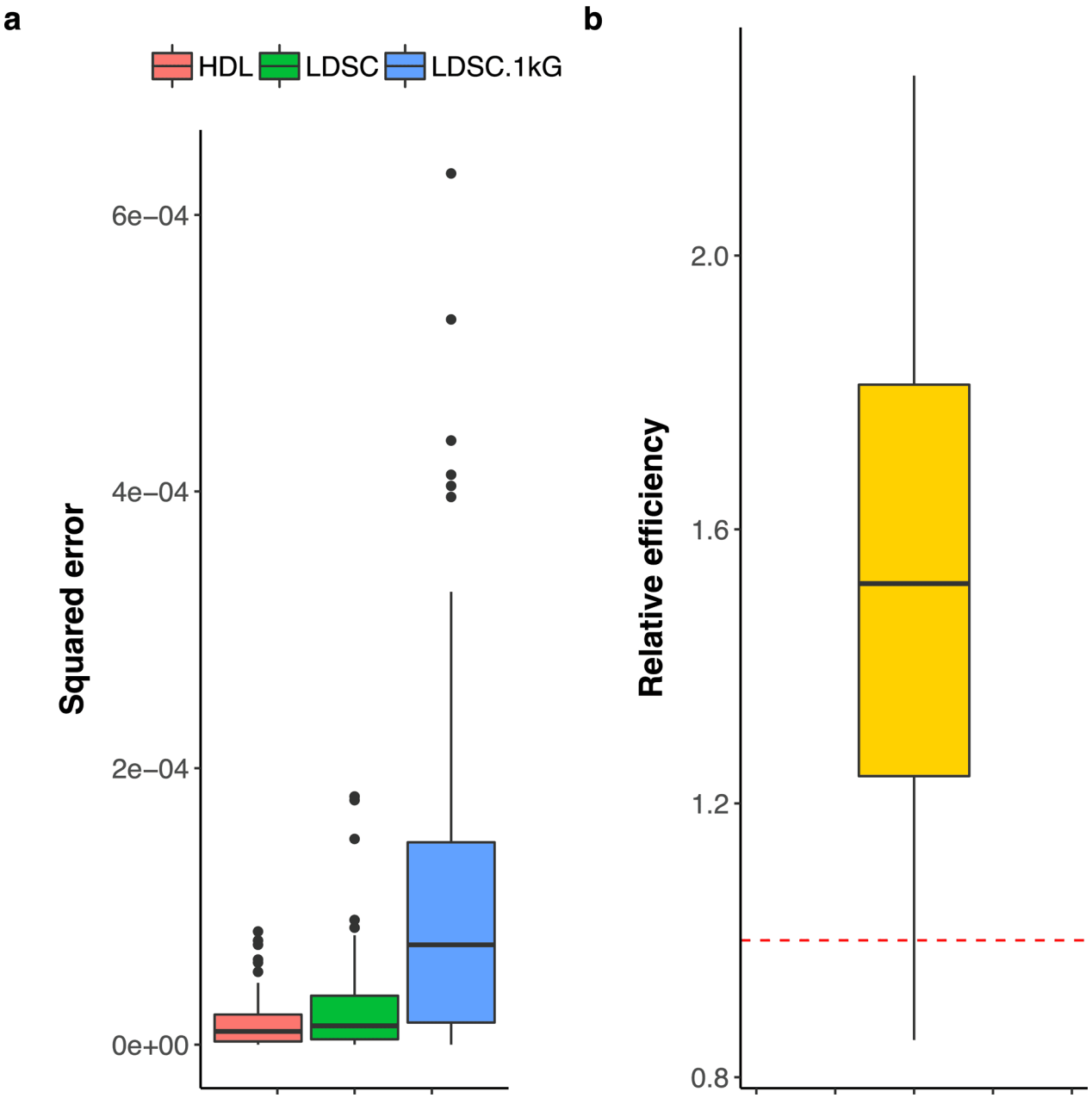
Extended Data Fig. 4 | Relative efficiency of HDL using imputed reference panel against LDSC. 100 pairs of traits were generated, where true heritabilities are 0.5, genetic correlation and phenotypic correlation are 0.5. The 1,029,876 imputed SNPs of ~336,000 UKBB genomic British individuals were used to simulate true phenotypes. LDSC and LDSC.1kG stand for the LDSC software using UKBB imputed reference panel and default 1000 Genomes reference panel, respectively. 102,988 (10% of 1,029,876) randomly sampled SNPs are set to be causal variants. The P-values are from Levene's test for variance heterogeneity. Inside each box, the line indicates the median value, the central box indicates the interquartile range (IQR), and whiskers extend up to 1.5 times the IQR.



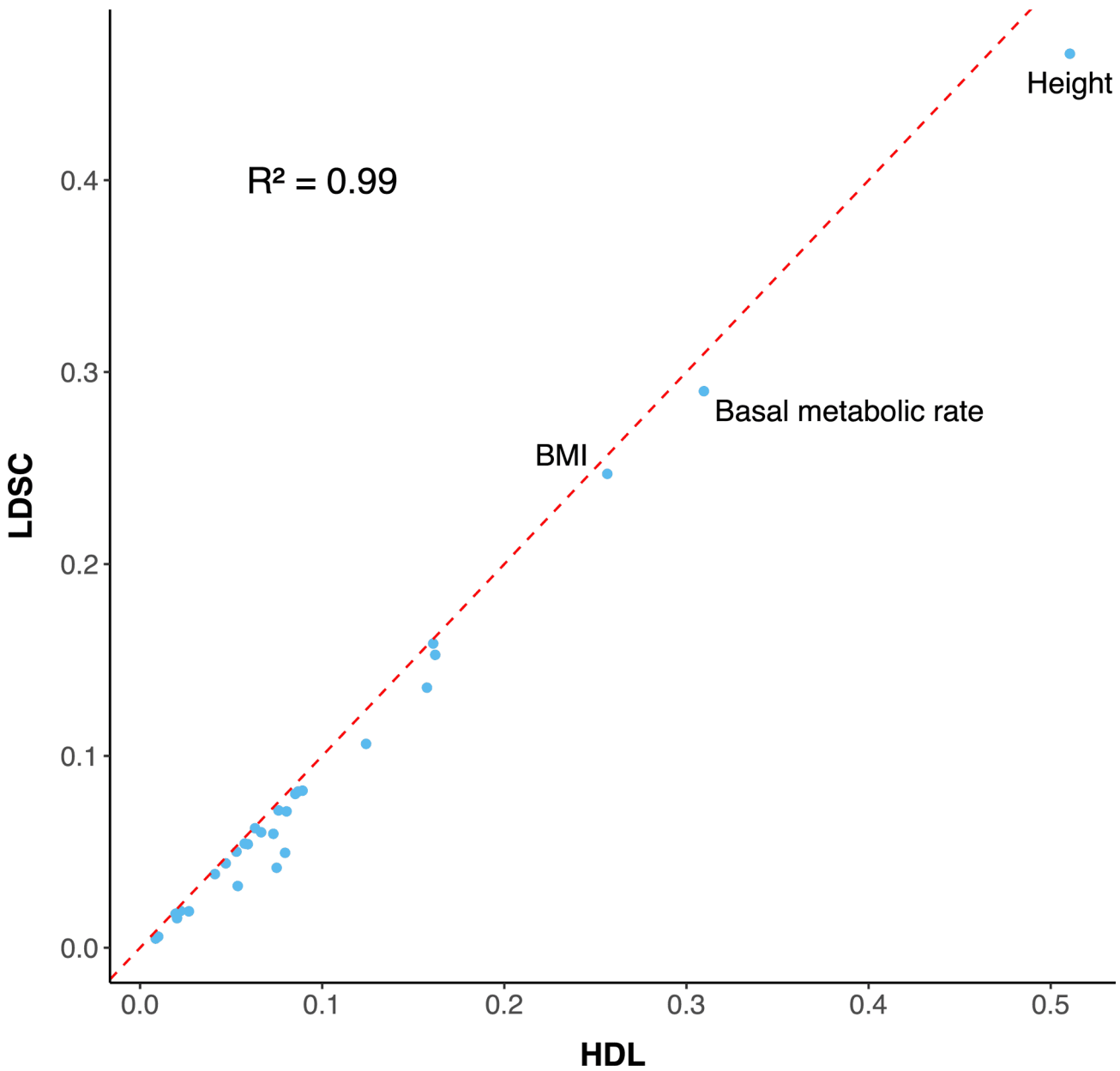
Extended Data Fig. 5 | Relative efficiency and standard error of LDSC estimate among 30 phenotypes in UK Biobank. Each dot represents genetic correlation results for one pair of traits among 435 pairs. The x-axis represents the standard error of the LDSC estimate. The y-axis represents the relative efficiency of HDL against LDSC. HDL reference panel: UKBB imputed SNPs; LDSC reference panel: 1000 Genomes (default). Colors indicate the number of binary traits in the pair.



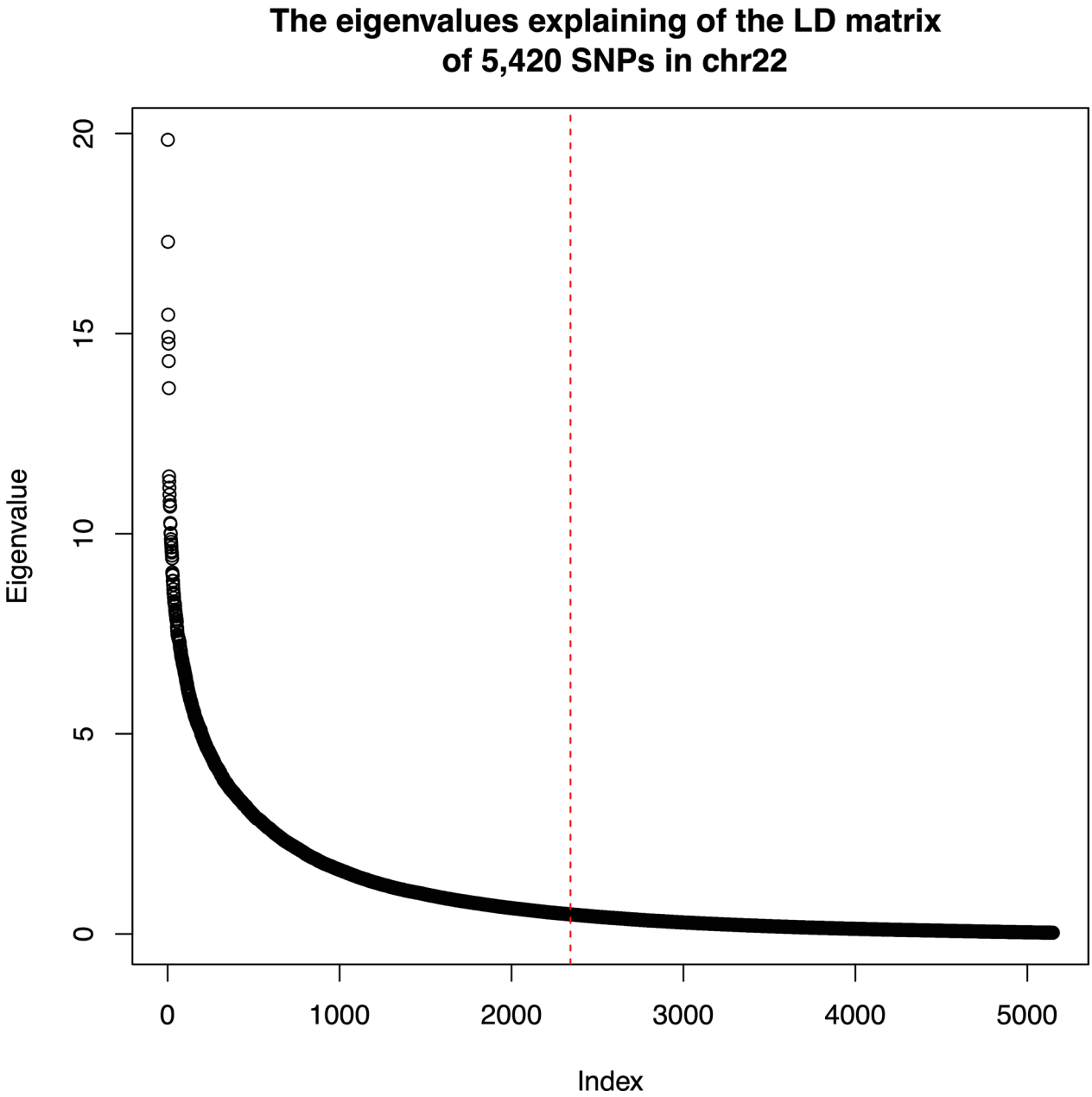
Extended Data Fig. 6 | Genetic correlation estimates from HDL and LDSC among 30 phenotypes in UK Biobank based on directly genotyped variants on the array. Lower triangle: HDL estimates; Upper triangle: LDSC estimates. The areas of the squares represent the absolute value of corresponding genetic correlations. After Bonferroni correction for 435 tests at 5% significance level, genetic correlations estimates that are significantly different from zero in both methods are marked with a dot; estimates that are significantly different from zero in only one method are marked with an asterisk and a black square. HDL reference panel: UKBB array SNPs; LDSC reference panel: UKBB array SNPs.



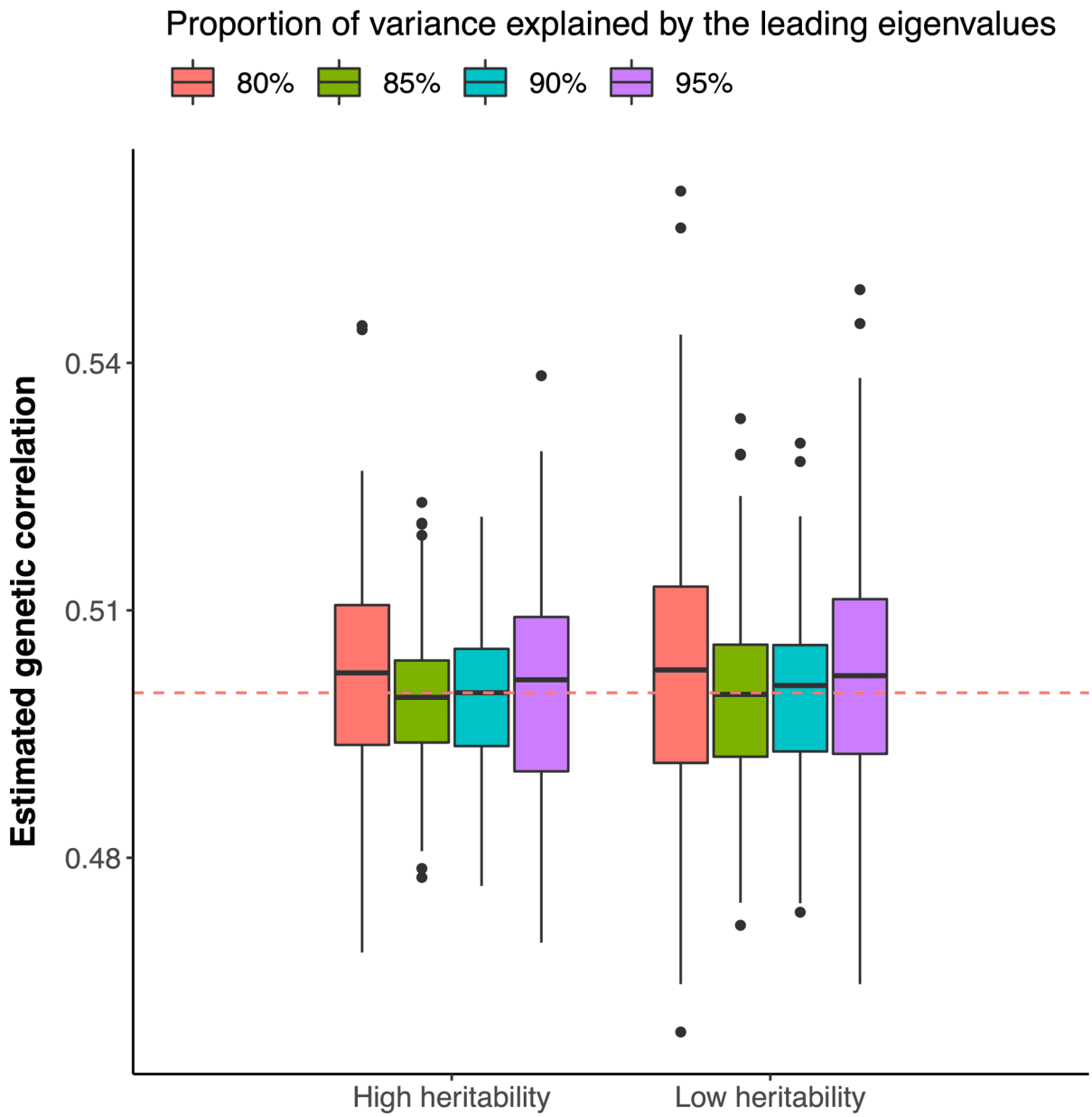
Extended Data Fig. 7 | Relative efficiency of HDL using imputed reference panel against LDSC for the estimation of heritability. a, 100 traits were generated using 14,867 imputed SNPs on chromosome 22 of ~336,000 UKBB genomic British individuals, where true heritability was set to 0.05. LDSC and LDSC.1kG stand for the LDSC software using UKBB imputed reference panel and default 1kG reference panel, respectively. 1,487 (10% of 14,867) randomly sampled SNPs are set to be causal variants. **b**, The relative efficiency, calculated as the ratio of the estimated variances of the LDSC estimates to those of the HDL estimates, was evaluated for 30 GWAS of real phenotypes in UKBB. HDL reference panel: UKBB imputed SNPs; LDSC reference panel: 1000 Genomes (default). Inside each box, the line indicates the median value, the central box indicates the interquartile range (IQR), and whiskers extend up to 1.5 times the IQR.



Extended Data Fig. 8 | Comparison of the heritability estimates from HDL and default LDSC across 30 UKBB phenotypes. The default LDSC uses the 1000 Genomes reference panel. HDL uses UKBB imputed markers as reference. *R* represents the correlation between the two sets of estimates. The red dashed line represents identity.



Extended Data Fig. 9 | Example of the eigenvalues of an LD matrix. 5,420 genotyped variants on chromosome 22 for UKBB genomic British individuals were used to generate the LD matrix. The red dashed line represents the cutoff where the leading eigenvalues and corresponding eigenvectors capture 90% of the information of the LD matrix.



Extended Data Fig. 10 | HDL results where the LD matrix is approximated by different numbers of leading eigenvalues and eigenvectors. After performing eigen-decomposition to the LD matrix, leading eigenvalues explaining different amount of variances of the LD matrix and their corresponding eigenvectors were taken to approximate the LD matrix. In each heritability group, we generated 100 pairs of traits, where true genetic correlation and phenotypic correlation are 0.5. In the high heritability group, the heritability of the pair of traits is 0.6 and 0.8 separately; in low heritability group, the heritability of the pair of traits is 0.2 and 0.4 separately. The 307,519 array SNPs of ~336,000 UKBB genomic British individuals were used to simulate true phenotypes and to compute the LD matrix for HDL. 30,752 SNPs are causal (10% of 307,519). Inside each box, the line indicates the median value, the central box indicates the interquartile range (IQR), and whiskers extend up to 1.5 times the IQR.

nature research

Corresponding author(s): Xia Shen

Last updated by author(s): Apr 28, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data were collected (UK Biobank genotype data and publicly available GWAS summary statistics for UK Biobank were used)
Data analysis	The data were analyzed with software HDL version 1.0 (https://github.com/zhenin/HDL), LDSC version 1.0.0 (https://github.com/bulik/ldsc) and LDAK version 5 (http://dougspeed.com/ldak/). PLINK version 1.9 (https://www.cog-genomics.org/plink/1.9) and 2.0 (https://www.cog-genomics.org/plink/2.0/) were used for data cleaning.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The simulations and LD computation used UK Biobank Axiom Array data and imputed genotype data, which are available from UK Biobank (https://www.ukbiobank.ac.uk/), accessible via applications. The GWAS summary statistics for UK Biobank and associated documentations are publicly available from http://www.nealelab.is/uk-biobank. The Linear Mixed Model results for UK Biobank by Canela-Xandri et al. can be downloaded from http://geneatlas.roslin.ed.ac.uk/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study develops a method, and an analysis of the UK Biobank publicly available summary-level data is given as an empirical example. We did not determine the sample size.
Data exclusions	In simulations and construction of LD reference panel using UK Biobank directly genotyped variants and imputed markers, we excluded individuals who are not genetically White British. For genetic variants, we excluded the MHC region and variants with sample MAF below 5% and performed LD pruning and missing call rate filtering. We then took the overlapped variants across (1) UKBB genotyping array, (2) variants list of LDSC and (3) variants in Neale's lab GWAS to make comparisons consistent. When imputed SNPs were used as reference panel, we took the overlapped SNPs between (1) SNP list of LDSC and (2) SNPs in the GWAS by Neale's lab. We excluded the SNPs which are (1) in the MHC region, (2) with sample MAF below 5%, (3) multi-allelic, (4) with imputation quality < 0.9, and (5) with call rate < 0.95.
Replication	Not applicable.
Randomization	Not applicable.
Blinding	Not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging