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

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

Zheng Ning[2], Yudi Pawitan[2] & Xia Shen[1,2,3*]

[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, Nobels väg 12A, SE-17 177, Stockholm, Sweden.

[3]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, UK.

*Correspondence should be addressed to: xia.shen@ed.ac.uk

Abstract

Genetic correlation is a central parameter for understanding the shared genetic architecture between complex traits and diseases. Making use of summary-level genome-wide association study (GWAS) data resources, LD Score regression (LDSC) was developed for unbiased estimation of genetic correlation. Though easy to use, LDSC only uses a small part of all the linkage disequilibrium (LD) information in the modeling of summary association statistics. In contrast, by fully accounting for LD information across the human genome, we develop a High-Definition Likelihood (HDL) method to improve the precision in genetic correlation estimation. Compared to LDSC, HDL reduces the variance of a genetic correlation estimate by about 60%, which is equivalent to a 2.5-fold increase in sample size. We implement HDL and LDSC to estimate 435 genetic correlations amongst 30 behavioral and disease-related phenotypes measured in UK Biobank. In addition to 154 genetic correlations significant for both methods, HDL identifies another 57 significant genetic correlations compared to only another 2 by LDSC. In summary, HDL brings more power to genome-wide analyses and can better reveal the underlying connections across human complex traits.

Estimating genetic correlation is a key step towards understanding the shared genetic architecture between complex traits and diseases. The genetic correlation parameter describes how the 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 (LMM)[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 appearance[5].

Though easy to use, the standard errors of genetic correlation estimates by LDSC are substantially larger than those from 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 LD Scores and consequently decreases the accuracy of estimation. However, it is worthy to note that 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 only uses a small part of the LD information in the modeling of summary association statistics. To fully exploit the information from GWAS summary-level data, we develop High-Definition Likelihood (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 UK Biobank (UKBB). We find that HDL is more accurate than LDSC with relative efficiency (ratio of estimator variance, which is equivalent to the ratio of sample size) 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, among the 435 tests for the genetic correlations across 30 behavioral and disease-related phenotypes, 57 were significant for HDL only versus 2 for 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[8] 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:

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

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 SNP j. LDSC is then developed using this relationship between single SNP LD Score and the variance of its test statistic.

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

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

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 shown in the **Supplementary Note**. To rewrite (2) into general matrix form, denoting the $M \times M$ full LD matrix as $\mathbf{R}$ with entries $\{r_{jj'}\}$, we define *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

$$\mathrm{Cov}\left[\mathbf{z}\right] = \frac{Nh^2}{M}\mathbf{L} + \mathbf{R}. \tag{3}$$

Note that the M diagonal elements of $\mathbf{L}$ are exactly the LD Scores of the M SNPs; and the M diagonal elements of $\mathrm{Cov}\left[\mathbf{z}\right]$ 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 joint normal distribution (**Supplementary Note**), LDSC can estimate their genetic covariance h_{12} based on

$$\mathrm{Cov}\left[z_{1j}, z_{2j}\right] = \mathrm{E}\left[z_{1j} z_{2j}\right] = \frac{\sqrt{N_1 N_2} h_{12}}{M} l_{jj} + \frac{N_0(h_{12} + \rho_{12})}{\sqrt{N_1 N_2}}, \tag{4}$$

where z_{1j} and z_{2j} are 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. Similar to the extension in the one-trait scenario, equation (4) can be extended to

$$\mathrm{Cov}\left[\mathbf{z}_1, \mathbf{z}_2\right] = \frac{\sqrt{N_1 N_2} h_{12}}{M}\mathbf{L} + \frac{N_0(h_{12} + \rho_{12})}{\sqrt{N_1 N_2}}\mathbf{R} \tag{5}$$

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 in LDSC, from the likelihood based on (3) and (5), HDL is developed to exploit the information within the whole $\mathbf{L}$ matrix and the covariance matrix of Z scores, not only their diagonal elements as used by LDSC.

Normalizing genetic covariance by heritabilities gives genetic correlation. Literature has suggested that, for LDSC, the estimates of genetic correlations are less susceptible to bias than the estimates of heritabilities[4,6,7,9]. Although HDL improves accuracy in estimating both heritability and genetic correlation, we shall also focus on the estimation of genetic correlation in this report. Similar 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 UK Biobank Axiom Array data from 336,000 ethnically British individuals in UKBB. To be consistent with literature[4,10], we took SNPs with minor allele frequency (MAF) above 5%. Further quality control steps resulted in 307,519 SNPs (**Online Methods**). For both HDL and LDSC, the LD matrix was computed using these 307,519 SNPs of 336,000 individuals. Among these SNPs, a proportion was 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 are 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 than LDSC. The relative efficiency was 2.58 (Levene's test P-value = 7.1×10^{-5}) for high-heritability traits (with heritability 0.6 and 0.8) and 2.93 (Levene's test P-value = 1×10^{-5}) for low-heritability traits (with heritability 0.2 and 0.4). The standard errors from block jackknifing were consistent with the observed standard deviations (**Supplementary Table 1**). To further compare HDL and LDSC, we performed simulations when (1) all of the SNPs were simulated to be causal (**Supplementary Fig. 1**); (2) model assumptions were violated (**Supplementary Fig. 2-3**). To compare HDL and LDSC when a large set of imputed SNPs were used as reference panel, we firstly built an imputed reference panel based on 1,029,876 quality-controlled HapMap3 SNPs (see **Online Methods**); then simulated true phenotypes using these SNPs; and implemented HDL and LDSC, both using imputed reference panel (**Supplementary Fig. 4**). Under all scenarios, the relative efficiency was around or above 2.

Application to summary statistics from UK Biobank

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 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 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 standard errors from HDL were in general (422 out of 435) smaller than those from LDSC, with median relative efficiency $= 2.35$. The relative efficiency was positively correlated with the standard error given by LDSC (**Supplementary Fig. 5**). The efficiency gains were larger among binary traits. Among the 435 tests for the genetic correlations (**Supplementary Table 2**), after Bonferroni correction ($P < 1.15 \times 10^{-4}$), 154 were significant for both methods, 57 were significant for only HDL (**Table 1**) and 2 were significant for only LDSC. Similar power gain can be found when both HDL and LDSC use UKBB array SNPs as reference panel (**Supplementary Fig. 6**).

Comparison with LMM results

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.[10], where LMM was fitted on UKBB individual-level data to estimate genetic correlations between hundreds of traits. Among our analyzed 30 traits, 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 the estimates from LMM ($R^2 = 0.80$), 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 is able to 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; (2) likelihood-based methods such as HDL are more efficient than the 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 (**Supplementary Fig. 7-8**), and the intercept evaluates population stratification; in double-trait HDL, the intercept implies pheno-

typic correlation and sample overlap. However, some concerns have been raised about estimating these quantities using LDSC[9,11–13]. Therefore, we are cautious about interpreting the intercept term and the single-trait HDL results, although HDL does improve heritability estimation (**Supplementary Fig. 7**). On the other hand, the LDSC estimates of genetic correlations are shown to be unbiased under different circumstances [4,6,7,9]. This robustness is mainly attributed to the ratio form of genetic correlation, and the biases on the numerator and the denominator are in the same direction, so they cancel out[4]. Given these considerations, we choose to focus the application of HDL on estimating genetic correlations.

In application, the efficiency gain by HDL was more substantial when LDSC generated large standard errors (**Supplementary Fig. 5**). This phenomenon was consistent with the simulation results that when the traits' heritabilities are low, LDSC standard errors were larger and the relative efficiency was higher. These results indicate 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 (**Supplementary Fig. 5**). As diseases are mostly recorded as binary traits and of interest in many GWAS projects and consortia, HDL would be more beneficial in such applications.

In some cases[14], 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. As a consequence, 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. We also use a constrained algorithm to prevent meaningless genetic correlation estimates. More details can be found in the **Supplementary Note**.

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 LMM to incorporate the full LD information, but for binary outcomes, fitting a normal linear mixed model misspecifies the likelihood function 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 using the original LD matrix directly, we perform eigen-decomposition on the LD matrix and pass its top eigenvalues and eigenvectors to HDL. The selected eigenvalues and eigenvectors capture most information in the LD matrix (**Supplementary Fig. 16**). There are three benefits of this decomposition step: (1) improving the efficiency of HDL (**Supplementary Fig. 9-10**); (2) saving computation time by avoiding matrix multiplication (**Supplementary Note**); (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 eigenval-

ues explaining 90% variance of the LD matrix has the highest estimation efficiency for array SNPs reference panel (**Supplementary Fig. 9**), and 99% has the highest estimation efficiency for imputed SNPs reference panel (**Supplementary Fig. 10**). Hence in this report, when array SNPs reference panel was used, we implemented HDL based on the leading eigenvalues explaining 90% variance and their corresponding eigenvectors; when imputed SNPs reference panel was used, we implemented HDL based on the leading eigenvalues explaining 99% variance and their corresponding eigenvectors. Note that for heritability estimation, as we 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 lose information for gaining estimation efficiency (**Supplementary Fig. 11**). Nevertheless, bias is not a concern for genetic correlation estimation (**Supplementary Fig. 10**).

In LDSC, 378 Europeans from the 1000 Genomes Project is 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 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 GWAS and reference samples are independent. In these simulations, we also evaluated the robustness of HDL under different scenarios where the LD matrix (1) was computed from different reference sample sizes (**Supplementary Figs. 12-13**), and (2) was approximated by its different numbers of top eigenvalues and corresponding eigenvectors (**Supplementary Figs. 9-11**). The results suggest that (1) HDL provides unbiased estimate of genetic correlation when a large independent reference sample is used; (2) the efficiency based on a large independent reference sample is almost equal to the efficiency when the GWAS sample and reference sample are identical; (3) HDL based on a large independent reference sample is robust against the choice of top eigenvalues and corresponding eigenvectors; (4) HDL based on the leading eigenvalues explaining 90% variance still gives the optimal efficiency for array SNPs panel; (5) HDL based on a small independent reference sample can still be unbiased but is less efficient and less robust against the choice of top eigenvalues and corresponding eigenvectors.

URLs. Software package for HDL inference using GWAS summary statistics, https://github.com/zhenin/HDL. LDSC, https://github.com/bulik/ldsc/; UKBB summary statistics, http://nealelab.is/uk-biobank; PLINK, http://zzz.bwh.harvard.edu/plink/; LDAK, http://dougspeed.com/ldak/.

To referees: The estimates across ∼4,000 UKBB phenotypes will be made publicly available on LD-Hub once this paper is published (Personal contact: Dr. Jie Zheng at the University of Bristol).

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Acknowledgements

XS was in receipt of a Swedish Research Council grant (No. 2017-02543) and funding from the Recruitment Program of Global Experts in China. We thank Benjamin Neale's lab for making their UK Biobank GWAS results publicly available. We thank Edinburgh Compute and Data Facility at the University of Edinburgh and National Supercomputer Center at Sun Yat-sen University for providing high-performance computing resources.

Author contributions

XS and YP initiated and coordinated the study; ZN performed data analysis; All authors contributed to method development and manuscript writing.

Competing interests statement

The authors declare no competing financial interests.

ONLINE METHODS

Modeling and estimation of genetic correlation. Suppose we have 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. Denoting the Z score vector of the M SNPs from study i of trait i as $\mathbf{z}_i$, then under a polygenic model without population stratification[8], we have

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

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

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

$$\begin{aligned}\boldsymbol{\Sigma}_{ii} &= \frac{N_i h_i^2}{M}\mathbf{L} + \mathbf{R}\\ \boldsymbol{\Sigma}_{12} &= \frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{L} + \frac{N_0(h_{12}+\rho_{12})}{\sqrt{N_1 N_2}}\mathbf{R},\end{aligned}$$

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

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

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

Following (8) and (9), we can use maximum likelihood to estimate h_1^2, h_2^2 and $r_g := h_{12}/\sqrt{(h_1^2 h_2^2)}$. Complete derivations can be found in **Supplementary Note**.

Literature has shown that LDSC with constrained intercept may produce substantially biased estimates[6,9]. However, LDSC with unconstrained intercept is much more robust. Therefore in (6) and (7), we introduced parameters $\{c_{11}, c_{22}, c_{12}\}$, which were analogous to the unconstrained intercept in LDSC:

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

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

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

the performance of HDL (**Supplementary Fig. 14**).

Quality control of UK Biobank genotype array data. In UK Biobank, ∼500,000 people aged between 40-69 years were recruited in 2006-2010 from across the country. By March 2018, most of them had been genotyped on an Affymetrix chip including ∼800,000 variants. Among the genotyped individuals, ∼336,000 were identified as unrelated genetically White British by the UK Biobank. These subjects and their genotypes were taken forward. Because we used GWAS summary statistics by Neale et al. (http://www.nealelab.is/uk-biobank/), and compared HDL with LDSC, we took the overlapped SNPs between (1) UKBB array SNPs, (2) SNP list of LDSC and (3) SNPs in Neale's GWAS to make fair comparison when array SNPs were used as reference panel. Following ref. 10 and LDSC, we excluded the MHC region and SNPs with sample MAF below 5%. We further performed LD pruning and missing call rate filtering using plink[15] software with flags –geno 0.1 –indep-pairwise 1000 5 0.95. We ended up with 307,519 autosomal SNPs for analysis related to array SNPs in this report. For both simulation and application where reference panel consists of array SNPs, the LD matrix used in HDL and LDSC were computed with these 307,519 SNPs of ∼336,000 unrelated genetically 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 UK Biobank imputed genotype data. 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. 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 analysis related to imputed markers in this report. This panel was applied in HDL for analyses related to real UKBB GWAS summary statistics in **Results**.

GWAS summary statistics of UK Biobank. The UKBB GWAS summary statistics used in this report were from the second wave of results released in July 2018 by Neale's group (http://www.nealelab.is/uk-biobank/). They performed association tests on the ∼336,000 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. Adjusted covariates are age, age^2, inferred sex, age $\times$ inferred sex, age^2 $\times$ inferred sex, and PCs 1-20.

LDSC settings. when reference panel consists of array SNPs, the LD scores based on the 307,519 SNPs were computed using flag –l2 –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 are highly consistent with those computed by 1 centimorgan (**Supplementary Fig. 15**). When the reference panel

consists 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 on https://github.com/bulik/ldsc/. For analyses related to real UKBB GWAS summary statistics in **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 consists of array SNPs, each chromosome was averagely cut into pieces with less than 10,000 SNPs, which led to 43 pieces for the whole genome. For each piece, we firstly banded its LD block with bandwidth $= 500$. Then we performed eigen-decomposition on the LD matrix and took the leading eigenvalues explaining 90% variance and their correspondent eigenvectors (see also **Supplementary Fig. 16**). When the reference panel consists of imputed SNPs, each chromosome was averagely cut into pieces with less than 20,000 SNPs, which led to 61 pieces for the whole genome. In eigen-decomposition, the leading eigenvalues explaining 99% variance and their correspondent eigenvectors were taken. After estimating heritabilities and genetic covariance for each piece, the piece-wise results were integrated into one estimate for the whole genome. The standard error of the genetic correlation estimate was computed via block jackknife with one piece out. More details can be found in the **Supplementary Note**.

Run times. When the leading eigenvalues and their corresponding eigenvectors of the LD matrices are available for loading, HDL takes around 1.5 minutes to get the point estimate using 307,519 array SNPs as reference on a single 2.8 GHz Intel©core i7, and another 4 minutes are needed to get the standard error via jackknifing. When using 1,029,876 imputed markers as reference, it takes around 7 minutes to get the point estimate and another 8 minutes to get the standard error via jackknifing. The overall computation requires about 1 GB memory.

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. *Nature Genetics* **47**, 1385 (2015).

[3] Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291 (2015).

[4] Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236 (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. *The American Journal of Human Genetics* **102**, 1185–1194 (2018).

[7] Yang, J. *et al.* Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Human Molecular Genetics* **24**, 7445–7449 (2015).

[8] Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565 (2010).

[9] Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics* **51**, 277 (2019).

[10] Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nature Genetics* **50**, 1593 (2018).

[11] Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics* **50**, 737 (2018).

[12] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature genetics* **50**, 906 (2018).

[13] Yengo, L., Yang, J. & Visscher, P. M. Expectation of the intercept from bivariate ld score regression in the presence of population stratification. *bioRxiv* 310565 (2018).

[14] Ganna, A. *et al.* Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science* **365**, eaat7693 (2019).

[15] Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575 (2007).

Figure Legends

Figure 1: **Relative efficiency of HDL against LDSC when 10% SNPs are causal.** 30,752 out of 307,519 SNPs were randomly selected as causal variants. In each group, 100 replicates were simulated, where for each pair of traits, the true genetic and phenotypic correlations are both set to 0.5. In the high-heritability group, the heritability of the two traits was set to 0.6 and 0.8, respectively; In the low-heritability group, the heritability of the two traits was set to 0.2 and 0.4, respectively. Both HDL and LDSC were based on the LD matrix computed from 307,519 array SNPs of 336,000 individuals in UKBB.

Figure 2: **Genetic correlation estimates from HDL and LDSC among 30 phenotypes in UK Biobank.** 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.

Figure 3: **Comparing genetic correlation estimates from HDL and LDSC with those from LMM across 11 phenotypes in UK Biobank.** HDL estimates are shown in dots; LDSC estimates are in crosses. For each pair of traits, the genetic correlation estimates are in the same color and connected by a gray dashed line. The black dashed line on the diagonal represents identity.

Table

Table 1: **Genetic correlation estimates that are 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}
Pulse rate, automated reading	Type 2 diabetes	0.21 (0.04)	0.23 (0.06)	1.8×10^{-8}	2.9×10^{-4}
Pulse rate, automated reading	Year ended full time education	-0.08 (0.02)	-0.1 (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}
Major coronary heart disease event	Type 2 diabetes	0.28 (0.06)	0.33 (0.1)	9.2×10^{-6}	7.5×10^{-4}
Lifetime number of sexual partners	Major coronary heart disease event	0.1 (0.02)	0.08 (0.04)	4.1×10^{-6}	2.2×10^{-2}
Birth weight	Major coronary heart disease event	-0.14 (0.03)	-0.15 (0.04)	7.4×10^{-8}	1.8×10^{-4}
Basal metabolic rate	Major coronary heart disease event	0.1 (0.02)	0.09 (0.03)	4.5×10^{-5}	2.6×10^{-3}
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	Lifetime number of sexual partners	-0.08 (0.02)	-0.06 (0.02)	3.9×10^{-6}	1.3×10^{-2}
Getting up in morning	Alcohol intake frequency	0.08 (0.02)	0.08 (0.02)	4.9×10^{-6}	4.8×10^{-4}
Alcohol intake frequency	Birth weight	-0.06 (0.01)	-0.06 (0.02)	3.9×10^{-6}	7.5×10^{-3}
Drinking water intake	Alcohol intake frequency	-0.15 (0.04)	-0.19 (0.06)	2.5×10^{-5}	2.6×10^{-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}
Body mass index (BMI)	Depression	0.13 (0.02)	0.11 (0.03)	8.7×10^{-9}	3.2×10^{-4}
Getting up in morning	Body mass index (BMI)	0.07 (0.02)	0.07 (0.02)	8.9×10^{-6}	9.0×10^{-4}
Smoking status: Current	Type 2 diabetes	0.16 (0.04)	0.19 (0.08)	8.4×10^{-5}	1.4×10^{-2}
Neoplasms	Depression	0.16 (0.04)	0.2 (0.07)	3.9×10^{-5}	3.1×10^{-3}
Lifetime number of sexual partners	Depression	0.14 (0.03)	0.1 (0.04)	5.3×10^{-7}	1.5×10^{-2}
Standing height	Depression	-0.07 (0.02)	-0.08 (0.02)	8.8×10^{-5}	1.5×10^{-3}
Year ended full time education	Depression	-0.19 (0.04)	-0.17 (0.05)	4.4×10^{-7}	9.3×10^{-4}
Mother's age at death	Depression	-0.22 (0.05)	-0.24 (0.09)	6.6×10^{-6}	7.6×10^{-3}
Risk taking	Bipolar disorder	0.19 (0.04)	0.25 (0.08)	3.5×10^{-6}	3.5×10^{-3}
Year ended full time education	Bipolar disorder	0.19 (0.04)	0.22 (0.09)	7.6×10^{-6}	1.2×10^{-2}
Risk taking	Neoplasms	0.13 (0.03)	0.16 (0.05)	2.5×10^{-5}	2.6×10^{-3}
Lifetime number of sexual partners	Neoplasms	0.14 (0.03)	0.16 (0.04)	2.8×10^{-7}	1.3×10^{-4}
Basal metabolic rate	Neoplasms	0.16 (0.02)	0.16 (0.04)	4.7×10^{-16}	1.3×10^{-4}
Standing height	Neoplasms	0.07 (0.02)	0.07 (0.04)	8.2×10^{-5}	6.0×10^{-2}
Mother's age at death	Neoplasms	-0.24 (0.05)	-0.25 (0.09)	2.0×10^{-6}	4.1×10^{-3}
Usual walking pace	Neoplasms	-0.12 (0.03)	-0.13 (0.04)	2.6×10^{-6}	9.9×10^{-4}
Drinking water intake	Length of mobile phone use	0.12 (0.03)	0.2 (0.06)	4.6×10^{-5}	6.6×10^{-4}
Length of mobile phone use	Salad / raw vegetable intake	0.09 (0.02)	0.1 (0.03)	3.4×10^{-5}	8.9×10^{-4}
Carbohydrate	Length of mobile phone use	-0.17 (0.03)	-0.24 (0.07)	1.2×10^{-6}	7.7×10^{-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}
Smoking status: Current	Wears glasses or contact lenses	-0.19 (0.03)	-0.18 (0.05)	5.1×10^{-10}	3.1×10^{-4}
Salad / raw vegetable intake	Risk taking	0.12 (0.02)	0.13 (0.03)	2.7×10^{-7}	1.3×10^{-4}
Risk taking	Mother's age at death	-0.15 (0.04)	-0.19 (0.07)	4.4×10^{-5}	5.1×10^{-3}
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}
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)	-0.2 (0.06)	3.5×10^{-6}	1.4×10^{-3}
Sleep duration	Lifetime number of sexual partners	-0.1 (0.02)	-0.09 (0.03)	2.3×10^{-8}	5.2×10^{-3}
Getting up in morning	Standing height	-0.05 (0.01)	-0.06 (0.02)	5.8×10^{-5}	3.8×10^{-4}
Sleep duration	General happiness	0.13 (0.03)	0.1 (0.04)	2.8×10^{-6}	1.5×10^{-2}
Fresh fruit intake	Birth weight	0.09 (0.02)	0.06 (0.03)	6.7×10^{-6}	2.0×10^{-2}
Birth weight	Year ended full time education	0.11 (0.02)	0.12 (0.03)	1.4×10^{-8}	1.5×10^{-4}
Frequency of friend/family visits	Basal metabolic rate	-0.08 (0.02)	-0.09 (0.02)	3.5×10^{-7}	1.4×10^{-4}
Drinking water intake	Standing height	0.13 (0.03)	0.14 (0.04)	3.6×10^{-7}	6.6×10^{-4}
Sleep duration	Standing height	0.07 (0.01)	0.05 (0.02)	2.4×10^{-8}	3.0×10^{-3}
Coffee consumed	Standing height	0.15 (0.03)	0.18 (0.06)	5.7×10^{-7}	2.9×10^{-3}
Frequency of friend/family visits	Standing height	0.06 (0.01)	0.07 (0.02)	6.9×10^{-6}	2.0×10^{-3}
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}
Snoring	Fresh fruit intake	0.1 (0.02)	0.08 (0.03)	3.8×10^{-7}	2.8×10^{-3}
Carbohydrate	Mother's age at death	0.26 (0.07)	0.43 (0.14)	1.0×10^{-4}	1.9×10^{-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	Usual walking pace	0.08 (0.01)	0.05 (0.02)	2.4×10^{-7}	2.8×10^{-2}
Frequency of friend/family visits	Wears glasses or contact lenses	0.16 (0.03)	0.18 (0.05)	3.4×10^{-6}	2.6×10^{-4}

Results that passed Bonferroni correction $0.05/435$ were reported as significant. r_g^{HDL}(s.e.), genetic correlation estimate and standard error given by HDL; r_g^{LDSC}(s.e.), genetic correlation estimate and standard error given by LDSC; P_{HDL}, P-value given by HDL; P_{LDSC}, P-value given by LDSC.

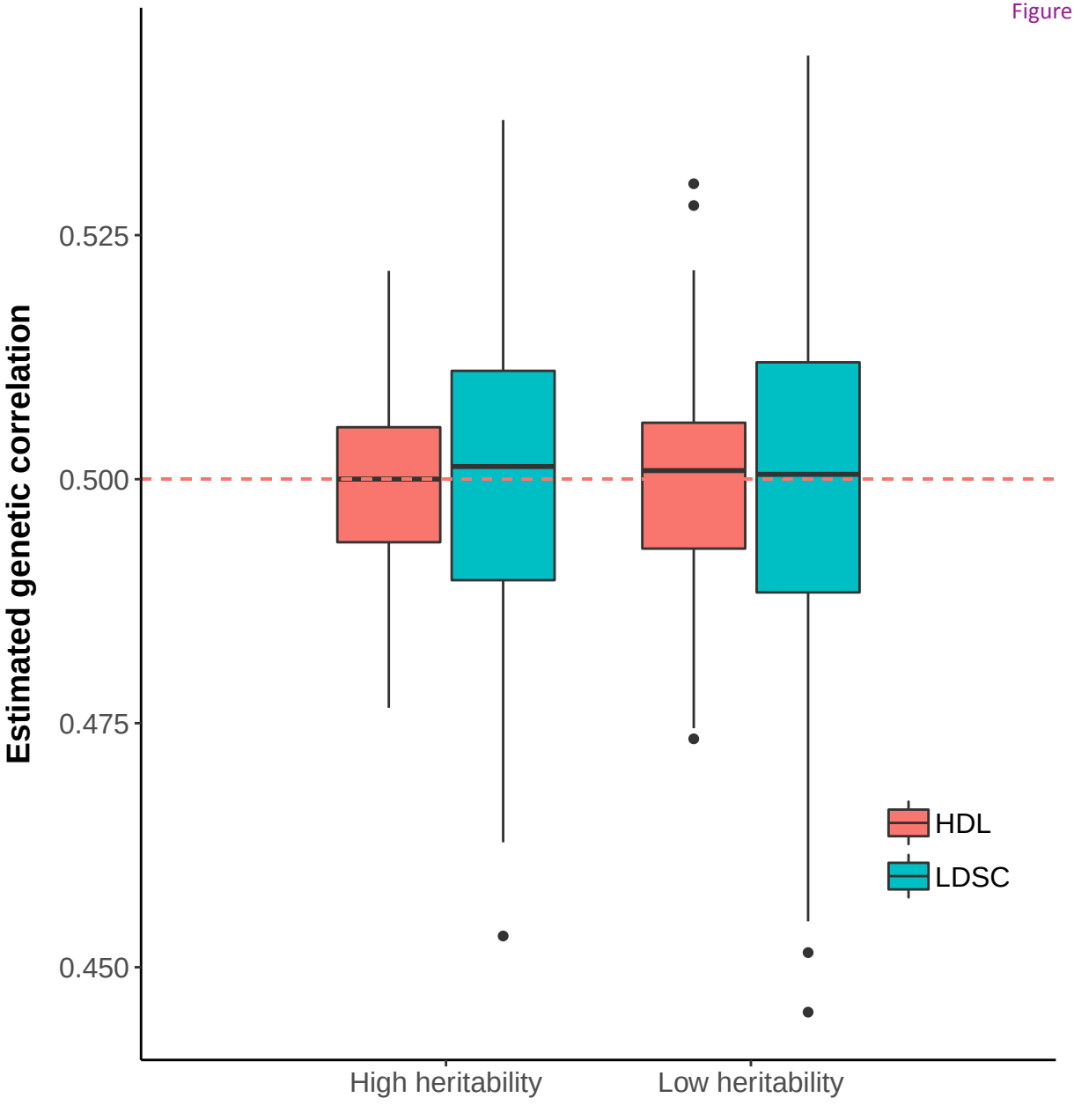

Estimated genetic correlation
0.525
0.500
0.475
0.450
High heritability
Low heritability
HDL
LDSC

Figure 2

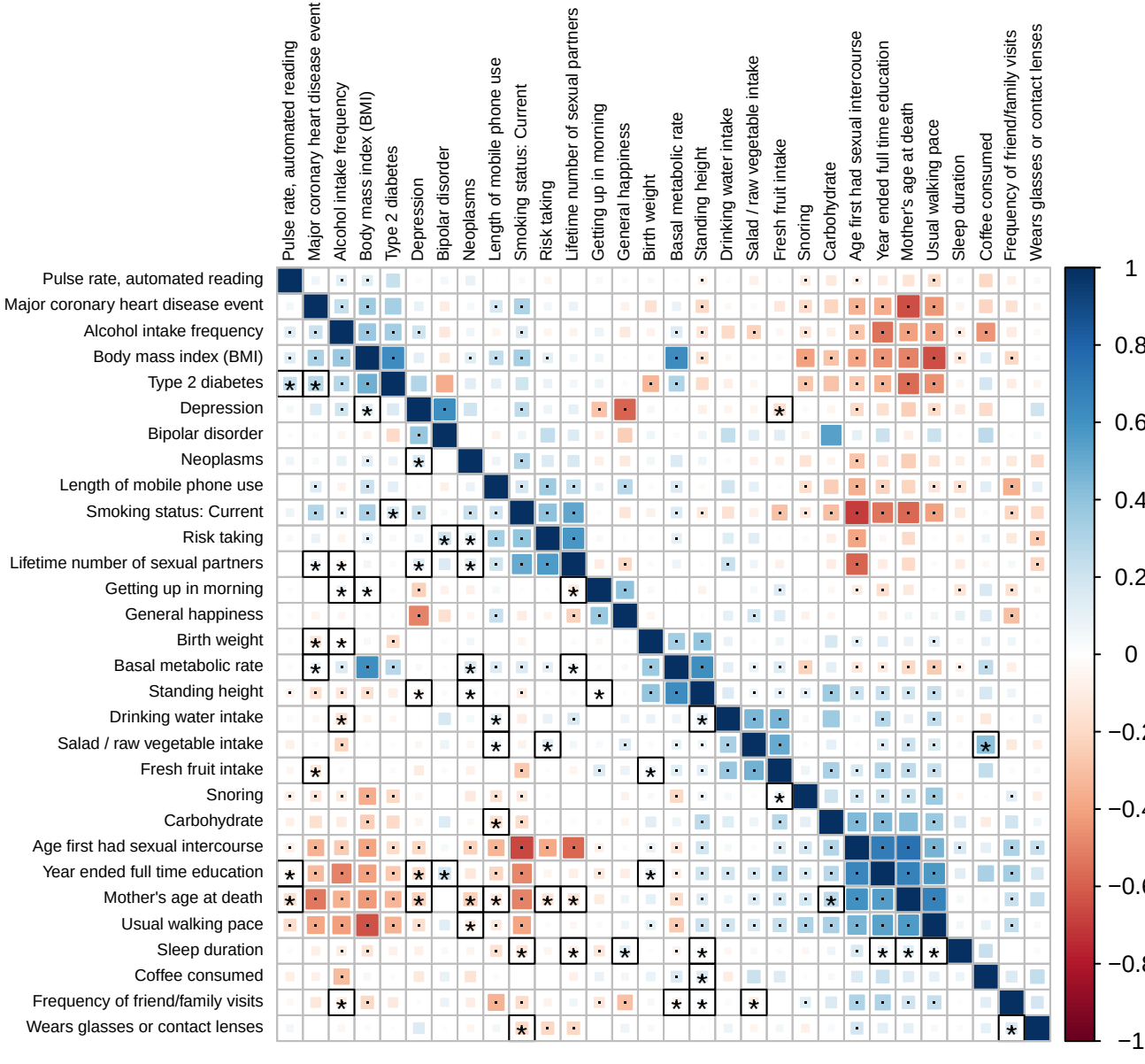
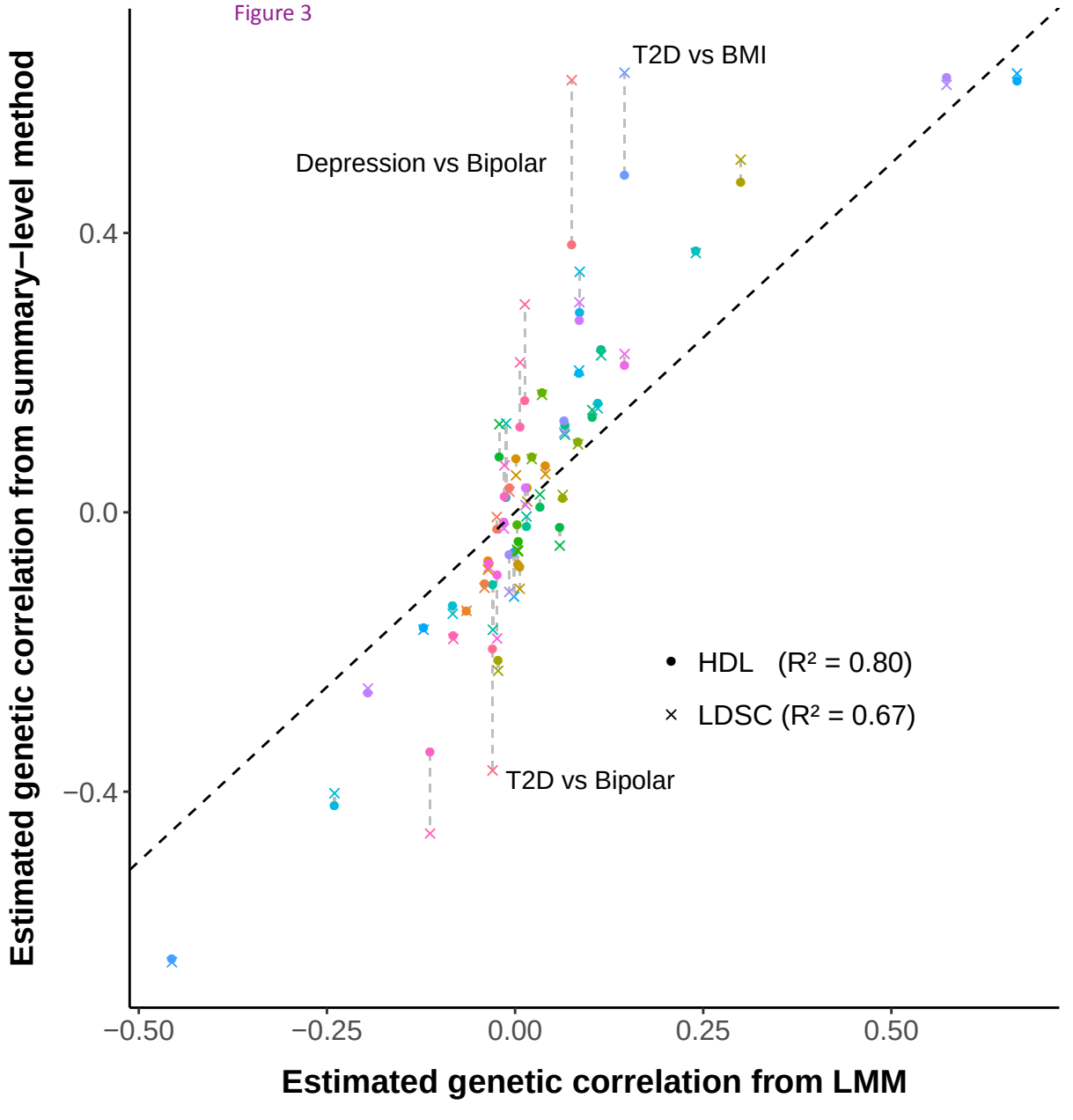

Pulse rate, automated reading
Major coronary heart disease event
Alcohol intake frequency
Body mass index (BMI)
Type 2 diabetes
Depression
Bipolar disorder
Neoplasms
Length of mobile phone use
Smoking status: Current
Risk taking
Lifetime number of sexual partners
Getting up in morning
General happiness
Birth weight
Basal metabolic rate
Standing height
Drinking water intake
Salad / raw vegetable intake
Fresh fruit intake
Snoring
Carbohydrate
Age first had sexual intercourse
Year ended full time education
Mother's age at death
Usual walking pace
Sleep duration
Coffee consumed
Frequency of friend/family visits
Wears glasses or contact lenses
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 3


T2D vs BMI
Depression vs Bipolar
T2D vs Bipolar
HDL (R² = 0.80)
LDSC (R² = 0.67)
Estimated genetic correlation from summary−level method
Estimated genetic correlation from LMM
0.4
0.0
−0.4
−0.50
−0.25
0.00
0.25
0.50

Supplementary Information

for

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

by Zheng Ning, Yudi Pawitan & Xia Shen

1 Supplementary Note

1.1 Estimating heritability of one trait using likelihood

Suppose a quantitative trait y is affected by a group of genetic variants $X_1, ..., X_M$ through a multi-variant linear model without population stratification

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1}$$

If we have N individuals, then $\boldsymbol{y} = \{y_i\}$ is a $N \times 1$ phenotype vector, $\mathbf{X} = \{x_{ij}\}$ is a $N \times M$ genotype matrix. Without loss of generality, we assume $\mathbf{X}$ is scaled to mean zero and variance one. $\boldsymbol{\beta}$ is an $M \times 1$ vector of standardized genetic effects with $\boldsymbol{\beta} \sim \mathcal{N}(0, (h^2/M)\mathbf{I})$, where h^2 represents narrow sense heritability. $\boldsymbol{\epsilon}$ is an $N \times 1$ vector of residual with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, (1-h^2)\mathbf{I})$. The genotypes are assumed to be independent across individuals with a $M \times M$ LD matrix $\mathbf{R} = \{r_{jk}\}$, where $r_{jk} = \mathbb{E}[X_j X_k]$. $\mathbf{X}, \boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are assumed to be independent with each other.

In GWAS, the estimated marginal effect of variant j is

$$\hat{b}_j = (\mathbf{X}_j^T \mathbf{X}_j)^{-1} \mathbf{X}_j^T \boldsymbol{y} \approx \frac{\mathbf{X}_j^T \boldsymbol{y}}{N}$$

and its variance

$$\sigma^2_{\hat{b}_j} = \sigma_r^2 (\mathbf{X}_j^T \mathbf{X}_j)^{-1} \approx \frac{1}{N},$$

where σ_r^2 represents the residual variance in univariate regression. As the variance explained by single variant is usually small, σ_r^2 can be approximated by phenotypic variance, which is assumed to be one in our derivation.

Therefore, the z-score of variant j is

$$z_j = \frac{\hat{b}_j}{\sqrt{\sigma^2_{\hat{b}_j}}} \approx \sqrt{N}\hat{b}_j \tag{2}$$

Lemma 1. *Let $l_{jj'} := \sum_{k=1}^{M} r_{jk} r_{kj'}$, the expected product of z_j and $z_{j'}$*

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

Specifically,

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

Proof. According to (2),

$$\mathrm{E}[z_j z_{j'} \mid \mathbf{X}] = N\mathrm{E}[\hat{b}_j \hat{b}_{j'} \mid \mathbf{X}].$$

The expected product of $\hat{b}_j$ and $\hat{b}_{j'}$ given $\mathbf{X}$

$$\begin{aligned}
\mathrm{E}\left[\hat{b}_j\hat{b}_{j'} \mid \mathbf{X}\right] &= \frac{1}{N^2}\mathrm{E}\left[\mathbf{X}_j^T\boldsymbol{y}\boldsymbol{y}^T\mathbf{X}_{j'} \mid \mathbf{X}\right] \\
&= \frac{1}{N^2}\mathrm{E}\left[\mathbf{X}_j^T\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\right)\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\right)^T\mathbf{X}_{j'} \mid \mathbf{X}\right] \\
&= \frac{1}{N^2}\mathrm{E}\left[\mathbf{X}_j^T\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T+\boldsymbol{\epsilon}\boldsymbol{\beta}^T\mathbf{X}^T+\mathbf{X}\boldsymbol{\beta}\boldsymbol{\epsilon}^T+\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right)\mathbf{X}_{j'} \mid \mathbf{X}\right] \\
&= \frac{1}{N^2}\mathrm{E}\left[\mathbf{X}_j^T\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T+\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right)\mathbf{X}_{j'} \mid \mathbf{X}\right] \\
&= \frac{1}{N^2}\Big(\mathbf{X}_j^T\mathbf{X}\mathrm{E}\left[\boldsymbol{\beta}\boldsymbol{\beta}^T \mid \mathbf{X}\right]\mathbf{X}^T\mathbf{X}_{j'}+\mathbf{X}_j^T\mathrm{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \mid \mathbf{X}\right]\mathbf{X}_{j'}\Big) \\
&= \frac{1}{N^2}\Big(\frac{h^2}{M}\mathbf{X}_j^T\mathbf{X}\mathbf{X}^T\mathbf{X}_{j'}+(1-h^2)\mathbf{X}_j^T\mathbf{X}_{j'}\Big).
\end{aligned}$$

Let $\hat{r}_{jk} = (\mathbf{X}_j^T\mathbf{X}_k)/N$, then

$$\mathrm{E}\left[\hat{b}_j\hat{b}_{j'} \mid \mathbf{X}\right] = \frac{h^2}{M}\sum_{k=1}^{M}\hat{r}_{jk}\hat{r}_{j'k}+\frac{1-h^2}{N}\hat{r}_{jj'}.$$

Take expectation over $\mathbf{X}$, we have

$$\begin{aligned}
\mathrm{E}\left[\hat{b}_j\hat{b}_{j'}\right] = \mathrm{E}\left[\mathrm{E}\left[\hat{b}_j\hat{b}_{j'} \mid \mathbf{X}\right]\right] &= \mathrm{E}\left[\frac{h^2}{M}\sum_{k=1}^{M}\hat{r}_{jk}\hat{r}_{j'k}+\frac{1-h^2}{N}\hat{r}_{jj'}\right] \\
&= \frac{h^2}{M}\sum_{k=1}^{M}\mathrm{E}\left[\hat{r}_{jk}\hat{r}_{j'k}\right]+\frac{1-h^2}{N}\mathrm{E}\left[\hat{r}_{jj'}\right] \qquad (3)
\end{aligned}$$

By the law of large numbers, $\mathrm{E}\left[\hat{r}_{jj'}\right] = r_{jj'}$. For the expected value of $\hat{r}_{jk}\hat{r}_{j'k}$

$$\begin{aligned}
\mathrm{E}\left[\hat{r}_{jk}\hat{r}_{j'k}\right] &= \mathrm{E}\left[\hat{r}_{jk}\right]\mathrm{E}\left[\hat{r}_{j'k}\right]+\mathrm{Cov}\left[\hat{r}_{jk},\hat{r}_{j'k}\right] \\
&= r_{jk}r_{j'k}+\mathrm{Cov}\left[\hat{r}_{jk},\hat{r}_{j'k}\right] \qquad (4)
\end{aligned}$$

According to Pearson and Filon [1, 2],

$$\mathrm{Cov}\left[\hat{r}_{jk},\hat{r}_{j'k}\right] = \frac{1}{N}\left[r_{jj'}(1-r_{jk}^2-r_{j'k}^2)-\frac{1}{2}r_{jk}r_{j'k}(1-r_{jk}^2-r_{j'k}^2+r_{jj'})\right]$$

Because long range LD is usually close to zero, when M is large, most r_{jk} and $r_{j'k}$ will be close to zero, which makes both $l_{jj} = \sum_{k=1}^{M} r_{jk}^2$ and $l_{j'j'} = \sum_{k=1}^{M} r_{j'k}^2$ much less than M. Therefore when M is large, we have

$$\begin{aligned}
\frac{h^2}{M}\sum_{k=1}^{M}\mathrm{Cov}\left[\hat{r}_{jk},\hat{r}_{j'k}\right] &= \frac{h^2}{MN}\left[\sum_{k=1}^{M}r_{jj'}(1-r_{jk}^2-r_{j'k}^2)-\sum_{k=1}^{M}\frac{1}{2}r_{jk}r_{j'k}(1-r_{jk}^2-r_{j'k}^2+r_{jj'})\right] \\
&= \frac{h^2}{MN}\left[r_{jj'}(M-l_{jj}-l_{j'j'})-\sum_{k=1}^{M}\frac{1}{2}r_{jk}r_{j'k}(1-r_{jk}^2-r_{j'k}^2+r_{jj'})\right] \\
&\approx \frac{h^2}{N}r_{jj'} \qquad (5)
\end{aligned}$$

Based on (4) and (5), in (3),

$$\begin{aligned}
\mathrm{E}\left[\hat{b}_j\hat{b}_{j'}\right] &= \frac{h^2}{M}\sum_{k=1}^{M}\mathrm{E}\left[\hat{r}_{jk}\hat{r}_{j'k}\right] + \frac{1-h^2}{N}\mathrm{E}\left[\hat{r}_{jj'}\right] \\
&= \frac{h^2}{M}\left[\sum_{k=1}^{M} r_{jk}r_{j'k} + \sum_{k=1}^{M}\mathrm{Cov}\left[\hat{r}_{jk},\hat{r}_{j'k}\right]\right] + \frac{1-h^2}{N}r_{jj'} \\
&\approx \frac{h^2}{M}l_{jj'} + \frac{h^2}{N}r_{jj'} + \frac{1-h^2}{N}r_{jj'} \\
&= \frac{h^2}{M}l_{jj'} + \frac{1}{N}r_{jj'}
\end{aligned}$$

Therefore,

$$\mathrm{E}[z_j z_{j'}] = N\mathrm{E}[\hat{b}_j\hat{b}_{j'}] = \frac{Nh^2}{M}l_{jj'} + r_{jj'}. \qquad \square$$

According to Lemma 1, we have

Theorem 1. *Let LD Score Matrix* $\mathbf{L} := \mathbf{R}^T\mathbf{R} = \mathbf{R}^2$ *with entries*

$$l_{jj'} = \sum_{k=1}^{M} r_{jk}r_{kj'}.$$

Denoting the z-score vector of the M variants as $\mathbf{z}$*, then*

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \textit{ where } \boldsymbol{\Sigma} = \frac{Nh^2}{M}\mathbf{L} + \mathbf{R}$$

Theorem 1 enables us to estimate h^2 by maximizing its simplified log-likelihood function:

$$\ell(h^2) = -\frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) + \mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}\right]. \tag{6}$$

1.2 Estimating genetic correlation between two traits using likelihood

Now we extend (1) to two traits scenario. Suppose we have two cohorts with sample sizes N_1 and N_2, where N_0 individuals are included in both cohorts. $\boldsymbol{y}_1 = \{y_{1i}\}$ is a $N_1 \times 1$ vector for phenotype 1 measured in cohort 1; and $\boldsymbol{y}_2 = \{y_{2i}\}$ is a $N_2 \times 1$ vector for phenotype 2 measured in cohort 2. $\mathbf{X}_1$ is a $N_1 \times M$ genotype matrix in cohort 1; and $\mathbf{X}_2$ is a $N_2 \times M$ genotype matrix in cohort 2. The genotype matrix for those individuals who are included in both cohorts is $\mathbf{X}_0$. Without loss of generality, we assume $\mathbf{X}_1$ and $\mathbf{X}_2$ are scaled to mean zero and variance one. Given the absence of population stratification, model (1) can be extended to

$$\begin{aligned}
\boldsymbol{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1 \\
\boldsymbol{y}_2 &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2,
\end{aligned}$$

where standardized genetic effects

$$\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \frac{1}{M}\begin{pmatrix}h_1^2\mathbf{I} & h_{12}\mathbf{I}\\ h_{12}\mathbf{I} & h_2^2\mathbf{I}\end{pmatrix}\right), \tag{7}$$

and residuals

$$\begin{pmatrix}\boldsymbol{\epsilon}_1\\ \boldsymbol{\epsilon}_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}(1-h_1^2)\mathbf{I} & \rho_{12}\mathbf{I}\\ \rho_{12}\mathbf{I} & (1-h_2^2)\mathbf{I}\end{pmatrix}\right). \tag{8}$$

In (7), h_{12} represents genetic covariance between the two traits; and ρ_{12} in (8) represents covariance of residuals between the two traits.

If we denote the estimated marginal effects of variant j as $\hat{b}_{1j}$ for trait 1 and $\hat{b}_{2j}$ for trait 2, then similar to (2), we have

$$z_{1j} \approx \sqrt{N_1}\hat{b}_{1j},\ z_{2j} \approx \sqrt{N_2}\hat{b}_{2j}. \tag{9}$$

Lemma 2. *If we define $l_{jj'} := \sum_{k=1}^{M} r_{jk}r_{kj'}$ as in Lemma 1, then the expected product of z_{1j} and $z_{2j'}$*

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

Specifically,

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

Proof. According to (9),

$$\mathrm{E}[z_{1j}z_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2] = \sqrt{N_1N_2}\mathrm{E}[\hat{b}_{1j}\hat{b}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2].$$

The expected product of $\hat{b}_{1j}$ and $\hat{b}_{2j'}$ given $\mathbf{X}_1$ and $\mathbf{X}_2$

$$\begin{aligned}
\mathrm{E}[\hat{b}_{1j}\hat{b}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2] &= \frac{1}{N_1N_2}\mathrm{E}\left[\mathbf{X}_{1j}^T\boldsymbol{y}_1\boldsymbol{y}_2^T\mathbf{X}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2\right]\\
&= \frac{1}{N_1N_2}\mathrm{E}\left[\mathbf{X}_{1j}^T\left(\mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1\right)\left(\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2\right)^T\mathbf{X}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2\right]\\
&= \frac{1}{N_1N_2}\mathrm{E}\left[\mathbf{X}_{1j}^T\left(\mathbf{X}_1\boldsymbol{\beta}_1\boldsymbol{\beta}_2^T\mathbf{X}_2^T + \boldsymbol{\epsilon}_1\boldsymbol{\beta}_2^T\mathbf{X}_2^T + \mathbf{X}_1\boldsymbol{\beta}_1\boldsymbol{\epsilon}_2^T + \boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_2^T\right)\mathbf{X}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2\right]\\
&= \frac{1}{N_1N_2}\mathrm{E}\left[\mathbf{X}_{1j}^T\left(\mathbf{X}_1\boldsymbol{\beta}_1\boldsymbol{\beta}_2^T\mathbf{X}_2^T + \boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_2^T\right)\mathbf{X}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2\right]\\
&= \frac{1}{N_1N_2}\Big(\mathbf{X}_{1j}^T\mathbf{X}_1\mathrm{E}\left[\boldsymbol{\beta}_1\boldsymbol{\beta}_2^T \mid \mathbf{X}_1, \mathbf{X}_2\right]\mathbf{X}_2^T\mathbf{X}_{2j'} + \mathbf{X}_{1j}^T\mathrm{E}\left[\boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_2^T \mid \mathbf{X}_1, \mathbf{X}_2\right]\mathbf{X}_{2j'}\Big)\\
&= \frac{1}{N_1N_2}\Big(\frac{h_{12}}{M}\mathbf{X}_{1j}^T\mathbf{X}_1\mathbf{X}_2^T\mathbf{X}_{2j'} + \rho_{12}\mathbf{X}_{0j}^T\mathbf{X}_{0j'}\Big).
\end{aligned}$$

Let $\hat{r}_{1,jk} = (\mathbf{X}_{1j}^T\mathbf{X}_{1k})/N_1$, $\hat{r}_{2,jk} = (\mathbf{X}_{2j}^T\mathbf{X}_{2k})/N_2$ and $\tilde{r}_{jk} = (\mathbf{X}_{0j}^T\mathbf{X}_{0k})/N_0$, then

$$\mathrm{E}[\hat{b}_{1j}\hat{b}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2] = \frac{h_{12}}{M}\sum_{k=1}^{M}\hat{r}_{1,jk}\hat{r}_{2,j'k} + \frac{N_0}{N_1N_2}\rho_{12}\tilde{r}_{jj'}.$$

Take expectation over $\mathbf{X_1}$ and , we have

$$\begin{aligned}\mathrm{E}\left[\hat{b}_{1j}\hat{b}_{2j'}\right] &= \mathrm{E}\left[\mathrm{E}\left[\hat{b}_{1j}\hat{b}_{2j'} \mid \mathbf{X}_1, \mathbf{X}_2\right]\right] = \mathrm{E}\left[\frac{h_{12}}{M}\sum_{k=1}^{M}\hat{r}_{1,jk}\hat{r}_{2,j'k} + \frac{N_0}{N_1N_2}\rho_{12}\tilde{r}_{jj'}\right] \\ &= \frac{h_{12}}{M}\sum_{k=1}^{M}\mathrm{E}\left[\hat{r}_{1,jk}\hat{r}_{2,j'k}\right] + \frac{N_0}{N_1N_2}\rho_{12}\mathrm{E}\left[\tilde{r}_{jj'}\right] \end{aligned} \tag{10}$$

By the law of large numbers, $\mathrm{E}\left[\tilde{r}_{jj'}\right] = r_{jj'}$. For the expected value of $\hat{r}_{1,jk}\hat{r}_{2,j'k}$

$$\begin{aligned}\mathrm{E}\left[\hat{r}_{1,jk}\hat{r}_{2,j'k}\right] &= \mathrm{E}\left[\hat{r}_{1,jk}\right]\mathrm{E}\left[\hat{r}_{2,j'k}\right] + \mathrm{Cov}\left[\hat{r}_{1,jk}, \hat{r}_{2,j'k}\right] \\ &= r_{jk}r_{j'k} + \mathrm{Cov}\left[\hat{r}_{1,jk}, \hat{r}_{2,j'k}\right] \end{aligned} \tag{11}$$

Similar to (5), when M is large, we have

$$\begin{aligned}\frac{h_{12}}{M}\sum_{k=1}^{M}\mathrm{Cov}\left[\hat{r}_{jk}, \hat{r}_{j'k}\right] &= \frac{h_{12}}{MN_1N_2}\sum_{k=1}^{M}\mathrm{Cov}\left[\mathbf{X}_{1j}^T\mathbf{X}_{1k}, \mathbf{X}_{2j'}^T\mathbf{X}_{2k}\right] \\ &= \frac{1}{N_1N_2}\frac{h_{12}}{M}\sum_{k=1}^{M}\mathrm{Cov}\left[\mathbf{X}_{0j}^T\mathbf{X}_{0k}, \mathbf{X}_{0j'}^T\mathbf{X}_{0k}\right] \\ &= \frac{N_0^2}{N_1N_2}\frac{h_{12}}{M}\sum_{k=1}^{M}\mathrm{Cov}\left[\tilde{r}_{jk}, \tilde{r}_{j'k}\right] \\ &\approx \frac{N_0^2}{N_1N_2}\frac{h_{12}}{N_0}r_{jj'} \\ &= \frac{N_0}{N_1N_2}h_{12}r_{jj'} \end{aligned} \tag{12}$$

Based on (11) and (12), in (10),

$$\begin{aligned}\mathrm{E}\left[\hat{b}_{1j}\hat{b}_{2j'}\right] &= \frac{h_{12}}{M}\sum_{k=1}^{M}\mathrm{E}\left[\hat{r}_{1,jk}\hat{r}_{2,j'k}\right] + \frac{N_0}{N_1N_2}\rho_{12}\mathrm{E}\left[\tilde{r}_{jj'}\right] \\ &= \frac{h_{12}}{M}\left[\sum_{k=1}^{M}r_{jk}r_{j'k} + \sum_{k=1}^{M}\mathrm{Cov}\left[\hat{r}_{1,jk}\hat{r}_{2,j'k}\right]\right] + \frac{N_0}{N_1N_2}\rho_{12}r_{jj'} \\ &\approx \frac{h_{12}}{M}l_{jj'} + \frac{N_0}{N_1N_2}h_{12}r_{jj'} + \frac{N_0}{N_1N_2}\rho_{12}r_{jj'} \\ &= \frac{h_{12}}{M}l_{jj'} + \frac{N_0(h_{12}+\rho_{12})}{N_1N_2}r_{jj'} \end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{E}[z_{1j}z_{2j'}] &= \sqrt{N_1N_2}\mathrm{E}[\hat{b}_{1j}\hat{b}_{2j'}] \\ &= \frac{\sqrt{N_1N_2}h_{12}}{M}l_{jj'} + \frac{N_0(h_{12}+\rho_{12})}{\sqrt{N_1N_2}}r_{jj'}. \end{aligned}$$

□

Following Lemma 2 and Theorem 1 we have

Theorem 2. *Denoting the z-score vectors of the M variants for phenotype 1 and phenotype 2 as* $\mathbf{z}_1$ *and* $\mathbf{z}_2$ *respectively, then*

$$\begin{pmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{11} &= \frac{N_1 h_1^2}{M}\mathbf{L} + \mathbf{R},\\ \boldsymbol{\Sigma}_{22} &= \frac{N_2 h_2^2}{M}\mathbf{L} + \mathbf{R},\\ \boldsymbol{\Sigma}_{12} &= \frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{L} + \frac{N_0(h_{12}+\rho_{12})}{\sqrt{N_1 N_2}}\mathbf{R}.\end{aligned}$$

Let $r_g := h_{12}/\sqrt{h_1^2 h_2^2}$. Theorem 2 enables us to estimate h_1^2, h_2^2 and r_g by maximizing the full joint likelihood. Because the likelihood is a smooth function, it can be maximized sequentially as follows:

$$\begin{aligned}\max_{h_1^2,h_2^2,r_g} \ell(h_1^2,h_2^2,r_g) &= \max_{r_g}\left\{\max_{h_1^2,h_2^2}\ell(h_1^2,h_2^2,r_g)\right\}\\ &= \max_{r_g}\left\{\ell(\tilde{h}_1^2(r_g),\tilde{h}_2^2(r_g),r_g)\right\}\\ &= \max_{r_g}\left\{\ell(\tilde{h}_1^2,\tilde{h}_2^2,r_g)\right\}. \qquad (13)\end{aligned}$$

In (13) we have used the fact that $\tilde{h}_1^2(r_g) = \tilde{h}_1^2$ and $\tilde{h}_2^2(r_g) = \tilde{h}_2^2$, which are the MLEs of the individual heritabilities. That is, knowing the correlation does not give us information about individual variances. Then, to reduce the dimension of the matrices, the final maximization be simplified using

$$\begin{aligned}\max_{r_g}\left\{\ell(\tilde{h}_1^2,\tilde{h}_2^2,r_g)\right\} &= \max_{r_g}\left\{\ell_m(\tilde{h}_1^2) + \ell_c(\tilde{h}_1^2,\tilde{h}_2^2,r_g)\right\}\\ &= \ell_m(\tilde{h}_1^2) + \max_{r_g}\left\{\ell_c(\tilde{h}_1^2,\tilde{h}_2^2,r_g)\right\}\end{aligned}$$

where $\ell_m(h_1^2)$ is the marginal log-likelihood based on $\mathbf{z_1}$; and $\ell_c(h_1^2,h_2^2,r_g)$ is the conditional log-likelihood based on $\mathbf{z_2} \mid \mathbf{z_1}$. So in summary, in the HDL algorithm, we firstly get $\tilde{h}_1^2$ and $\tilde{h}_2^2$ from the marginal likelihood of h_1^2 and h_2^2 separately. Then estimate r_g by maximizing the conditional likelihood ℓ_c at the estimated heritability values. The conditional likelihood can be found from the following:

Corollary 1. *The conditional distribution for* $\mathbf{z}_2$ *given* $\mathbf{z}_1$ *is*

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

This gives the conditional log-likelihood

$$\begin{aligned}\ell_c(h_1^2, h_2^2, r_g) = &-\frac{1}{2}\log(|\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}|) \\ &-\frac{1}{2}\left(\mathbf{z_2} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1\right)^T\left(\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right)^{-1}\left(\mathbf{z_2} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1\right).\end{aligned}$$

The standard error of $\hat{r}_g$ is computed using a block-jackknife procedure described in Section 1.5.

1.3 $\mathbf{\Sigma}_{11}$, $\mathbf{\Sigma}_{22}$ and $\mathbf{\Sigma}_{12}$ in working algorithm

Literature has shown that LDSC with unconstrained intercept is much more robust against incorrect model [3,4]. Similarly, in the application of HDL, we introduce parameters $\{c_{11}, c_{22}, c_{12}\}$ into $\mathbf{\Sigma}_{11}$, $\mathbf{\Sigma}_{22}$ and $\mathbf{\Sigma}_{12}$:

$$\begin{aligned}\mathbf{\Sigma}_{11} &= \frac{N_1 h_1^2}{M}\mathbf{L} + c_{11}\mathbf{R}, \\ \mathbf{\Sigma}_{22} &= \frac{N_2 h_2^2}{M}\mathbf{L} + c_{22}\mathbf{R}, \\ \mathbf{\Sigma}_{12} &= \frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{L} + c_{12}\frac{N_0}{\sqrt{N_1 N_2}}\mathbf{R},\end{aligned}$$

which were analogous to the unconstrained intercept in LDSC. Therefore the working log-likelihoods in HDL are

$$\ell(h_i^2, c_{ii}) = -\frac{1}{2}\left[\log(|\mathbf{\Sigma}_{ii}|) + \mathbf{z}_i^T\mathbf{\Sigma}_{ii}^{-1}\mathbf{z}_i\right] \tag{14}$$

and

$$\begin{aligned}&\ell_c(\tilde{h}_1^2, \tilde{c}_{11}, \tilde{h}_2^2, \tilde{c}_{22}, r_g, c_{12}) \\ =&-\frac{1}{2}\log(|\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}|) \\ &-\frac{1}{2}\left(\mathbf{z_2} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1\right)^T\left(\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right)^{-1}\left(\mathbf{z_2} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1\right).\end{aligned} \tag{15}$$

1.4 Using eigen-decomposition to simplify computation

As a real symmetrix matrix, the $M \times M$ LD matrix $\mathbf{R}$ can be decomposed as

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T,$$

where $\mathbf{Q}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{R}$, and $\mathbf{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{R}$. As a way of regularization and facilitating computation, instead of taking all M eigenvalues, we can take the leading p eigenvalues and their corresponding eigenvectors. Denoting the $p \times p$ diagonal matrix consists of the leading p eigenvalues $(\lambda_1, ..., \lambda_p)$ as $\mathbf{\Lambda}_p$, and the $M \times p$ eigenvectors matrix as $\mathbf{Q}_p$, the LD matrix $\mathbf{R}$ can be approximated as

$$\mathbf{R} \approx \mathbf{Q}_p\mathbf{\Lambda}_p\mathbf{Q}_p^T.$$

Then $\mathbf{\Sigma}_{ii}$ and $\mathbf{\Sigma}_{12}$ can be reformed to

$$\begin{aligned}
\mathbf{\Sigma}_{ii} &= \frac{N_i h_i^2}{M}\mathbf{L} + c_{ii}\mathbf{R} \\
&\approx \frac{N_i h_i^2}{M}\mathbf{Q}_p\mathbf{\Lambda}_p^2\mathbf{Q}_p^T + c_{ii}\mathbf{Q}_p\mathbf{\Lambda}_p\mathbf{Q}_p^T \\
&= \mathbf{Q}_p\left(\frac{N_i h_i^2}{M}\mathbf{\Lambda}_p^2 + c_{ii}\mathbf{\Lambda}_p\right)\mathbf{Q}_p^T \\
\mathbf{\Sigma}_{12} &= \frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{L} + c_{12}\frac{N_0}{\sqrt{N_1 N_2}}\mathbf{R} \\
&\approx \mathbf{Q}_p\left[\frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{\Lambda}_p^2 + c_{12}\frac{N_0}{\sqrt{N_1 N_2}}\mathbf{\Lambda}_p\right]\mathbf{Q}_p^T
\end{aligned}$$

Then (14) can be transformed to

$$\begin{aligned}
\ell(h_i^2, c_{ii}) &= -\frac{1}{2}\left[\log(|\mathbf{\Sigma}_{ii}|) + \mathbf{z}_i^T\mathbf{\Sigma}_{ii}^{-1}\mathbf{z}_i\right] \\
&\approx -\frac{1}{2}\left[\sum_{j=1}^{p}\log\left(\frac{N_i h_i^2}{M}\lambda_j^2 + c_{ii}\lambda_j\right) + \mathbf{z}_i^T\mathbf{Q}_p\left(\frac{N_i h_i^2}{M}\mathbf{\Lambda}_p^2 + c_{ii}\mathbf{\Lambda}_p\right)^{-1}\mathbf{Q}_p^T\mathbf{z}_i\right]
\end{aligned}$$

.

Denoting $\mathbf{u}_i = \mathbf{Q}_p^T\mathbf{z}_i$ with entries $\{u_{ij}\}$, we have

$$\begin{aligned}
\ell(h_i^2, c_{ii}) &\approx -\frac{1}{2}\left[\sum_{j=1}^{p}\log\left(\frac{N_i h_i^2}{M}\lambda_j^2 + c_{ii}\lambda_j\right) + \mathbf{u}_i^T\left(\frac{N_i h_i^2}{M}\mathbf{\Lambda}_p^2 + c_{ii}\mathbf{\Lambda}_p\right)^{-1}\mathbf{u}_i\right] \\
&= -\frac{1}{2}\left[\sum_{j=1}^{p}\log\left(\frac{N_i h_i^2}{M}\lambda_j^2 + c_{ii}\lambda_j\right) + \sum_{j=1}^{p}\frac{u_{ij}^2}{\frac{N_i h_i^2}{M}\lambda_j^2 + c_{ii}\lambda_j}\right].
\end{aligned}$$

Equation (15) can be transformed similarly. To simplify notation, we denote

$$\begin{aligned}
&\mathbf{\Lambda}_{ii}^* = \frac{N_i h_i^2}{M}\mathbf{\Lambda}_p^2 + c_{ii}\mathbf{\Lambda}_p, \text{ with diagonal entries } \lambda_{ii,j}^* = \frac{N_i h_i^2}{M}\lambda_j^2 + c_{ii}\lambda_j \\
&\mathbf{\Lambda}_{12}^* = \frac{\sqrt{N_1 N_2}h_{12}}{M}\mathbf{\Lambda}_p^2 + c_{12}\frac{N_0}{\sqrt{N_1 N_2}}\mathbf{\Lambda}_p, \\
&\quad \text{with diagonal entries } \lambda_{12,j}^* = \frac{\sqrt{N_1 N_2}h_{12}}{M}\lambda_j^2 + c_{12}\frac{N_0}{\sqrt{N_1 N_2}}\lambda_j, \\
&\mathbf{u}^* = \mathbf{Q}_p^T\left[\mathbf{z_2} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1\right] = \mathbf{Q}_p^T\mathbf{z_2} - \mathbf{\Lambda}_{12}^*\left(\mathbf{\Lambda}_{11}^*\right)^{-1}\mathbf{Q}_p^T\mathbf{z_1}, \text{ with entries } \{u_j^*\}.
\end{aligned}$$

Then

$$\begin{aligned}
\ell_c(\tilde{h}_1^2, \tilde{c}_{11}, \tilde{h}_2^2, \tilde{c}_{22}, r_g, c_{12}) &= -\frac{1}{2}\log(|\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}|) \\
&\quad -\frac{1}{2}\left(\mathbf{z_2} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1\right)^T\left(\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right)^{-1}\left(\mathbf{z_2} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{11}^{-1}\mathbf{z}_1\right) \\
&= -\frac{1}{2}\left[\sum_{j=1}^{p}\log\left(\lambda_{22,j}^* - \frac{\lambda_{12,j}^*}{\lambda_{11,j}^*}\right) + \sum_{j=1}^{p}\frac{\left(u_j^*\right)^2}{\lambda_{22,j}^* - \frac{\lambda_{12,j}^*}{\lambda_{11,j}^*}}\right].
\end{aligned}$$

1.5 Integration of piece-wise likelihood

To improve the computational performance of HDL, each chromosome was cut into pieces, which led to m pieces for the whole genome. Because long-distance LD is rare and the LD blocks around the cutting positions are a small proportion among the overall LD, we assume these m pieces are independent of each other. Denoting the LD matrix of piece k as $\mathbf{R}_k$, then

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & & & \\ & \mathbf{R}_2 & & \\ & & \ddots & \\ & & & \mathbf{R}_m \end{pmatrix}$$

Therefore,

$$\mathbf{L} = \mathbf{R}^T\mathbf{R} = \begin{pmatrix} \mathbf{L}_1 & & & \\ & \mathbf{L}_2 & & \\ & & \ddots & \\ & & & \mathbf{L}_m \end{pmatrix}, \text{ and } \mathbf{\Sigma}_{ii} = \begin{pmatrix} \mathbf{\Sigma}_{ii,1} & & & \\ & \mathbf{\Sigma}_{ii,2} & & \\ & & \ddots & \\ & & & \mathbf{\Sigma}_{ii,m} \end{pmatrix},$$

where $\mathbf{\Sigma}_{ii,k} = \frac{N_i h_i^2}{M}\mathbf{L}_k + c_{ii}\mathbf{R}_k$. Noticing that

$$|\mathbf{\Sigma}_{ii}| = \prod_{k=1}^{m} |\mathbf{\Sigma}_{ii,k}|, \text{ and } \mathbf{\Sigma}_{ii}^{-1} = \begin{pmatrix} \mathbf{\Sigma}_{ii,1}^{-1} & & & \\ & \mathbf{\Sigma}_{ii,2}^{-1} & & \\ & & \ddots & \\ & & & \mathbf{\Sigma}_{ii,m}^{-1} \end{pmatrix},$$

the likelihood in (14) is therefore additive across pieces as

$$\begin{aligned} \ell(h_i^2, c_{ii}) &= -\frac{1}{2}\left[\log(|\mathbf{\Sigma}_{ii}|) + \mathbf{z}_i^T\mathbf{\Sigma}_{ii}^{-1}\mathbf{z}_i\right] \\ &= -\frac{1}{2}\left[\sum_{k=1}^{m}\log(|\mathbf{\Sigma}_{ii,k}|) + \sum_{k=1}^{m}\mathbf{z}_{i,k}^T\mathbf{\Sigma}_{ii,k}^{-1}\mathbf{z}_{i,k}\right]. \end{aligned}$$

Similarly, (15) is also additive.

Another benefit of cutting genome into pieces is to allow block-jackknife by leaving one piece out. The block-jackknife procedure provides robust estimates of standard errors for parameters.

References

[1] Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185**, 71–110 (1894).

[2] Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychological bulletin* **87**, 245 (1980).

[3] Ni, G. *et al.* Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *The American Journal of Human Genetics* **102**, 1185–1194 (2018).

[4] Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics* **51**, 277 (2019).

2 Supplementary Tables

Supplementary Table 1: Simulations with different heritability groups when 10% 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. 30,752 SNPs are causal (10% of 307,519). True value: true genetic correlation; Estimate: estimate of genetic correlation; s.d.: standard deviation of the estimates across 100 simulations; s.e: median standard error across 100 simulations.

Heritability group	Method	True value	Estimate	s.d.	s.e.
High	HDL	0.50	0.50	0.010	0.010
	LDSC	0.50	0.50	0.016	0.014
Low	HDL	0.50	0.50	0.011	0.012
	LDSC	0.50	0.50	0.019	0.017

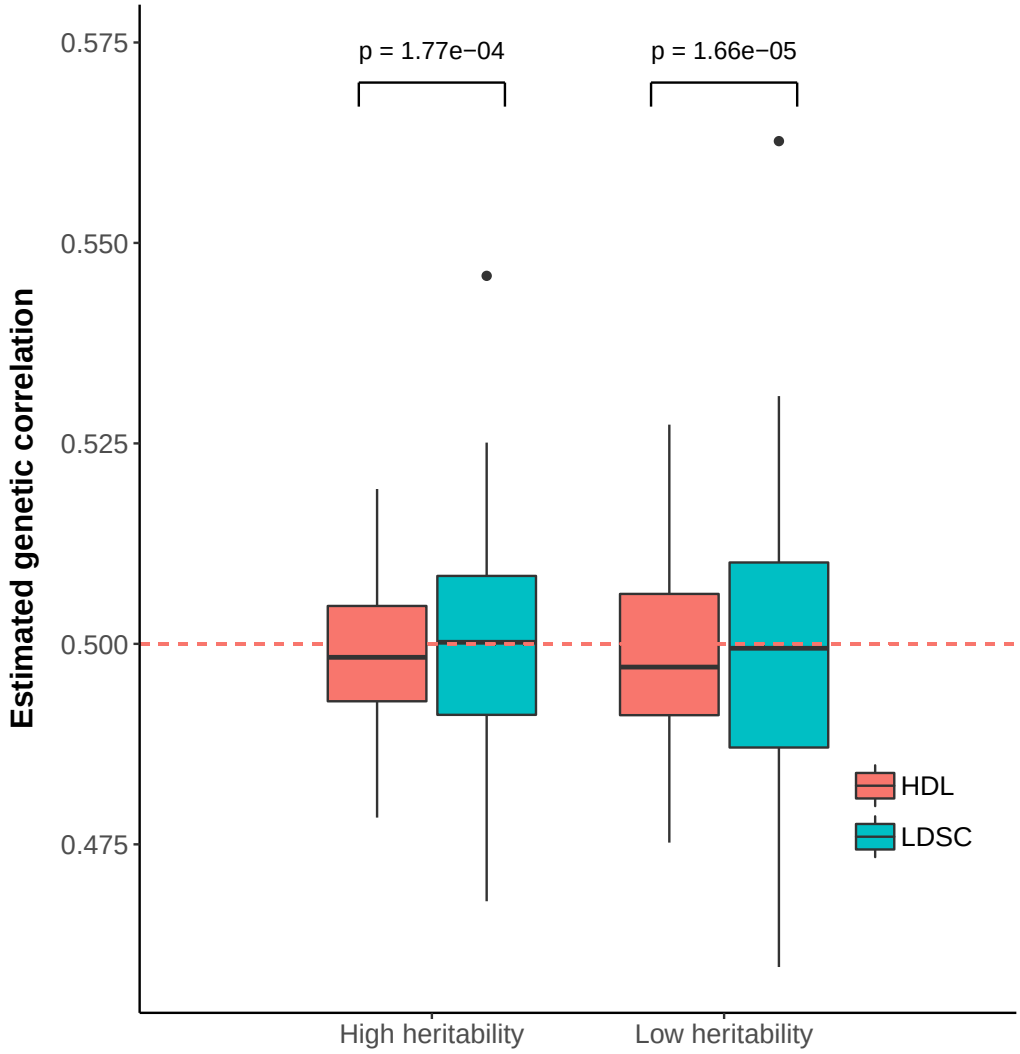
Supplementary Table 2: 435 genetic correlations among 30 phenotypes in UK Biobank. rg.HDL (s.e.), genetic correlation estimate and standard error given by HDL using UKBB imputed SNPs as reference panel; rg.LDSC (s.e.), genetic correlation estimate and standard error given by LDSC using default 1000 Genomes reference panel; p.HDL, P-value given by HDL; p.LDSC, P-value given by LDSC.

[See the Excel File]

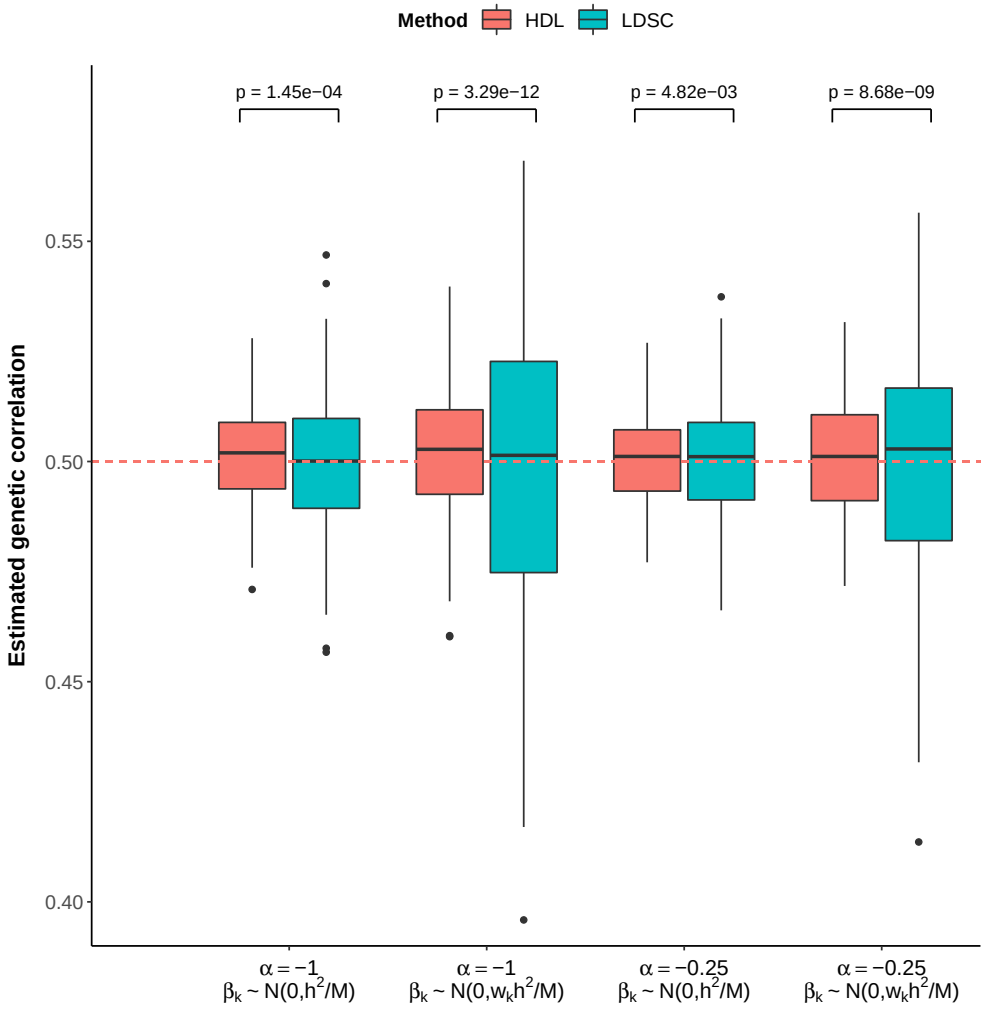
Supplementary Table 3: HDL, LDSC and LMM estimates of 55 genetic correlations among 11 phenotypes in UK Biobank. rg.LMM, genetic correlation estimate given by LMM; rg.HDL (s.e.), genetic correlation estimate and standard error given by HDL using UKBB imputed SNPs as reference panel; rg.LDSC (s.e.), genetic correlation estimate and standard error given by LDSC using default 1000 Genomes reference panel.

[See the Excel file]

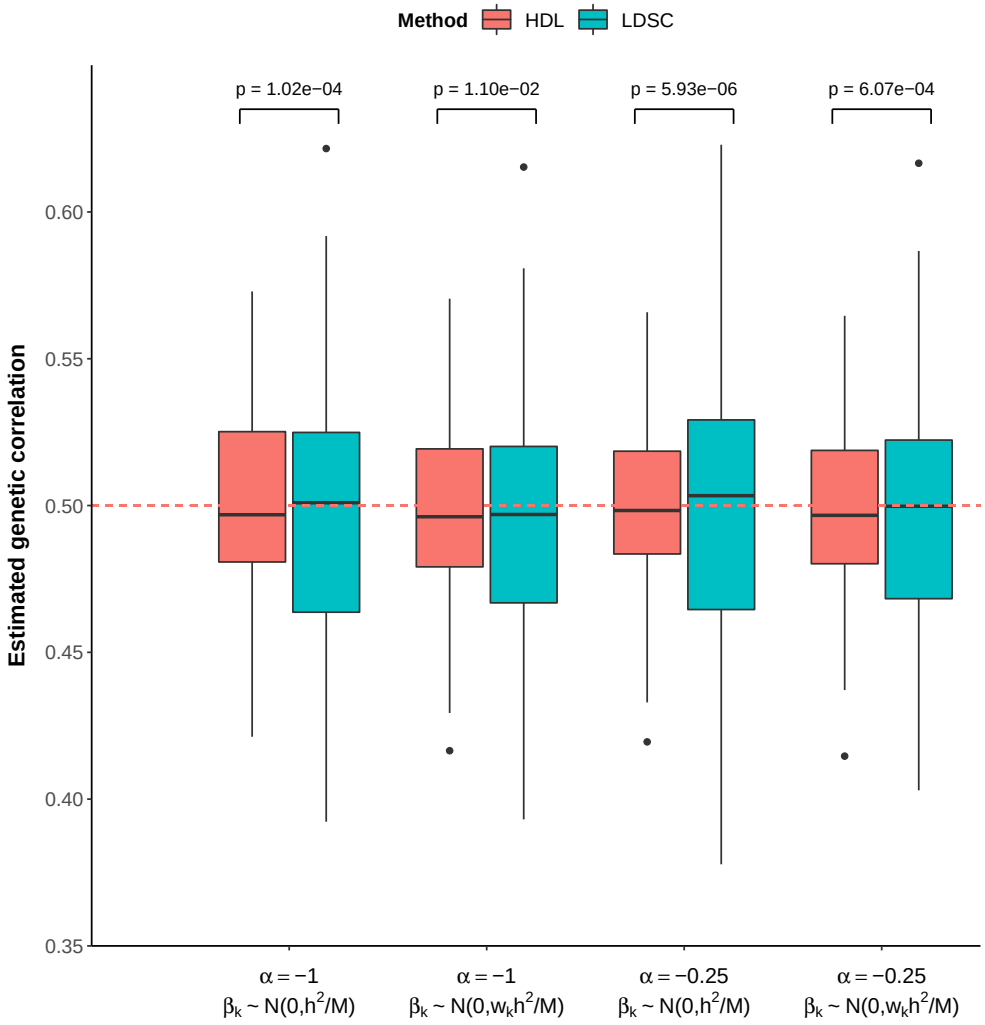
3 Supplementary Figures



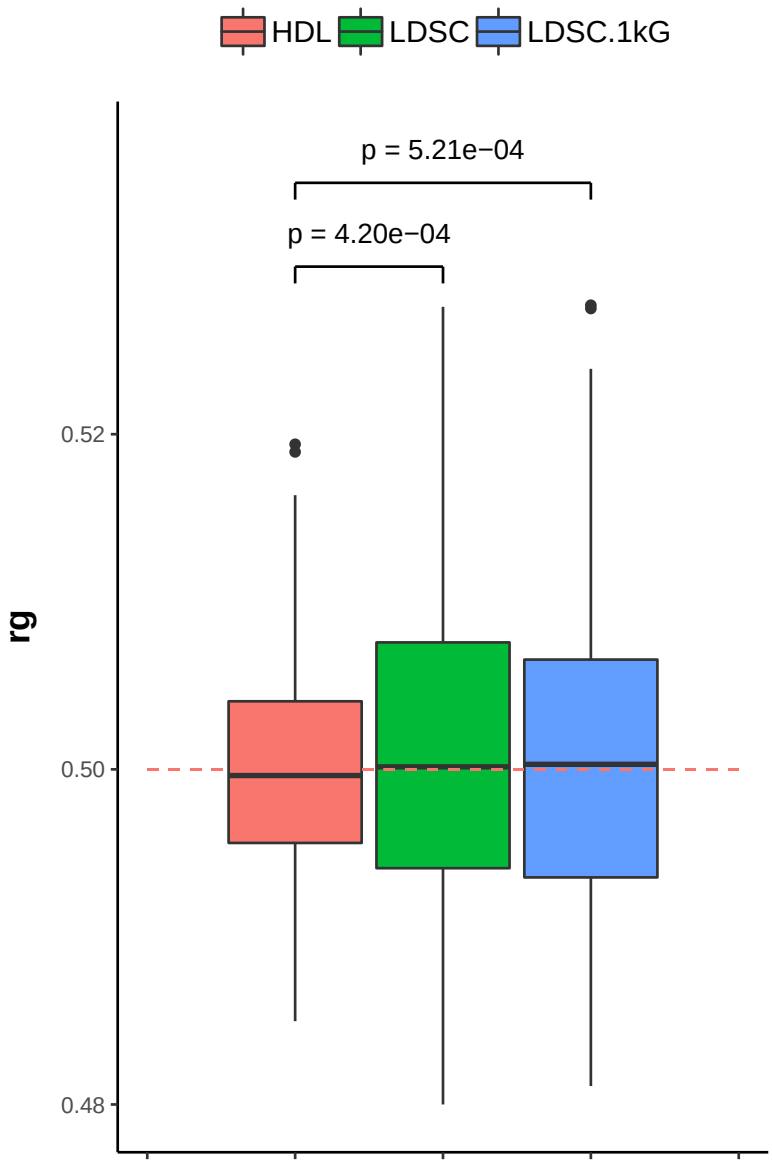
Supplementary Figure 1: Relative efficiency of HDL against LDSC when 100% SNPs are causal. In each heritability group, we generated 100 pairs of traits, where true genetic correlation and phenotypic correlation are 0.5. In the high heritability group, the heritability of the pair of traits is 0.6 and 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.



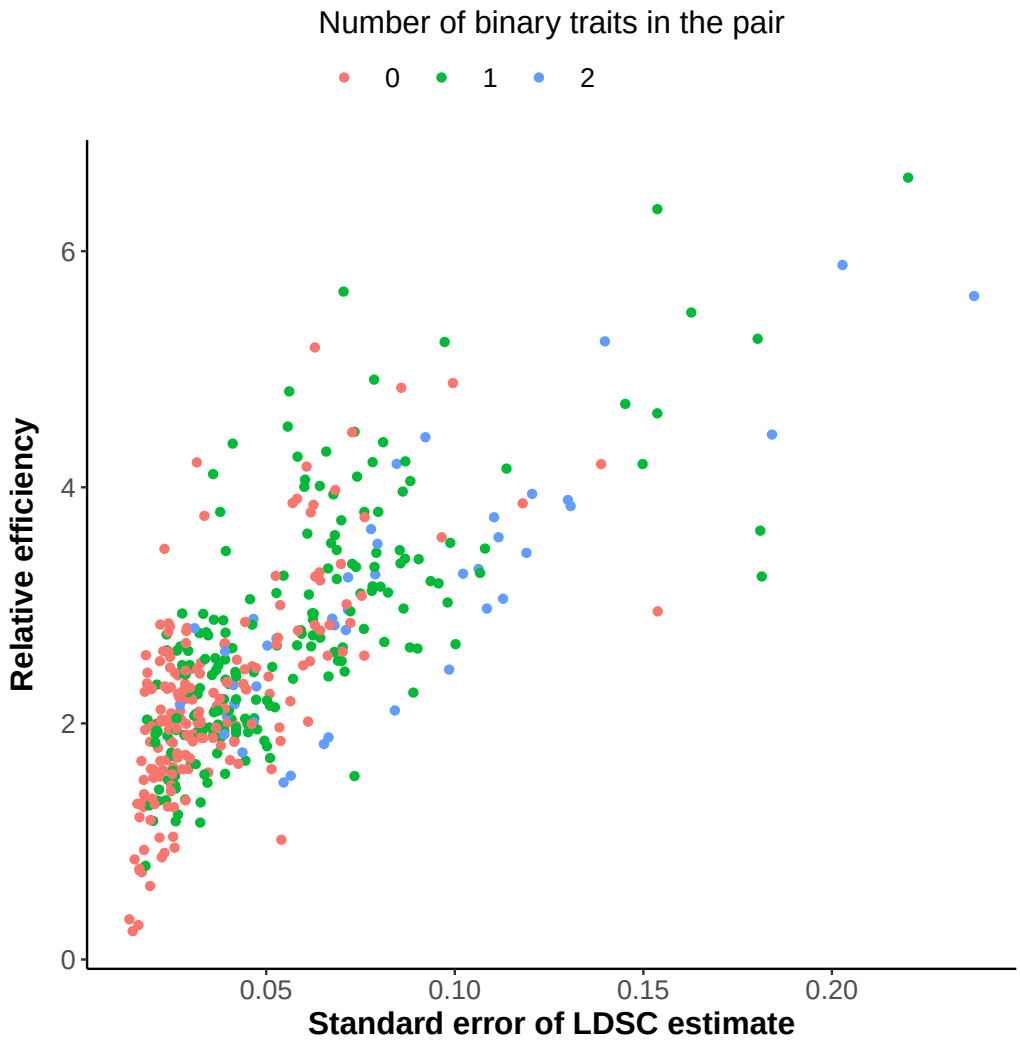
Supplementary Figure 2: Relative efficiency of HDL against LDSC under different model setups when 10% SNPs with MAF > 1% are causal. 52,914 out of 529,139 array SNPs with MAF > 1% were randomly selected as causal variants. 100 pairs of traits were generated, where true genetic correlation and phenotypic correlation are 0.5. The true phenotypes of trait i is generated from model $\boldsymbol{y}_i = \sum_{k=1}^{M} \mathbf{X}_{ik}\beta_{ik} + \boldsymbol{\epsilon}_i$, where $\mathbf{X}_{ik} = (\mathbf{Z}_{ik} - 2p_k\mathbf{1})[2p_k(1-p_k)]^{\alpha/2}$; $\mathbf{Z}_{ik}$ 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.



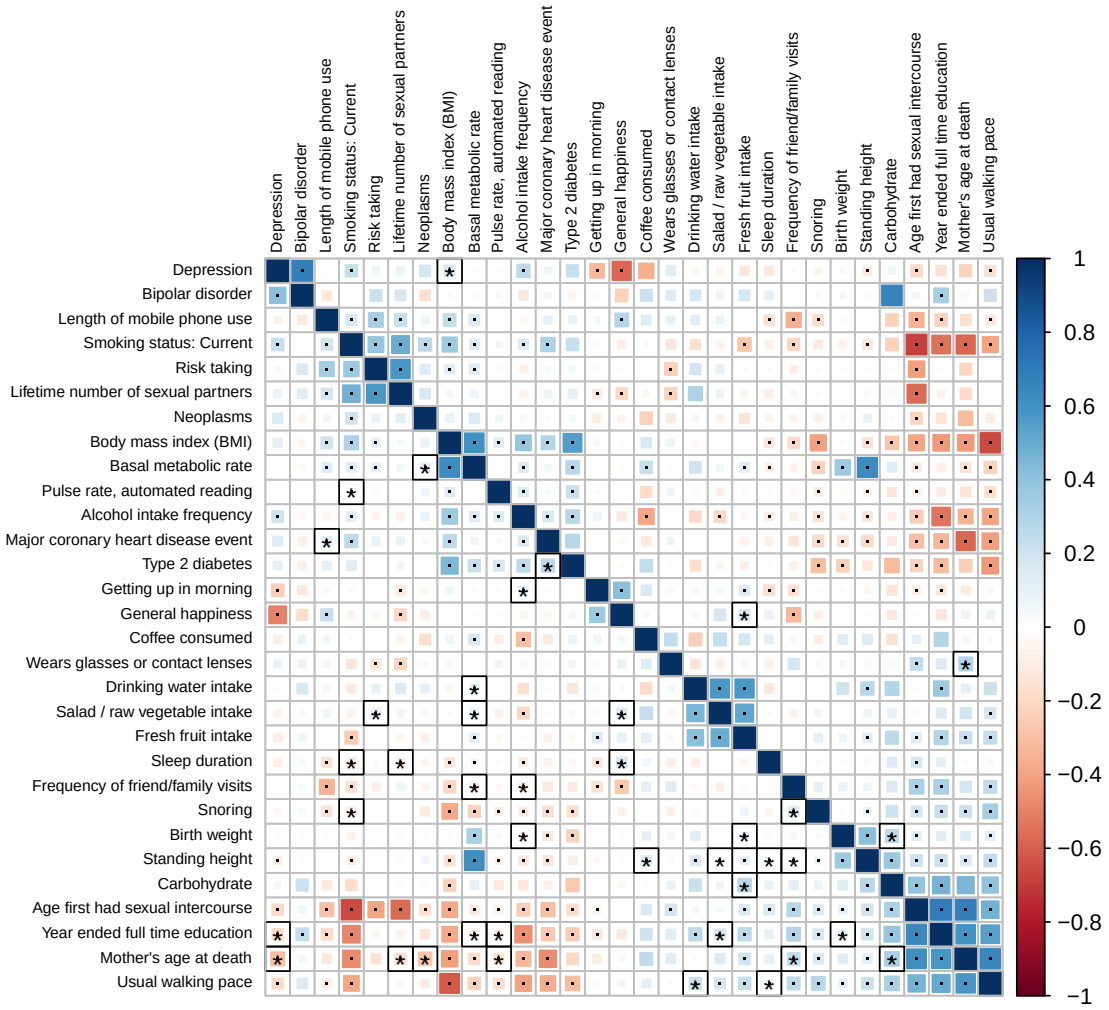
Supplementary Figure 3: Relative efficiency of HDL against LDSC under different model setups when 10% SNPs with 5% > MAF > 1% are causal. 52,914 out of 221,620 array SNPs with 5% > MAF > 1% were randomly selected as causal variants. 100 pairs of traits were generated, where true genetic correlation and phenotypic correlation are 0.5. The true phenotypes of trait i is generated from model $\boldsymbol{y}_i = \sum_{k=1}^{M} \mathbf{X}_{ik}\beta_{ik} + \boldsymbol{\epsilon}_i$, where $\mathbf{X}_{ik} = (\mathbf{Z}_{ik} - 2p_k\mathbf{1})[2p_k(1-p_k)]^{\alpha/2}$; $\mathbf{Z}_{ik}$ 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.



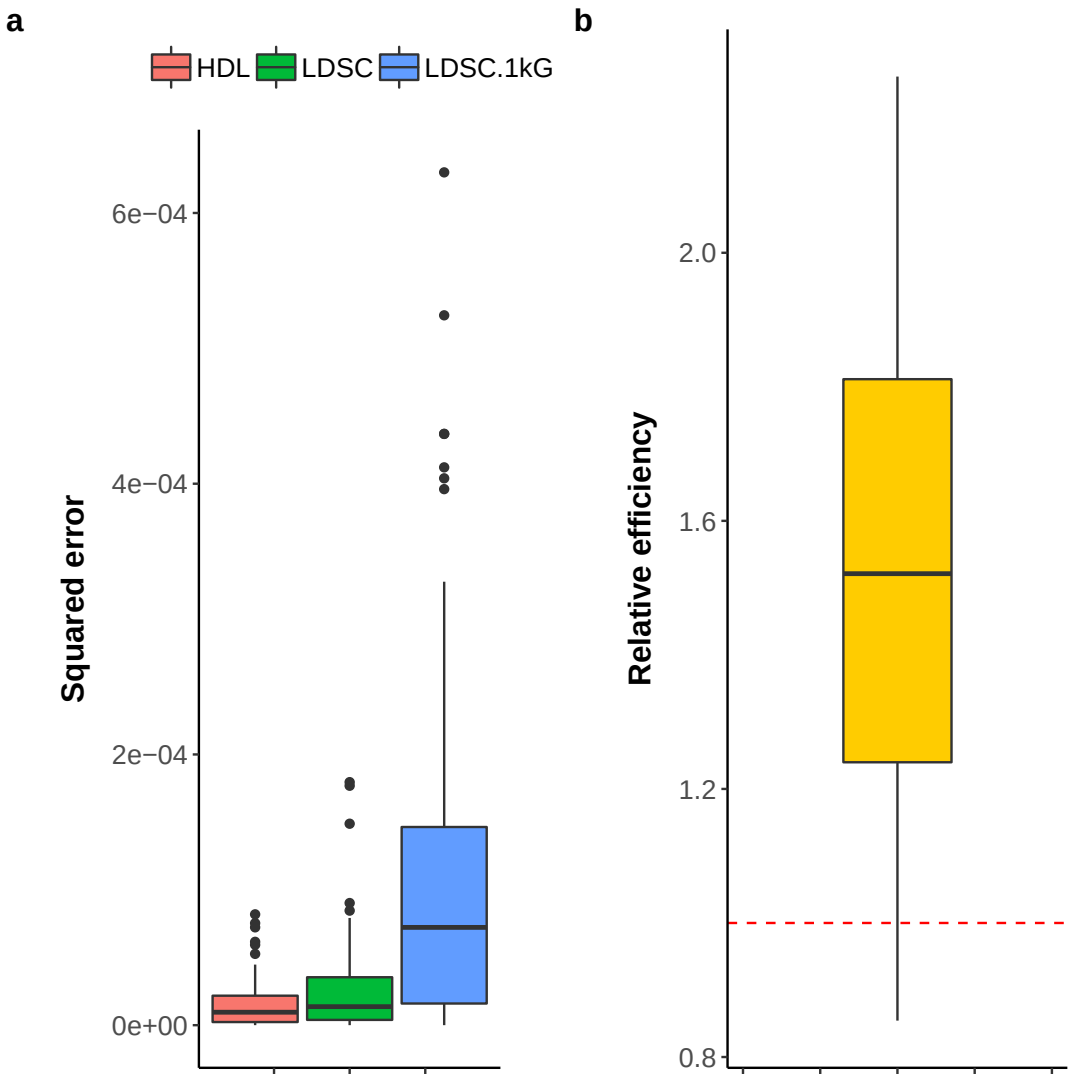
Supplementary Figure 4: Relative efficiency of HDL using imputed reference panel against LDSC. 100 pairs of traits were generated, where true heritabilities are 0.5, genetic correlation and phenotypic correlation are 0.5. The 1,029,876 imputed SNPs of ∼336,000 UKBB genomic British individuals were used 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.



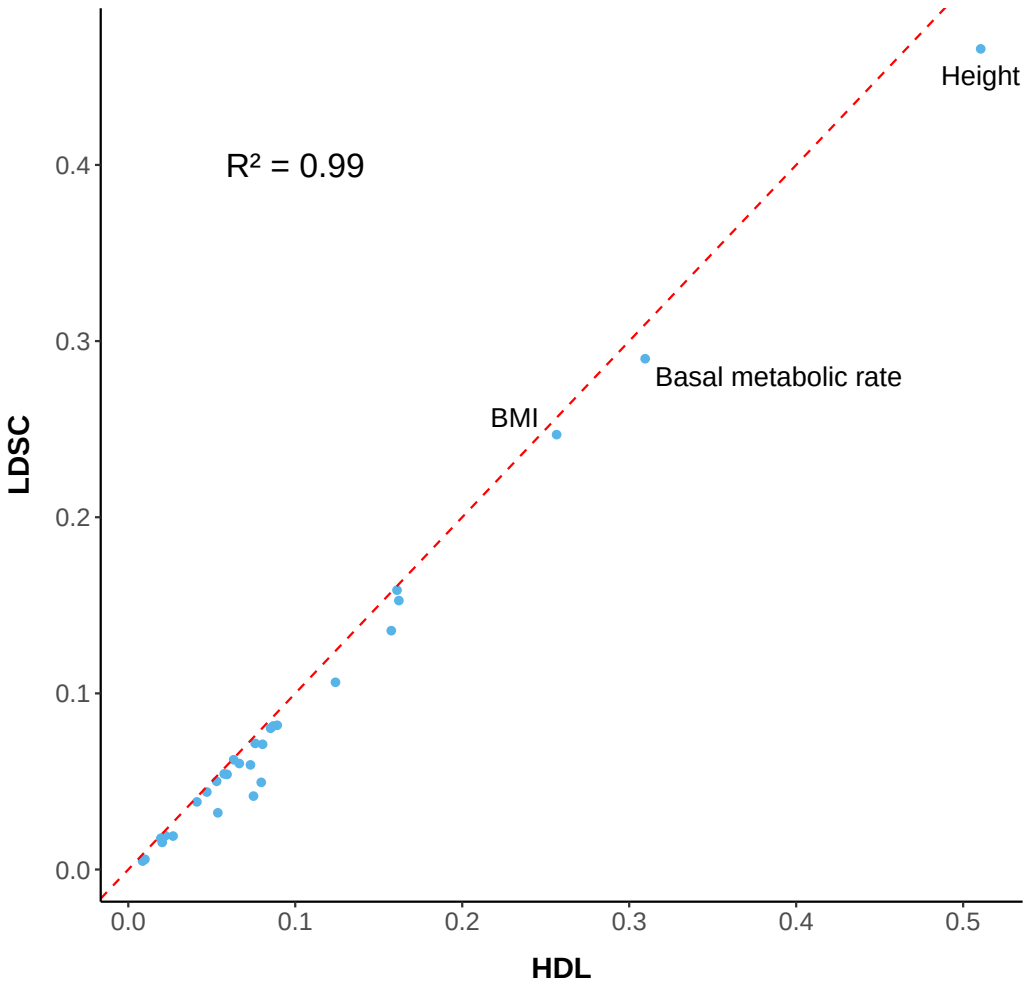
Supplementary Figure 5: Relative efficiency and standard error of LDSC estimate among 30 phenotypes in UK Biobank. Each dot represents genetic correlation results for one pair of traits among 435 pairs. The x-axis represents the standard error of the LDSC estimate. The y-axis represents the relative 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.



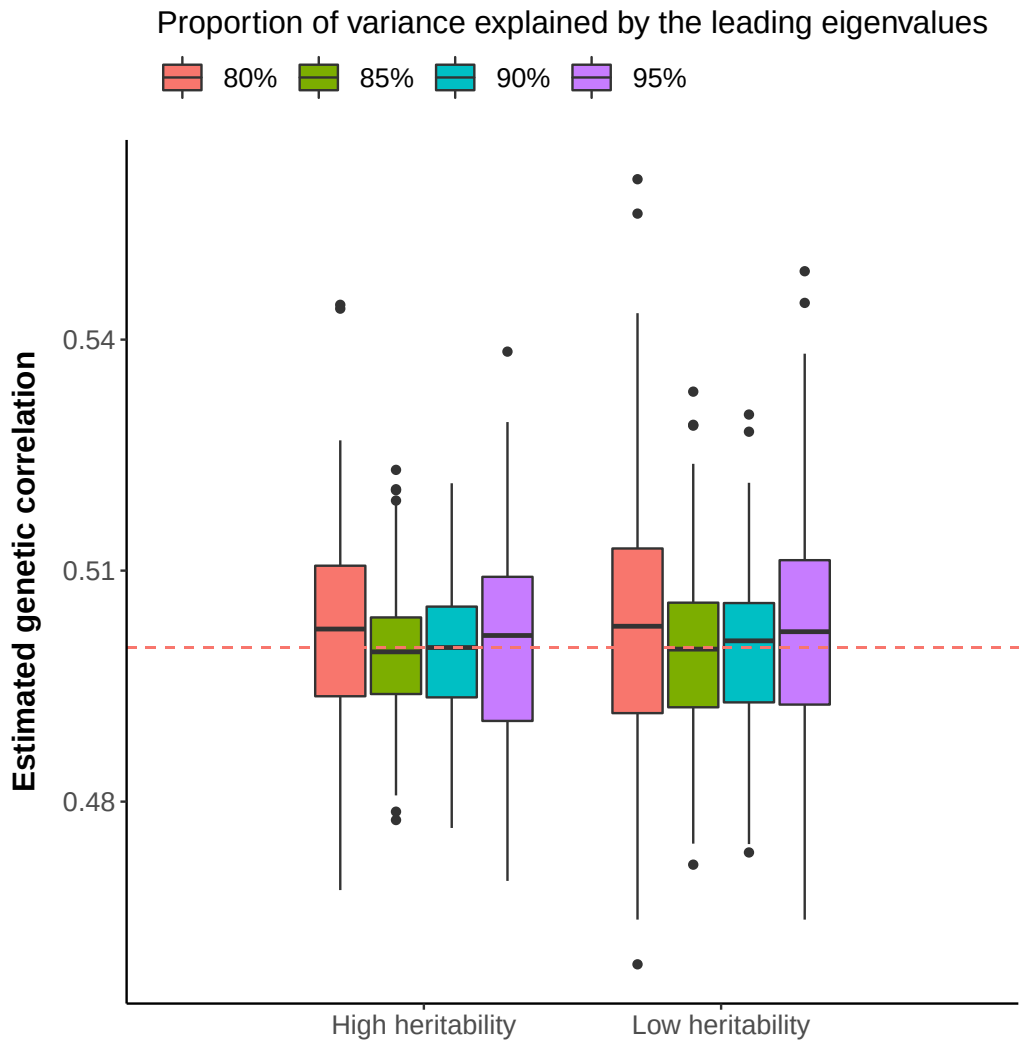
Supplementary Figure 6: Genetic correlation estimates from HDL and LDSC among 30 phenotypes in UK Biobank based on directly genotyped variants on the array. Lower triangle: HDL estimates; Upper triangle: LDSC estimates. The areas of the squares represent the absolute value of corresponding genetic correlations. 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.



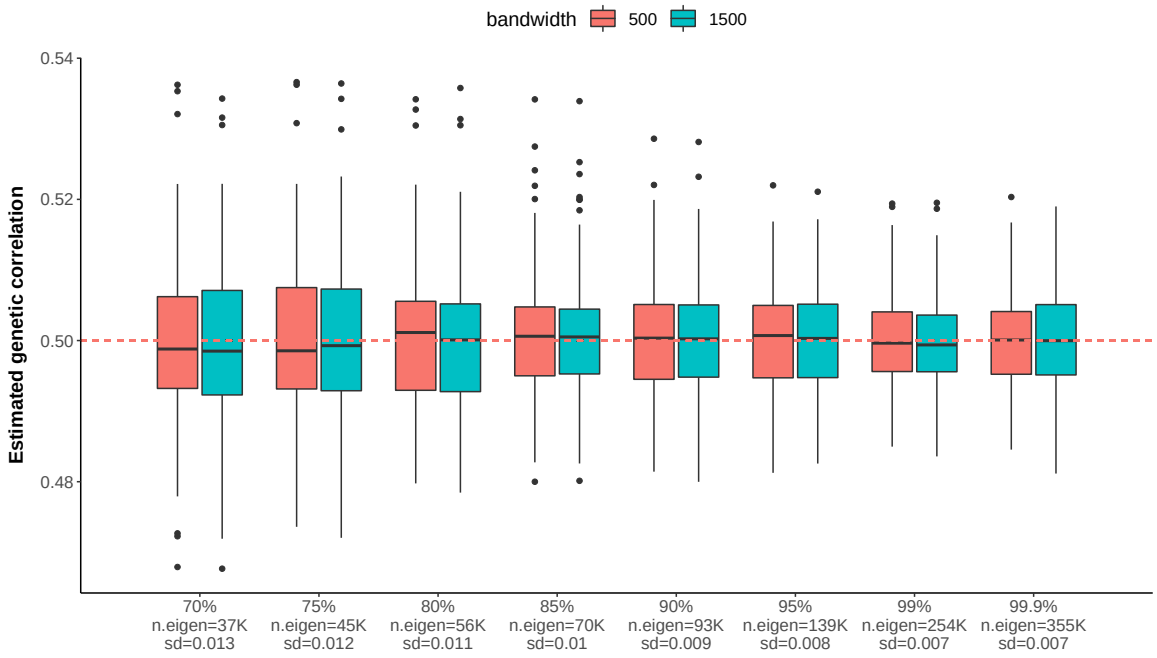
Supplementary Figure 7: Relative efficiency of HDL using imputed reference panel against LDSC for the estimation of heritability. a) 100 traits were generated using 14,867 imputed SNPs on chromosome 22 of ∼336,000 UKBB genomic British individuals, where true heritability was set to 0.05. LDSC and LDSC.1kG 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).



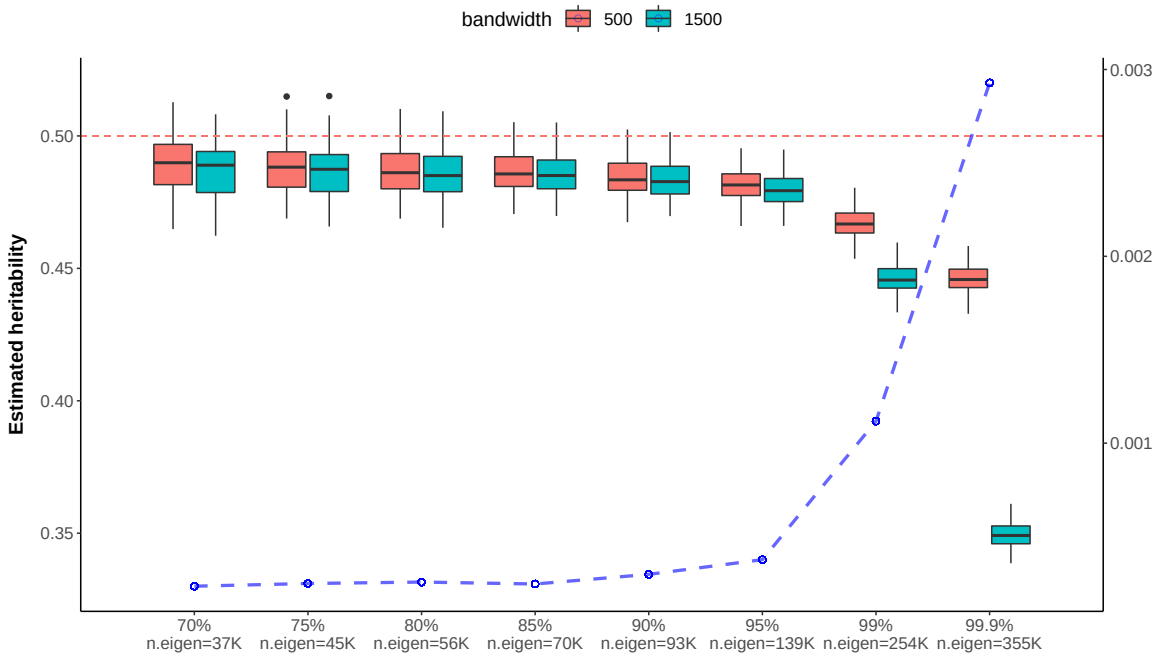
Supplementary Figure 8: Comparison of the heritability estimates from HDL and default LDSC across 30 UKBB phenotypes. The default LDSC uses the 1000 Genomes reference panel. HDL uses UKBB imputed markers as reference. R represents the correlation between the two sets of estimates. The red dashed line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



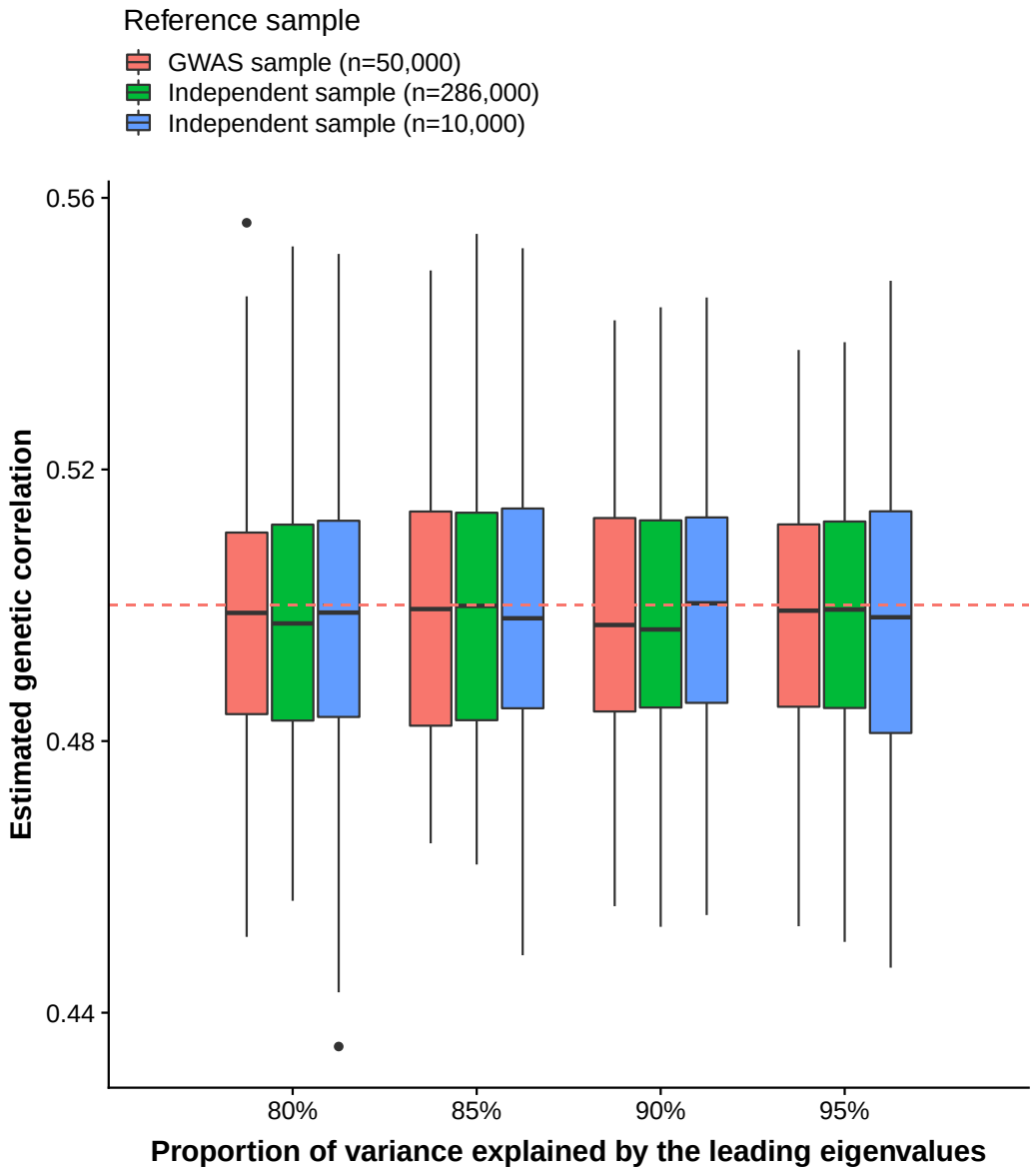
Supplementary Figure 9: HDL results where the LD matrix is approximated by different numbers of leading eigenvalues and eigenvectors. After performing eigen-decomposition to the LD matrix, leading eigenvalues explaining different amount of variances of the LD matrix and their corresponding eigenvectors 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).



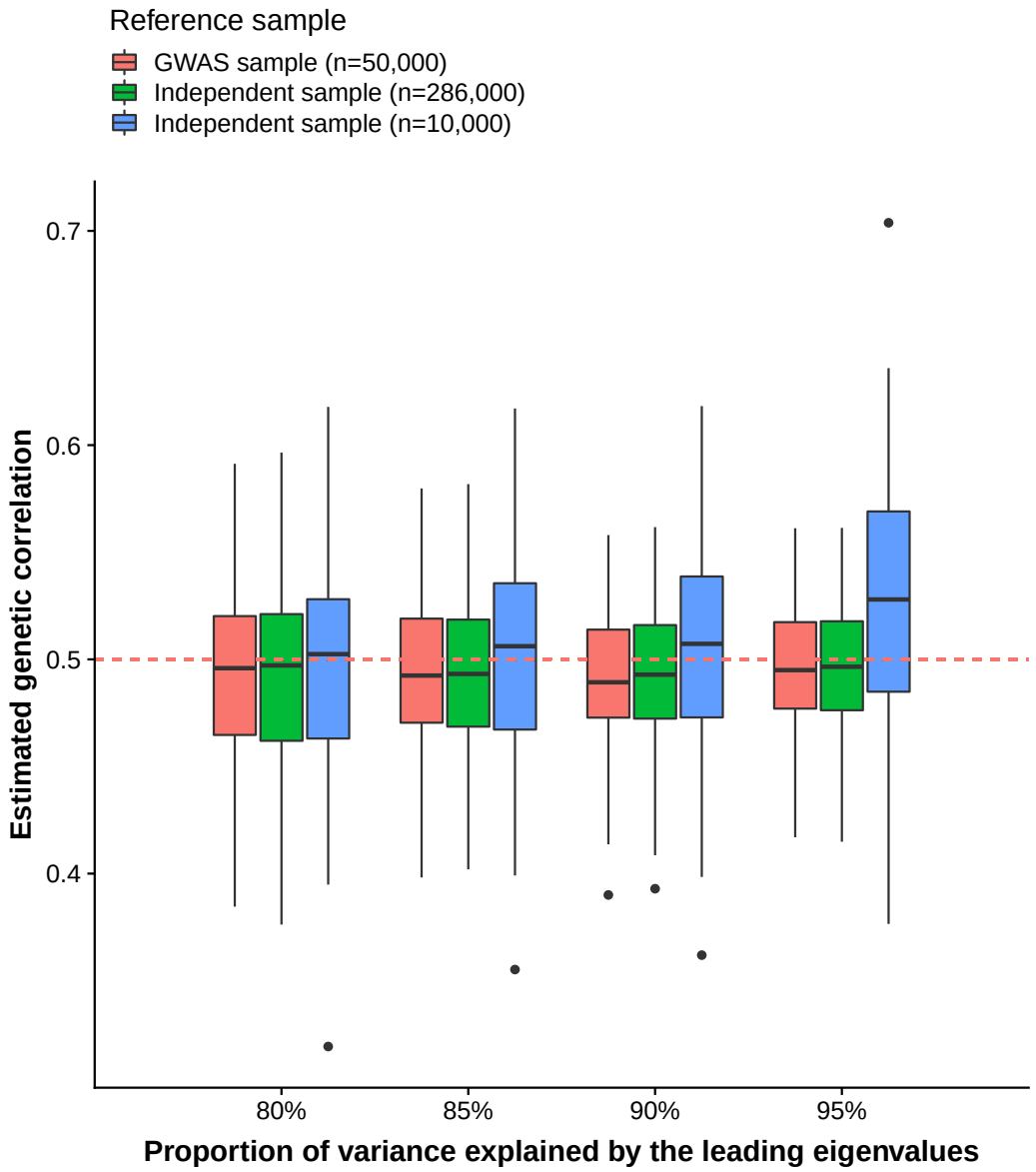
Supplementary Figure 10: Genetic correlation estimated by HDL using imputed reference panel, where the LD matrix is approximated by different numbers of leading eigenvalues and eigenvectors under two different bandwidths of the LD blocks. After performing eigen-decomposition to the LD matrix, leading eigenvalues explaining different amount of variances of the LD matrix and their corresponding eigenvectors were taken to approximate the LD matrix. 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. 102,988 (10% of 1,029,876) randomly sampled SNPs are set to be causal variants. The x-axis shows the proportion of variances explained by the leading eigenvalues, the corresponding number of leading eigenvalues and the corresponding standard deviation of genetic correlation estimates under 500 bandwidth.



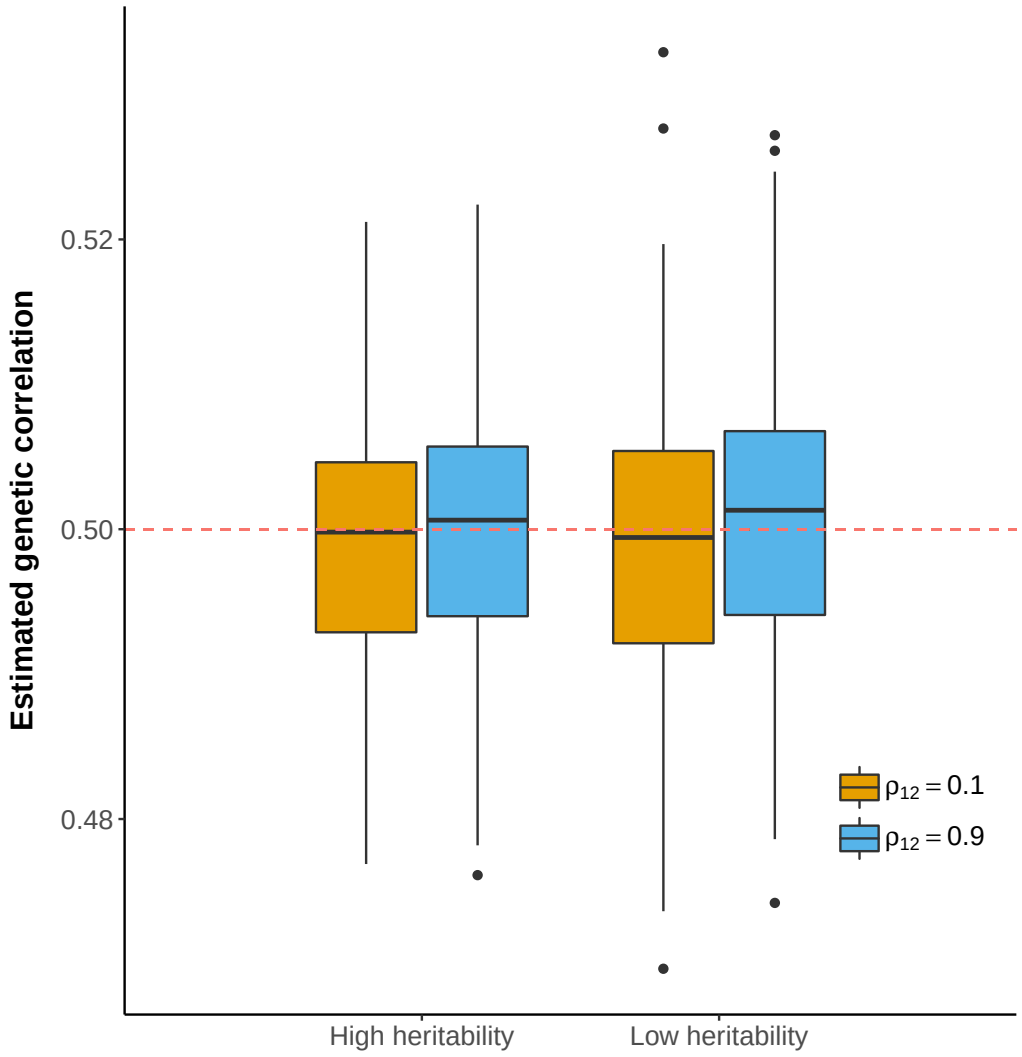
Supplementary Figure 11: Heritability estimated by HDL using imputed reference panel, where LD matrix is approximated by different numbers of leading eigenvalues and eigenvectors under two different bandwidths of the LD blocks. After performing eigen-decomposition to the LD matrix, leading eigenvalues explaining different amount of variances of the LD matrix and their corresponding eigenvectors were taken to approximate the LD matrix. 100 traits were generated using the 1,029,876 imputed SNPs of ∼336,000 UKBB genomic British individuals, where true heritability was set to 0.5. 102,988 (10% of 1,029,876) randomly sampled SNPs are set to be causal variants. The x-axis shows the proportion of variances explained by the leading eigenvalues and the corresponding number of leading eigenvalues. The blue dashed line and circles are the corresponding mean squared errors of heritability estimates under 500 bandwidth (y-axis on the right).



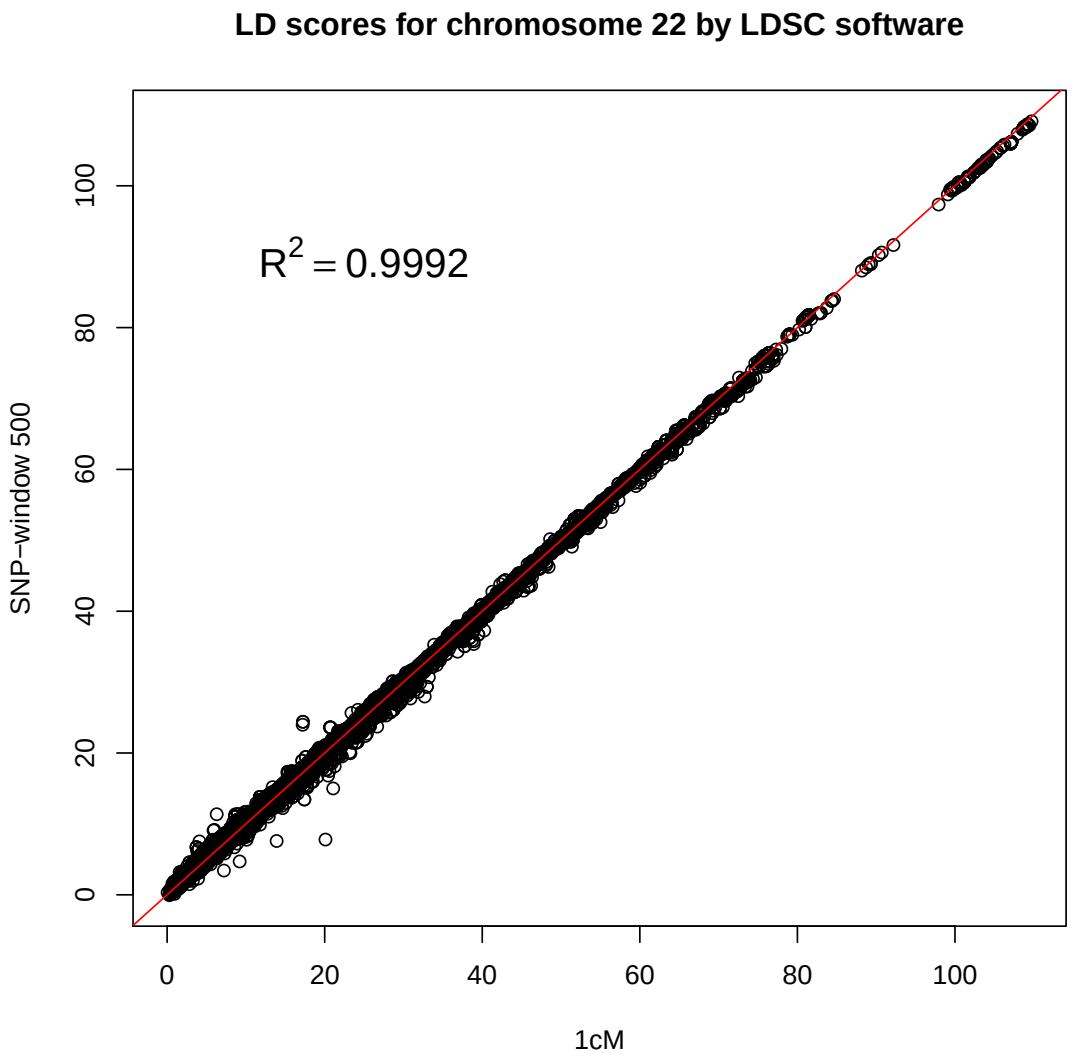
Supplementary Figure 12: HDL results based on different reference samples for high heritability group. 50,000 individuals were randomly sampled from 336,000 UKBB as the GWAS sample to generate GWAS summary statistics. The LD matrix is computed from the 307,519 array SNPs of (1) the GWAS sample; (2) the rest 286,000 individuals; (3) a 10,000 individuals random sample of the rest 286,000 individuals. After performing eigen-decomposition to the LD matrix, different numbers of leading eigenvalues and eigenvectors were taken to approximate the LD matrix. In this simulation, we generated 100 pairs of traits for the 50,000 individuals in the GWAS sample. True genetic correlation and phenotypic correlation are 0.5. The heritability of the pair of traits is 0.6 and 0.8 separately. 30,752 SNPs are causal (10% of 307,519).



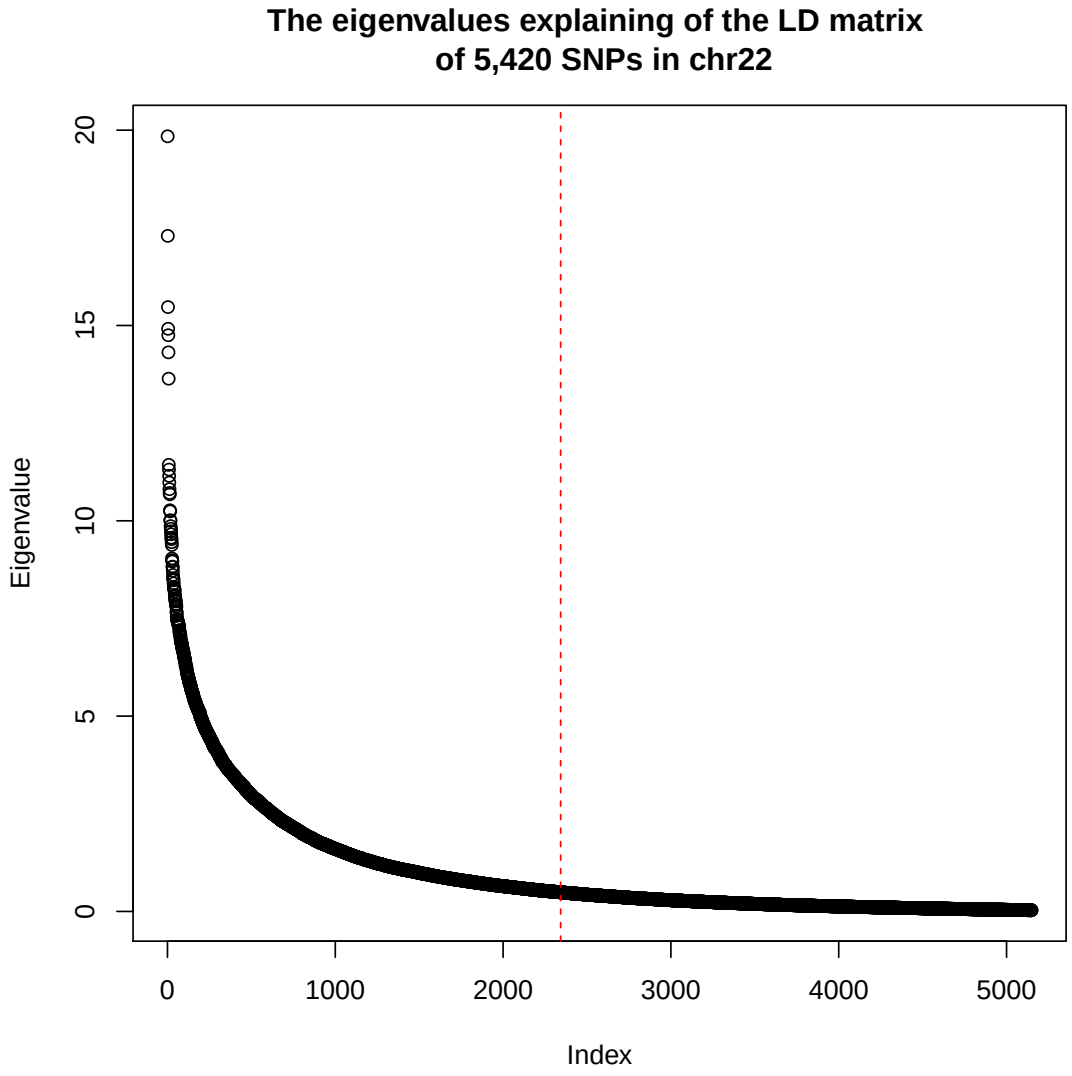
Supplementary Figure 13: HDL results based on different reference samples for low heritability group. 50,000 individuals were randomly sampled from 336,000 UKBB as the GWAS sample to generate GWAS summary statistics. The LD matrix is computed from the 307,519 array SNPs of (1) the GWAS sample; (2) the rest 286,000 individuals; (3) a 10,000 individuals random sample of the rest 286,000 individuals. After performing eigen-decomposition to the LD matrix, different numbers of leading eigenvalues and eigenvectors were taken to approximate the LD matrix. In this simulation, we generated 100 pairs of traits for the 50,000 individuals in the GWAS sample. True genetic correlation and phenotypic correlation are 0.5. The heritability of the pair of traits is 0.2 and 0.4 separately. 30,752 SNPs are causal (10% of 307,519).



Supplementary Figure 14: Genetic correlation estimated by HDL under different levels of residual correlation (ρ_{12}). In each heritability group and residual correlation level, we generated 100 pairs of traits. The true genetic correlation is set to 0.5. The level of residual correlation is either 0.1 or 0.9. 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. 30,752 SNPs are causal (10% of 307,519).



Supplementary Figure 15: Comparison of LD scores estimated based on 1cM windows and 500-SNP windows. LD scores were computed using the example 1000 Genomes genotype data included in the LDSC software.



Supplementary Figure 16: 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.