BRIEF RESEARCH REPORT
published: 24 August 2021
doi: 10.3389/fgene.2021.665252

Improved Estimation of Phenotypic Correlations Using Summary Association Statistics

Ting Li[1†], *Zheng Ning*[2†] and *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, Stockholm, Sweden, [3] Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom

OPEN ACCESS

Edited by:
Hugues Aschard,
Institut Pasteur, France

Reviewed by:
Doug Speed,
Aarhus University, Denmark
Tao He,
San Francisco State University, United States
Yuning Chen,
Boston University, United States

***Correspondence:**
Xia Shen
xia.shen@ed.ac.uk

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *07 February 2021*
Accepted: *25 June 2021*
Published: *24 August 2021*

Citation:
Li T, Ning Z and Shen X (2021) Improved Estimation of Phenotypic Correlations Using Summary Association Statistics. Front. Genet. 12:665252. doi: 10.3389/fgene.2021.665252

Estimating the phenotypic correlations between complex traits and diseases based on their genome-wide association summary statistics has been a useful technique in genetic epidemiology and statistical genetics inference. Two state-of-the-art strategies, Z-score correlation across null-effect single nucleotide polymorphisms (SNPs) and LD score regression intercept, were widely applied to estimate phenotypic correlations. Here, we propose an improved Z-score correlation strategy based on SNPs with low minor allele frequencies (MAFs), and show how this simple strategy can correct the bias generated by the current methods. The low MAF estimator improves phenotypic correlation estimation, thus it is beneficial for methods and applications using phenotypic correlations inferred from summary association statistics.

Keywords: phenotypic correlation, genome-wide association, low MAF estimator, LD score regression, genetic correlation, minor allele frequency

1. INTRODUCTION

Phenotypic correlation is an essential parameter that helps understand observational correlations between complex traits and the etiological perspectives underlying complex diseases. Conventionally, estimation of the phenotypic correlation between a pair of phenotypes, by definition, is straightforward in a sample where both phenotypes are measured. Depending on the distribution of each phenotype, the estimated phenotypic correlation serves as a sufficient statistic for many linear statistical models, such as ordinary linear and logistic regressions, allowing us to assess parameters such as odds ratios of risk factors on disease outcomes.

Since a large number of genome-wide association studies (GWAS) were conducted, many GWASed phenotypes had measurements in an overlapping set of individuals, where many were from more than one participating cohort in GWAS meta-analysis. In practice, inference of the phenotypic correlations across these phenotypes would be complicated if estimating using the conventional way, which requires individual-level phenotypic data and subsequent meta-analysis. Fortunately, the phenotypic correlations can be estimated based on established GWAS summary statistics, especially when the proportion of sample overlap between two GWASed phenotypes is large. Two state-of-the-art strategies were proposed:

1. *"Z-cut" estimator*: The phenotypic correlation can be estimated by the correlation between the two sets of GWAS estimated effects or Z-scores, assuming the genetic effect per SNP is tiny or even null (Stephens, 2013; Zhu et al., 2015; Cichonska et al., 2016; Shen et al., 2017).
2. *LDSC intercept.* The phenotypic correlation can be estimated by the intercept of a bivariate linkage disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015; Turley et al., 2018; Zheng et al., 2018).

Both estimators have reasonable performance in practice, however, bias exists for both strategies. Stephens (2013) reasoned that the correlation between Z-scores for the two phenotypes under the null is the same as the phenotypic correlation, thus "a set of putative null SNPs" were selected by taking SNPs with $|z| < 2$. The same idea was also adopted by later studies (Zhu et al., 2015; Shen et al., 2017). The tool metaCCA (Cichonska et al., 2016) neglected the null effect requirement, as the genetic effect per variant is tiny, and computed the correlation between Z-scores across as many SNPs as possible. However, the Z-cut estimator can generate bias due to its constrain on the summary statistics of the SNPs (Zheng et al., 2018). LDSC intercept performs better and thus was adopted in statistical methods that requires pre-calculated phenotypic correlations (Turley et al., 2018; Zheng et al., 2018), but the intercept collects noise generated by population substructure, which may also lead to biased estimates of phenotypic correlations (Yengo et al., 2018).

Here, we revisit the correlation between GWAS summary statistics of two phenotypes and propose an alternative approach to select variants for the Z-score correlation estimation strategy. We show that selecting SNPs with low minor allele frequencies (MAFs) can lead to simple and consistent estimation of phenotypic correlations based on multi-SNP Z-score correlations. Via simulations, we show that the "low MAF" estimator can overcome bias generated by the Z-cut estimator and the LDSC intercept. With higher estimation efficiency, when applied to UK Biobank GWAS results, the low MAF estimator could discover 30% more significant phenotypic correlations than using the LDSC intercept.

2. METHODS

We start by deriving a general mathematical form of the correlation between the summary statistics of two phenotypes $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, centered at a zero mean. The sample sizes for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are N_1 and N_2, respectively, and the overlapping part of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ has a length of N_0. For a single genetic variant in an association analysis, the model is $\boldsymbol{y}_i = \boldsymbol{g}_i\beta_i + \boldsymbol{e}_i$ $(i = 1, 2)$, where $\boldsymbol{g}_i$ is the vector of genotypic values coded as 0, 1, and 2, and $\boldsymbol{e}_i$ are the residuals. Only β_i and $\boldsymbol{e}_i$ are random in the model. Assuming Hardy–Weinberg equilibrium (HWE), for SNP j, the genotypic values of $\boldsymbol{g}_i$ has a sample mean of $2f_j$ and a sample standard deviation of $\sqrt{2f_j(1-f_j)}$, where f_j is the allele frequency of the coding allele. $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ may differ due to different levels of sample overlap between the two phenotypes. At the single SNP j (omitted the subscripts j for simplicity),

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \sim \text{dist}\left(\mathbf{0}, \begin{bmatrix}\sigma^2_{\beta_1} & r_G\sigma_{\beta_1}\sigma_{\beta_2}\\ r_G\sigma_{\beta_1}\sigma_{\beta_2} & \sigma^2_{\beta_2}\end{bmatrix}\right), \tag{1}$$

and

$$\begin{pmatrix}\boldsymbol{e}_1\\ \boldsymbol{e}_2\end{pmatrix} \sim \text{dist}\left(\mathbf{0}, \begin{bmatrix}\sigma_1^2\boldsymbol{I}_{N_1\times N_1} & r_E\sigma_1\sigma_2\begin{pmatrix}\boldsymbol{I}_{N_0\times N_0} & \\ & \mathbf{0}_{N_1'\times N_2'}\end{pmatrix}\\ r_E\sigma_1\sigma_2\begin{pmatrix}\boldsymbol{I}_{N_0\times N_0} & \\ & \mathbf{0}_{N_2'\times N_1'}\end{pmatrix} & \sigma_2^2\boldsymbol{I}_{N_2\times N_2}\end{bmatrix}\right), \tag{2}$$

where $N_i' = N_i - N_0$, r_G is the underlying genetic correlation at SNP j, and r_E is the residual correlation. dist() denotes a multivariate distribution with a given mean vector and a variance–covariance matrix. In an association study, r_G is unidentifiable at a single SNP. The estimated genetic effects are $\hat{\beta}_i = \boldsymbol{g}_i'\boldsymbol{y}_i/\boldsymbol{g}_i'\boldsymbol{g}_i$, then

$$\begin{aligned}\text{var}(\hat{\beta}_i) &= \frac{\text{var}(\boldsymbol{g}_i'\boldsymbol{y}_i)}{(\boldsymbol{g}_i'\boldsymbol{g}_i)^2}\\ &= \frac{\text{var}(\boldsymbol{g}_i'\boldsymbol{g}_i\beta_i + \boldsymbol{g}_i'\boldsymbol{e}_i)}{(\boldsymbol{g}_i'\boldsymbol{g}_i)^2}\\ &= \sigma^2_{\beta_i} + \sigma_i^2(\boldsymbol{g}_i'\boldsymbol{g}_i)^{-1}.\end{aligned} \tag{3}$$

Denote $\boldsymbol{g}$ as the overlapping part of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, and let $\boldsymbol{x}$ and $\boldsymbol{z}$ be the rest parts of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, respectively. We have $\boldsymbol{g}_i'\boldsymbol{g}_i \approx 2f(1-f)N_i$ $(i = 1, 2)$ and $\boldsymbol{g}'\boldsymbol{g} \approx 2f(1-f)N_0$. So that, defining $z_i = \hat{\beta}_i/\sqrt{\text{var}(\hat{\beta}_i)}$, we have

$$\begin{aligned}&\text{cor}(\hat{\beta}_1, \hat{\beta}_2)\\ &= \text{cor}(z_1, z_2)\\ &= \frac{\text{cov}(\boldsymbol{g}_1'\boldsymbol{y}_1, \boldsymbol{g}_2'\boldsymbol{y}_2)}{\sqrt{(\boldsymbol{g}_1'\boldsymbol{g}_1)^2\sigma^2_{\beta_1} + \boldsymbol{g}_1'\boldsymbol{g}_1\sigma_1^2}\sqrt{(\boldsymbol{g}_2'\boldsymbol{g}_2)^2\sigma^2_{\beta_2} + \boldsymbol{g}_2'\boldsymbol{g}_2\sigma_2^2}}\\ &= \frac{(\boldsymbol{g}_1'\boldsymbol{g}_1)(\boldsymbol{g}_2'\boldsymbol{g}_2)\text{cov}(\beta_1, \beta_2) + \boldsymbol{g}_1'\text{cov}(\boldsymbol{e}_1, \boldsymbol{e}_2)\boldsymbol{g}_2}{\sqrt{(\boldsymbol{g}_1'\boldsymbol{g}_1)^2\sigma^2_{\beta_1} + \boldsymbol{g}_1'\boldsymbol{g}_1\sigma_1^2}\sqrt{(\boldsymbol{g}_2'\boldsymbol{g}_2)^2\sigma^2_{\beta_2} + \boldsymbol{g}_2'\boldsymbol{g}_2\sigma_2^2}}\\ &= \frac{(\boldsymbol{g}_1'\boldsymbol{g}_1)(\boldsymbol{g}_2'\boldsymbol{g}_2)r_G\sigma_{\beta_1}\sigma_{\beta_2} + r_E\sigma_1\sigma_2(\boldsymbol{g}', \boldsymbol{x}')\begin{pmatrix}\boldsymbol{I}_{N_0\times N_0} & \\ & \mathbf{0}_{N_1'\times N_2'}\end{pmatrix}\begin{pmatrix}\boldsymbol{g}\\ \boldsymbol{z}\end{pmatrix}}{\sqrt{(\boldsymbol{g}_1'\boldsymbol{g}_1)^2\sigma^2_{\beta_1} + \boldsymbol{g}_1'\boldsymbol{g}_1\sigma_1^2}\sqrt{(\boldsymbol{g}_2'\boldsymbol{g}_2)^2\sigma^2_{\beta_2} + \boldsymbol{g}_2'\boldsymbol{g}_2\sigma_2^2}}\\ &= \frac{(\boldsymbol{g}_1'\boldsymbol{g}_1)(\boldsymbol{g}_2'\boldsymbol{g}_2)r_G\sigma_{\beta_1}\sigma_{\beta_2} + \boldsymbol{g}'\boldsymbol{g}r_E\sigma_1\sigma_2}{\sqrt{(\boldsymbol{g}_1'\boldsymbol{g}_1)^2\sigma^2_{\beta_1} + \boldsymbol{g}_1'\boldsymbol{g}_1\sigma_1^2}\sqrt{(\boldsymbol{g}_2'\boldsymbol{g}_2)^2\sigma^2_{\beta_2} + \boldsymbol{g}_2'\boldsymbol{g}_2\sigma_2^2}}\\ &= \frac{2f(1-f)\sqrt{N_1N_2}r_G\sigma_{\beta_1}\sigma_{\beta_2} + N_0/\sqrt{N_1N_2}r_E\sigma_1\sigma_2}{\sqrt{2f(1-f)N_1\sigma^2_{\beta_1} + \sigma_1^2}\sqrt{2f(1-f)N_2\sigma^2_{\beta_2} + \sigma_2^2}},\end{aligned} \tag{4}$$

When $\sigma_{\beta_i} = 0$ $(i = 1, 2)$, i.e., for any variant with null genetic effect, the above equation simplifies to

$$\text{cor}(\hat{\beta}_1, \hat{\beta}_2) = \text{cor}(z_1, z_2) = \frac{N_0}{\sqrt{N_1N_2}}r_E = \frac{N_0}{\sqrt{N_1N_2}}r(\boldsymbol{y}_1, \boldsymbol{y}_2) \tag{5}$$

where $r(\boldsymbol{y}_1, \boldsymbol{y}_2)$ is the phenotypic correlation based on completely overlapped individual-level data. Thus, in order to estimate

$r(\boldsymbol{y}_1, \boldsymbol{y}_2)$, we can estimate $\text{cor}(z_1, z_2)$ instead. Particularly, for perfectly overlap samples, i.e., $N_0 = N_1 = N_2$, we have $\text{cor}(z_1, z_2) = r(\boldsymbol{y}_1, \boldsymbol{y}_2)$, which is the phenotypic correlation estimator derived by Zhu et al. (Zhu et al., 2015). Our theory above in Equation (4) is an extension of Zhu et al.'s theory, covering the substantial amount of genetic correlation across the genome. Only when σ_{β_i} or f is zero and the samples perfectly overlap between the two traits, Equation (4) reduces to Zhu et al.'s result. Equation (4) shows the reasoning behind the low MAF estimator, i.e., in practice, one can hardly control σ_{β_i} but f to be close to zero, so that the correlation between Z-scores becomes close to the phenotypic correlation, subject to a shrinkage factor if the samples do not perfectly overlap.

The result suggests that the phenotypic correlation between the two phenotypes $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, subject to a shrinkage factor corresponding to sample overlap, can be estimated by the sample correlation of the summary statistics across any sufficient number of null variants. This leads to a commonly adopted strategy of estimating the phenotypic correlation from summary association statistics by taking a subset with, e.g., $|z_i| < 2\ (i = 1, 2)$. However, we will show that such thresholding may introduce bias into the correlation estimate.

According to Equation (4), null genetic effect for the variant is a sufficient but not necessary condition for $\text{cor}(z_1, z_2)$ to reduce to Equation (5). When $f = 0$, Equation (4) also becomes (5). In practice, the phenotypic correlation can be estimated by the correlation of the summary statistics across a sufficient number of variants with very low MAFs, *regardless* of whether the genetic effects are null. The thresholding on the MAF does not directly introduce a threshold on β_i or z_i so that not prone to bias in the phenotypic correlation estimation.

2.1. Simulation Settings

We conducted a series of simulations to compare the low MAF estimators with the Z-cut estimators. Based on the real UK Biobank genotypes, two phenotypes were simulated based on the 784,256 genotyped SNPs in the UK Biobank and the model:

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i \qquad (6)$$

where $i = 1, 2$ is the phenotype index, $\boldsymbol{X}_i$ is the matrix of genotypic values, $\boldsymbol{\beta}_i$ is the vector of genetic effects, and $\boldsymbol{e}_i$ are the residuals. Each column of $\boldsymbol{X}_i$ was standardized to have a zero mean and unit variance. Two heritability values (h^2) for the phenotypes were considered: 0.3 and 0.6. The genetic effects and residuals were drawn from a Gaussian distribution with corresponding variance components: $\boldsymbol{\beta}_i \sim N(\boldsymbol{0}, (h^2/M)\boldsymbol{I})$, where M represents the number of causal variants, and $\boldsymbol{e}_i \sim N(\boldsymbol{0}, (1 - h^2)\boldsymbol{I})$. Each phenotype was simulated for 168,000 genomic British individuals. Two different scenarios of the proportion of causal SNPs were considered: 10% randomly selected SNPs and 100%. Three scenarios of the true genetic correlation (correlation between the $\boldsymbol{\beta}_i$ vectors) were considered: 0, 0.5, and 1. Three scenarios of sample overlap proportions were considered: 0 (no overlap), 0.5 (half overlapped), and 1 (perfectly matched).

Nine different methods for phenotypic correlation estimation were considered, including the true phenotypic correlation estimator was based on the individual-level phenotype data and the other eight that use the correlation between Z-scores for the estimation. For the illustration purpose, an estimator based on the Z-scores of 500 simulated SNPs with random genotypes and zero genetic effects was considered (referred to as "random" estimator here on). For the low MAF and Z-cut estimators, without loss of generality, Z-scores of the 12,966 SNPs on chromosome 22 were used for the estimation. Four MAF cutoffs for the low MAF estimator, 0.5 (all SNPs), 0.05, 0.005, and 0.0005, were considered. Three absolute value cutoffs for the Z-cut method, 0.5, 1, and 2, were considered.

For each method, the number of SNPs used for estimation in the simulation is given in **Supplementary Table 3**. Each scenario of the simulation was repeated for 30 times. The estimates were saved for evaluating the consistency of the phenotypic correlation estimators and their corresponding standard errors. All the above simulation analyses were performed using the R language (version 3.6.3).

In order to compare the low MAF estimator with the LDSC-intercept estimator, we conducted another simulation that used the real UK Biobank genotypes for 336,000 genomic British individuals across the 1,029,876 quality-controlled HapMap3 SNPs selected by the high-definition likelihood (HDL) software (Ning et al., 2020). Similar to the above, we draw the genetic effects across 10% of these SNPs from a normal distribution with zero mean. The heritability, phenotypic, genetic, and residual correlations all had a true value of 0.5. GWAS Z-scores of 70,042 SNPs with MAF $< 5 \times 10^{-4}$ were used for the low MAF estimator. Two reference panels were evaluated for LDSC, including the ldsc software inbuilt 1000 Genomes reference and the UK Biobank reference based on the HDL software reference data.

3. RESULTS

3.1. The Low MAF Estimator Corrects the Bias of the Z-Cut Estimator

In **Figure 1**, we provide some representative simulation results when the genetic correlation between the traits is 0.5. The general conclusion is that the low MAF estimator with a low enough MAF cutoff is able to overcome the bias of the Z-cut estimator. The complete simulation results comparing the low MAF and Z-cut estimators were summarized and given in **Supplementary Figures 1–8** and **Supplementary Tables 1**, **2**. Here, we summarize the key points as follows.

- When the samples do not overlap between the two traits, the phenotypic correlation is by definition zero. When no genetic correlation exists, all the methods that use the correlation between Z-scores give consistent zero estimates for the phenotypic correlation. However, bias in the estimation could happen when the genetic correlation is non-zero, which agrees with our theory in section 2. When there is a non-zero genetic correlation spread across the genome, only those methods that use the SNPs capturing little genetic variance would yield a consistent estimate for the phenotypic correlation, e.g., the

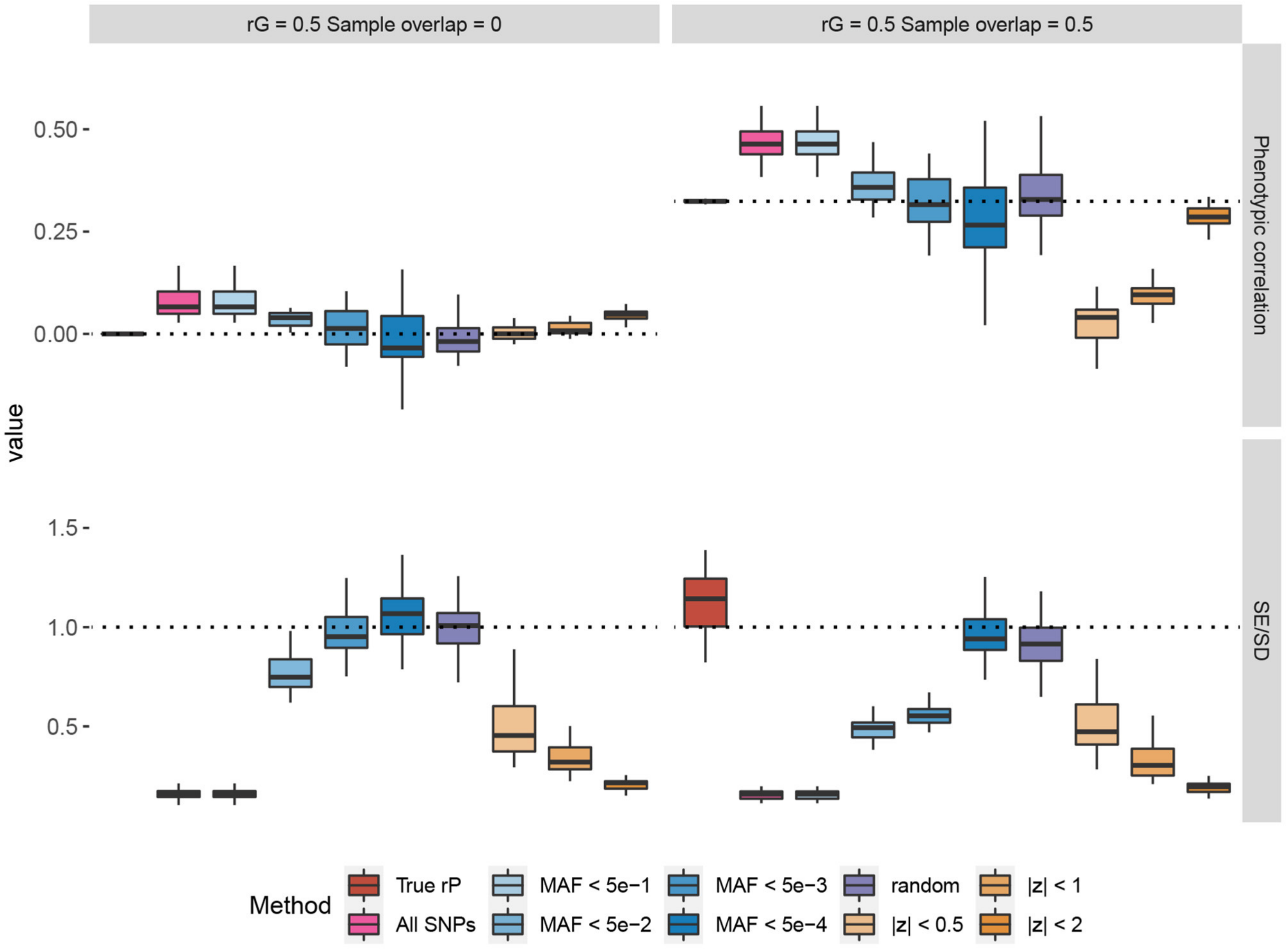

FIGURE 1 | Simulations comparing the Z-cut and low minor allele frequency (MAF) estimators for phenotypic correlation. The box plots show the simulation results from 30 replicates, where in each replicate, two phenotypes with heritability of 0.3 were simulated based on 10% randomly selected causal single nucleotide polymorphisms (SNPs) from the 784,256 genotyped SNPs in the UK Biobank. The true genetic correlation (rG) were set to 0.5 and the residual correlations were set to 0.25 for each pair of traits. Two scenarios of sample overlap proportions are shown. In the top panels, each phenotype was simulated for 168,000 genomic British individuals. The box plots compare the estimated phenotypic correlations using different estimators. The dash lines represent the true values. The bottom panels compare the estimated standard errors (SE) across the replicates to the standard deviation (SD) of the phenotypic correlation estimates across the 30 replicates, and each phenotype was simulated for 1,000 genomic British individuals. The dash lines at 1 represent that the estimated SE matches the empirical sampling SD. The "random" method represents the estimator based on 500 simulated SNPs with random genotypes and zero genetic effects. The true phenotypic correlation (rP) estimator was based on the individual-level phenotype data. The other estimators were based on the SNPs on chromosome 22 (numbers given in **Supplementary Table 3**).

random estimator where the SNPs capture absolutely zero genetic variance and the low MAF estimator with low enough MAF cutoffs.

- In overlapping samples, when the genetic correlation is zero, slight bias can be observed when using the Z-scores of common SNPs for phenotypic correlation estimation. Such bias can be corrected when a sufficiently low MAF cutoff is applied to the low MAF estimator. In partially overlapping scenarios, the observational phenotypic correlation in individual-level data can be estimated by adjusting the shrinkage factor $N_0/\sqrt{N_1 N_2}$.
- For the UK Biobank real genotype data, a 0.005 cutoff is low enough to yield a consistent estimate of the phenotypic correlation. Nevertheless, unless too few SNPs exist, using SNPs with MAF $< 5 \times 10^{-4}$ is recommended, as the standard error of the estimator can be consistently obtained in a simple way. The LD between low MAF SNPs is so small that we may consider the SNP genotypes as independent. Therefore, when simply obtaining the test statistic for the Pearson's correlation coefficient, the standard error for the Wald test can be back-calculated from the nominal p-value. The simulation results showed that the SEs of the low MAF method calculated in this way are consistent with the empirical standard deviation across the simulation repeats.
- Applying a low MAF cutoff on the pre-filtered SNPs based on the Z-cut method could reduce the bias in some

cases, but the bias cannot be completely overcome as the Z-cut method itself is a biased sampling strategy of the SNPs.

3.2. The Low MAF Estimator Corrects the Bias of the LDSC Intercept Estimator

For the second simulation, we observed downward bias in the LDSC intercept when the default 1000 Genomes reference was applied (**Figure 2**). Such a bias was overcome by the UK Biobank reference, nevertheless, the estimates were slightly inflated possibly due to the population substructure in the UKB genomic British individuals (Yengo et al., 2018). These biases were all absent when applying the low MAF estimator for the phenotypic correlation. Furthermore, the low MAF estimator had a substantially higher estimation efficiency than the LDSC intercept (**Supplementary Table 4**).

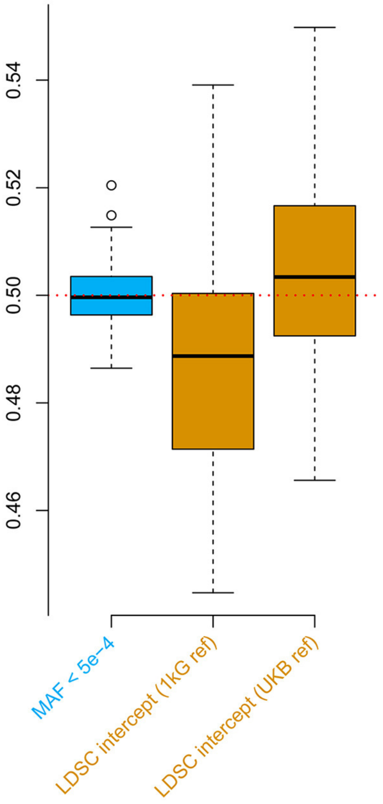

FIGURE 2 | Simulations comparing the low minor allele frequency (MAF) estimator and LD score regression (LDSC) intercept using the UK Biobank genotype data. The box plots show the results from 100 replicates, where in each replicate, two phenotypes with heritability of 0.5 were simulated for 336,000 genomic British individuals. The true phenotypic, genetic, and residual correlations were all set to 0.5. The low MAF estimates were based on 70,042 single nucleotide polymorphisms (SNPs) with MAF $< 5 \times 10^{-4}$. 1 kG ref: LD scores calculated based on the 1000 Genomes reference panel; UKB ref: LD scores calculated based on the UK Biobank reference panel.

3.3. Example Based on UK Biobank GWAS Summary Statistics

As a real data example, we applied the different estimation methods on the same 30 UK Biobank phenotypes used in Ning et al.'s study in genetic correlation estimation (Ning et al., 2020), where the GWAS summary statistics are publicly available (see Data Availability Statement section). The low MAF estimates were based on 70,042 SNPs with MAF $< 5 \times 10^{-4}$, and the LD scores were calculated based on the 1000 Genomes reference panel (default). At a 5% Bonferroni-corrected p-value threshold for 435 pairs of traits, the low MAF method discovered 223 significant phenotypic correlations, and LDSC intercept discovered 171. Among these, 61 phenotypic correlations were only significant in the low MAF method, vs. 9 only significant using the LDSC intercept (**Figure 3A**). The point estimates of the phenotypic correlations by the low MAF method and bivariate LDSC intercept were nearly the same (**Figure 3B**). As expected, when a Z-cut method is applied, the estimates became severely biased toward zero (**Figure 3C**). For seven of these phenotypes that we have individual-level data in our UK Biobank project (No. 14302), including body mass index, basal metabolic rate, usual walking pace, standing height, birth weight, coffee consumed, and year ended full time education, we extracted the initial measurement values. In order to be more consistent with the GWAS quality control procedure, we took away the effects of sex and age on these phenotypes by taking the residuals from linear regressions. The residuals were subsequently inverse-Gaussian transformed. After computing the individual-level observational phenotypic correlations and adjusted for the shrinkage factor $N_0/\sqrt{N_1N_2}$, the estimates were close to the low MAF estimates for these 21 pairs of traits (**Figure 3D**).

4. DISCUSSION

We have proposed the low MAF estimator of phenotypic correlations based on GWAS summary statistics, as an improvement of the Z-score correlation strategy based on all SNPs or SNPs that pass a particular Z-score cutoff. The estimator overcomes the bias generated when thresholding on summary association statistics and even that generated in the bivariate LDSC intercept. We suggest the use of the low MAF phenotypic correlation estimator in future practice. The more consistent and efficient estimation can improve our understanding of connections across human complex traits and diseases.

Although the low MAF method also introduces a filter on the tested SNPs, it is a threshold-free technique for the genetic effect parameter. Thus, the low MAF estimator does not constrain the estimated genetic effects of selected SNPs, equivalent to sampling a set of null effect SNPs from the genome. This explains why "putative null effect" SNPs with, e.g., $|z| < 2$ generate bias, whereas the low MAF estimator does not. Even if all the SNPs are null, some of them will generate z-score with $|z| > 2$ due to randomness. Removing them would lead to bias.

As the low MAF estimator is equivalent to sampling a set of null effect SNPs from the genome, the resulted phenotypic correlation estimates are close to those estimated

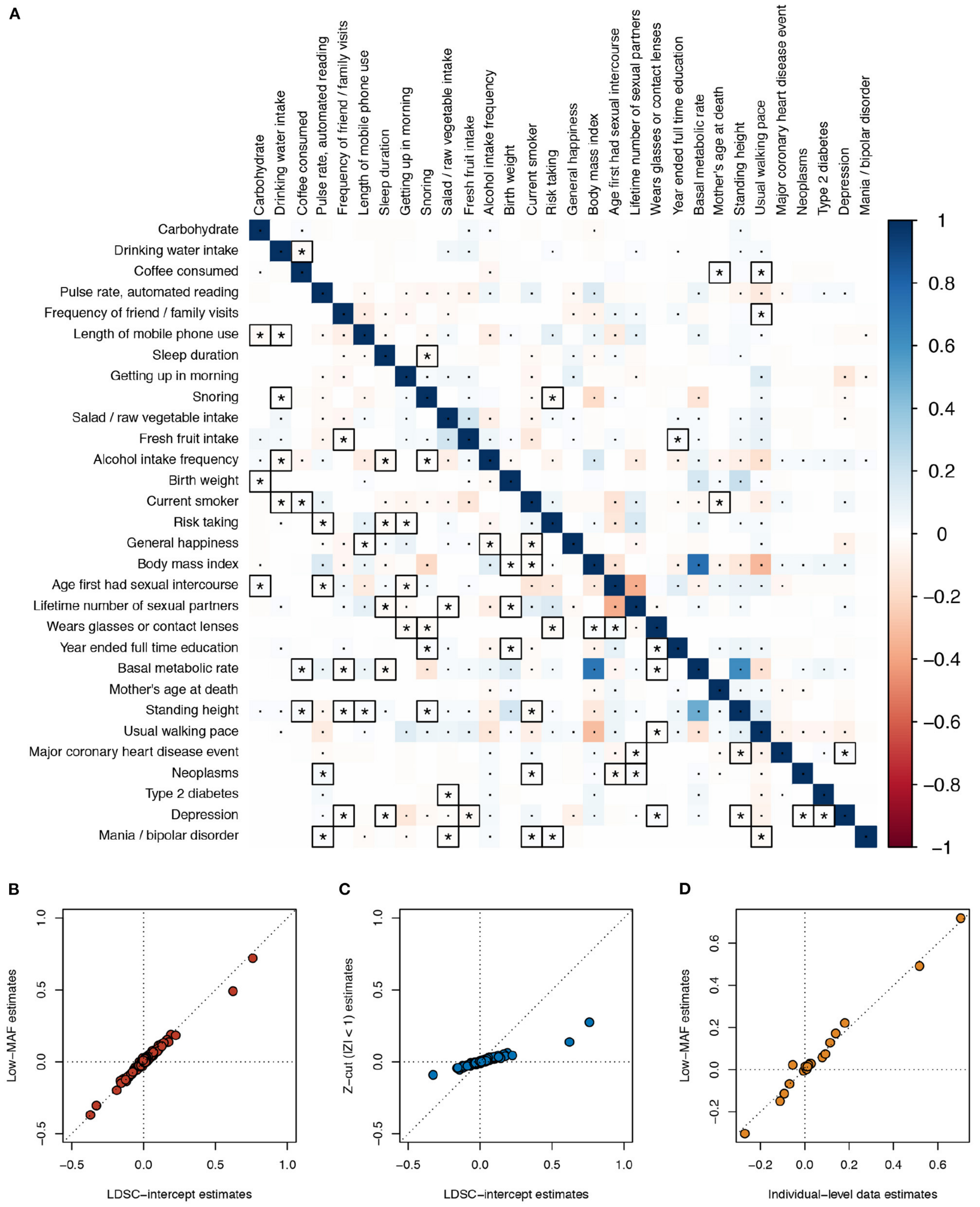

FIGURE 3 | Phenotypic correlation estimates across 30 UK Biobank traits. **(A)** Estimates using the low minor allele frequency (MAF) estimator (lower triangle) and LD score regression (LDSC) intercept (upper triangle) are compared. The low MAF estimates were based on 70,042 single nucleotide polymorphisms (SNPs) with MAF

(Continued)

FIGURE 3 | $< 5 \times 10^{-4}$. The default 1000 Genomes reference panel was used in LDSC. Bonferroni-corrected significant correlations with $P < 0.05/435$ are marked with asterisks or dots, where those correlations that are only significant using one of the two methods are marked with asterisks and squares. **(B)** Scatterplot comparing the LDSC intercept and low MAF estimates. **(C)** Scatterplot comparing the LDSC intercept and the Z-cut ($|Z| < 1$) estimates. **(D)** Scatterplot comparing the individual-level observational data and low MAF estimates; for the seven traits, we have data for, i.e., body mass index, basal metabolic rate, usual walking pace, standing height, birth weight, coffee consumed, and year ended full time education.

using individual-level phenotypic data. In the real UKB genotype data simulation, we showed that the LDSC intercept could not produce consistent estimates of the phenotypic correlation due to population substructure. Such a complication in LDSC was overcome by the low MAF estimator; although the GWAS summary statistics were used, the estimator approximates observed phenotypic correlation and is irrelevant to genetic data structure. For example, the genotypic data are treated as nuisance in the low MAF estimator.

As a comparative reference, we considered the "random" estimator using the Z-scores of 500 completely "irrelavant" SNPs. These SNPs were simply randomly generated, with random genotypes and zero genetic effects. These "bad" SNPs in GWAS appeared to be perfect for estimating the phenotypic correlation. The reason is simple according to our theory: they explain no phenotypic variance, so correlating their Z-scores for two traits becomes equivalent to correlating the phenotypic values themselves. This also explains why using the low MAF SNPs almost does the same: the low MAF SNPs explain little phenotypic variance. As LD between low MAF SNPs is rather low, using the low MAF SNPs is also helpful for getting the standard errors of the phenotypic correlation estimates. In the real data, low MAF SNPs are usually prone to genotyping errors or imputation failures if imputed. For the phenotypic correlation estimation purpose, even such errors are good, as they add more noise to the genotype data so that the SNP genotypes are even closer to noise.

Different sample overlap scenarios can be adjusted to obtain a consistent estimate of the observational phenotypic correlation. As long as N_0, N_1, and N_2 are known, the shrinkage factor $N_0/\sqrt{N_1 N_2}$ can be adjusted in the low MAF estimator. It should be noted that the adjustment becomes bad when N_0 is too small. In the extreme case, when $N_0 = 0$, i.e., in non-overlapping samples, there is no information we can learn about the phenotypic correlation from the two sets of GWAS summary statistics.

For binary phenotypes, an advantage of summary-statistics-based estimators, such as the low MAF estimator, is that it estimates the underlying phenotypic correlations on the liability scale. The liabilities follow an unobservable logistic distribution therefore the estimates are not exactly the same as the observed phenotypic correlations directly computed using the 0–1 outcome data. The phenotypic correlation estimates on the liability scale is mathematically easier to interpret and can be transformed into odds ratios from logistic regressions. Although for low MAF SNPs (rare variants) the GWAS test statistics would be generally inflated when the case–control data are unbalanced (Ma et al., 2013), the correlation between the Z-scores of two traits across the genome is still a valid estimator for the phenotypic correlation, which is not affected by low allele frequencies (**Supplementary Figure 9**).

DATA AVAILABILITY STATEMENT

The individual-level genotype and phenotype data are available by application from the UK Biobank (http://www.ukbiobank.ac.uk/). The UKBB GWAS summary statistics by the Neale's lab can be obtained from http://www.nealelab.is/uk-biobank/.

AUTHOR CONTRIBUTIONS

XS initiated and coordinated the study and drafted the manuscript. TL and ZN contributed to data analysis. All authors contributed to manuscript writing and gave final approval to publish.

ACKNOWLEDGMENTS

XS was in receipt of a Swedish Research Council (VR) Starting Grant (No. 2017-02543). We thank Edinburgh Compute and Data Facility at the University of Edinburgh for providing high-performance computing resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.665252/full#supplementary-material

REFERENCES

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., et al. (2015). An Atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47, 1236–1241. doi: 10.1038/ng.3406

Cichonska, A., Rousu, J., Marttinen, P., Kangas, A. J., Soininen, P., Lehtimäki, T., et al. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* 32, 1981–1989. doi: 10.1093/bioinformatics/btw052

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., and investigators, G. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* 37, 539–50. doi: 10.1002/gepi.21742

Ning, Z., Pawitan, Y., and Shen, X. (2020). High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.* 52, 859–864. doi: 10.1038/s41588-020-0653-y

Shen, X., Klarić, L., Sharapov, S., Mangino, M., Ning, Z., Wu, D., et al. (2017). Multivariate discovery and replication of five novel loci associated with immunoglobulin G N-glycosylation. *Nat. Commun.* 8:447. doi: 10.1038/s41467-017-00453-3

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* 8:e65245. doi: 10.1371/journal.pone.0065245

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* 50, 229–237. doi: 10.1038/s41588-017-0009-4

Yengo, L., Yang, J., and Visscher, P. M. (2018). Expectation of the intercept from bivariate LD score regression in the presence of population stratification. *bioRxiv [Preprint]*. doi: 10.1101/310565

Zheng, J., Richardson, T. G., Millard, L. A. C., Hemani, G., Elsworth, B. L., Raistrick, C. A., et al. (2018). PhenoSpD: an integrated toolkit for phenotypic correlation estimation and multiple testing correction using GWAS summary statistics. *GigaScience* 7, 1–10. doi: 10.1093/gigascience/giy090

Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* 96, 21–36. doi: 10.1016/j.ajhg.2014.11.011

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.