Estimation of genetic correlation using linkage disequilibrium score regression and genomic restricted maximum likelihood

Guiyan Ni[1], Gerhard Moser[1], Schizophrenia Working Group of the Psychiatric Genomics Consortium[2], Naomi R. Wray [3,4], and S. Hong Lee [1,4,5*]

[1]School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351, Australia

[2] A list of authors in the Supplementary file

[3]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, 4072, Australia.

[4]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, 4072, Australia.

[5]Centre for Population Health Research, School of Health Sciences and Sansom Institute of Health Research, University of South Australia, Adelaide, Australia

Guiyan Ni: Guiyan.ni@une.edu.au

Gerhard Moser: gmoser@aaco.com.au

Naomi R. Wray: naomi.wray@uq.edu.au

*S. Hong Lee: hong.lee@une.edu.au

ABSTRACT

Genetic correlation is a key population parameter that describes the shared genetic architecture of complex traits and diseases. It can be estimated by current state-of-art methods, i.e. linkage disequilibrium score regression (LDSC) and genomic restricted maximum likelihood (GREML). The massively reduced computing burden of LDSC compared to GREML makes it an attractive tool for analyses of many traits, achieved through use of GWAS summary statistics rather than individual level genotype data. While both methods generate unbiased estimates of the genetic correlation from genome-wide single nucleotide polymorphism (SNP) data, the accuracy of the LDSC estimates has not been thoroughly studied. In simulation, we show that the accuracy of GREML is generally higher than that of LDSC, which is more obvious when using a smaller number of SNPs (< 500K). When there is a genetic heterogeneity between the actual sample and reference data from which LD scores are estimated, the accuracy of LDSC decreases further. For genomic partitioning analyses, LDSC estimates can be biased. This is also confirmed with real data analyses, showing that GREML estimates based on ~150,000 individuals give a higher accuracy and power than LDSC estimates based on ~400,000 individuals (from combined meta-data) in estimating genetic correlation between SCZ and BMI (-0.136 SE=0.017, p-value=4.54E-15 for GREML vs. -0.087 SE=0.019 and p-value=4.91E-06 for LDSC). We show a GREML genomic partitioning analysis reveals that the genetic correlation between SCZ and height is significantly negative for regulatory regions, which cannot be found by non-partitioned GREML or by LDSC (whole genome or partitioned). We conclude that LDSC estimates should be carefully interpreted as there can be uncertainty about homogeneity among combined meta-data sets. We suggest that any interesting findings from massive LDSC analysis for a large number of complex traits should be followed up, where possible, with more detailed analyses with

individual genotype data and with GREML methods, even though sample sizes with individual genotype data may be smaller.

INTRODUCTION

Genetic correlation is a key population parameter that describes the shared genetic architecture of complex traits and diseases [1-3]. Methods to estimate genetic correlation based on genetic covariance structure are well established for both quantitative and disease traits, e.g. (restricted) maximum likelihood for linear mixed models (LMM) [4-7]. Genetic covariance structure can be derived from phenotypic records using pedigree information in twin or family-based designs [8]. Recently, genome-wide single nucleotide polymorphism (SNP) data have been used to construct a genomic relationship matrix for the genetic covariance structure in LMM that captures the contribution of causal variants that are in linkage disequilibrium (LD) with the genotyped SNPs [4,9].

In contrast to the genomic restricted maximum likelihood (GREML) approach, a linkage disequilibrium score regression (LDSC) [10,11] method does not require individual-level genotype data but instead uses GWAS summary statistics, regressing association test statistics of SNPs on their LD score. The LD score of a SNP is the sum of LD r^2 measured with its neighboring SNPs, and can be calculated in a reference sample of the same ethnicity when individual genotype data are not available for the GWAS sample. The method exploits the relationship between association test statistic and LD score expected under polygenicity. Because of this simplicity, and the massively reduced computing burden in terms of memory, time and literacy, it is feasible for LDSC to be applied to a large number of multiple traits, e.g. Bulik-Sullivan et al. [11],Zheng et al. [12], Finucane et al. [13] .

Given the attractiveness of LDSC for a massive analysis of many sets of GWAS summary statistics, it has been widely used in the community. However, genetic correlations estimated by

LDSC are often reported without caution although the approach is known to be less accurate, compared to GREML[11]. In fact, the accuracy of the approach has not been thoroughly studied.

In this study, we report that the accuracy of GREML (i.e. standard error of estimates, SE) is at least more than 1.5-fold higher than that of LDSC regardless of the number of samples in the analyses of estimating genetic correlation. When decreasing the number of SNPs, the accuracy of LDSC decreases further. When increasing the degree of genetic heterogeneity between the actual sample and reference data from which LD scores are estimated, the difference of the accuracy between GREML and LDSC becomes larger (up to 3-fold). We show that GREML is superior in more detailed analyses, such as genomic partitioning, over LDSC. In genomic partitioning analyses the genetic parameters are calculated for genomic subsets defined by user-specified annotations. Here we show that LDSC can give a biased estimate of enrichment for each partition. In analyses of real data, we show that GREML is more accurate and powerful, e.g. GREML estimates based on ~ 150,000 individuals give a higher accuracy and power than LDSC estimates based on 400,000 individuals in estimating genetic correlation between SCZ and BMI (-0.136 (SE=0.017) and p-value=4.54E-15 for GREML vs. -0.087 (SE=0.019) and p-value=4.91E-06 for LDSC). In the analyses, GREML estimate is based on UK sample only whereas LDSC estimate is based on combined meta-data sets among which there is uncertainty about homogeneity. Furthermore, a GREML genomic partitioning analysis reveals that the genetic correlation between SCZ and height is significantly negative for regulatory regions, which is not found by LDSC whether using whole-genome or partitioned estimates of genetic correlation.

RESULTS

Overview

We compared GREML and LDSC for estimates of genetic correlation using simulated as well as real data. Simulations used UK Biobank (UKBB) imputed genotype data after stringent quality control (QC) (see Methods). We calculated a ratio of empirical standard error (SE) and its 95% confidence interval (CI) to assess the accuracy of the methods for each set of simulated data. The 95% CI of SE were estimated based on the delta method [14]. For real data analyses, we obtained theoretical SE to assess the accuracy of the methods.

To explore the stability of the accuracy for both methods, we used two additional genotype data sets without imputation, Wellcome trust case control consortium 2 (WTCCC2) and genetic epidemiology research on adult health and aging cohort (GERA), which are publicly available. We also used UKBB raw (non-imputed) genotype data (UKBBr). We estimated the correlation between LD scores calculated from 1KG reference data and the actual target data (i.e. UKBB, WTCCC2, GERA and UKBBr), and compared it to the accuracy of LDSC or GREML estimates to determine if there is any association.

Genome partitioning analyses are an emerging tool to estimate the genetic variance and covariance explained by functional categories (e.g. DHS and non-DHS). Currently, genomic partitioning analyses focus on SNP-heritability enrichment analyses, formally testing for enrichment of signal compared to expectation proportional to the number of SNPs allocated to each annotation. Considering genomic partitioning in cross-disorder analyses is a natural extension to identify regions where genetic correlations between disorders are highest and

lowest. Here, we assessed the performance of GREML and LDSC in the context of genome partitioning analyses using simulated phenotypes based on UKBB genotype data.

Finally, we used real phenotype and individual genotype data from the Psychiatric Genomics Consortium (PGC) and UKBB to estimate genetic variance and covariance using LDSC and GREML, and compared their accuracy. We also used publicly available GWAS summary stats for LDSC to see how much the accuracy could be increased by increasing the number of samples and number of SNPs.

Simulations

When estimating genetic correlation using simulated phenotypes based on UKBB genotype data we found that the estimates were unbiased for both GREML and LDSC (Figure S1), but the accuracy of GREML was at least 1.5 times higher than that of LDSC (Figure 1). The ratio of the empirical SE from LDSC to GREML was increased up to 3.5-fold when using a smaller number of SNPs (Figure 1). All values of the ratio were significantly different from 1. It is notable that the SE of GREML estimates showed almost no difference across different numbers of SNPs whereas that of LDSC estimates gradually increased with a smaller number of SNPs (Figure S2). The ratio was invariant to sample size (Figure S3). As expected, when using the intercept constrained as zero, LDSC estimates were substantially biased when there were overlapping samples (Figure S4).

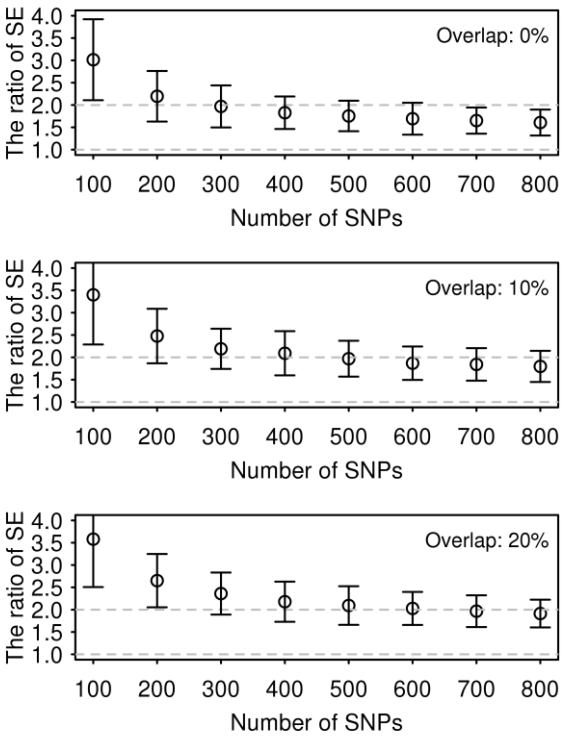

Figure 1. The ratio of SE of LDSC estimate to that of GREML estimate using simulated phenotypes based on UK Biobank genotypes. Bars are 95% CI based on 100 replicates. The unit for the number of SNPs is thousand. This result was based on 858K SNPs (after QC) and 10,000 individuals that were randomly selected from UK Biobank. SNPs in each bin were randomly drawn from the 858k SNPs independently. The number of causal SNPs was 10,000 that were randomly selected in each bin. The true simulated value for the genetic correlation was 0.6 and that for the heritability was 0.5 for both traits. Overlap (0%, 10% and 20%) stands for the percentage of overlapping individuals in the first and second traits.

Table 1. Correlation between LD scores estimated based on the Hapmap3 SNPs using the 1KG CEU reference sample and that from different target populations

	Correlation	Nr.SNPs
UKBB[a]	0.946	858,991
UKBBr[b]	0.720	123,615[c]
WTCCC2	0.899	421,035[c]
GERA	0.661	238,089[c]

[a]UKBB was imputed to the combined data of the 1KG reference and UK10K data.

[b]UKBBr was based on the raw genotype data of UK Biobank data

[c]The number of SNPs reduced further from the set of the QCed SNPs because of using only SNPs matched with the Hapmap3 SNPs used in calculating LD scores.

To explore the stability of the accuracy, we used other available data sets, i.e., WTCCC2, GERA and UKBBr. We firstly calculated the correlation between LD scores estimated based on the Hapmap3 SNPs using the 1KG CEU reference sample (downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/) and LD scores estimated based on in-sample genotype, i.e. UKBB, WTCCC2, GERA and UKBBr data set (Table 1), indicating that the WTCCC2, GERA or UKBBr (raw) genotypes were less similar to the 1KG reference genotypes, compared to the UKBB (imputed) genotypes (noting that UKBB samples had been imputed to the combined data of 1KG reference and UK10K data). Table 2 shows that the ratio of SE of LDSC estimate to that of GREML estimate was higher for WTCCC2, GERA or UKBBr than that for UKBB. Figure 2 shows that the accuracy of GREML was consistent across different

data sets, whereas that of LDSC was decreased for WTCCC2, GERA or UKBBr data set. This was probably due to the fact that a higher (or lower correlation) between LD scores based on the 1KG reference and the target data sets (Table 1) might positively or (negatively) affect on the accuracy of LDSC estimates. With WTCCC2, GERA and UKBBr data, the ratio of SE of LDSC to that of GREML based on different number of individuals is shown in Figures S5 and S6 and S7.

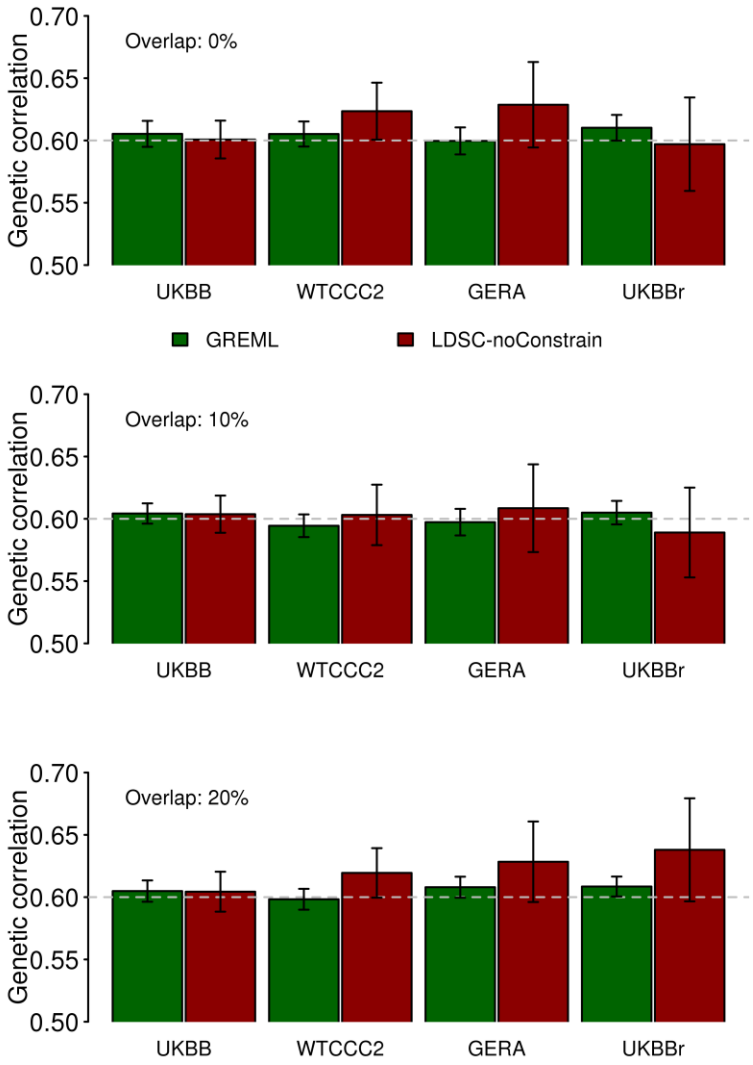

Figure 2. Estimated genetic correlation with GREML and LDSC (without constrain to the intercept) based on different genetic data sets. Simulation was based on 10,000 individuals that were randomly selected from UKBB, WTCCC2, GERA and UKBBr (the raw genotype of UKBB), with 858K, 432K, 239K, and 124K SNPs, respectively. Bars are 95% CI based on 100 replicates. Overlap (0%, 10% and 20%) stands for the percentage of overlapping individuals in the first and second traits. The grey dashed line stands for the true simulated genetic correlation 0.6.

Table 2. The ratio of SE of LDSC estimate to that of GREML estimate using simulated phenotypes based on UKBB, WTCCC2, GERA and UKBBr genotypes in the scenarios without overlapping individuals

	800k	400k	200k	100k
UKBB	1.60(0.15)	1.70 (0.18)	1.85 (0.25)	2.04 (0.33)
WTCCC2	NA	2.15 (0.31)	2.35 (0.43)	2.68 (0.61)
GERA	NA	NA	2.87 (0.56)	3.31 (1.17)
UKBBr	NA	NA	NA	3.74 (0.79)

For genome partitioning analyses, we showed that LDSC estimates of genetic correlation were biased while GREML estimates gave unbiased estimates for each functional category (Figure 3). LDSC estimates were -0.187 (SE 0.008) for DHS region and -0.232 (SE 0.007) for non-DHS region, which were far different from the true values (dotted lines in Figure 3). However, GREML estimates were close to the true values (0.619 (SE 0.016) for DHS region and -0.603 (SE 0.009) for non-DHS region). This was probably due to the fact that the different distribution of causal variants and their effects between DHS and non-DHS regions was better captured by an explicit covariance structure fitted in GREML. Whether samples are overlapping or not, the estimates and SE are not much different (Figure 3).

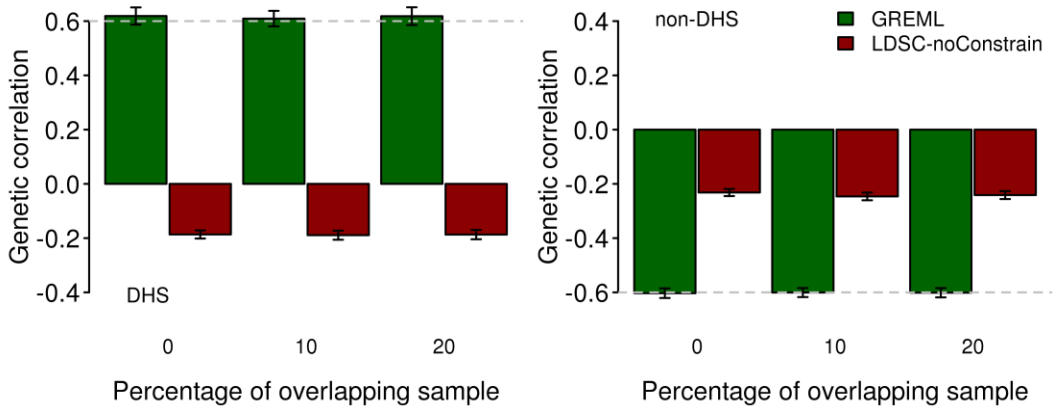

Figure 3. Estimated genetic correlation of simulated data based on a genomic partitioning model. Simulation was based on 10,000 individuals that were randomly selected from UKBB with 858K SNP. Based on Gusev et al.[39], the genome were stratified as two categories: DHS (194K SNPs with 2268 causal SNPs) and non-DHS (664K SNPs with 7732 causal SNPs). The genetic correlation for two simulated phenotypes was 0.6 in DHS region, and -0.6 in non-DHS region. Bars are 95% CI based on 100 replicates. Overlap (0%, 10% and 20%) stands for the percentage of overlapping individuals in the first and second traits.

While focusing on the accuracy of genetic correlation estimates, there is an important implication for the biasedness of SNP-heritability estimates for both GREML and LDSC (Figure S8). When using the WTCCC2, GERA and UKBBr data, which were less similar to the 1KG reference genotypes, compared to the UKBB data, LDSC estimates were substantially biased whereas GREML estimates were close to the true value in estimation of SNP heritability (Figure S8). When decreasing the number of SNPs, estimated SNP-heritabilities from LDSC were consistently unbiased, however, those from GREML were proportionally underestimated (Figure S9). When using non-hapmap3 SNPs, LDSC estimates were constantly biased (Figure S10). Although the genetic correlation is robust to such biasedness [4,11], SNP-heritability itself should be carefully interpreted for both GREML and LDSC. It is also noted that, LDSC (and stratified LDSC) estimates for SNP-heritability were biased in the genome partitioning analysis (Figure S11). Bias in LDSC for SNP-heritability is already recognized, which is why the method focuses on relative enrichment rather than absolute values.

Real data

We used real data to estimate SNP-heritability and genetic correlation between SCZ and BMI (Table 3 and Figure S12). GREML and LDSC estimates for the SNP-heritability were 0.192 (SE 0.004) and 0.280 (SE 0.016) for SCZ and 0.184 (SE 0.004) and 0.255 (SE 0.014) for BMI. The notable difference between GREML and LDSC was probably because of a relatively small number of SNPs (500K) that might result in underestimated GREML SNP-heritability (see Figure S9). This is one of the caveats of using GREML with real data that usually comprise

multiple cohorts genotyped on different platforms, such that even with imputation the overlapping SNP set of SNPs imputed with high confidence may be limited. The estimated genetic correlation for GREML and LDSC was -0.136 (SE 0.017) and -0.173 (SE 0.031). This indicated that the GREML estimate was 3.5 and 1.8 times more precise than LDSC estimates for the SNP-heritability and genetic correlation, respectively. For LDSC, we also considered using additional GWAS summary statistics from publicly available resources[15,16]. The sample sizes used for additional LDSC analyses (LDSC-meta) are summarized in Table 3. The estimated SNP-heritability was 0.259 (SE 0.019) for SCZ and 0.121 (SE 0.007) for BMI, and the estimated genetic correlation was -0.087 (SE 0.019). Although sample size was increased by 2.7-fold, the accuracy (SE) for LDSC was not better than those for GREML (SE = 0.017 vs. 0.019, and p-value = 4.54E-15 vs. 4.91E-06 for GREML vs. LDSC) (Table 3). It should be noted that GREML estimates used a homogeneous population (within UK and after stringent QC excluding population outliers) whereas LDSC-meta1 and -meta2 were based on combined meta-data sets consisting of ~ 80 different studies for which there is much more uncertainty about homogeneity than when using a single study cohort such as UKBB. The large difference of the estimates between LDSC and LDSC-meta1 (or -meta2) was probably due to the fact that heterogeneity among the 80 different studies resulted in underestimating the common genetic variance and covariance, and that the difference of LD scores between the target and 1KG reference data would bias the LDSC estimates as shown in Figure S8. We also analysed height data[17] and found a similar pattern in that GREML estimates were more accurate than LDSC estimates whether using the same data or using additional GWAS summary statistics for LDSC (Figure S13 and Table S1).

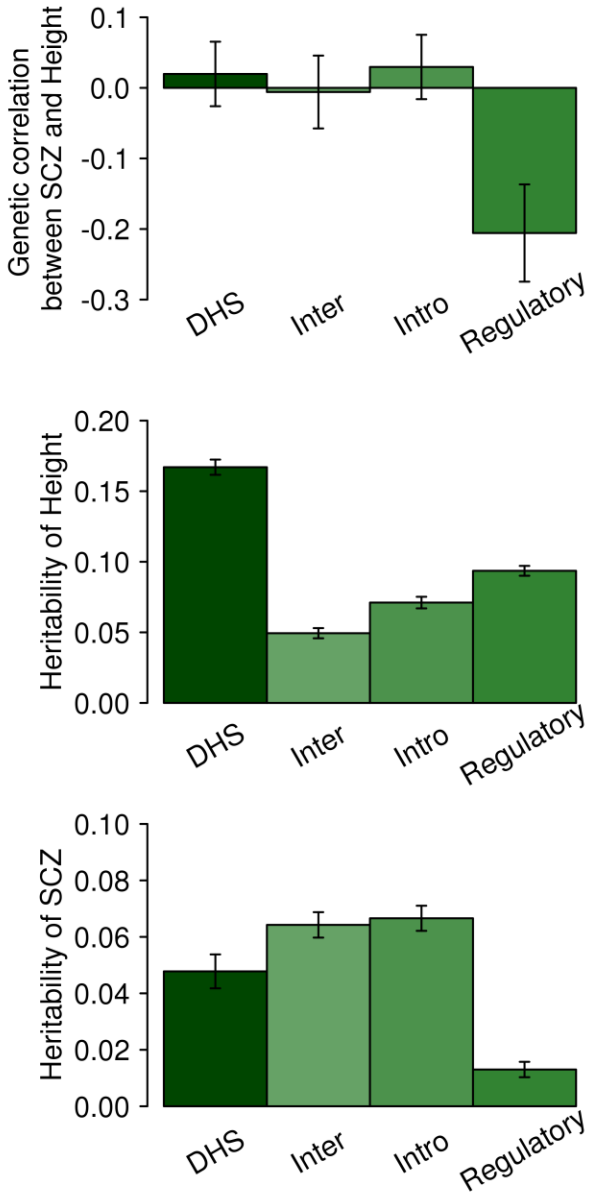

Figure 4. Genetic correlation between SCZ and height and heritability based on SNPs in partitioned genomic regions estimated with GREML. A joint model was applied by fitting four genomic relationship matrices simultaneously, each estimated based on the set of SNPs belong to each of the functional categories (regulatory, intron, intergene and DHS). The bars are standard errors. P-value for the estimate significantly different from 0 was 0.0028, 0.52, 0.91 and 0.67 for regulatory, intronic, intergenic and DHS region, respectively.

Table 3. Heritability and genetic correlation based on different data sets

Method	#SNPs	Data	#individuals		h^2 BMI		h^2 SCZ (liability scale)		Genetic correlation		
			Mean	SD	Estimate	SE	Estimate	SE	Estimate	SE	P
GREML	518,992	UKBB+ SCZ(qced)	152,961		0.184	3.80E-03	0.192	4.39E-03	-0.136	1.74E-02	4.54E-15
LDSC	516,519	UKBB+ SCZ(qced)	151,262	1432.7	0.255	1.38E-02	0.280	1.63E-02	-0.173	3.08E-02	1.91E-08
LDSC-meta1	477,163	UKBB+ GIANT+ PGCSCZ	422,499	20226.0	0.111	8.10E-03	0.259	1.28E-02	-0.091	2.44E-02	1.95E-04
LDSC-meta2	1,011,748	UKBB+ GIANT+ PGCSCZ	414,707	32697.8	0.121	6.50E-03	0.261	1.03E-02	-0.087	1.90E-02	4.91E-06

GREML: Analysis was based on quality controlled genetic data for BMI (from UK Biobank with 111,019 individuals and 518,992 SNPs) and schizophrenia (from PGC with 41,630 individuals and 518,992 SNPs).

LDSC: The data sets used in LDSC were the same as in GREML.

LDSC-meta1: GWAS summary statistics for BMI were based on meta-analysed GWAS results of UKBB individual-level genetic data (with 111,019 individuals and 518,992 SNPs) and of GIANT (245,051 individuals and 477,163 SNPs). For SCZ, the GWAS summary statistics from the full PGC sample based on 77,096 individuals were used.

LDSC-meta2: The data sets used in LDSC-meta2 were the same as in LDSC-meta1 except the increased number of SNPs (1,011,748) with which its performance was to check.

Due to different call rates of each SNP, number of individuals for each SNP used in GWAS were different, which shows in the column of the number of individuals with mean and SD.

We carried out a functional category analysis partitioning the genome into regulatory, DHS, intronic and intergenic region using GREML (Figure 4 for SCZ/height and Figure S14 for SCZ/BMI). For SCZ and height, the genetic correlation for the regulatory region was negative and significantly different from 0 (p-value = 0.0028) (Figure 4). We also compared the results with those from stratified LDSC particularly in SNP-heritability estimation (Table S2). The SE of the estimate for each category was much lower for GREML than stratified LDSC, ranging from 2.2 to 5.9-fold.

Box 1. Summary points

1. Although both LDSC and GREML provide estimates that are mostly unbiased on average for the genetic correlation, it is possible that two estimates differ in a single analysis of the same data (Figure S15 and S16). When LDSC and GREML estimates are dissimilar, it is recommended to report the estimate with a lower SE. The theoretical SE of the estimates is a reliable indicator for determining the better estimator, which agrees well with the empirical SE (from simulation replicates) (Figure S17).

2. When combining multiple data sets to estimate genetic correlations between multiple traits, it is possible, in practice, that the number of SNPs remaining after QC is relatively small. When the number of available SNPs is small (e.g. < 500K), the accuracy of LDSC to estimate genetic correlation can be notably decreased compared to that of GREML (Figure S2).

3. SNP-heritability has a different property, compared to genetic correlation since the latter is robust to biased estimation of genetic variance and covariance (presumably the biases occur in

the numerator and denominator and hence must cancel out)[4,11]. Especially when using a small number of SNPs (< 500K) for GREML or when using multiple meta-data sets for LDSC, estimated SNP-heritability itself should be reported with caution as both methods can give biased estimates.

4. When using a study cohort, it is desirable to measure heterogeneity between the cohort and 1KG reference data (e.g. measuring the correlation between LD scores estimated based on the cohort and 1KG reference data as in Table 1). If the correlation is not close to one, LDSC estimates should be carefully interpreted. We recommend that when GWAS summary statistics are provided, cohort specific LD scores are provided also. It is also warranted that an optimal approach to meta-analyse LD scores across multiple cohorts should be developed to improve LDSC performance [18].

5. When using extensive meta-data that possibly include heterogeneous sources, there are two problems. Firstly, the LD scores estimated from a reference samples such 1KG reference may be a poor representation of the LD scores of the heterogeneous meta-data, such that the accuracy of LDSC decreases. Second, the distribution of causal variants and pleiotropic effects may be different between heterogeneous sources such that the estimates can be biased (capturing only common effects between heterogeneous sources). This implies that LDSC estimates should be reported with caution when using extensive meta-data sets (Table 3).

6. One of advantages of having access to individual-level genotype data comes when more detailed analyses are required, such as genomic partitioning analyses. As shown in Figure 4, a GREML genomic partitioning analysis reveals a significant negative genetic correlation between SCZ and height for the regulatory region, which cannot be found by genome-wide GREML or LDSC approach.

DISCUSSION

LDSC and GREML are the methods that have been widely used in estimating genetic correlation, shedding light on the shared genetic architecture of complex traits, based on genome-wide SNPs. Two critical parameters for assessing methods are bias (whether the estimates over replicated analyses differ from the true value) and accuracy (reflected by the standard error of the estimate). Although the property of the accuracy of GREML has been thoroughly studied and tested [19,20], that of LDSC has not been sufficiently investigated. In this report, we compare the accuracy of GREML and LDSC estimates based on various scenarios using simulated as well as real data sets, and draw simple but useful guidelines (Box 1).

The reduced computing burden of LDSC over GREML makes it the method of choice for generating a quick overview of the genetic relationship between disorders, however, our results suggest that important associations could be overlooked. For example, Bulik-Sullivan et al.[11] reported a negative genetic correlation between BMI and SCZ estimated by LDSC (estimate = -0.095, SE = 0.025 with p-value = 1.75E-4) which was not significant after Bonferroni correction for the multiple testing. Because of the limited power from LDSC analysis, the shared genetic architecture between BMI and SCZ, perhaps, has had less attention than it is due. We confirmed the negative genetic correlation between BMI and SCZ with a greater confidence (p-value = 4.54E-15) using GREML. A second example is in analyses investigating the shared genetic architecture between height and SCZ, in which epidemiological evidence points to a negative association [21], supported by genetic analyses [22]. However, there was no evidence of genetic correlation between height and SCZ in whole-genome level analyses of Bulik-Sullivan et al. [11] (estimate = -0.002, SE = 0.022). We used a GREML genomic partitioning analysis and found a

significant negative genetic correlation between height and SCZ for the regulatory region (Figure 4). It was noted that the regulatory region was highly enriched for height (p-value = 1.02E-116 in Table S2), which intuitively supports a significant genetic correlation with SCZ for the region. As shown in Figure 3 and Figure S11, GREML estimate was closer to the true values with a lower SE than LDSC or stratified LDSC estimate in simulated data. For the real data analyses (Table S2), GREML had more accurate SNP-heritability estimates (lower SE) than stratified LDSC. Moreover, the sum of each category matched well with the estimate of the whole-genome for GREML whereas this was not the case for stratified LDSC (Tables S2).

Here we focused on genetic correlation estimates, and did not consider a number of alternative approached that have been explored in detail for estimation of SNP-heritability, e.g. LDAK approach[23], Weighted genomic relationship matrix[24], MAF stratified[19] and LD-MAF stratified approaches [25] . It was beyond the scope of our study to check if biasedness and accuracy can be improved with these methods, although a general observation is that biases in SNP-heritability estimation can 'cancel' in estimates of genetic correlations, as biases impact both the numerator and denominator of the genetic correlation quotient[4,11].

In conclusion, LDSC may be the best tool for a massive analysis of multiple sets of GWAS summary statistics in estimating genetic correlation between complex traits, because of its low computing burden and because summary statistics may be available for much larger sample sizes than those with individual genotype data. However, LDSC estimates should be carefully interpreted, considering the summary points (Box 1). Any interesting findings from LDSC

analyses should be followed up, where possible, with more detailed analyses using individual genotypes data and with GREML methods, even though sample sizes with individual genotype data may be smaller.

METHODS

Data

Schizophrenia (SCZ) data

The SCZ GWAS data were from the second phase of the Psychiatric Genomic Consortium [15] (PGC). Quality control (QC) and imputation of raw autosomal SNPs was performed by PGC with imputation program IMPUTE2/SHAPEIT [26,27] using CEU samples of the 1000 Genomes Project dataset [28] as the reference set. Post-imputation quality control was performed in each cohort as in Mehta et al. [1], which were merged across all cohorts afterward. Then, based on the merged genotype data, we utilized HapMap3 SNPs with call rate ≥ 0.9 and individuals with call rate ≥ 0.9. In addition, one individual in a pair were randomly excluded if their genomic relationship was larger than or equal to 0.05. After QC, 688,145 SNPs and 41,630 individuals (18,987 cases and 22,673 controls) remained.

UK Biobank (UKBB)

Based on the data released in the first version of UKBB [29], which were collected from a community sample, there initially were 152,249 individuals and 72,355,667 imputed SNPs available. Non-ambiguous strand SNPs identified on HapMap3 with imputation INFO ≥ 0.6, minor allele frequency (MAF) ≥ 0.01, call rate ≥ 0.95, and Hardy-Weinberg equilibrium P-value

≥ 10-7 were kept. Individuals with call rate ≥ 0.95 and clustered as Caucasian which were within 6 standard deviation from the mean of the EUR reference sample [30] based on the first and second principal components of genetic relationship matrix were kept. Similar to SCZ sample, one individual in a pair was randomly excluded if their genomic relationship was larger than 0.05 within UKBB sample. In addition, UKBB samples were excluded if their genomic relationship with SCZ sample > 0.05. After QC, 111,330 individuals, 858,991 SNPs were available. In addition to the imputed UKBB data, the raw genotype data of UKBB (UKBBr) were used for a comparison analysis. The raw genotypes were available for 805,426 SNPs from the UK Biobank and UK BiLEVE Axiom array. After the same QC process as above and matching with the hapmap3 SNPs, 111,330 individuals and 123,921 SNPs were used for simulation study.

Wellcome Trust Case Control Consortium 2 (WTCCC2)

The WTCCC2 data [31-34] used in current study were combined from four case datasets (Ischaemic Stroke, Multiple Sclerosis, Primary Biliary Cirrhosis, and Psoriasis) and two control datasets (1958 Birth cohort and UK Blood Service) which were genotyped with Illumina Human 1M-Duo BeadChip and QCed separately as following: SNPs were excluded due to either call rete ≤ 0.95 MAF < 0.01, HWE P-value <1E-4 or significantly (P < 1E-5) different call rate between cases and controls and individuals were excluded due to either individual call rate ≤ 0.97,being duplicated individuals, genomic relationship > 0.185 or from non-CEU ancestry. In the combined genotype data, we further excluded those SNPs with call rate ≤ 0.95, HWE P-value ≤ 10E-7, or MAF ≤ 0.01 and one random individual in a pair with high relatedness (> 0.05). After QC, 20,659 individuals each with 432,663 SNPs were available for simulation study.

Genetic epidemiology research on adult health and aging cohort (GERA)

The detailed description regarding GERA cohort data and its QC process was in Lee et al. [35]. In brief, those genetic data were from the European ancestry participants of the Kaiser Permanente GERA cohort [36]. In QC, SNPs were excluded due to either call rate ≤ 0.95, HWE p-value ≤ 10E-4, or MAF ≤ 0.01. Only Hapmap3 SNPs were kept for the simulation and analyses. Individuals were excluded due to either call rate ≤ 0.95 or being population outliers (i.e. out of 6 SD from the first and second principal components). Again, one individual in a pair with genomic relationship larger than 0.05 was randomly exclude. After QC, 46,345 individuals each with 239,976 SNPs were available for simulation study.

Simulation

The simulation process was based on the individual-level genotype data of UKBB (111,330 individuals and 858,991 imputed or 123,921 raw genotyped SNPs), WTCCC2 (20,659 individuals and 432,663 SNPs), and GERA (46,345 individuals and 239,976 SNPs), respectively.

Simulation of phenotype. For the bivariate model, 10,000 SNPs were randomly selected as causal variants, and each of them was assigned two sets of causal effects following a multivariate normal distribution with mean [0, 0] and (co)variance matrix as $\begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$. Thus, the genetic correlation between the first and second traits was 0.6. True breeding values or genetic profile scores were obtained from the product of SNP genotype and the corresponding SNP effects. The simulated phenotype were generated as the sum of the true genetic profile scores and the residual effects that were drawn from a multivariate normal distribution with mean [0, 0] and the

covariance matrix $\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$. Therefore, the simulated true heritability was 0.5 for both first and second traits.

For the sample available in each dataset, a random set of 10,000 individuals was made available for the first and second trait, respectively, such that the percentage of the overlapping individuals between the first and second traits was 0, 10 or 20%. Genetic correlation between two simulated traits was estimated using GREML and LDSC. The number of replicates was 100. The simulation process was conducted using MTG2 [37,38].

For sensitively analyses using UKBB data, a different number of individuals (2500, 5000 or 15000) or a different number of SNPs (100, 200 … 700k) was used. For testing a different number of SNPs, a subset of SNPs was randomly selected from 858k SNPs from which 10,000 SNPs were randomly selected as causal variants to simulate phenotypes as above. We also tested a situation that the number of causal SNPs reduced proportionally in a subset of randomly selected SNPs that used the same causal variants as in the original 858k SNPs.

Simulation for genomic partitioning analyses. This simulation was based on the QCed UKBB genotype data with 858,991 SNPs. According to Gusev et al [39], the genomic regions were divided into two categories: DHS (194,778 SNPs) and non-DHS (664,213 SNPs). For the categories, phenotypes were simulated such that the heritability was 0.25 and 0.5 for DHS region and non-DHS region, respectively, and the genetic correlation between the first and second traits was 0.6 and -0.6 for DHS region and non-DHS region, respectively. For the genetic correlation, we gave substantially different values between DHS and non-DHS region to make the contrast clear in the estimation, which also increased the power access the performance of the methods.

The residual correlation between two traits was 0.8 as above. For the sample, 10,000 individuals were randomly selected for each trait to perform GREML and LDSC with different levels (0, 10, or 20%) of overlapping individuals between two traits.

Real data

Traits of interest are height, body mass index (BMI), and SCZ. After QC as described above, the number of phenotypic records was 111,143 for height and 111,019 for BMI from the UKBB, and 41,630 for SCZ from PGC. The number of SNPs was 518,992. The reason for the reduced number of SNPs compared to the simulation study was that only the common SNPs between UKBB and SCZ can be used to build a genomic relationship in GREML. Phenotypic records of height, BMI were adjusted for age at interview, assessment center at which participant consented, genotype batch, year of birth and the first 15 principal components (PCs). SCZ were adjusted for sex, cohort and the first 15 PCs. The pre-adjusted phenotypes were used for GREML and LDSC estimations.

GREML and LDSC were compared based on the real phenotypic records and QCed genotype data. In addition, publicly available GWAS summary statistics for height, BMI and SCZ were used for LDSC (LDSC-meta). The GWAS results for height and BMI from GIANT [16,17] and those for SCZ from the full PGC samples [15] were meta-analysed, where SNPs were treated as fixed effects. In addition, an increased number of SNPs was used for LDSC to check if the performance was improved.

Genomic partitioning analysis was applied to the real data using GREML and stratified LDSC. Based on Gusev et al.[39], we stratified the genome into four categories, i.e. regulatory, DHS, intronic and intergenic region.

Analysis methods

We used LDSC and GREML to estimate SNP-heritability and genetic correlation for the simulated as well as real data sets. The detailed algorithms for the methods are well documented in somewhere else[10-12,37,40]. For a GREML genomic partitioning analysis, genomic relationship matrices (GRMs) [31] estimated based on the information of the functional categories were jointly fitted to estimate SNP-heritability and genetic correlation for each category. For a LDSC partitioning analysis, we used the stratified LDSC[13] that could estimate SNP-heritability for each category, but was not applicable to estimate genetic correlation. For stratified LDSC, we followed Finucane et al. [13] in the case of non-overlapping categories.

Acknowledgements

This research is supported by the Australian National Health and Medical Research Council (1080157, 1087889), and the Australian Research Council (DP160102126, FT160100229). This research has been conducted using the UK Biobank Resource. UK Biobank (http://www.ukbiobank.ac.uk) Research Ethics Committee (REC) approval number is 11/NW/0382. Our reference number approved by UK Biobank is 14575. GERA data came from a grant, the Resource for Genetic Epidemiology Research in Adult Health and Aging (RC2 AG033067; Schaefer and Risch, PIs) awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics. The RPGEH was supported by grants from the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente Northern California, and the Kaiser Permanente National and Northern California Community Benefit Programs. The RPGEH and the Resource for Genetic Epidemiology Research in Adult Health and Aging are described in the following publication, Schaefer C., et al., The Kaiser Permanente Research Program on Genes, Environment and Health: Development of a Research Resource in a Multi-Ethnic Health Plan with Electronic Medical Records, In preparation, 2013. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the WTCCC data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

Author contributions

S.H.L. conceived the idea. S.H.L. and N.R.W. directed the study. G.N., G.M. and S.H.L. performed the analyses. S.H.L., G.N. and N.R.W. drafted the manuscript. All authors contributed to editing and approval of the final manuscript.

REFERENCE

1. Mehta, D. *et al.* Evidence for genetic overlap between schizophrenia and age at first birth in women. *JAMA Psychiatry* **73**, 497-505 (2016).
2. Lee, S.H. *et al.* New data and an old puzzle: the negative association between schizophrenia and rheumatoid arthritis. *International Journal of Epidemiology* **44**, 1706-1721 (2015).
3. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet* **44**, 247-250 (2012).
4. 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).
5. Harville, D.A. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320-338 (1977).
6. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38 (1977).
7. Patterson, H.D. & Thompson, R. Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika* **58**, 545-554 (1971).
8. Neale, M. & Cardon, L. *Methodology for genetic studies of twins and families*, (Springer Science & Business Media, 2013).
9. VanRaden, P.M. Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91**, 4414-4423 (2008).
10. Bulik-sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. **47**, 291-295 (2015).
11. Bulik-Sullivan, B. *et al.* An Atlas of Genetic Correlations across Human Diseases and Traits. *Natrure genetics* **47**, 1236-1241 (2015).

12. 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).

13. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47**, 1228-1235 (2015).

14. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*, (Sinauer Sunderland, MA, 1998).

15. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).

16. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).

17. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186 (2014).

18. Pasaniuc, B. & Price, A.L. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* (2016).

19. Lee, S.H. *et al.* Estimation of SNP heritability from dense genotype data. *American Journal of Human Genetics* **93**, 1151-1155 (2013).

20. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**, 1011-1021 (2012).

21. Gunnell, D. *et al.* The association of fetal and childhood growth with risk of schizophrenia. Cohort study of 720,000 Swedish men and women. *Schizophrenia research* **79**, 315-322 (2005).

22. Bacanu, S.-a., Chen, X. & Kendler, K.S. The genetic overlap between schizophrenia and height. *Schizophrenia Research* **151**, 226-228 (2013).

544 23. Speed, D. *et al.* Re-evaluation of SNP heritability in complex human traits. *bioRxiv*, 074310
545 (2016).
546 24. de los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C. & Sorensen, D. Prediction of
547 complex human traits using the genomic best linear unbiased predictor. *PLoS genetics* **9**,
548 e1003608-e1003608 (2013).
549 25. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing
550 heritability for human height and body mass index. *Nature genetics* **47**, 1114 (2015).
551 26. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of
552 genomes. *Nature methods* **9**, 179-81 (2012).
553 27. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3*
554 *(Bethesda, Md.)* **1**, 457-70 (2011).
555 28. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between
556 five psychiatric disorders estimated from genome-wide SNPs. *Natrue Genetics* **45**, 984-994
557 (2013).
558 29. Collins, R. What makes UK Biobank special? *The Lancet* **379**, 1173-1174 (2012).
559 30. The Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**,
560 68-74 (2015).
561 31. International Stroke Genetics, C. & Wellcome Trust Case Control, C. Genome-wide association
562 study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nature genetics*
563 **44**, 328-333 (2012).
564 32. The International Multiple Sclerosis Genetics, C. *et al.* Genetic risk and a primary role for cell-
565 mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-219 (2011).
566 33. Mells, G.F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary
567 biliary cirrhosis. *Nature genetics* **43**, 329-332 (2011).

34. Tsoi, L.C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* **44**, 1341-1348 (2012).

35. Lee, S.H., Weerasinghe, W.M.S.P., Wray, N.R., Goddard, M.E. & van der Werf, J.H.J. Using information of relatives in genomic prediction to apply effective stratified medicine. *Scientific Reports* **7**, 42091 (2017).

36. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-1295 (2015).

37. Lee, S.H. & van der Werf, J. MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* **32**, 1420-1422 (2016).

38. Maier, R. *et al.* Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. *The American Journal of Human Genetics* **96**, 283-294 (2015).

39. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics* **95**, 535-552 (2014).

40. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76-82 (2011).