ARTICLES
https://doi.org/10.1038/s41588-019-0465-0
nature genetics

Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture

Kangcheng Hou[1,2,9], Kathryn S. Burch[3,9]*, Arunabha Majumdar[1], Huwenbo Shi[3,4], Nicholas Mancuso[1,5], Yue Wu[6], Sriram Sankararaman[3,6,7,8] and Bogdan Pasaniuc[1,3,7,8]*

SNP-heritability is a fundamental quantity in the study of complex traits. Recent studies have shown that existing methods to estimate genome-wide SNP-heritability can yield biases when their assumptions are violated. While various approaches have been proposed to account for frequency- and linkage disequilibrium (LD)-dependent genetic architectures, it remains unclear which estimates reported in the literature are reliable. Here we show that genome-wide SNP-heritability can be accurately estimated from biobank-scale data irrespective of genetic architecture, without specifying a heritability model or partitioning SNPs by allele frequency and/or LD. We show analytically and through extensive simulations starting from real genotypes (UK Biobank, $N = 337$ K) that, unlike existing methods, our closed-form estimator is robust across a wide range of architectures. We provide estimates of SNP-heritability for 22 complex traits in the UK Biobank and show that, consistent with our results in simulations, existing biobank-scale methods yield estimates up to 30% different from our theoretically-justified approach.

SNP-heritability, the proportion of phenotypic variance attributable to the additive effects of a given set of SNPs, is a fundamental quantity in genetics[1]; it provides an upper bound on risk prediction from a linear model[2] and, when defined as a function of all SNPs on an array, yields insights into the 'missing heritability' of complex traits[3–5]. Traditionally, SNP-heritability is estimated by fitting variance components models with restricted maximum likelihood (REML)[3,6–9]. With some exceptions[8], REML-based methods are not scalable to biobanks that assay hundreds of thousands of individuals (for example, UK Biobank[10]). SNP-heritability can also be estimated by assessing the deviation in marginal association statistics as a function of LD scores[11–14]; such methods can scale to millions of individuals. More recently, a randomized extension of Haseman–Elston regression[15] was shown to estimate a single genetic variance component from individual-level data as accurately as REML methods but in a fraction of the runtime[16].

To facilitate inference, all existing methods for genome-wide SNP-heritability inference make assumptions on genetic architecture, which is typically parametrized by polygenicity (the number of variants with effects larger than some small constant δ) and minor allele frequency (MAF)/LD-dependence (the coupling of effects with MAF, local LD or other functional annotations)[17]. Since the true genetic architecture of any given trait is unknown, existing methods are susceptible to bias and often yield vastly different estimates even when applied to the same data[9,14,18]. Although multi-component methods that stratify SNPs by MAF/LD ameliorate some of these robustness issues[7,18,19], fitting multiple variance components to biobank-scale data with REML is highly resource-intensive[8] and it is unclear whether multi-component methods based on summary statistics produce accurate estimates of total SNP-heritability. Alternate methods that explicitly model MAF/LD-dependency[6,9,14] are also sensitive to model misspecification[6,9,14,18,19]. In addition, genetic architecture varies across traits and populations due to, for example, variable degrees of negative selection acting on different traits in different populations[17,20–25]. Methods that jointly infer SNP-heritability and parameters such as the strength of negative selection or polygenicity[14,23,26] are computationally intensive and/or sensitive to LD-dependency. Thus, it remains unclear which estimates of SNP-heritability computed from biobank-scale data are reliable.

In this study, we investigate whether genome-wide SNP-heritability can be accurately estimated under a generalized random effects (GRE) model that makes minimal assumptions on genetic architecture. Under this model, every causal effect has an arbitrary SNP-specific variance and SNP-heritability is defined as the sum of the SNP-specific variances (Methods). To the best of our knowledge, all existing methods make additional assumptions on top of the GRE model (Table 1). For example, GREML[3] (and several other methods[8,16,27]) imposes an inverse relationship between MAF and allelic effect size whereas Linkage Disequilibrium Adjusted Kinships (LDAK) assumes that each SNP-specific variance is inversely proportional to both MAF and LD tagging[6,9]. We derive a closed-form estimator for SNP-heritability as a function of marginal association statistics and in-sample LD and show that this estimator is consistent (approaches the true SNP-heritability as sample size increases) and unbiased (its expectation is equal to the true SNP-heritability) when the number of individuals exceeds the number of

[1]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [2]College of Computer Science and Technology, Zhejiang University, Hangzhou, China. [3]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA. [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [5]Biostatistics Division, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [6]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA. [7]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [8]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [9]These authors contributed equally: Kangcheng Hou, Kathryn S. Burch. *e-mail: kathrynburch@ucla.edu; pasaniuc@ucla.edu

Table 1 | Existing methods to estimate SNP-heritability impose additional assumptions on top of the GRE model

Model	Assumptions on β_i	Description
Generalized random effects	$E[\beta_i]=0$, $Var[\beta_i]=\sigma_i^2$, $\sigma_i^2 \geq 0$	Each SNP i has a non-negative SNP-specific variance σ_i^2. Total SNP-heritability is $h_g^2 \equiv \sum_{i=1}^{M} \sigma_i^2$.
GREML-SC[3,8,16]	$\beta_i \sim N(0, h_g^2/M)$	Each SNP explains an equal portion of h_g^2. In other words, $\sigma_i^2 = h_g^2/M$ for all $i=1, \ldots, M$.
GREML-MC[7,8,18,42,43]	$\beta_i \sim N\left(0, \sum_{c \in C} [\text{SNP}_i \in c] h_c^2/m_c\right)$	h_g^2 is partitioned by a set of disjoint SNP partitions C that span all M SNPs. Partition $c \in C$ contains m_c SNPs that have per-SNP variances h_c^2/m_c. Total SNP-heritability is $h_g^2 = \sum_{c \in C} h_c^2$.
LDAK[6,9]	$\beta_i \sim N(0, \sigma_i^2)$, $\sigma_i^2 \propto w_i [f_i(1-f_i)]^{1+\alpha}$	Each SNP-specific variance is proportional to a function of f_i (the MAF of SNP i) and to w_i (a SNP-specific weight that is a function of the inverse of the LD score of SNP i). α controls the relationship between σ_i^2 and f_i. The most recent recommendation by ref. [9] is to assume $\alpha=-0.25$.
LDSC[11]	$E[\beta_i]=0$, $Var[\beta_i]=h_g^2/M$	Each SNP explains an equal portion of h_g^2 (similar to the GREML-SC model when h_g^2 is defined with respect to the same set of M SNPs).
S-LDSC[12,13,30]	$E[\beta_i]=0$, $Var[\beta_i]=\sum_{a \in A} \tau_a a(i)$	Each SNP-specific variance is a linear function of a set of annotations A where each $a \in A$ represents a binary or continuous-valued annotation. $a(i)$ is the value of annotation a at SNP i. τ_a is the expected contribution of a one-unit increase in annotation a to each SNP-specific variance.
SumHer[14]	$E[\beta_i]=0$, $Var[\beta_i] \propto w_i [f_i(1-f_i)]^{1+\alpha}$	An extension of the LDAK model to operate on summary-level data; can also efficiently partition h_g^2 by multiple annotations. The most recent recommendations by refs. [9,14] is to set $\alpha=-0.25$.

Under the GRE model, the causal effects at any two SNPs are assumed to be independent ($E[\beta_i \beta_j]=0$ for all $i \neq j$) and genome-wide SNP-heritability is defined as $h_g^2 \equiv \sum_{i=1}^{M} \sigma_i^2$, where each σ_i^2 can be an arbitrary non-negative real number as long as $0 \leq h_g^2 \leq 1$ (Methods). All existing methods make assumptions on the distribution of β_i and/or the form of σ_i^2 that can be subsumed under the GRE model. To simplify notation, we assume for each model that phenotypes are standardized in the population ($Var[y_n]=1$ for every individual n).

SNPs. Most importantly, the accuracy of this estimator is invariant to genetic architecture. While the GRE estimator is similar in form to previously proposed 'fixed effect estimators'[28,29], our approach differs from previous work in two main ways. First, SNP-heritability defined under a fixed effect model is different from the estimand of interest here (Methods). Second, previous methods have applied the estimator locally to identify regions contributing disproportionately to the genome-wide signal[28,29]; here we define a different genome-wide estimator (equation (1)) that requires large-scale genotype data. In addition, previous work has applied a singular value decomposition (SVD)-based regularization to account for errors in LD estimation from reference panels[29], which was unnecessary in the present work (Methods).

Through extensive simulations across a range of MAF/LD-dependent architectures starting from real genotypes from the UK Biobank[10] (337,205 individuals, 593,300 SNPs), we find that the GRE estimator is nearly unbiased across all architectures whereas existing methods are sensitive to model misspecification. For example, across 126 distinct architectures, the maximum bias of the GRE estimator is 2% of the simulated SNP-heritability whereas stratified LD score regression (S-LDSC)[12,13] and SumHer[14] yield biases between −64% and 28%. For completeness, we also contrast the GRE estimator with several REML-based methods in simulations at lower sample sizes (due to the computational burden of most REML methods) and find that, consistent with recent reports[18], all REML-based methods are biased when their model assumptions are violated, and multi-component REML methods that stratify SNPs by MAF and LD score (GREML-LDMS-I[18]) are more accurate than single-component REML methods. The performance of the GRE estimator is similar to that of GREML-LDMS-I, thereby confirming that SNP-heritability can be accurately estimated without stratifying SNPs or specifying a heritability model[6,9,14].

Finally, we use marginal association statistics and in-sample LD from 290,641 unrelated British individuals and 459,792 SNPs (MAF > 1%) to estimate SNP-heritability for 22 complex traits in the UK Biobank[10]. Consistent with simulations, estimates from S-LDSC and SumHer differ from the GRE estimates by a median of −9% and 11%, respectively, across the 18 traits with SNP-heritability estimates exceeding 0.05. For example, for height, estimates from S-LDSC (0.56) and SumHer (0.63) are approximately 7% lower and 5% higher, respectively, than our estimate of 0.60. Similarly, for hypertension, estimates from S-LDSC (0.14) and SumHer (0.18) are ±12.5% different from our estimate of 0.16. Taken together, our results demonstrate that SNP-heritability can be accurately estimated from biobank-scale data without the prior knowledge of the genetic architecture of the trait, motivating the development of scalable methods with fewer modeling assumptions.

Results

Overview of the approach. We investigate the utility of an estimator derived under a model that makes minimal assumptions on genetic architecture. We model the standardized phenotype of an individual as $y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon$, where $\mathbf{x}$ is an M-vector of standardized genotypes, $\boldsymbol{\beta}$ is the corresponding vector of standardized effects and $\epsilon \sim N(0, \sigma_\epsilon^2)$ is environmental noise (Methods). The effect size of each SNP is assumed to have mean zero and a finite SNP-specific variance (σ_i^2 for SNP i) that is allowed to be 0; the covariance between all pairs of effects is assumed to be zero. We term this model the 'GRE' model as, to the best of our knowledge, all existing methods impose additional assumptions on top of this model. For example, the single-component GREML model[3] assumes $\sigma_i^2 = h_g^2/M$ for $i = 1, \ldots, M$, whereas the most recent LDAK model[9] assumes $\sigma_i^2 \propto w_i [f_i(1-f_i)]^{0.75}$ (where w_i is a SNP-specific LD weight and f_i is MAF) (Table 1). Under the GRE model, the SNP-heritability explained by the M SNPs is the sum of the SNP-specific variances: $h_g^2 \equiv \mathrm{Var}[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}]/\mathrm{Var}[y] = \sum_{i=1}^{M} \sigma_i^2$ (Methods).

Given genotype measurements across N individuals at M SNPs and assuming $N > M$, the estimator $\hat{h}_g^2 = \frac{N\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\mathbf{V}}^{\dagger}\hat{\boldsymbol{\beta}} - q}{N-q}$, where $\hat{\boldsymbol{\beta}}$ is the vector of estimated marginal effects, $\hat{\mathbf{V}}^{\dagger}$ is the pseudoinverse of the in-sample LD matrix and q is the rank of the in-sample LD, is an unbiased estimator of SNP-heritability under the GRE model. That is, $E[\hat{h}_g^2] = \sum_{i=1}^{M} \sigma_i^2 = h_g^2$ (Methods). Unfortunately, even the largest biobanks currently have $N < M$ (for instance, UK Biobank has genotyped $M \approx 593{,}000$ SNPs in $N \approx 337{,}000$ unrelated white British individuals), which limits the utility of the above estimator.

We therefore extend our approach by partitioning the genome by chromosome:

$$\hat{h}^2_{\mathrm{GRE}} = \sum_{k=1}^{22} \frac{N\hat{\boldsymbol{\beta}}_k^{\mathrm{T}} \hat{\mathbf{V}}_k^{\dagger} \hat{\boldsymbol{\beta}}_k - q_k}{N - q_k} \quad (1)$$

where for chromosome k with p_k SNPs, $\hat{\boldsymbol{\beta}}_k$ is the p_k-vector of estimated effects, $\hat{\mathbf{V}}_k^{\dagger}$ is the pseudoinverse of the in-sample LD matrix and q_k is the rank of the in-sample LD. Although this estimator introduces bias, we show through extensive simulations that the magnitude of the bias is extremely small when N is sufficiently larger than p_k.

The GRE estimator is robust in simulations. To investigate the bias and variance of $\hat{h}^2_{\mathrm{GRE}}$, we perform simulations starting from real genotypes (N = 337,205, UK Biobank[10]). First, we simulate 64 MAF/LD-dependent quantitative trait architectures from chromosome 22 (M = 9,654 typed SNPs) by varying the SNP-heritability (h^2_g), proportion of causal variants (p_{causal}), distribution of causal variant MAF (CV MAF) and strength of coupling between effect size and MAF/LD; we use 'LDAK-LD-dependent' to describe causal effects that are coupled with 'LDAK weights' (Methods). To compare estimates across different values of h^2_g, we assess bias as a percentage of the simulated value of h^2_g (relative bias). Errors of individual estimates are also expressed as percentages of h^2_g. Consistent with analytical derivations, the GRE estimator restricted to chromosome 22 is unbiased across the 64 architectures (bias P value < 0.05/16 is considered significant to correct for 16 tests (architectures) at each value of h^2_g; Methods) (Fig. 1a,c and Supplementary Table 1). The average relative bias across the 64 architectures is 0.00015% × h^2_g and the largest bias under any single architecture is approximately ±0.2% × h^2_g (Supplementary Fig. 1a and Supplementary Table 1). In simulations of unascertained case-control studies (Methods), the GRE estimator is approximately unbiased across a range of disease prevalences (for $h^2_g = 0.10$, relative bias range is [−0.20%, 0.30%]) and has larger variance for lower prevalences (Supplementary Fig. 2a, Supplementary Table 2). For ascertained case-control studies, estimates are downward-biased but invariant to architecture (when $h^2_g = 0.10$, prevalence = 0.10 and $N_{\mathrm{case}} = N_{\mathrm{control}}$, relative bias is approximately −4%) (Supplementary Table 3). Masking 0%, 50% or 100% of causal SNPs from the observed summary statistics induces downward-bias when CV MAF = [0.01, 0.05] due to lower average LD between the observed SNPs and masked causal SNPs (Supplementary Fig. 3). The analytical estimator of the standard error of the mean (Methods) is well-calibrated (Supplementary Fig. 4a and Supplementary Table 4). As expected, partitioning chromosome 22 into disjoint, non-independent blocks induces upward bias that increases as block size decreases (Supplementary Fig. 5 and Supplementary Table 5).

Next, we performed genome-wide simulations (N = 337,205 individuals, M = 593,300 SNPs) to assess $\hat{h}^2_{\mathrm{GRE}}$ with the 22-block approximation (equation (1)). Despite the approximation, $\hat{h}^2_{\mathrm{GRE}}$ is highly accurate and robust across all 64 MAF- and LDAK-LD-dependent quantitative trait architectures (Fig. 1b,c). Across the 64 architectures, the bias ranges from 0.07% to 2.1% × h^2_g (average = 0.97% × h^2_g) (Supplementary Fig. 1b, Supplementary Table 6). Across all 6,400 simulations (64 genetic architectures × 100 simulation replicates), the largest error of any single estimate is approximately 17% × h^2_g (Fig. 1c). As N/M increases, the variance of $\hat{h}^2_{\mathrm{GRE}}$ decreases while the relative bias appears to be approximately fixed, ranging between 0.91% (N = 100,000) and 0.99% (N = 200,000) (Fig. 1d). These trends hold for a range of p_{causal} (Supplementary Fig. 6 and Supplementary Table 6), for unascertained case-control studies (Supplementary Fig. 2b and Supplementary Table 7), and in a smaller set of simulations with N = 7,685 individuals of South Asian ancestry and M = 1,642 SNPs (Supplementary Table 8; Methods). Most importantly, the accuracy of the GRE estimator is invariant to the underlying architecture (Fig. 1b). The analytical estimator for the standard error of the mean is downward-biased (and invariant to genetic architecture) with respect to the empirical standard deviation of $\hat{h}^2_{\mathrm{GRE}}$ estimates (Supplementary Fig. 4b and Supplementary Table 9). For example, across 16 architectures where $h^2_g = 0.25$, the empirical standard deviation of 100 independent estimates ranges from 0.0049 to 0.0064, whereas our estimated standard errors of the mean are approximately 0.0036 across all architectures (Supplementary Fig. 4b and Supplementary Table 9).

We investigate the effects of unmodeled substructure and/or cryptic relatedness by filtering individuals at different kinship coefficient thresholds (Methods) and find that using stricter relatedness thresholds increases the variance of the estimates (due to smaller sample size) while reducing bias, albeit not significantly (Supplementary Fig. 7, Supplementary Table 10). To assess the impact of population stratification, we simulated an effect of the first genetic principal component (PC) on phenotype and computed ordinary least squares (OLS) association statistics both with and without adjusting for the first PC (Methods). As expected, OLS without PC adjustment yields inflated estimates while OLS with PC adjustment yields approximately unbiased estimates (Supplementary Fig. 8 and Supplementary Table 11). However, even when a relatively large proportion of phenotypic variance is explained by the first PC (for example, $h^2_g = 0.25$, $\sigma^2_s = 0.05$), the maximum bias we observe using unadjusted association statistics is 5% × h^2_g (bias P value = 2.7×10^{-9}). Together, these results indicate that the GRE estimator is robust to modest amounts of unmodeled substructure and/or stratification. In all subsequent analyses, we compute $\hat{h}^2_{\mathrm{GRE}}$ with the 22-block approximation as this provides sufficiently accurate estimates and a fair comparison to other methods.

Comparison of methods to estimate SNP-heritability. We compare $\hat{h}^2_{\mathrm{GRE}}$ with existing state-of-the-art methods that are easily scalable to the full UK Biobank data (N = 337,205): LD score regression (LDSC), which assumes $\alpha = -1$ and no coupling of effects with LD[11]; stratified LD score regression (S-LDSC), which partitions h^2_g by a set of annotations of interest[12,13]; and SumHer, a scalable extension of LDAK that explicitly models MAF/LD-dependency through a specific form of the SNP-specific variances[14] (Table 1). To ensure a fair comparison, LD scores for all methods are computed using in-sample LD among the M SNPs, and in all simulations we aim to estimate the SNP-heritability explained by the same M SNPs (Methods).

As expected, $\hat{h}^2_{\mathrm{GRE}}$ is robust across all architectures while LDSC, S-LDSC and SumHer are sensitive to model misspecification. For example, when $h^2_g = 0.25$ (Fig. 2), LDSC is approximately unbiased under the 'single-component GREML model' (relative bias = 0.04%, P = 0.86) but is sensitive to CV MAF and the degree of coupling between effect size and MAF/LD (for example, when $p_{\mathrm{causal}} = 1\%$, relative bias ranges from −44% to 50%) (Supplementary Table 12). Similarly, SumHer is accurate under the 'LDAK model' (relative bias = 5.3%) but highly sensitive to other architectures (when $p_{\mathrm{causal}} = 1\%$, relative bias ranges from −19% to 22%) (Fig. 2 and Supplementary Table 13). S-LDSC (MAF), which partitions h^2_g by 10 MAF bins (Supplementary Table 14; Methods), is less biased than LDSC when effects are coupled with only MAF, but is significantly downward-biased when effects are also coupled with LDAK weights (for $h^2_g = 0.25$, relative bias range is [1.9%, 7.0%] when $\gamma = 0$ and [−58%, −37%] when $\gamma = 1$) (Fig. 2 and Supplementary Table 15). S-LDSC with 10 MAF bins and an additional 'level of LD' annotation, denoted S-LDSC (MAF + LLD) (Methods), produces similar results (for $h^2_g = 0.25$, relative bias range is [1.8%, 6.5%] when $\gamma = 0$ and [−80%, −33%] when $\gamma = 1$) (Supplementary Table 16). In contrast, the relative bias of $\hat{h}^2_{\mathrm{GRE}}$ ranges from 0.45% to 1.3%

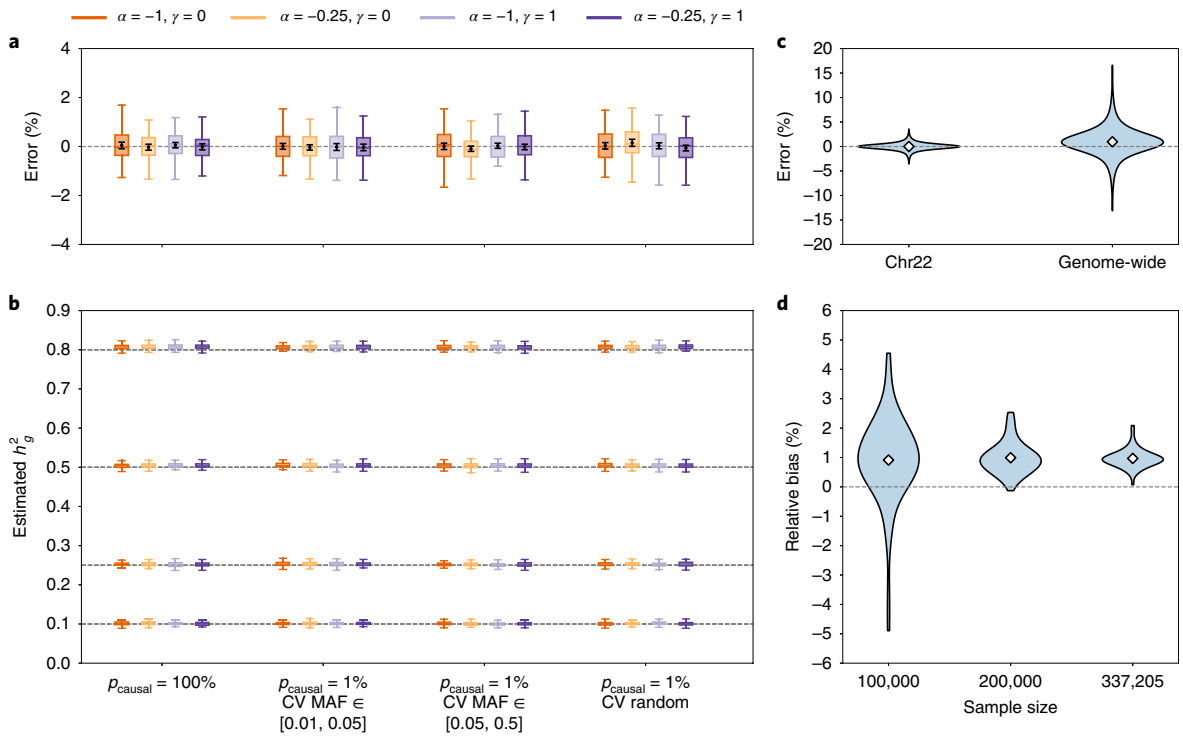

Fig. 1 | Simulations under 64 distinct MAF/LD-dependent architectures ($N = 337{,}205$). For each value of h_g^2, phenotypes were drawn according to 1 of 16 genetic architectures defined by p_{causal}, CV MAF, α and γ (Methods). **a**, Distribution of errors $\hat{h}^2_{GRE}(i) - h_g^2$ (as a percentage of h_g^2), where $\hat{h}^2_{GRE}(i)$ is the estimate from the ith simulation under a given genetic architecture, in simulations on chromosome 22 ($M = 9{,}654$ SNPs). $\hat{h}^2_{GRE}$ was computed with 1 chromosome-wide LD block. Black points and error bars represent the mean and ±2 s.e.m. that were used to test whether the bias under a single architecture is significant (Methods). **b**, Distribution of $\hat{h}^2_{GRE}$ in genome-wide simulations ($M = 593{,}300$ SNPs) where $\hat{h}^2_{GRE}$ was computed with 22 chromosome-wide LD blocks. In **a** and **b**, each boxplot represents estimates from 100 simulations. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × interquartile range (IQR) from the first and third quartiles, respectively. **c**, Distribution of errors for chromosome 22 and genome-wide simulations. Each violin plot represents the errors of 6,400 estimates (64 genetic architectures × 100 simulation replicates). **d**, Distribution of relative bias (as a percentage of h_g^2) as a function of sample size ($N = 100{,}000$, 200,000 or 337,205) in genome-wide simulations. Each violin plot represents 64 estimates of relative bias. In **c** and **d**, the white diamonds mark the mean of each distribution.

across the same 16 architectures where $h_g^2 = 0.25$ and $p_{causal} = 1\%$ (Fig. 2 and Supplementary Table 6). These trends hold for a range of h_g^2 and p_{causal}: across 112 LDAK-LD- and/or MAF-dependent architectures, the average and range of the relative bias of each method are 0.96% [−0.06%, 2.1%] (GRE), −2.2% [−71%, 70%] (LDSC), −22% −62%, 8.7%] (S-LDSC (MAF)), −29% [−89%, 9.0%] (S-LDSC (MAF + LLD)) and 2.8% [−27%, 28%] (SumHer) (Figs. 1b and 2, Supplementary Figs. 9–12 and Supplementary Tables 6, 12, 13, 15 and 16). Across 14 alternative LD-dependent architectures where SNP-specific variances are coupled with inverse LD scores instead of LDAK weights ('LD score-dependent' architectures; Methods and Supplementary Fig. 13), $\hat{h}^2_{GRE}$ remains nearly unbiased (relative bias range [0.52%, 1.3%]) whereas S-LDSC (MAF), S-LDSC (MAF + LLD) and SumHer are generally downward-biased (Supplementary Fig. 14 and Supplementary Table 17).

For completeness, we compare to four widely used REML-based methods: GREML, which assumes $\alpha = -1$ and no coupling of effects with LD[3]; GREML-LDMS-I, a multi-component extension of GREML that partitions SNPs by MAF and LD score[18]; BOLT-REML, a computationally efficient variance components estimation method with assumptions similar to those of GREML[8]; and LDAK, which assumes a specific form of the SNP-specific LD weights and recommends setting $\alpha = -0.25$ (refs. [6,9]) (Table 1). Because it is computationally intractable to apply the REML-based methods to thousands of genome-wide simulations with 337,205 individuals, we perform simulations using a reduced number of individuals ($N = 8{,}430$) and SNPs ($M = 14{,}821$) (Methods). As expected, the single-component methods (GREML, BOLT-REML and LDAK) are sensitive to MAF/LD-dependency whereas the GRE estimator is robust across all architectures. For example, when $h_g^2 = 0.25$ (Fig. 3), GREML and BOLT-REML are accurate under the GREML model (GREML: relative bias = −14%, $P = 6.0 \times 10^{-3}$, Supplementary Table 18; BOLT-REML: relative bias = −0.16%, $P = 0.75$, Supplementary Table 19) and LDAK is approximately unbiased under the LDAK model (relative bias = 0.16%, $P = 0.77$, Supplementary Table 20), but all three are sensitive to CV MAF, α and γ. Across 12 architectures where $p_{causal} = 1\%$ (Fig. 3), the relative biases are within [−15%, 7.9%] (GREML), [−14%, 9.1%] (BOLT-REML) and [−34%, 8.2%] (LDAK) (Supplementary Tables 18–20). In contrast, for the same 12 architectures, $\hat{h}^2_{GRE}$ yields relative biases in the range [-2.1%, 1.7%], which is comparable to the relative bias of GREML-LDMS-I (range [−2.9%, 1.5%]) using eight GRMs (4 LD quartiles × 2 MAF bins) that align with CV MAF (Fig. 3 and Supplementary Tables 21 and 22). These trends hold over a range of h_g^2 and p_{causal}: across 112 LDAK-LD- and/or MAF-dependent architectures (Supplementary Figs. 15–19), the average and range of the relative bias are 0.09% [−4.9%, 6.4%] (GRE), −0.6% [−5.9%, 2.3%] (GREML-LDMS-I), −2.9% [−27%, 15%] (GREML), −1.8% [−25%, 18%] (BOLT-REML) and −8.2% [−44%, 13%] (LDAK) (Supplementary Tables 18–22). Similar trends are observed for LD score-dependent architectures (Supplementary Fig. 20 and Supplementary Table 23). In an extreme example where CV MAF is tightly concentrated near 1%, GREML-LDMS-I with the same eight GRMs as before is downward-biased whereas the GRE estimator remains robust (Supplementary

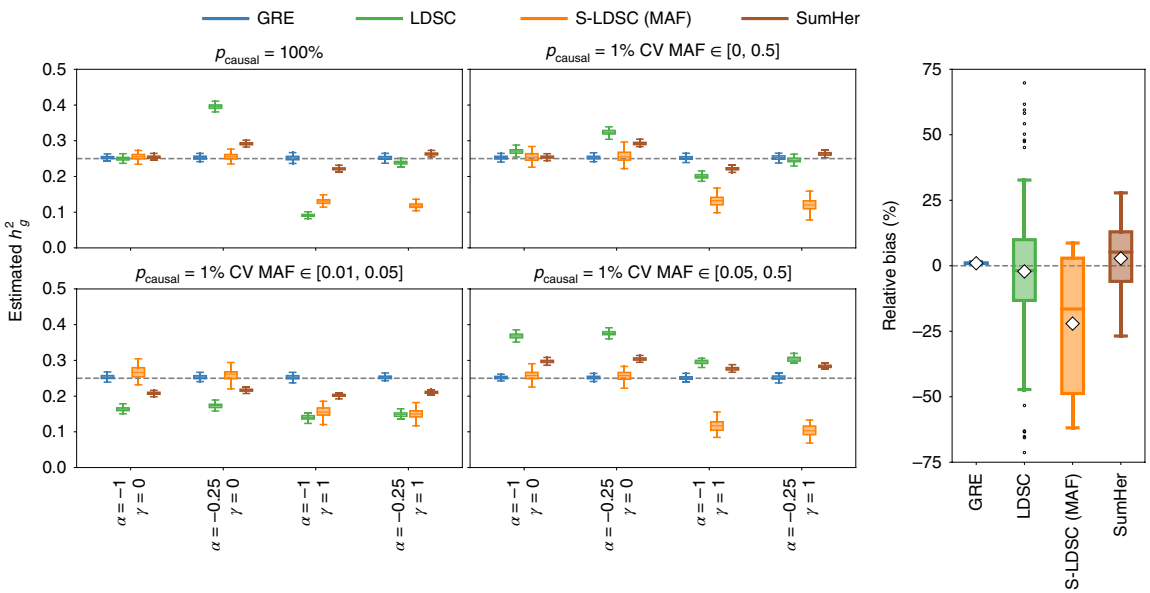

Fig. 2 | Comparison of $\hat{h}^2_{GRE}$ with LDSC, S-LDSC (MAF) and SumHer in genome-wide simulations ($N = 337{,}205$, $M = 593{,}300$). Left: Phenotypes were drawn under 1 of 16 MAF- and/or LDAK-LD-dependent architectures by varying p_{causal}, α, γ and CV MAF (Methods). Each boxplot contains estimates of h_g^2 from 100 simulations. Right: Relative bias of each method (as a percentage of h_g^2) across 112 distinct MAF- and LDAK-LD-dependent architectures (Methods). Each boxplot contains 112 points; each point is the relative bias estimated from 100 simulations under a single genetic architecture. The white diamonds mark the average of each distribution. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

Fig. 21 and Supplementary Tables 18–22). While the variance of our estimator is larger than the variances of the REML-based methods (Fig. 3), our approach is designed for sample sizes several orders of magnitude larger than what we used in these simulations. In summary, our results confirm that it is possible to accurately estimate h_g^2 under the GRE model.

SNP-heritability of 22 complex traits in the UK Biobank. Finally, we compute $\hat{h}^2_{GRE}$ for 22 complex traits in the UK Biobank (290,641 unrelated British individuals, 459,792 SNPs; Methods)[10]. For comparison, we also provide estimates from LDSC, S-LDSC (controlling for the baseline-LD model[13,30]) and SumHer. Of the 22 traits analyzed (6 quantitative, 16 binary), we focus on 18 traits for which $\hat{h}^2_{GRE} > 0.05$ (Table 2). For the six quantitative traits, $\hat{h}^2_{GRE}$ ranges from 0.12 (smoking status) to 0.60 (height). Across the 12 binary traits, $\hat{h}^2_{GRE}$ ranges from 0.064 (autoimmune disorders) to 0.16 (hypertension) (Table 2). These estimates are robust to the filtering of individuals based on relatedness (Supplementary Table 24). We also computed $\hat{h}^2_{GRE}$ from two additional sets of SNPs (MAF > 0.1% and MAF > 0.01%) and found that the estimates increase slightly for lower MAF thresholds (Supplementary Table 25), which is expected due to the increased number of SNPs. To enable a direct comparison between $\hat{h}^2_{GRE}$ and the quantities estimated by LDSC, S-LDSC and SumHer, we ran the summary-statistics-based methods with LD scores and regression weights computed from in-sample LD and estimate h_g^2 defined as a function of the same set of SNPs (Methods). Across the 18 traits, S-LDSC (baseline-LD/in-sample) and SumHer (in-sample) differ from $\hat{h}^2_{GRE}$ by a median of −9% and 11%, respectively (expressed as a percentage of $\hat{h}^2_{GRE}$) (Fig. 4 and Table 2). As expected[11], LDSC (in-sample) yields inflated estimates.

To compare $\hat{h}^2_{GRE}$ to estimates reported in the literature, we also ran the summary-statistics-based methods with their recommended parameter settings[11,12,14,30] and with LD scores and regression weights computed from the 1000 Genomes Phase 3 reference panel (489 Europeans)[31]—we note that when running these methods as recommended, their estimands are not equivalent to our definition of h_g^2 (see Methods and refs. [11,12,14,19] for details). Across the 18 traits for which $\hat{h}^2_{GRE} > 0.05$, the median differences with respect to $\hat{h}^2_{GRE}$ are −11% for LDSC (1KG), −14% for S-LDSC (baseline-LD/1KG) and 38% for SumHer (1KG) (Supplementary Fig. 22 and Supplementary Table 26). For nine of these traits, a previous study reported single-component BOLT-REML estimates (computed from a similar UK Biobank cohort[27]) that differ from our estimates by a median of 8% (Supplementary Table 26).

Runtime and memory requirement. We report the runtime and memory requirements for computing $\hat{h}^2_{GRE}$ with the 22-block approximation from 337,000 individuals and 593,000 SNPs. First, computing chromosome-wide LD has complexity $O(Np_k^2)$ for chromosome k with p_k SNPs. In practice, this step does not impose a computational bottleneck because the computations can be parallelized over SNPs. Second, the pseudoinverse of each LD matrix is computed via truncated SVD, which has complexity $O(p_k^3)$ for chromosome k. For 50,000 typed SNPs this takes about 3 h and 60 GB of memory. Lastly, given the pseudoinverse LD matrices and OLS association statistics, computing $\hat{h}^2_{GRE}$ has complexity $O(p_1^2 + \cdots + p_{22}^2)$. For any of the traits analyzed in this study, this takes less than 1 h and requires 24 GB of memory; most of this time is spent loading the data into memory. For comparison, running LDSC, S-LDSC or SumHer consists of precomputing LD scores and SNP-specific weights and performing linear regression to estimate the variance parameters. Precomputing LD scores and SNP-specific weights can be parallelized over blocks of SNPs. The second step (least squares regression) is $O(C^2M)$ for M SNPs in the regression and C variance parameters.

Discussion

In this study, we show that SNP-heritability can be accurately estimated under minimal assumptions on genetic architecture. Our proposed estimator allows the SNP-specific variances to capture arbitrary relationships between effect size and MAF/LD, and we

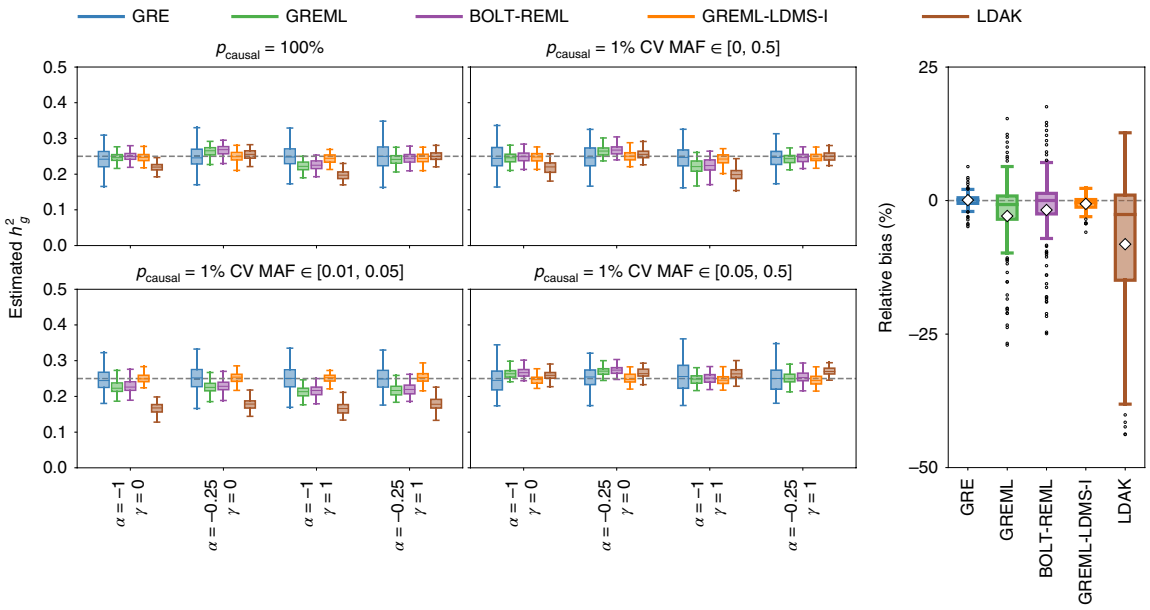

Fig. 3 | Comparison of $\hat{h}^2_{GRE}$ with GREML, BOLT-REML, GREML-LDMS-I and LDAK in small-scale simulations ($N = 8{,}430$, $M = 14{,}821$ SNPs). Left: phenotypes were drawn under 1 of 16 MAF- and/or LDAK-LD-dependent architectures by varying p_{causal}, α, γ and CV MAF (Methods). Each boxplot contains estimates of h_g^2 from 100 simulations. Right: relative bias of each method (as a percentage of the true h_g^2) across 112 distinct MAF- and LDAK-LD-dependent architectures (Methods). Each boxplot represents the distribution of 112 points; each point is the relative bias estimated from 100 simulations under a single genetic architecture. The white diamonds mark the average of each distribution. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

Table 2 | Estimates of h_g^2 from the GRE approach, LDSC (in-sample), S-LDSC (baseline-LD/in-sample) and SumHer (in-sample) for 22 complex traits and diseases in the UK Biobank ($N = 290{,}641$ unrelated British individuals, $M = 459{,}792$ typed SNPs)

Trait	GRE	s.e.m.	LDSC	s.e.m.	S-LDSC	s.e.m.	SumHer	s.e.m.
Smoking status	0.122	3.90×10^{-3}	0.178	7.70×10^{-3}	0.110	8.50×10^{-3}	0.132	4.30×10^{-3}
Height	0.602	4.70×10^{-3}	0.730	2.70×10^{-2}	0.555	3.10×10^{-2}	0.634	2.70×10^{-2}
BMI	0.285	4.20×10^{-3}	0.436	1.20×10^{-2}	0.289	1.70×10^{-2}	0.315	9.00×10^{-3}
WHR	0.173	4.00×10^{-3}	0.256	1.20×10^{-2}	0.184	1.60×10^{-2}	0.198	9.40×10^{-3}
Systolic blood pressure	0.159	4.20×10^{-3}	0.243	9.00×10^{-3}	0.134	9.70×10^{-3}	0.177	5.70×10^{-3}
Diastolic blood pressure	0.154	4.20×10^{-3}	0.233	8.60×10^{-3}	0.130	9.70×10^{-3}	0.170	6.40×10^{-3}
Eczema	0.116	4.20×10^{-3}	0.165	1.10×10^{-2}	0.107	1.20×10^{-2}	0.130	8.80×10^{-3}
Asthma	0.116	4.90×10^{-3}	0.163	1.20×10^{-2}	0.116	1.70×10^{-2}	0.131	1.20×10^{-2}
Hypertension	0.162	4.00×10^{-3}	0.244	9.40×10^{-3}	0.142	1.10×10^{-2}	0.180	6.10×10^{-3}
High cholesterol	0.082	5.10×10^{-3}	0.127	1.30×10^{-2}	0.138	5.80×10^{-2}	0.088	8.30×10^{-3}
Diabetes (any)	0.070	3.70×10^{-3}	0.093	5.90×10^{-3}	0.062	8.70×10^{-3}	0.074	5.00×10^{-3}
Type 2 diabetes	0.071	3.80×10^{-3}	0.090	6.10×10^{-3}	0.057	8.80×10^{-3}	0.071	4.00×10^{-3}
Hypothyroidism	0.088	5.20×10^{-3}	0.142	1.30×10^{-2}	0.078	1.20×10^{-2}	0.110	1.70×10^{-2}
Thyroid disorders	0.084	5.20×10^{-3}	0.141	1.30×10^{-2}	0.080	1.20×10^{-2}	0.110	2.00×10^{-2}
Endocrinopathies	0.069	5.10×10^{-3}	0.084	7.00×10^{-3}	0.058	9.90×10^{-3}	0.068	5.00×10^{-3}
Cardiovascular Diseases	0.143	5.30×10^{-3}	0.228	1.10×10^{-2}	0.140	1.40×10^{-2}	0.164	6.00×10^{-3}
Respiratory and ENT diseases	0.086	5.20×10^{-3}	0.120	1.20×10^{-2}	0.079	1.40×10^{-2}	0.090	9.50×10^{-3}
Psoriasis	0.019	5.00×10^{-3}	0.071	3.10×10^{-2}	0.035	1.20×10^{-2}	0.059	4.20×10^{-2}
Dermatologic disorders	0.023	5.00×10^{-3}	0.049	1.40×10^{-2}	0.034	9.90×10^{-3}	0.031	1.10×10^{-2}
Rheumatoid arthritis	0.008	5.00×10^{-3}	0.041	2.10×10^{-2}	0.010	7.90×10^{-3}	0.021	1.20×10^{-2}
Autoimmune disorders (broad)	0.063	5.10×10^{-3}	0.105	1.20×10^{-2}	0.050	9.50×10^{-3}	0.079	1.70×10^{-2}
Autoimmune disorders (certain)	0.015	5.00×10^{-3}	0.052	2.60×10^{-2}	0.005	7.60×10^{-3}	0.047	3.40×10^{-2}

demonstrate through simulations that its accuracy is invariant to genetic architecture. We show that all existing methods impose additional assumptions on the GRE model, and we confirm through simulations that these methods can be sensitive to model misspecification. One practical advantage of our approach over summary-statistics-based methods is that the estimand is always the same for

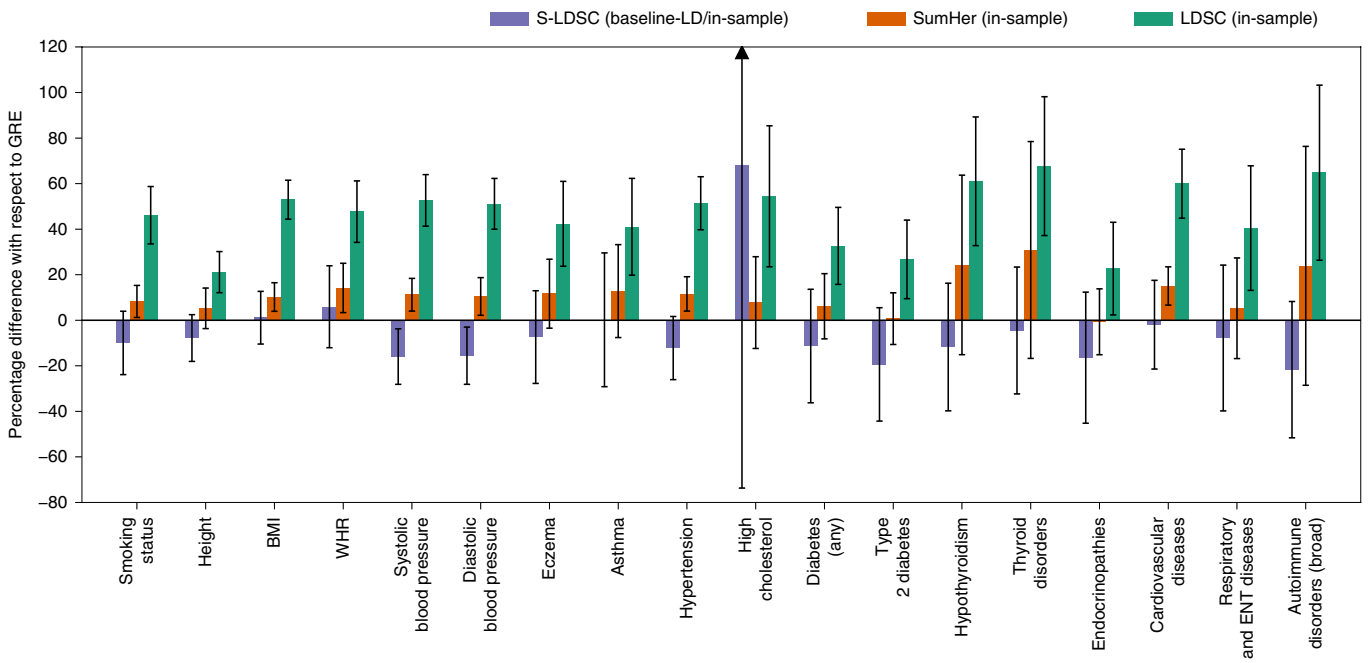

Fig. 4 | Percentage difference of h_g^2 estimates from LDSC (in-sample), S-LDSC (baseline-LD/in-sample) and SumHer (in-sample) with respect to $\hat{h}^2_{GRE}$ for 18 complex traits and diseases in the UK Biobank for which $\hat{h}^2_{GRE} > 0.05$ ($N = 290{,}641$ unrelated British individuals, $M = 459{,}792$ typed SNPs; Methods). Each bar represents the difference between the estimated h_g^2 from one of the methods (LDSC, S-LDSC or SumHer) and $\hat{h}^2_{GRE}$ as a percentage of $\hat{h}^2_{GRE}$. Black bars mark ±2 s.e.m. WHR, waist-to-hip ratio; ENT, ear, nose and throat.

a given genotype matrix, whereas the definitions and interpretations of the estimands of LDSC, S-LDSC and SumHer depend on which SNPs are used in each step of inference (for example, the SNPs used to compute LD scores need not be the same SNPs defining the estimand)[11,12,19]. Overall, our results show that while existing methods can yield biases, for the purpose of estimating total SNP-heritability, most methods are relatively robust.

We conclude with several caveats and future directions. First, the utility of $\hat{h}^2_{GRE}$ critically depends on the ratio between the number of SNPs (M) and the number of individuals (N)—as M/N increases, the eigenstructure of the in-sample LD matrix becomes increasingly distorted (larger eigenvalues are overestimated; smaller eigenvalues are underestimated)[32]. We mitigate this by assuming that chromosomes are approximately independent; as long as N exceeds the number of array SNPs per chromosome, $\hat{h}^2_{GRE}$ provides meaningful estimates of SNP-heritability. While the utility of our approach is limited by the availability of individual-level biobank-scale data, this concern will abate as more biobanks are established[33–35]. A major limitation remains with respect to imputed/sequencing data as M will continue to be orders of magnitude larger than N for the foreseeable future. We defer an investigation of regularized estimation of LD in high-dimensional settings ($M > N$) to future work.

Second, the theoretical guarantees of $\hat{h}^2_{GRE}$ rely on the assumption that OLS association statistics and LD are estimated from the same genotypes. While summary statistics have been made publicly available for hundreds of large-scale genome-wide association studies (GWAS), in-sample LD is usually unavailable for these studies since most are meta-analyses[36]. In addition, summary statistics are often computed using linear mixed models to control for confounding, and previous works have noted that the LD computation must be adjusted to accommodate mixed model association statistics[36,37]. Thus, the sensitivity of $\hat{h}^2_{GRE}$ to reference panel LD (with or without regularized LD estimation) and/or mixed model association statistics remains unclear[29,38]. Furthermore, we simulate phenotypes from typed SNPs because imputed genotypes have highly irregular LD patterns[9,18]. Although it would be more realistic to simulate from sequencing data[18], our simulation design required individual-level genotype measurements in biobank-scale sample sizes.

Third, $\hat{h}^2_{GRE}$ does not correct for population structure/stratification. In real data, we mitigate this by considering only unrelated individuals (> third-degree relatives) and including age, sex and the top 20 PCs as covariates when computing association statistics. While recent work has found evidence of assortative mating for some traits in the UK Biobank (for example, height)[39], our estimates are robust to different relatedness thresholds, suggesting that adjusting for the top 20 PCs sufficiently controls for population stratification. Still, it remains unclear how to quantify the bias of our genome-wide estimator due to structure or assortative mating in real data. Future work is needed to extend the GRE approach to control for ascertainment bias[15,16,40,41].

Finally, while previous works applied similar estimators (defined under fixed effects models) to estimate local SNP-heritability within small regions[28,29], additional work is needed to extend our approach to perform partitioning of SNP-heritability by functional annotations. Existing methods for partitioning SNP-heritability make various assumptions on genetic architecture[8,12–14,30], motivating the development of new methods in this area.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0465-0.

Received: 18 December 2018; Accepted: 13 June 2019;
Published online: 29 July 2019

References

1. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
2. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

3. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
4. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
5. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
6. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
7. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
8. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
9. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
10. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
11. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
12. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).
13. Gazal, S. et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
14. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2018).
15. Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19 (1972).
16. Wu, Y. & Sankararaman, S. A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics* **34**, i187–i194 (2018).
17. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2017).
18. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
19. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK models and functional enrichment estimates. Preprint at *bioRxiv* https://doi.org/10.1101/256412 (2018).
20. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl Acad. Sci. USA* **107**, 1752–1756 (2010).
21. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).
22. Schoech, A. P. et al. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).
23. Zeng, J. et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
24. O'Connor, L. J. *et al*. Polygenicity of complex traits is explained by negative selection. Preprint at *bioRxiv* https://doi.org/10.1101/420497 (2018).
25. Uricchio, L. H., Kitano, H. C., Gusev, A. & Zaitlen, N. A. An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evol. Lett.* **3**, 69–79 (2019).
26. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).
27. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
28. Gamazon, E. R., Cox, N. J. & Davis, L. K. Structural architecture of SNP effects on complex traits. *Am. J. Hum. Genet.* **95**, 477–489 (2014).
29. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary associationdata. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
30. Gazal, S. et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
31. Consortium, T. 1000 G. P. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
32. Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **88**, 365–411 (2004).
33. Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
34. Leitsalu, L. et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
35. Gaziano, J. M. et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
36. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2016).
37. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B. & Eskin, E. Identification of causal genes for complex traits. *Bioinformatics* **31**, i206–i213 (2015).
38. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).
39. Yengo, L. et al. Imprint of assortative mating on the human genome. *Nat. Hum. Behav.* **2**, 948–954 (2018).
40. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl Acad. Sci. USA* **111**, E5272–E5281 (2014).
41. Weissbrod, O., Flint, J. & Rosset, S. Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. *Am. J. Hum. Genet.* **103**, 89–99 (2018).
42. Lee, S. H. et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
43. Lee, S. H. et al. Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).

Acknowledgements

This research was conducted using the UK Biobank Resource under applications 33297 and 33127. We thank the participants of UK Biobank for making this work possible. We also thank R. Johnson, M. Freund, M. Major, S. Gazal, A. Price and D. Balding for helpful discussions. This work was funded by the National Institutes of Health (NIH) under awards R01HG009120, R01MH115676, R01HG006399, U01CA194393, R35GM125055, T32NS048004, T32MH073526 and T32HG002536 and the National Science Foundation (NSF) under award III-1705121.

Author contributions

K.H., K.S.B., H.S. and B.P. conceived and designed the experiments. K.H. and K.S.B. performed the experiments and statistical analyses. A.M., H.S., N.M. and S.S. provided statistical support. K.H., K.S.B. and Y.W. collected and managed the data. K.S.B. and B.P. wrote the manuscript with the participation of all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0465-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.S.B. or B.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

The generalized random effects model. We model the phenotype for an individual n randomly sampled from the population as $y_n = \mathbf{x}_n^T\boldsymbol{\beta} + \epsilon_n$ where $\mathbf{x}_n = (x_{n1} \ldots x_{nM})^T$ is a vector of standardized genotypes measured at M SNPs for individual n, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^T$ is an M-vector of the corresponding standardized SNP effects and $\epsilon_n \sim N(0, \sigma_e^2)$ is environmental noise. We assume $\mathrm{Var}[y_n] = 1$ and that the genotype at each SNP i is centered and scaled in the population such that $\mathrm{E}[x_{ni}] = 0$ and $\mathrm{Var}[x_{ni}] = 1$; that is, $x_{ni} = (g_{ni} - 2f_i)/\sqrt{2f_i(1-f_i)}$, where $g_{ni} \in \{0, 1, 2\}$ is the number of copies of the effect allele at SNP i for individual n and f_i is the population frequency of the effect allele at SNP i. We define the population LD between two SNPs i and j to be $v_{ij} \equiv \mathrm{E}[x_{ni}x_{nj}]$ for all $i \neq j$. The population LD matrix among the M SNPs is therefore $\mathbf{V} \equiv \mathrm{Cov}[\mathbf{x}_n^T]$. For simplicity, we use 'SNP effects' in lieu of 'standardized SNP effects' to refer to $\boldsymbol{\beta}$. We assume that $\mathbf{x}_n$ and $\boldsymbol{\beta}$ are independent given allele frequencies $(f_1, \ldots, f_M)$ and $\mathbf{V}$.

Under the GRE model, the first two moments of β_i are $\mathrm{E}[\beta_i] = 0$ and $\mathrm{Var}[\beta_i] = \sigma_i^2$, where σ_i^2 can be any arbitrary non-negative finite number. We assume the covariance between the effects of different SNPs is 0 (that is, $\mathrm{Cov}[\beta_i, \beta_j] = \mathrm{E}[\beta_i\beta_j] = 0$ for all $i \neq j$). Because the SNP-specific variances can capture any degree of polygenicity and any relationship between genomic features (for example, MAF and LD) and effect size, the GRE model encompasses most realistic genetic architectures (Table 1).

We define total SNP-heritability (h_g^2) to be the proportion of phenotypic variance attributable to the additive effects of a set of M SNPs whose genotypes are directly measured:

$$\begin{aligned} h_g^2 &\equiv \frac{\mathrm{Var}[\mathbf{x}_n^T\boldsymbol{\beta}]}{\mathrm{Var}[y_n]} \\ &= \mathrm{E}[\mathrm{Var}[\mathbf{x}_n^T\boldsymbol{\beta} \mid \boldsymbol{\beta}]] + \mathrm{Var}[\mathrm{E}[\mathbf{x}_n^T\boldsymbol{\beta} \mid \boldsymbol{\beta}]] \\ &= \mathrm{E}[\boldsymbol{\beta}^T\mathrm{Var}[\mathbf{x}_n^T]\boldsymbol{\beta}] + \mathrm{Var}[\mathrm{E}[\mathbf{x}_n^T]\boldsymbol{\beta}] \\ &= \mathrm{E}[\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}] + 0 \\ &= \mathrm{E}[\mathrm{tr}(\mathbf{V}\boldsymbol{\beta}\boldsymbol{\beta}^T)] \\ &= \mathrm{tr}(\mathbf{V}\mathrm{E}[\boldsymbol{\beta}\boldsymbol{\beta}^T]) \\ h_g^2 &= \sum_{i=1}^{M}\sigma_i^2 \end{aligned} \quad (2)$$

Thus, h_g^2 is defined with respect to a given population and a given set of SNPs. By definition, $0 \leq h_g^2 \leq 1$. Similarly, we define regional SNP-heritability (h_k^2) to be the proportion of phenotypic variance due to the additive effects of the genotyped SNPs in region k. We assume that the set of SNPs that defines h_k^2 is a subset of the M SNPs that define h_g^2 (thus, $0 \leq h_k^2 \leq h_g^2$). If region k is the whole genome, $h_k^2 = h_g^2$.

Estimating SNP-heritability under the GRE model. We are interested in estimating h_g^2 under the GRE model (equation (2)). In a GWAS with N individuals genotyped at M SNPs, let $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T$ be the $N \times M$ matrix of standardized genotypes (each column of $\mathbf{X}$ has been standardized to have mean 0 and variance 1), $\mathbf{y} = (y_1, \ldots, y_N)^T$ be the N-vector of standardized phenotypes and $\hat{\mathbf{V}} = (1/N)\mathbf{X}^T\mathbf{X}$ be the $M \times M$ in-sample LD matrix (an estimate of population LD, $\mathbf{V}$) with rank q, where $1 \leq q \leq M$. Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$ be the genotype matrices for K independent regions spanning all M SNPs (for example, chromosomes). For region k containing p_k SNPs, $\mathbf{X}_k$ is the $N \times p_k$ standardized genotype matrix and $\hat{\mathbf{V}}_k$ is the corresponding $p_k \times p_k$ in-sample LD matrix with rank q_k ($1 \leq q_k \leq p_k$). We propose the following estimator for genome-wide SNP-heritability:

$$\hat{h}^2_{\mathrm{GRE}} = \sum_{k=1}^{K}\frac{N\hat{\boldsymbol{\beta}}_k^T\hat{\mathbf{V}}_k^{\dagger}\hat{\boldsymbol{\beta}}_k - q_k}{N - q_k}$$

where $\hat{\boldsymbol{\beta}}_k = (1/N)\mathbf{X}_k^T\mathbf{y}$ is the p_k-vector of marginal SNP effects estimated by OLS for region k and $\hat{\mathbf{V}}_k^{\dagger}$ is the pseudoinverse of $\hat{\mathbf{V}}_k$. Detailed derivations for $\hat{h}^2_{\mathrm{GRE}}$ can be found in the Supplementary Note.

Analytical variance of $\hat{h}^2_{\mathrm{GRE}}$. Following quadratic form theory[29,44], the variance of $\hat{h}^2_{\mathrm{GRE}}$ in the single-block case is

$$\mathrm{Var}[\hat{h}^2_{\mathrm{GRE}}] = \left(\frac{N}{N-q}\right)^2\left(2q\left(\frac{1-h_g^2}{N}\right) + 4h_g^2\right)\left(\frac{1-h_g^2}{N}\right) \quad (3)$$

When using the K-block approximation, which assumes that the blocks are independent, we approximate equation (3) as the sum of the variances of the local SNP-heritabilities:

$$\mathrm{Var}[\hat{h}^2_{\mathrm{GRE}}] = \sum_{k=1}^{K}\left(\frac{N}{N-q_k}\right)^2\left(2q_k\left(\frac{1-h_k^2}{N}\right) + 4h_k^2\right)\left(\frac{1-h_k^2}{N}\right) \quad (4)$$

Equation (3) is estimated by plugging in $\hat{h}^2_{\mathrm{GRE}}$ and equation (4) is estimated by plugging in $(\hat{h}_1^2, \ldots, \hat{h}_K^2)$, the estimates of the regional SNP-heritabilities.

Simulation framework. We simulated phenotypes from real genotype array data (UK Biobank[10]) under a range of genetic architectures. We obtained a set of $N = 337{,}205$ unrelated British individuals by extracting individuals with self-reported British ancestry who are > third-degree relatives (pairs of individuals with kinship coefficient $< 1/2^{(9/2)}$) (ref. [10]) and excluding individuals with putative sex chromosome aneuploidy. In all simulations, we standardize the genotypes before drawing phenotypes. That is, for each SNP i and individual n, we compute $x_{ni} = (g_{ni} - 2f_i)/\sqrt{2f_i(1-f_i)}$, where $g_{ni} \in \{0, 1, 2\}$ is the number of minor alleles and f_i is the in-sample MAF.

Simulations of quantitative traits with no population stratification. Given $\mathbf{X}$ and a fixed value of h_g^2, phenotypes are drawn according to the following model. The proportion of causal variants, p_{causal}, is set to 1, 0.01 or 0.001. Let $c_i \in \{0, 1\}$ be the causal status of SNP i. If $p_{\mathrm{causal}} = 1$, $c_i = 1$ for $i = 1, \ldots, M$. If $0 \leq p_{\mathrm{causal}} < 1$, we draw $p_{\mathrm{causal}} \times M$ SNPs from the set of SNPs with MAF in one of three ranges: (0, 0.5], (0.01, 0.05] or (0.05, 0.5]. We use 'CV MAF' to refer to the MAF range from which the causal variants are drawn. Standardized effects and phenotypes are then drawn according to the model

$$\sigma_i^2 \propto c_i w_i^{\gamma}[2f_i(1-f_i)]^{1+\alpha} \quad (5)$$

$$(\beta_1, \ldots, \beta_M)^T \sim N(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_M^2)) \quad (6)$$

$$(y_1, \ldots, y_N)^T \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, (1-h_g^2)\mathbf{I}_N) \quad (7)$$

where α controls the coupling of MAF and effect size, w_i is a SNP-specific LD weight and $\gamma \in \{0, 1\}$ specifies whether effects are coupled with the LD weights. We simulated two types of LD-dependent architectures by defining $w_1, \ldots, w_M$ to be either (1) the default 'LDAK weights' computed by the LDAK software[6] or (2) the inverse unpartitioned 'LD score' of each SNP computed within a 2-Mb window ($w_i^{-1} = \sum_j v_{ij}^2$ where j indexes the set of SNPs within a 2-Mb window centered on SNP i)[11]. When $\gamma = 1$, both the LDAK weights and inverse LD score weights cause SNPs in regions of higher LD to have smaller effects than SNPs in regions of lower LD. We set α to one of two values: $\alpha = -1$ (a relatively strong inverse relationship between MAF and effect size) or $\alpha = -0.25$ (a weaker inverse relationship between MAF and effect size). Each per-SNP variance is multiplied by a scaling factor so that $\sum_{i=1}^{M}\sigma_i^2 = h_g^2$. Note that $\sigma_i^2 = 0$ if $c_i = 0$.

Finally, given phenotypes $\mathbf{y} = (y_1, \ldots, y_N)^T$ and genotypes $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T$, we computed marginal association statistics through OLS: $\hat{\boldsymbol{\beta}} = (1/N)\mathbf{X}^T\mathbf{y}$.

Simulations of case-control phenotypes with no population stratification. To simulate case-control studies, we first draw each individual's continuous liability (l_n for individual n) according to equation (7). For a given population prevalence ($0 \leq d_{\mathrm{pop}} \leq 1$), we computed the corresponding liability threshold $L = \Phi^{-1}(1 - d_{\mathrm{pop}})$, where Φ is the cumulative distribution function of the standard normal distribution. Each l_n was then converted into a case-control status: $y_n = 1$ if $l_n \geq L$ or $y_n = 0$ if $l_n < L$. For unascertained case-control studies, we assumed that the proportion of cases in the study is equal to the population prevalence ($d_{\mathrm{GWAS}} = d_{\mathrm{pop}}$). For ascertained case-control studies ($d_{\mathrm{GWAS}} > d_{\mathrm{pop}}$), we set $d_{\mathrm{GWAS}} = 0.5$ and selected a random set of controls to satisfy $N_{\mathrm{case}} = N_{\mathrm{control}}$.

We computed association statistics by regressing the binary case-control statuses on genotypes. The GRE estimator produces an estimate of SNP-heritability on the observed scale ($\hat{h}^2_{\mathrm{obs}}$). Assuming we know the population prevalence, we converted $\hat{h}^2_{\mathrm{obs}}$ to the liability scale with the transformation $\hat{h}^2_{\mathrm{liab}} = \hat{h}^2_{\mathrm{obs}}d^2_{\mathrm{pop}}(1-d_{\mathrm{pop}})^2/([f(L)]^2 d_{\mathrm{GWAS}}(1-d_{\mathrm{GWAS}}))$, where f is the standard normal probability density function[45].

Simulations with population stratification. To simulate GWAS with population stratification, we draw phenotypes from a model where a covariate that is correlated to genotypes has a non-zero effect on phenotype. To this end, we simulated an effect of the first genetic principal component ($\mathbf{PC}_1$). Letting σ_s^2 be the proportion of total phenotypic variance explained by $\mathbf{PC}_1$, phenotypes were drawn from the model

$$(y_1, \ldots, y_N)^T \mid \beta \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{PC}_1\beta_s, (1-h_g^2-\sigma_s^2)\mathbf{I}_N)$$

where $\mathrm{Var}[\mathbf{PC}_1\beta_s]/\mathrm{Var}[\mathbf{y}] = \beta_s^2\mathrm{Var}[\mathbf{PC}_1] = \sigma_s^2$. We compute association statistics from one of two models: $\mathbf{y} = \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\epsilon}$, which ignores population stratification and other sources of confounding, or $\mathbf{y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{PC}_1\beta_s + \boldsymbol{\epsilon}$, which controls for the effect of $\mathbf{PC}_1$.

Comparison of methods in simulations. Unless otherwise specified, in all genome-wide simulations, we used real genotypes of $N = 337{,}205$ unrelated British individuals measured at $M = 593{,}300$ array SNPs to draw causal effects for all M SNPs and phenotypes for all N individuals. OLS summary statistics are computed for all M SNPs using the simulated phenotypes and real genotypes of all N individuals. We compare to three methods that operate on summary statistics and are computationally tractable for these simulations: LD score regression (LDSC)[11], stratified LD score regression (S-LDSC)[12,13] and SumHer[14].

For LDSC and S-LDSC, we computed the unpartitioned LD score of each SNP as a function of its LD to all other SNPs in a 2-Mb window centered on the SNP. For each annotation included in S-LDSC, the partitioned LD score of each SNP is a function of its LD to all SNPs within a 2-Mb window that are in the annotation. For both LDSC and S-LDSC, LD scores were computed with the LDSC software (https://github.com/bulik/ldsc/) from a random sample of 40,000 individuals to reduce the amount of memory required by the software. We ran the regression with an unconstrained intercept, using all M SNPs as observations in the response variable. Each SNP was weighted to account for heteroscedasticity and correlations between association statistics[11]. For both methods, h_g^2 was estimated as a function of all M SNP-specific variances by using the flags --not-M-5-50 and --chisq-max 99999 (the latter option prevents the LDSC software from dropping high-effect SNPs).

We ran S-LDSC in two ways to account for MAF/LD-dependent architectures. S-LDSC (MAF) refers to S-LDSC with 10 binary MAF bin annotations (each bin contains exactly 10% of the typed SNPs), which is intended to mirror the ten MAF annotations in the 'baseline-LD model'[13] (see Supplementary Table 14 for precise MAF bin ranges for the UK Biobank Axiom Array). S-LDSC (MAF + LLD) refers to S-LDSC with the same 10 MAF bins and an additional continuous 'level of LD' (LLD) annotation computed by quantile-normalizing the unpartitioned LD scores within each MAF bin to a standard normal distribution[13]. While our definition of LLD is intended to mirror the LLD annotation in the baseline-LD model, we did not set the LLD of variants with MAF < 0.05 to 0 because our estimand of interest includes the effects of SNPs with MAF < 0.05 (ref. [13]).

To run SumHer, we used the LDAK software (http://dougspeed.com/ldak/) to compute the default 'LDAK weights' using in-sample LD[6,9,14]. We then computed 'LD tagging' (LD scores) using 1-Mb windows centered on each SNP and setting $\alpha = -0.25$ as recommended[14]. The LDAK software is memory-efficient, allowing us to use all 337,205 individuals to compute LDAK weights and LD tagging. Unless otherwise specified, all default parameter settings were used to run SumHer in simulations.

We also performed simulations with $N = 8{,}430$ unrelated individuals at $M = 14{,}821$ array SNPs. These individuals and SNPs are a subset of the data used in the genome-wide simulations, chosen by selecting approximately 2.5% of individuals and the first 2.5% of SNPs from the beginning of each chromosome to preserve the LD structure among the SNPs. We ran single-component GREML[3,46] (GCTA software https://cnsgenomics.com/software/gcta/) and single-component BOLT-REML[8] (https://data.broadinstitute.org/alkesgroup/BOLT-LMM/) with default parameters. We ran GREML-LDMS-I[18,46] using eight GRMs created from two MAF bins (MAF ≤ 0.05 and MAF > 0.05) and four LD score quartiles; LD scores were computed using the GCTA software with the default window size of 200-kb. We ran LDAK using the default LDAK weights, setting $\alpha = -0.25$ as recommended[6,9].

A third set of simulations was performed using 7,685 individuals of South Asian ancestry in the UK Biobank. This group was composed of individuals of Indian ($n = 5{,}716$), Pakistani ($n = 1{,}748$) and Bangladeshi ($n = 221$) ancestry. Due to the small sample size, we used a reduced set of 803 SNPs from chromosome 21 and 839 SNPs from chromosome 22 (1,642 SNPs in total) that were chosen so that N/p_k for each chromosome k was similar to N/p_k in the 'white British' cohort.

For a given genetic architecture, we generate 100 simulation replicates and obtain 100 estimates of h_g^2 from each method. We estimate the bias of an estimator $\hat{h}_g^2$ under a given architecture as $\text{bias}[\hat{h}_g^2] = \text{E}[\hat{h}_g^2] - h_g^2 \approx (1/100)\sum_{i=1}^{100}\hat{h}_g^2(i) - h_g^2$ where $\hat{h}_g^2(i)$ is the estimate from the ith simulation. To test whether the bias is statistically significant (null hypothesis: bias $[\hat{h}_g^2] = 0$), we assess the z-score of the bias ($z_{\text{bias}} = \text{bias}[\hat{h}_g^2]/\text{s.e.m.}[\hat{h}_g^2]$, where s.e.m.$[\hat{h}_g^2]$ is the standard error of the mean of 100 estimates), which follows a $N(0, 1)$ distribution under the null hypothesis. The P value of the bias is computed with a two-tailed test. To enable a comparison of estimators across different values of h_g^2, we assess the relative bias of an estimator under a single architecture $(\text{bias}[\hat{h}_g^2]/h_g^2)$ as a percentage of h_g^2. In Fig. 1a,c, we compute the error of a single estimate as $(\hat{h}_g^2(i) - h_g^2)/h_g^2$; errors are also reported as percentages of h_g^2.

Analysis of UK Biobank phenotypes. We estimated SNP-heritability for 22 complex traits (6 quantitative, 16 binary) in the UK Biobank[10]. We used PLINK[47] (https://www.cog-genomics.org/plink2) to exclude SNPs with MAF < 0.01 and genotype missingness > 0.01 as well as SNPs that fail the Hardy–Weinberg test at significance threshold 10^{-7}. We retained only the individuals with self-reported British white ancestry and no kinship (that is, greater than third-degree relatives, defined as pairs of individuals with kinship coefficient < $1/2^{(9/2)}$) (ref. [10]). After removing individuals who are outliers for genotype heterozygosity and/or missingness, we obtained a set of $N = 290{,}641$ individuals to use in the real data analyses. For all traits, marginal association statistics were computed through OLS in PLINK, using age, sex and the top 20 genetic PCs as covariates in the regression; these 20 PCs were precomputed by UK Biobank from a superset of 488,295 individuals. Additional covariates were used for waist-to-hip ratio (adjusted for body mass index (BMI)) and diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy and oral contraceptives). We computed $\hat{h}_{\text{GRE}}^2$ for each trait using in-sample LD estimated from all N individuals.

When using LDSC, S-LDSC or SumHer to estimate SNP-heritability, it is necessary to define and distinguish between the following sets of SNPs: the set of SNPs containing all possible causal SNPs of interest (used to compute LD scores and LDAK weights), the set of SNPs used as observations in the regression and the set of SNPs that defines the SNP-heritability estimand of interest. We ran two versions of LDSC, S-LDSC (controlling for the most recent baseline-LD model[12,13,30]) and SumHer[14]. First, to enable a direct comparison between $\hat{h}_{\text{GRE}}^2$ and the estimands of LDSC, S-LDSC and SumHer, we ran an 'in-sample LD' version of each method where the M typed SNPs are used to compute LD scores and LDAK weights, perform the regression and define the SNP-heritability estimand of interest. We refer to these as LDSC (in-sample), S-LDSC (baseline-LD/in-sample) and SumHer (in-sample). To run LDSC (in-sample) and S-LDSC (baseline-LD/in-sample), we used the LDSC software to compute LD scores and regression weights within 2-Mb windows centered on each SNP, using a random sample of 40,000 individuals to reduce the memory requirement. To run SumHer (in-sample), we used the LDAK software to compute LD tagging from the genotypes of all N individuals, using 1-Mb windows centered on each SNP and setting $\alpha = -0.25$ as recommended[9,14]. Unless otherwise specified, all other parameters were set to the default settings.

To enable comparisons between $\hat{h}_{\text{GRE}}^2$ and estimates reported in the literature, we also ran each method with its recommended parameter settings and LD estimated from reference panel sequencing data. We refered to these methods as LDSC (1KG), S-LDSC (baseline-LD/1KG) and SumHer (1KG) to indicate that LD is estimated from 489 Europeans in the 1000 Genomes Phase 3 reference panel[31]. We ran LDSC (1KG) and S-LDSC (baseline-LD/1KG) with LD scores and regression weights computed from 9,997,231 SNPs with minor allele count greater than 5 in the reference panel (1-cM windows), and we define the SNP-heritability estimand to be a function of the array SNPs with MAF > 0.05 (refs. [11,12]). We ran SumHer (1KG) using 8,569,062 SNPs with MAF > 0.01 in the reference panel to compute LDAK weights and LD tagging (1-cM windows) and to define the SNP-heritability estimand; we control for a multiplicative inflation of test statistics as recommended[14]. See refs. [11,12,14,19] for details about the definitions and interpretations of the estimands of LDSC, S-LDSC and SumHer.

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

Data availability

The baseline-LD annotations used in Fig. 4 are available at https://data.broadinstitute.org/alkesgroup/LDSCORE/. All individual-level genotypes and phenotypes were obtained from the UK Biobank (https://www.ukbiobank.ac.uk); we do not have permission to release this data. The 1000 Genomes Phase 3 reference panel can be downloaded at http://www.internationalgenome.org/data.

Code availability

Open source code implementing the GRE estimator and our simulation framework is available on Github at https://github.com/bogdanlab/h2-GRE.

References

44. Elman, R. S., Karpenko, N. & Merkurjev, A. *The Algebraic and Geometric Theory of Quadratic Forms* Vol. 56 (American Mathematical Society, 2008).
45. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
46. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
47. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

nature research

Corresponding author(s): Kathryn S. Burch and Bogdan Pasaniuc

Last updated by author(s): May 29, 2019

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	The code for the GRE estimator is available at https://github.com/bogdanlab/h2-GRE. We used LDSC v1.0.0 (https://github.com/bulik/ldsc), GCTA v1.91.6beta (https://cnsgenomics.com/software/gcta/), BOLT-LMM v2.3.2 (https://data.broadinstitute.org/alkesgroup/BOLT-LMM/), LDAK v5 (http://dougspeed.com/ldak/), PLINK v1.90 beta (https://www.cog-genomics.org/plink2), and Python 2.7.

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

Data

Policy information about availability of data

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

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

The baseline-LD annotations used in Fig. 4 are available at https://data.broadinstitute.org/alkesgroup/LDSCORE/. All individual-level genotypes (used in Figs. 1-4) and phenotypes (Fig. 4) were obtained from the UK Biobank (https://www.ukbiobank.ac.uk); we did not have permission to release this data. The 1000 Genomes Phase 3 reference panel can be downloaded at http://www.internationalgenome.org/data.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No data were generated for this study. We report sample sizes for all publicly available data used in our study.
Data exclusions	No data were excluded for our study.
Replication	No replication was performed.
Randomization	Randomization was not necessary for our study design or statistical analyses.
Blinding	Blinding was not necessary for our study design.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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