Evaluating and improving heritability models using summary statistics

Doug Speed,[1,2,3] John Holmes,[4] and David J Balding[3,4]
Corresponding author: doug@aias.au.dk

1 Aarhus Institute of Advanced Studies (AIAS), Aarhus University, Denmark.
2 Bioinformatics Research Centre, Aarhus University, Denmark.
3 UCL Genetics Institute, University College London, United Kingdom.
4 Melbourne Integrative Genomics, School of BioSciences and School of Mathematics & Statistics, University of Melbourne, Australia.

Abstract
There is currently much debate regarding the best way to model how heritability varies across the genome. The authors of GCTA recommend the GCTA-LDMS-I Model, the authors of LD Score Regression recommend the Baseline LD Model, while we have instead recommended the LDAK Model. Here we provide a statistical framework for assessing heritability models using summary statistics from genome-wide association studies. Using data from studies of 31 complex human traits (average sample size 136,000), we show that the Baseline LD Model is the most realistic of the existing heritability models, but that it can be improved by incorporating features from the LDAK Model. Our framework also provides a method for estimating the selection-related parameter α from summary statistics. We find strong evidence (P<1e-6) of negative genome-wide selection for traits including height, systolic blood pressure and college education, and that the impact of selection is stronger inside functional categories such as coding SNPs and promoter regions.

Previously,[1] we explained how the softwares GCTA[2] and LD Score Regression[3] (LDSC) derive their estimates from an underlying heritability model, and demonstrated how changing this model can lead to very different estimates of SNP heritability, confounding bias and heritability enrichments.[4] Originally, both GCTA and LDSC assumed that each SNP is expected to contribute equal heritability, which we refer to as the GCTA Model.[5] We showed that it is more realistic to assume the LDAK Model, where expected heritability varies with both linkage disequilibrium (LD) and minor allele frequency (MAF).[1,6] Now the authors of GCTA recommend using the GCTA-LDMS-I Model,[7,8] which partitions the GCTA Model based on MAF and LD, while the authors of LDSC recommend using the Baseline LD Model,[9,10] which adds to the GCTA Model 74 SNP annotations based on LD, MAF and functional classifications (when used with this model, LDSC is referred to as S-LDSC).

Our earlier work[1] compared the GCTA and LDAK Models based on the likelihood from restricted maximum likelihood[11] (REML). However, this approach requires access to individual-level data and is only feasible for relatively simple heritability models. We now propose an approximate model likelihood that can be computed from genome-wide association study (GWAS) summary statistics and for highly complex heritability models.

Results
For our main analysis, we use summary statistics from 31 GWAS to compare twelve heritability models (Table 1). We perform this analysis using SumHer, part of our software package LDAK (see URLs).[4,6] Here we briefly describe the heritability models, our proposed model likelihood and the data we use; for full details see Online Methods.

Heritability models. The heritability model specifies how $E[h^2_j]$, the expected heritability contributed by SNP j, varies across the genome. We consider nine existing heritability models. The one-parameter GCTA Model[5] assumes $E[h^2_j]$ is constant. The 20-parameter GCTA-LDMS-R[7] and GCTA-LDMS-I[8] Models both partition the genome based on MAF and LD, then assume $E[h^2_j]$ is constant within each bin. The 53-parameter Baseline Model[9] extends the GCTA Model by adding 52 functional annotations; these include 24 function indicators that can be used to estimate functional enrichments (the heritability enrichments of functional categories of SNPs). The 75-parameter Baseline LD Model[10] adds to the GCTA Model six LD-related annotations, ten MAF indicators and 58 functional annotations (including the 52 functional annotations of the Baseline Model). The two-parameter GCTA+1Fun Model[12] adds to the GCTA Model one function indicator from the Baseline LD Model (there are 24 versions of this model, depending on which indicator is added). The one-parameter LDAK Model[1,6] assumes $E[h^2_j]$ is proportional to $w_j[f_j(1-f_j)]^{0.75}$, where w_j is the LDAK weighting of SNP j (w_j tends to be higher for SNPs in low-LD regions) and f_j is its MAF. The two-parameter LDAK+1Fun[1] and 25-parameter LDAK+24Fun[4] Models extend the LDAK Model by adding either one or all 24 function indicators of the Baseline Model.

We construct three novel heritability models. The 66-parameter BLD-LDAK Model combines features of the Baseline LD and LDAK Models; first we add to the Baseline LD Model the LDAK weighting w_j, then we remove the ten MAF indicators and scale the remaining 66 annotations by $[f_j(1-f_j)]^{0.75}$. The 67-parameter BLD-LDAK+Alpha Model is the same, except it scales the annotations by $[f_j(1-f_j)]^{1+\alpha}$, where α is estimated from the data. The one-parameter LDAK-Thin Model is a simplified version of the LDAK Model, obtained by setting the LDAK weightings to either zero or one.

Measuring model fit. Suppose we have summary statistics from a GWAS; let S_j denote the $\chi^2(1)$ test statistic from regressing the phenotype on SNP j. Suppose also we have genotype data from an ancestrally-matched reference panel, from which we can estimate r^2_{jl}, the squared correlation between SNPs j and l. The authors of LDSC[3] derived that the marginal distribution of each S_j is approximately gamma with shape ½ and scale $2E[S_j]$, where $E[S_j]$ is the expectation of S_j (a function of the parameters of the chosen heritability model). Based on this, we propose the approximate joint log likelihood

$$logl_{SS} = \sum_j \frac{1}{u_j} \log\left(\Gamma\left(S_j \middle| \frac{1}{2}, 2E[S_j]\right)\right)$$

where

$$u_j = \sum_{l \text{ near } j} r^2_{jl},$$

and $\Gamma(X|a,b)$ is the probability density function of a gamma distribution with shape a and scale b. The weights $1/u_j$ are the same as those used by LDSC when regressing S_j onto $E[S_j]$, and are included to allow for correlations between local SNPs.[3]

We perform three analyses to support the use of $logl_{SS}$ to compare heritability models. Firstly, Supplementary Fig. 1 shows that for scenarios where both $logl_{SS}$ and the REML likelihood can be computed, they are concordant. Secondly, Supplementary Fig. 2 shows that when we add a non-informative annotation to a heritability model, twice the increase in $logl_{SS}$ is approximately $\chi^2(1)$ distributed (the expected distribution were $logl_{SS}$ an exact likelihood). Thirdly, Table 1 shows that the ranking of heritability models based on $logl_{SS}$ is consistent with the ranking based on leave-one-chromosome-out prediction of test statistics. Additionally, Supplementary Table 1 shows that the ranking of models is unchanged if we instead compute an unweighted version of $logl_{SS}$ using only SNPs in approximate linkage equilibrium and u_j=1.

Data. We use summary statistics from two sets of GWAS. The first set are 14 traits from UK Biobank (UKBb):[13,14] eight continuous (body mass index, forced vital capacity, height, impedance, neuroticism score, pulse rate, reaction time and systolic blood pressure), four binary (college education, ever smoked, hypertension and snorer) and two ordinal (difficulty falling asleep and preference for evenings). We performed these GWAS ourselves; after stringent quality control, 130k samples and 4.7M SNPs remained. We additionally use summary statistics from 17 Public GWAS:[4,15] ten continuous (including anthropometric measures and psychiatric scores) and seven binary (mostly complex diseases). The average sample size is 141k (range 21-329k). As a reference panel, we use 489 European individuals from the 1000 Genome Project,[16] recorded for 10.0M SNPs (MAF>0.005).

Performance of heritability models. Table 1 reports $logl_{SS}$ for the twelve heritability models, averaged across either the 14 UKBb or 17 Public GWAS (values for individual GWAS are in Supplementary Table 2). We rank models based on the Akaike Information Criterion[17] (AIC), equal to $2K-2logl_{SS}$, where K is the number of parameters.

When we restrict to the nine existing heritability models, the Baseline LD Model performs best; it has average AIC 221 lower than the next best model and is the top-ranked model for 28 of the 31 GWAS. However, when we consider all twelve heritability models, the BLD-LDAK and BLD-LDAK+Alpha Models are the best; they both have average AIC 124 lower than the Baseline LD Model, and now these are the top two models for 28 of the 31 GWAS. These two models would remain the best if instead of the AIC, we ranked models based on $-2\ logl_{SS}$ or $4K-2logl_{SS}$ (i.e., either removed or doubled the penalty on parameters).

		Averages over 14 UKBb GWAS			Averages over 17 Public GWAS		
Heritability Model	**K**	**$logl_{SS}$**	**AIC**	**ρ**	**$logl_{SS}$**	**AIC**	**ρ**
GCTA	1	0	0	0.059	0	0	0.038
LDAK	1	56	-111	0.066	37	-74	0.048
LDAK-Thin	1	174	-349	0.073	136	-273	0.056
GCTA+1Fun	2	257	-513	0.068	215	-427	0.056
LDAK+1Fun	2	145	-288	0.067	127	-252	0.055
GCTA-LDMS-R	20	173	-309	0.071	160	-281	0.052
GCTA-LDMS-I	20	179	-321	0.069	145	-252	0.048
LDAK+24Fun	25	220	-391	0.068	224	-400	0.057
Baseline	53	430	-756	0.074	429	-754	0.061
BLD-LDAK	66	611	-1092	0.085	618	-1106	0.071
BLD-LDAK+Alpha	67	612	-1092	0.086	619	-1107	0.071
Baseline LD	75	561	-974	0.082	563	-977	0.070

Table 1 | Performance of heritability models. K is the number of parameters. $logl_{SS}$ is the approximate model log likelihood (relative to the GCTA Model), AIC is the Akaike Information Criterion (also relative to the GCTA Model), and ρ is the weighted correlation between observed and predicted test statistics from leave-one-chromosome-out prediction; values are averaged across either the 14 UKBb or 17 Public GWAS. The s.d. of average ρ is always 0.001. See main text for details of the heritability models. There are 24 versions of the GCTA+1Fun and LDAK+1Fun Models (one for each function indicator); values here correspond to the versions with highest average $logl_{SS}$.

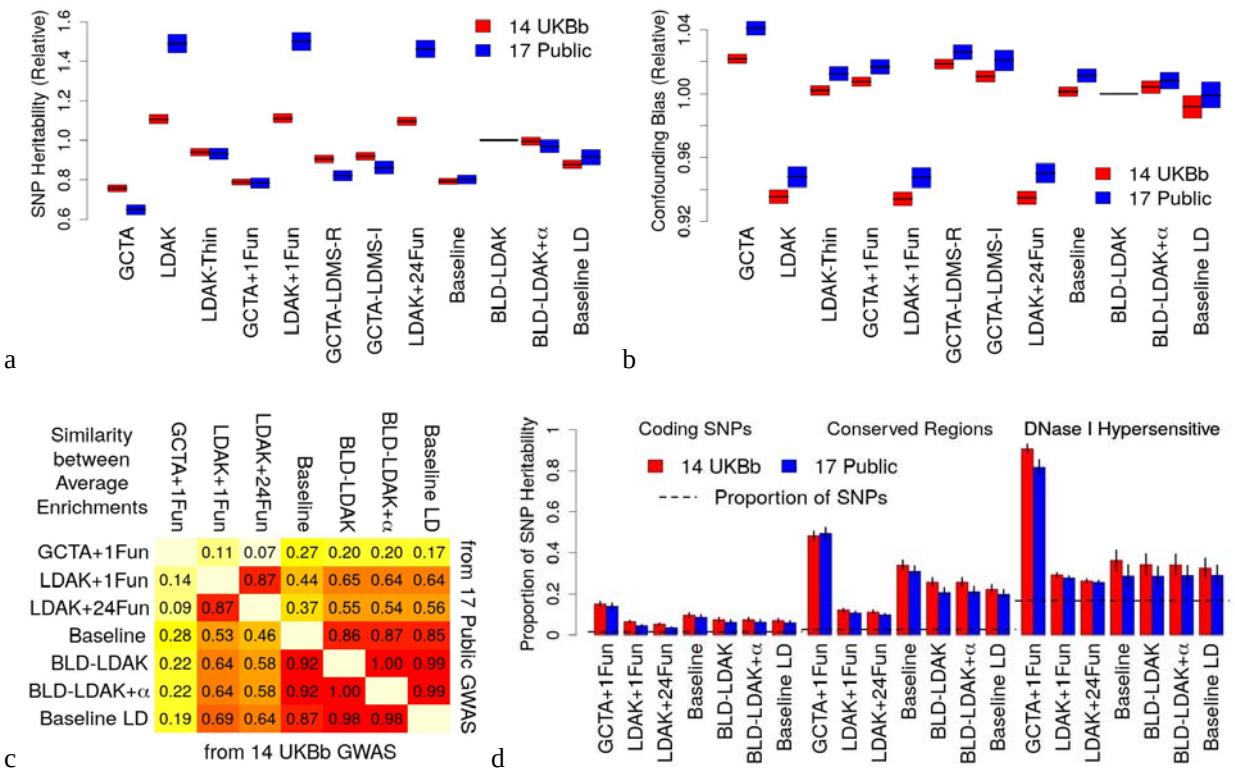
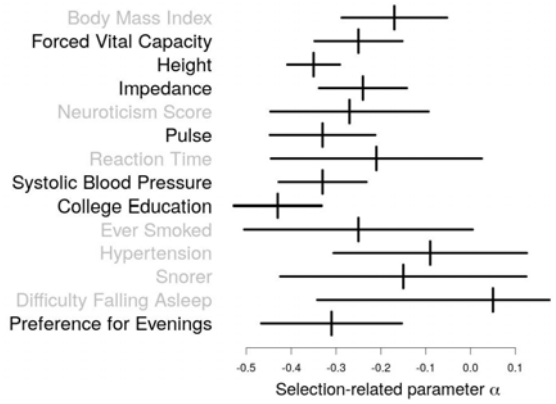


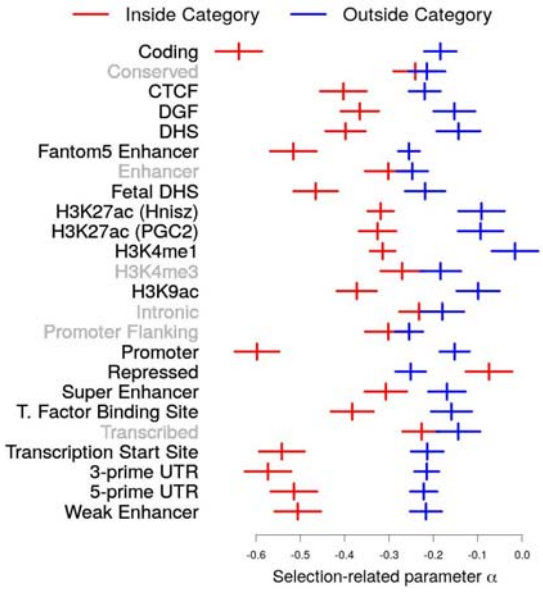
Fig. 1 | Genetic architecture estimates from different heritability models. a & **b**, Average estimates of SNP heritability and confounding bias from twelve heritability models; values are relative to those from the BLD-LDAK Model, and calculated using either the 14 UKBb or 17 Public GWAS (bar heights indicate 95% confidence intervals). There are 24 versions of the GCTA+1Fun and LDAK+1Fun Models (one for each function indicator); values here correspond to the versions with highest average $logl_{SS}$. **c**, Concordance correlation coefficient between average estimates of functional enrichments from seven heritability models; values are calculated using either the 14 UKBb or 17 Public GWAS. **d**, Average estimates of the proportion of SNP heritability contributed by coding SNPs, conserved regions and DNase I hypersensitive sites (three of the 24 functional categories of SNPs) from seven heritability models; values are calculated using either the 14 UKBb or 17 Public GWAS (vertical segments indicate 95% confidence intervals). The estimates of functional enrichments are obtained by dividing these estimates by the proportion of SNPs in each category (dashed horizontal lines).

Genetic architecture estimates. Figure 1, Supplementary Fig. 3 and Supplementary Tables 3-6 compare estimates of SNP heritability, confounding bias, and functional enrichments from the twelve heritability models (note that only the seven models that include function indicators can be used to estimate functional enrichments). As heritability models have become more complex, estimates have tended to converge (the more complex models produce estimates of SNP heritability and confounding bias intermediate between those from the GCTA and LDAK Models, and estimates of functional enrichments intermediate between those from the GCTA+1Fun and LDAK+1Fun Models). Based on model fit (Table 1), we consider the BLD-LDAK and BLD-LDAK+Alpha Models to be most reliable; their estimates of confounding bias and functional enrichments are close to those from the Baseline LD Model, while their estimates of SNP heritability tend to be between those from the LDAK Model and those from the Baseline LD or GCTA-LDMS-I Models.

Evidence of selection. Figure 2a and Supplementary Table 7 report estimates of α in the BLD-LDAK+Alpha Model. This parameter specifies the assumed relationship between heritability and MAF,[1,4] and has been used to measure selection[10,18] (negative α indicates that less-common SNPs tend to have larger effect sizes than more-common SNPs, and vice versa). Across the 31 GWAS, the average estimate of α is -0.25 (s.d. 0.01). Estimates of α are significantly negative (P<0.05/31) for seven of the 14 UKBb GWAS and three of the 17 Public GWAS; the three most significant are for height (α=-0.35, s.d. 0.03), college education (α=-0.43, s.d. 0.05) and systolic blood pressure (α=-0.33, s.d. 0.05).



a



b

Fig. 2 | Evidence of selection. a, Estimates of α for the 14 UKBb GWAS, obtained using the BLD-LDAK+Alpha Model (horizontal lines indicate 95% confidence intervals). α is significantly negative (P<0.05/31) for the seven **bold** traits. **b,** Estimates of α for SNPs inside and outside of the 24 functional categories, obtained using the BLD-LDAK-Lite+Alpha Model and averaged across the 14 UKBb GWAS (horizontal lines indicate 95% confidence intervals). α varies significantly for the 18 **bold** categories.

We also investigate whether α varies across functional categories of SNPs. For computational reasons, we use the seven-parameter BLD-LDAK-Lite+Alpha Model, a reduced version of the BLD-LDAK+Alpha Model (Supplementary Table 8 explains how we constructed the BLD-LDAK-Lite+Alpha Model by identifying the most important annotations of BLD-LDAK+Alpha Model, while Supplementary Fig. 4 shows that estimates of α from the BLD-LDAK-Lite+Alpha Model are consistent with those from the BLD-LDAK+Alpha Model). Figure 2b and Supplementary Table 9 show that α differs significantly ($P<0.05/24$) for 18 of the 24 functional categories. The largest differences in α are observed for coding SNPs, promoter regions, 3'UTR and transcription start sites (inside each of these four categories, the average estimate of α is below -0.5).

Discussion

When software for estimating SNP heritability were first developed, little attention was given to the heritability model, and it was standard to assume that all SNPs are expected to contribute equal heritability.[2,3] It is now recognized that this assumption is sub-optimal,[1,4] and the best way to model how heritability varies across the genome has become a topic of debate.[7,8,10,19,20] Heritability models have previously been compared based on REML likelihood,[1,20] prediction accuracy[4,20] and performance on simulated data,[7,8] however, all three approaches have shortcomings; the REML likelihood requires individual-level data and can not be computed for complex heritability models, to measure prediction accuracy requires two independent datasets for each trait (one for training and one for testing) and there is no consensus regarding the best prediction method, while comparisons of heritability models based on simulated data are sensitive to the assumptions of the simulation model.[1]

We have proposed $\log l_{SS}$, an approximate model likelihood that can be computed from summary statistics and for complex heritability models. $\log l_{SS}$ can be used both to evaluate heritability models (e.g., if two models produce contrasting estimates, then those from the model with highest AIC should be preferred), and to improve them (for example, by combining pairs of models when this significantly improves $\log l_{SS}$). Using $\log l_{SS}$, we have shown that the Baseline LD Model is the best of the existing heritability models, but that it can be substantially improved by incorporating the SNP weightings and MAF scaling used by the LDAK Model. Estimates of confounding bias and functional enrichments from the resulting BLD-LDAK Model are close to those from the Baseline LD Model, while its estimates of SNP heritability are between those of the Baseline LD, GCTA-LDMS-I and LDAK Models.

Our results support the conclusion of Gazal et al.,[20] that estimates of functional enrichments from the Baseline LD Model are more accurate than those from the LDAK+1Fun Model. They provide partial support for Evans et al.,[8] who argued that the GCTA-LDMS-I Model produces the most accurate estimates of SNP heritability (although we found that the GCTA-LDMS-I Model performs well compared to the other existing models, our analysis indicates that the BLD-LDAK Model should now be preferred). Our results support our previous finding[4] that the LDAK Model is more realistic than the GCTA Model, but not that estimates of functional enrichments from the LDAK+24Fun Model should be preferred to those from the Baseline and Baseline LD Models. We discuss these three papers in detail in the Supplementary Note.

Recently, Hou et al.[21] proposed GRE, a method for estimating SNP heritability without specifying a heritability model. While we agree with the benefits of performing heritability analysis without requiring a heritability model, this is not feasible for most analyses. For example, GRE can not be used on our Public GWAS, because it requires individual-level data, nor can it be used on our (full) UKBb GWAS, because it requires that the number of samples is larger than the number of SNPs on the largest chromosome. For Supplementary Fig. 5, we run GRE using a subset of our UKBb GWAS (we restrict to 623,000 directly-genotyped SNPs). Estimates of SNP heritability from the BLD-LDAK Model are consistent ($P>0.05/14$) with those from GRE for all 14 traits. However, this partly reflects that reducing the number of SNPs reduces the impact of the heritability model (the LDAK-Thin, GCTA-LDMS-R, GCTA-LDMS-I and Baseline LD Model also produce consistent estimates for all 14 traits).

The BLD-LDAK+Alpha Model is a generalization of the BLD-LDAK Model. The two models have similar fit and produce similar estimates of SNP heritability, confounding bias and functional enrichments. For computational reasons, we generally recommend the BLD-LDAK Model. However, the advantage of the BLD-LDAK+Alpha is that it provides estimates of the selection-related parameter α. Our results broadly agree with those of Zeng et al.;[18] using the software BayesS, they found significantly negative α for 23 out of 28 UK Biobank traits, while their average estimate of α was -0.38 (s.d. 0.01). To our knowledge, SumHer is the only software to estimate α from summary statistics, and therefore can be viewed as a more computationally-efficient alternative to BayesS (for a full comparison of the two methods and their results, see Supplementary Fig. 6).

Although its shortcomings have been well documented, the GCTA Model continues to be widely used in statistical genetics. It remains the default model of both the GCTA and LDSC software, and is the model used by LD Hub.[22] More widely, the GCTA Model is implicitly assumed by any penalized or Bayesian regression method that

standardizes genotypes then assigns the same penalty or prior distribution to each SNP, or in simulations when causal SNPs are picked at random then their standardized effects sizes drawn from the same distribution. Ideally, the GCTA Model should be replaced by the BLD-LDAK or BLD-LDAK+Alpha Model whenever it occurs. However, we recognize that for many methods, introducing a multi-parameter heritability model would require substantial algorithmic changes and dramatically increase computational demands. When this is the case, we instead recommend using the LDAK-Thin Model; this is a one-parameter model, so computation demands should not be affected, that can be incorporated in any existing method simply by changing which predictors are included in the regression and how these are standardized.

We finish by highlighting three areas for future work. Firstly, we have only considered common SNPs; with the increasing availability of sequence data, it will be necessary to examine whether the BLD-LDAK and BLD-LDAK+Alpha Models remain the best performing model when rare SNPs are included. Secondly, the ability to measure model fit for very large sample sizes means that we now have sufficient power to construct heritability models specific to either individual traits or groups of traits (Supplementary Table 2). Thirdly, we have only considered the genomic annotations contained within the Baseline LD Model. We expect it will be possible to find new annotations predictive of how heritability varies across the genome (i.e., whose inclusion in the heritability model significantly increases $\log l_{SS}$). Identifying these will both improve the performance of the heritability model and our understanding of the genetic architecture of complex traits.

Online Methods

Let h^2_j denote the heritability contributed (uniquely) by SNP j. Suppose we have summary statistics from a GWAS of n individuals; let S_j denote the $\chi^2(1)$ test statistic from regressing the phenotype on SNP j. Suppose also we have access to an ancestrally-matched reference panel, from which we can estimate r^2_{jl}, the squared correlation between SNPs j and l (genotypes coded additively).

Linear heritability models. The heritability model describes how the expectation of h^2_j varies across the genome. We first assume the model takes the form

$$E[h_j^2] = \sum_k a_{jk} \tau_k$$

where the a_{jk} are pre-specified SNP annotations and the parameters τ_k are estimated in the analysis. Assuming no confounding bias[3]

$$E[S_j] \approx 1 + \sum_k \left(n \sum_{l \in \text{Ref j}} r_{jl}^2 a_{jk} \right) \tau_k \quad (1)$$

where Ref j indexes the reference panel SNPs 'near' SNP j (a working definition of near is within 1cM[4]). The term within the parentheses is known, and thus we can estimate the τ_k by regressing S_j on $E[S_j]$ (details below). To allow for confounding bias, the authors of LDSC[3] recommend increasing each $E[S_j]$ by A, the average amount each test statistic is inflated additively (it is standard to then report 1+A, referred to as the intercept), while we[4] instead recommended scaling each $E[S_j]$ by C, the average amount each test statistic is inflated multiplicatively. Allowing for confounding bias (whether additive or multiplicative) results in a revised form for Eq. (1), but it remains that $E[S_j]$ can be expressed as a linear combination of the model parameters[4] (now the τ_k plus either A or C), and so we can continue to estimate the parameters by regressing S_j on $E[S_j]$.

Approximate model likelihood. The authors of LDSC[3] derived that each S_j approximately follows a scaled $\chi^2(1)$ distribution with scale factor $E[S_j]$ (or equivalently, a gamma distribution with shape ½ and scale $2E[S_j]$). Let

$$u_j = \sum_{l \in \text{Reg j}} r_{jl}^2,$$

where Reg j indexes the regression SNPs (those used when regressing S_j on $E[S_j]$) near SNP j. We propose the approximate log likelihood

$$\log l_{SS} = \sum_j \frac{1}{u_j} \left(-\frac{S_j}{2E[S_j]} - \frac{1}{2}\log(S_j) - \frac{1}{2}\log(2E[S_j]) - \frac{1}{2}\log(\pi) \right)$$

~~To ensure $logl_{SS}$ can be computed, we replace non-positive $E[S_j]$ with 10-6 (this is rarely an issue, because $E[S_j]$ generally remains positive even if some $E[h^2_j]$ are negative).~~ The term within the large parentheses is the log likelihood for a single SNP; therefore $logl_{SS}$ computes a weighted sum of these, where the weights $1/u_j$ reflect local correlations. Supplementary Figs. 1 & 2 show that $logl_{SS}$ is concordant with the exact likelihood computed from REML and can be used for likelihood ratio testing.

Estimating parameters. LDSC estimates the parameters of the heritability model using weighted least-squares regression,[3] with regression weights $1/u_j$ x $1/(2E[S_j]^2)$; weighting by $1/u_j$ allows for correlations between nearby SNPs (motivating our use of $1/u_j$ in the definition of $logl_{SS}$), while weighting by $1/(2E[S_j]^2)$ allows for heteroscedasticity ($2E[S_j]^2$ is the variance of a gamma distribution with shape ½ and scale $2E[S_j]$).

When we proposed SumHer, we followed LDSC and estimated parameters using weighted least-squares regression.[4] However, now that we have an expression for the model likelihood, we can instead use maximum likelihood estimation. To identify the values that maximize $logl_{SS}$, we use (multi-dimensional) Newton-Raphson.[23] Let θ denote the vector of parameters (the τ_k and, if allowing for confounding, either A or C). Starting from the null model (τ_k=0, A=0, C=1), we update θ iteratively until convergence using $\theta \rightarrow \theta - E^{-1}D$, where the vector D and matrix E contain, respectively, the first and second derivatives of $logl_{SS}$ evaluated using the current parameter values (the required derivatives can be computed using the chain rule; for example, $\delta logl_{SS} / \delta\theta_k = \delta logl_{SS} / \delta E[S_j] \times \delta E[S_j] / \delta\theta_k$). Sometimes, a move causes a (non-negligible) reduction in $logl_{SS}$. When this happens, we cancel the move, then for the next iteration (only) update each parameter once individually using (one-dimensional) Newton-Raphson.

Except where stated, our analyses use the maximum likelihood solver, regardless of the heritability model. In Supplementary Fig. 7, we compare the impact of changing the solver. This shows that for simple heritability models, the weighted least-squares and maximum likelihood solver result in identical $logl_{SS}$, but that for complex models, the maximum likelihood solver often results in substantially higher $logl_{SS}$

Non-linear heritability models. To date LDSC and SumHer have required that the heritability model is linear (this ensures that $E[S_j]$ can be expressed as a linear combination of the model parameters). However, SumHer can now accommodate a small number of non-linear parameters. We first crudely estimate the non-linear parameters using a grid-search, selecting the values that result in highest $logl_{SS}$. We then increase resolution and obtain standard deviations by fitting a Gaussian likelihood to the realizations of $logl_{SS}$. Supplementary Fig. 8 demonstrates how we use this approach to estimate the selection-related parameter α.

Leave-one-chromosome-out prediction of test statistics. For each SNP we compute $E[S_j]$ in Eq. (1) using parameter estimates obtained from the other 21 chromosomes. In Table 1 we report a weighted correlation between predicted and observed test statistics

$$\rho = \frac{C(S_j, E[S_j])}{\sqrt{C(S_j, S_j)}\sqrt{C(E[S_j], E[S_j])}},$$

where

$$C(a_j, b_j) = \left(\sum_j \frac{a_j b_j}{u_j}\right)\left(\sum_j \frac{1}{u_j}\right) - \left(\sum_j \frac{a_j}{u_j}\right)\left(\sum_j \frac{b_j}{u_j}\right)$$

We estimated the standard deviation of ρ using block jackknifing with 200 blocks.[3,24] We consider it appropriate to include the weights $1/u_j$, as otherwise ρ will overweight high-LD regions, however, Supplementary Table 1 shows that the ranking of models is the same if we instead compare unweighted correlations.

GWAS. We accessed UK Biobank[13,14] (UKBb) data via Project 21432. In total we identified 20 phenotypes that were recorded for the majority of individuals: the 14 we retained were body mass index (data field 21001), forced vital capacity (3062), height (50), impedance (23106), neuroticism score (20127), pulse rate (102), reaction time (20023), systolic blood pressure (4080), college education (6138), ever smoked (20160), hypertension (20002), snorer (1210), difficulty falling asleep (1200) and preference for evenings (1180); the six we discarded were asthma, wears glasses, handedness, any mouth problem, basal metabolic rate and diastolic blood pressure (each either had estimated heritability less than 0.1 or was highly correlated with one of the retained phenotypes). The imputed

dataset contains 487k individuals recorded for 93M SNPs. However, after quality control, which included filtering individuals based on ancestry and relatedness (for the latter, we ensured that no pair remained with allelic correlation >0.02), and excluding SNPs with MAF<0.01, info score <0.99 or within the major histocompatibility complex (Chr6:25-34Mb), only 130,080 individuals and 4,725,151 SNPs remained.[4,5,25] As we have individual-level data, we could use our previously-published protocols to confirm that confounding due to residual population structure, relatedness or genotyping errors was slight[1]. For the association analysis, we tested each SNP using linear regression (regardless of whether the phenotype was continuous, categorical or binary), having first regressed the phenotype on 13 covariates: age (data field 21022), sex (31), Townsend Deprivation Index (189) and ten principal components. For more details, see Supplementary Figs. 9 & 10.

The 17 Public GWAS are coronary artery disease,[26] Crohn's Disease,[27] ever smoked,[28] inflammatory bowel disease,[27] rheumatoid arthritis,[29] schizophrenia,[30] type 2 diabetes,[31] bone mineral density,[32] body mass index,[33] depressive symptoms,[34] height,[35] menarche age,[36] menopause age,[37] neuroticism,[34] subjective well-being,[34] waist-hip ratio[38] and years education.[39] These are a subset of the 24 GWAS we considered previously;[4] we excluded the remaining seven GWAS as the authors of LDSC[9,40] recommend only using traits with a heritability Z-score above seven. For these GWAS we have to rely on the quality control choices of the original authors (Supplementary Table 10), which are generally less strict than ours, and without access to individual-level data, we could not test for confounding due to population structure, relatedness or genotyping errors.

Software settings. When running an analysis using LDSC or SumHer it is necessary to choose the heritability and confounding models, provide a reference panel, select the regression and heritability SNPs, and if estimating enrichments, specify the expected proportion of SNP heritability contributed by each category (for an explanation of each option, see Supplementary Table 11). We describe the different heritability models we consider below. For the other options, our main analysis follows the recommendations of LDSC.[3,9] When analyzing the UKBb GWAS, we assume there is no confounding bias (the exception is for Figure 1b, when we allow for additive confounding bias, then report the estimate of 1+A); when analyzing the Public GWAS, we always allow for additive confounding bias. Our reference panel is the 1000 Genome Project[16] dataset provided on the LDSC website (see URLs), which contains 489 European individuals recorded for 10.0M autosomal SNPs with MAF>0.005. When analyzing the UKBb GWAS, the regression SNPs are all 4.7M GWAS SNPs; when analyzing the Public GWAS, the regression SNPs are the GWAS SNPs present in HapMap3[41] but not in the major histocompatibility complex (on average 1.1M SNPs per GWAS). The heritability SNPs are the 6.0M reference panel SNPs with MAF≥0.05. When computing estimates of enrichments, we divide the estimated proportion of SNP heritability contributed by a category by the proportion of SNPs it contains.

Sensitivity analyses. Supplementary Table 1 shows that the ranking of heritability models is the same if we use a UK Biobank[13,14] reference panel (instead of the 1000 Genomes Project panel), if we reduce the reference panel to the 4.7M SNPs in our UKBb GWAS, or if we reduce the regression SNPs to a subset in approximate linkage equilibrium. Supplementary Table 12 shows that enrichment estimates from the BLD-LDAK Model are similar if the heritability SNPs are all 10.0M reference panel SNPs (rather than just the 6.0M with MAF≥0.05).

Heritability models. Full details of all heritability models are provided in Supplementary Tables 13 & 14. The one-parameter GCTA Model[2] (the default model of both the GCTA and LDSC software) assumes $E[h^2_j]=\tau_1$. The GCTA-LDMS-R[7] and GCTA-LDMS-I[8] Models both take the form

$$E\left[h_j^2\right]=\sum_k I_{jk}\tau_k$$

where I_{jk} indicates whether SNP j is in Bin k. The bins are obtained by first dividing the genome four ways based on LD, then M ways based on MAF, (when dividing based on LD, the GCTA-LDMS-R Model ranks SNPs based on regional LD scores, while the GCTA-LDMS-I Model ranks based on per-SNP LD scores). We opt for M=5, with the boundaries at 0.1, 0.2, 0.3 and 0.4 (in total 20 bins); this choice is based on the first application of GCTA-LDMS-R,[7] which used seven MAF tranches with boundaries at 0.001, 0.01, 0.1, 0.2, 0.3 and 0.4 (we exclude the bottom two tranches as we only consider common SNPs). Supplementary Table 15 shows that this version performs better than using M=2, with the boundary at 0.05 (in total 8 bins), a choice based on the first application of GCTA-LDMS-I,[8] which used four MAF tranches with boundaries at 0.0025, 0.01 and 0.05.

The 53-parameter Baseline Model[9] adds to the GCTA Model 52 functional annotations; 24 of these are function indicators (e.g., indicating SNPs within coding regions), while 28 are 'buffer' indicators (e.g., indicating SNPs within 500bp of a coding region). The 75-parameter Baseline LD Model[10] adds to the GCTA Model 74 SNP

annotations: six are LD-related annotations (e.g., estimated allele age), ten are MAF indicators (e.g., indicating SNPs with $0.05 \leq MAF < 0.07$) and 58 are functional annotations (including the 52 used in the Baseline Model). The two-parameter GCTA+1Fun Model[12] adds to the GCTA Model one function indicator from the Baseline Model (there are 24 versions of this model, one for each indicator).

The one-parameter LDAK Model[1,4] assumes $E[h^2_j]=w_j p_j^{0.75}\tau_1$, where w_j is the LDAK weighting of SNP j, f_j is its MAF and $p_j=f_j(1-f_j)$. We recommend that the weightings are only computed over high-quality SNPs[1,6] (so low- and moderate-quality SNPs automatically get $w_j=0$). We do not have SNP info scores for the 1000 Genome Project reference panel, so when computing weightings, we restrict to the 4.7M SNPs in our UKBb GWAS (i.e., we assume that SNPs well-genotyped in the UK Biobank are well-genotyped in the 1000 Genome Project). The two-parameter LDAK+1Fun Model[1] assumes $E[h^2_j]=w_j p_j^{0.75}(\tau_1+b_{ji}\tau_2)$, where b_{ji} is the ith function indicator from the Baseline LD Model (there are 24 versions of this model, one for each indicator), while the 25-parameter LDAK+24Fun Model[1] assumes

$$E[h_j^2]=w_j p_j^{0.75}\left(\tau_1+\sum_{k=2}^{25} b_{j(k-1)}\tau_k\right)$$

The novel 66-parameter BLD-LDAK and 67-parameter BLD-LDAK+Alpha Model both take the form

$$E[h_j^2]=p_j^{1+\alpha}\left(\tau_1+\sum_{k=2}^{65} c_{jk}\tau_k+w_j\tau_{66}\right)$$

where the c_{jk} are the 64 LD-related and functional annotations from the Baseline LD Model;[10] the BLD-LDAK Model fixes $\alpha=-0.25$, while the BLD-LDAK+Alpha estimates α from the data. To construct the BLD-LDAK Model we first added the LDAK weighting to the Baseline LD Model (this increased average $logl_{SS}$ by 11), then scaled all annotations by $p_j^{0.75}$ (this increased average $logl_{SS}$ by a further 50). At this point we noted that the 10 MAF indicators had limited value (excluding them reduced average $logl_{SS}$ by only 8) so we removed them. We were unable to improve the model further by adding features from the GCTA-LDMS-R and GCTA-LDMS-I Models (for example, if we incorporated the 4 LD or the 20 MAF-LD bins from the GCTA-LDMS-I Model, this increased the number of parameters by 3 and 19, respectively, but increased average $logl_{SS}$ by only 2 and 13). For more details, see Supplementary Table 16.

The seven-parameter BLD-LDAK-Lite and eight-parameter BLD-LDAK-Lite+Alpha Models are reduced versions of the BLD-LDAK and BLD-LDAK+Alpha Models, respectively, obtained by removing two of the nine continuous annotations and all 57 binary annotations (Supplementary Table 8 explains how we used forward stepwise selection to decide which of the continuous annotations to retain). Supplementary Fig. 4 shows that estimates of SNP heritability and confounding bias from the BLD-LDAK-Lite Model are consistent with those from the BLD-LDAK Model, and that estimates of α from the BLD-LDAK-Lite+Alpha Model are consistent with those from the BLD-LDAK+Alpha Model. To allow α to vary based on functional annotations, we use a 16-parameter model obtained by concatenating two versions of the BLD-LDAK-Lite+Alpha Model (one where only SNPs within the category contribute heritability, and one where only SNPs outside the category contribute).

When computing the LDAK weightings, the first step is to thin SNPs so that no pair remains within 100kb with $r^2_{jl}>0.98$ (excluding duplicate SNPs substantially improves the efficiency of the solver used to compute the weightings[1]). The LDAK-Thin Model assumes $E[h^2_j]=I_j p_j^{0.75}\tau_1$, where I_j indicates whether SNP j remains after the thinning. To implement the LDAK-Thin Model within an existing penalized or Bayesian regression method requires two changes: firstly, thin the SNPs (for the UKBb data, this reduced the number from 4.7M to 1.4M); secondly, center and scale the genotypes so that SNP j has variance $p_j^{0.75}$.

Computational demands. To run SumHer with a linear heritability model involves two steps.[4] The first is to compute the tagging file; this takes at most one day on a single CPU and requires 20Gb memory. The second is to estimate the parameters of the heritability model; this takes at most two hours and requires 20Gb memory. To run SumHer with a non-linear heritability model, it is necessary to repeat the above process multiple times (e.g., Supplementary Fig. 8 explains how we fit the BLD-LDAK+Alpha Model by calculating tagging files and estimating parameters for 31 versions of BLD-LDAK Model). The LDAK website (see URLs) provides pre-computed tagging files for the BLD-LDAK and BLD-LDAK-Lite+Alpha Models, suitable for analyzing European, South Asian, East Asian or African populations.

URLs. LDAK, http://www.ldak.org; LDSC, http://www.github.com/bulik/ldsc; UK Biobank https://www.ukbiobank.ac.uk

Acknowledgements
We thank B. Shaban for help with the LDAK website, and A. Price, S. Gazal and H. Finucane for helpful discussions. D.S. is funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 754513, by Aarhus University Research Foundation (AUFF) and by the Independent Research Fund Denmark under Project no. 7025-00094B. D.S. and D.J.B. are funded by the Australian Research Council under Project no. DP190103188.

Author contributions
D.S. and J.H. performed the analyses, D.S. and D.J.B. wrote the manuscript.

Competing interests
The authors declare no competing interests.

Data availability
We performed the UKBb GWAS using data applied for and downloaded via the UK Biobank website (see URLs). We obtained summary statistics for the Public GWAS from the websites of the corresponding studies. We downloaded the 1000 Genome Project data from the LDSC website (see URLs).

References

1. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).

2. 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).

3. Bulik-Sullivan, B. *et al.* LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nat. Genet.* **47**, 291–295 (2015).

4. Speed, D. & Balding, D. Better estimation of SNP heritability from summary statistics provides a new understanding of the genetic architecture of complex traits. *Nat. Genet.* **51**, 277–284 (2019).

5. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

6. Speed, D., Hemani, G., Johnson, M. & Balding, D. Improved heritability estimation from genome-wide SNP data. *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. 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**, (2018).

9. Finucane, H. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

10. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

11. Corbeil, R. R. & Searle, S. R. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* **18**, 31–38 (1976).

12. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

13. Sudlow, C. *et al.* UK Biobank□: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, e1001779 (2015).

14. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. (2018) doi:10.1038/s41586-018-0579-z.

15. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).

16. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

17. Akaike, H. A new look at the statistical model identification. *Trans. Autom. Contr* **19**, 716–723 (1974).

18. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, (2018).

19. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49**, 1304–1310 (2017).

20. Gazal, S., Marquez-luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK models and functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).

21. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* **51**, 1244–1251 (2019).

22. 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 (2016).

23. Ypma, T. Historical development of the Newton-Raphson method. *SIAM Rev.* **37**, 531–551 (1995).

24. Efron, B. & Stein, C. The Jackknife estimate of variance. *Ann. Stat.* **9**, 586–596 (1981).

25. Speed, D. *et al.* Describing the genetic architecture of epilepsy through heritability analysis. *Brain* **137**, 2680–2689 (2014).

26. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).

27. Liu, J. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

28. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).

29. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

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

31. Scott, R. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).

32. Zheng, H. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).

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

34. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 626–633 (2016).

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

36. Perry, J. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).

37. Day, F. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**, 1294–1303 (2015).

38. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nat. Genet.* **518**, 187–196 (2015).

39. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).

40. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

41. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).