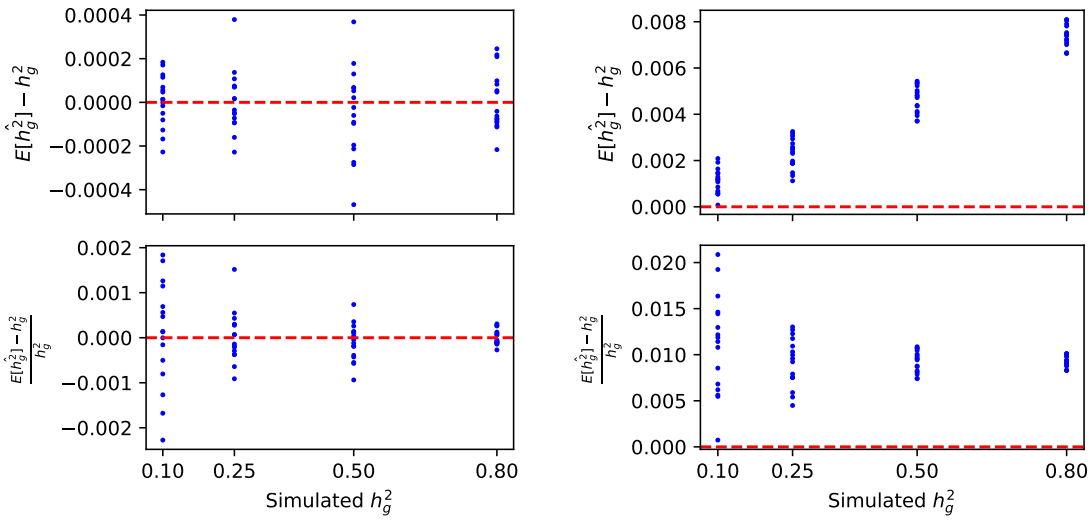
Supplementary Information for "Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture"

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

[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, Zhejiang, 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]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA
[6]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[7]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[*]These authors contributed equally to this work.
[†]Correspondence should be addressed to K.S.B. (kathrynburch@ucla.edu) or B.P. (pasaniuc@ucla.edu)

January 21, 2019



(a) chromosome 22 simulations ($M = 9654$) (b) genome-wide simulations ($M = 593300$)

Figure S1: Bias and relative bias of $\widehat{h}^2_{\text{GRE}}$ in simulations under 64 MAF- and LDAK-LD-dependent architectures ($N = 337$K). (a) Phenotypes were drawn from $M = 9654$ SNPs on chromosome 22; h_g^2 was estimated with a single LD block spanning chromosome 22. (b) Phenotypes were drawn from $M = 593300$ SNPs genome-wide; h_g^2 was estimated using 22 chromosome-wide LD blocks. Each point represents the magnitude of the bias of $\widehat{h}^2_{\text{GRE}}$ (top row) or the bias of $\widehat{h}^2_{\text{GRE}}$ relative to the simulated h_g^2 (bottom row) estimated from 100 simulations under a single genetic architecture.

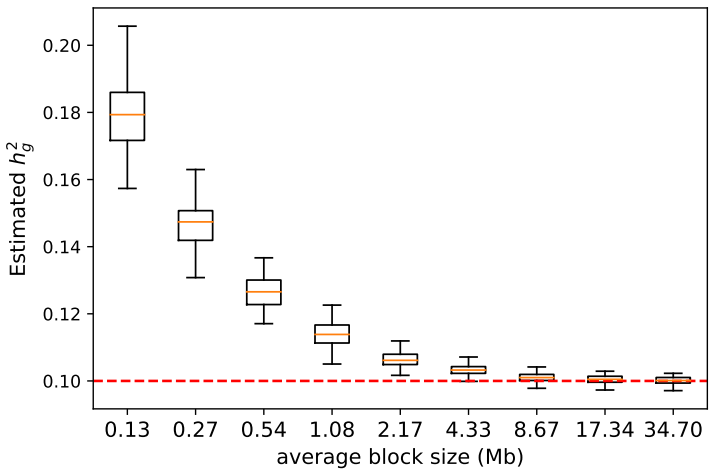


Figure S2: Distribution of $\widehat{h}^2_{\text{GRE}}$ in simulations on chromosome 22 ($N = 337205$, $M = 9564$ array SNPs) as a function of the average size (Mb) of the LD blocks that were used to compute $\widehat{h}^2_{\text{GRE}}$. The largest block size (34.70 Mb) corresponds to using a single chromosome-wide LD block. All simulations were performed $h_g^2 = 0.1$, $p_{\text{causal}} = 0.01$, $\alpha = -1$, and $\gamma = 0$ (no LD weights). Each boxplot represents 100 estimates.

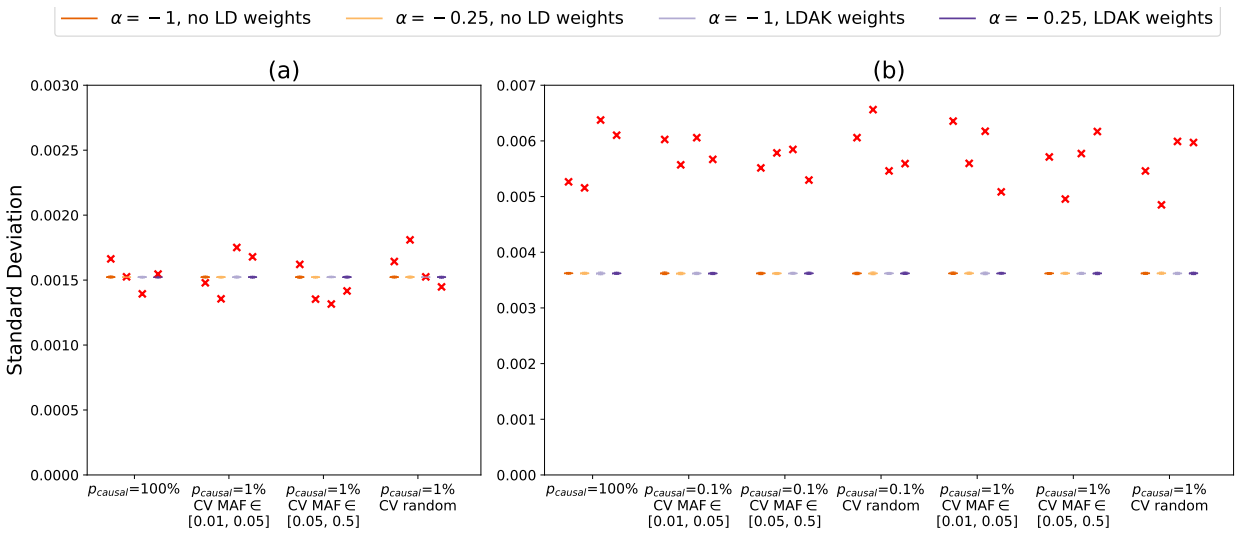


Figure S3: Comparison of the analytical standard error of $\widehat{h}^2_{\text{GRE}}$ with the standard deviation of $\widehat{h}^2_{\text{GRE}}$ computed from 100 simulations ($h^2_g = 0.25$). (a) Phenotypes were simulated from SNPs on chromosome 22 ($N = 337205$, $M = 9564$ array SNPs) under one of 16 LDAK-LD- and/or MAF-dependent architectures and $\widehat{h}^2_{\text{GRE}}$ was computed with a single chromosome-wide LD block. (b) Phenotypes were simulated from all genome-wide SNPs ($N = 337205$, $M = 593300$ array SNPs) under one of 28 LDAK-LD- and/or MAF-dependent architectures and $\widehat{h}^2_{\text{GRE}}$ was computed with 22 chromosome-wide LD blocks. The colored bars represent the distribution of standard error estimates from 100 simulations. The red crosses mark the empirical standard deviation of the 100 estimates of h^2_g.

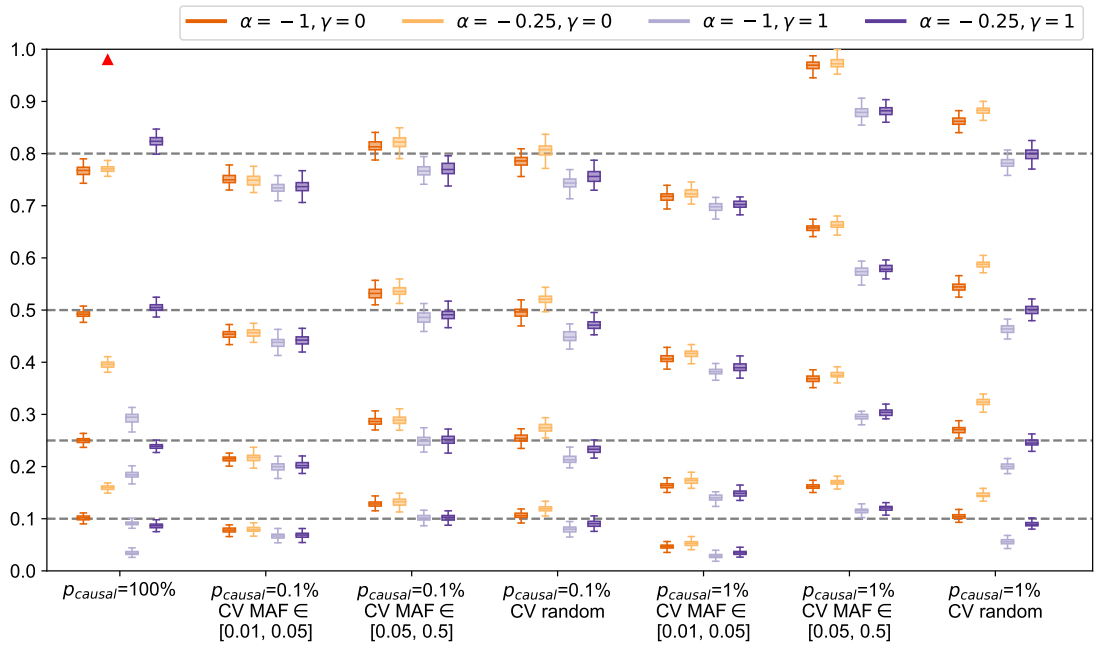


Figure S4: Distribution of h_g^2 estimates from LDSC (no annotations) in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 337205$ individuals, $M = 593300$ array SNPs). Each boxplot represents the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

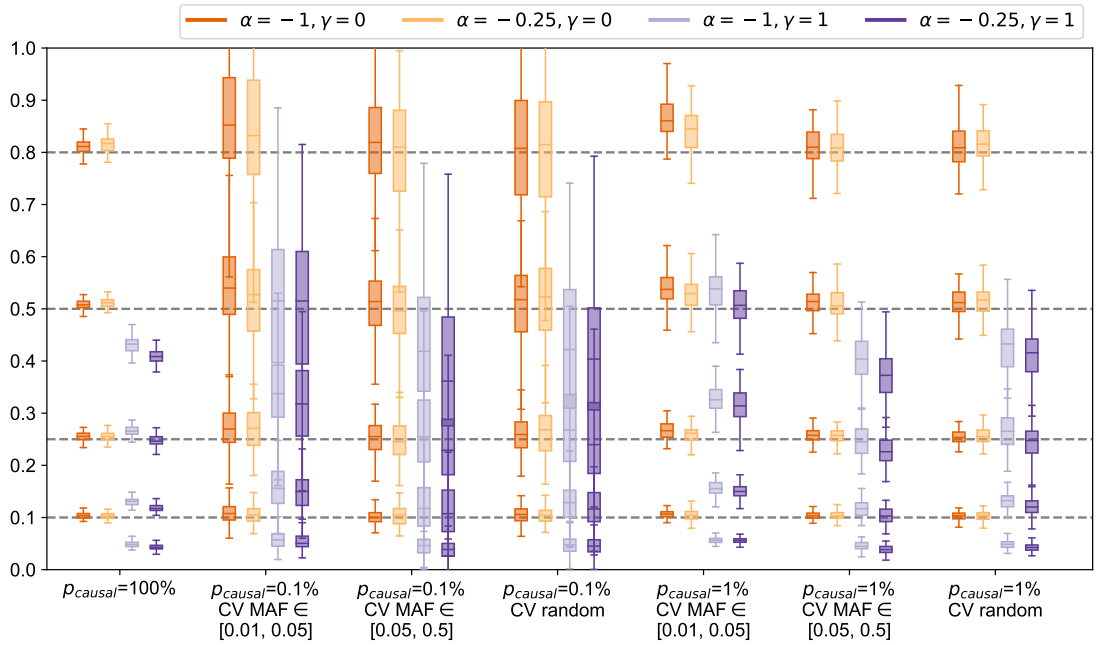


Figure S5: Distribution of h_g^2 estimates from S-LDSC (10 MAF bins) in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 337205$ individuals, $M = 593300$ array SNPs). Each boxplot represents the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

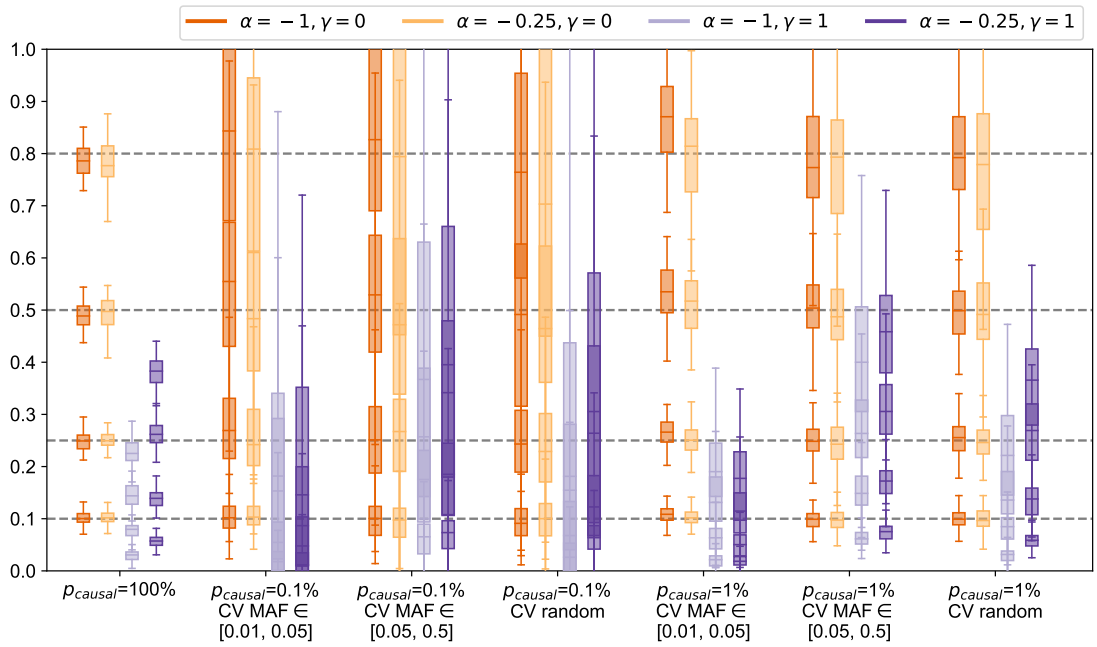


Figure S6: Distribution of h_g^2 estimates from S-LDSC (10 MAF bins + LLD) in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 337205$ individuals, $M = 593300$ array SNPs). Each boxplot shows the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

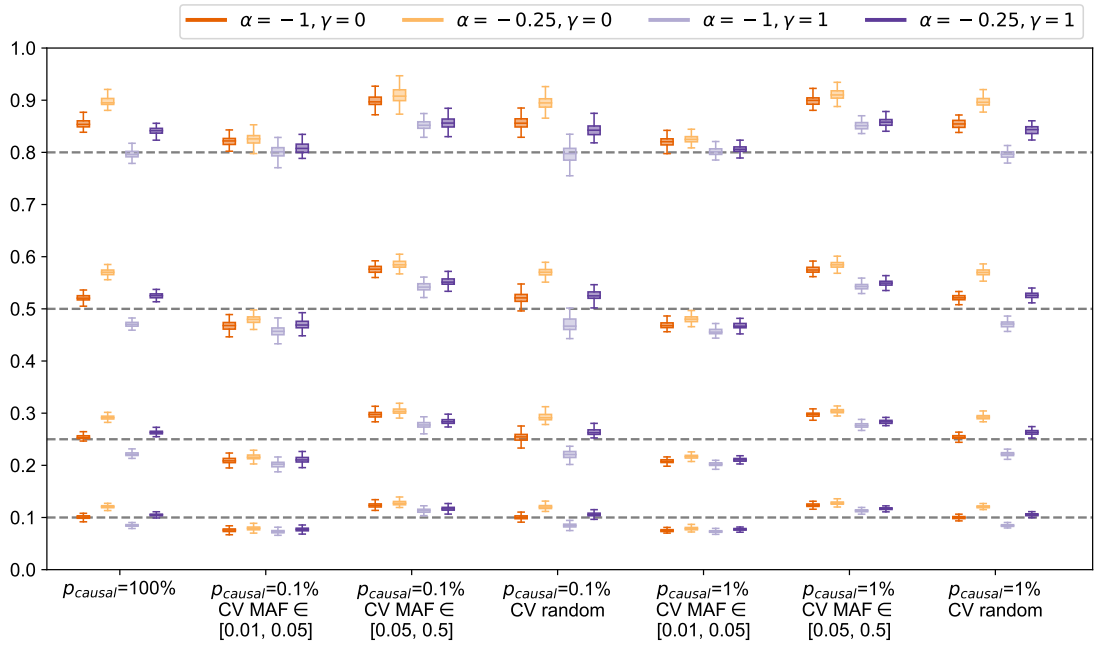


Figure S7: Distribution of h_g^2 estimates from SumHer in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 337205$ individuals, $M = 593300$ array SNPs). Each boxplot represents the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

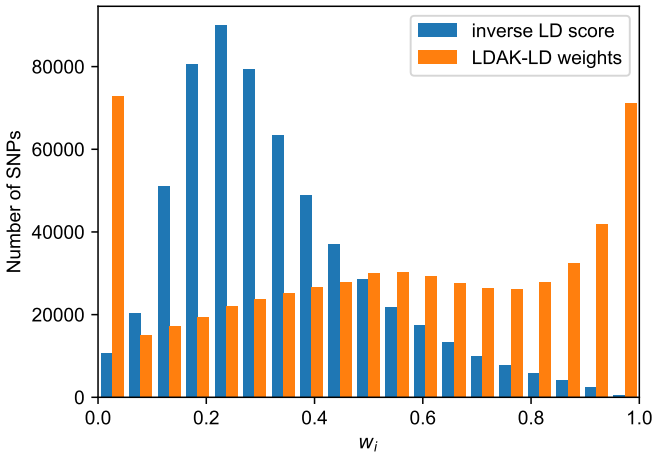


Figure S8: Histograms of LDAK weights and inverse LD score weights used in genome-wide simulations ($M = 593$K SNPs).

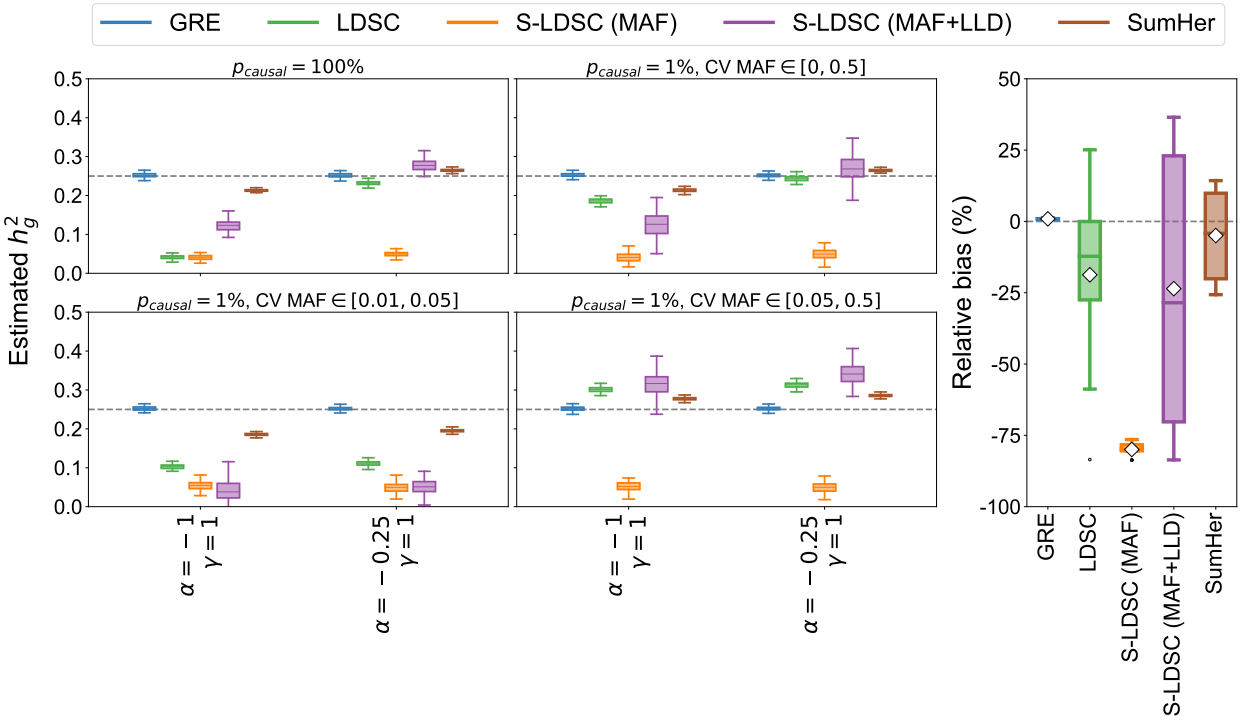


Figure S9: Comparison of methods across 14 MAF- and LD-score-dependent architectures ($N = 337205$ individuals, $M = 593300$ array SNPs, $h_g^2 = 0.25$). LD-score-dependent architectures are simulated by coupling the variance of each SNP to the inverse of its LD score (Methods). **Left:** Each boxplot represents 100 estimates under a single architecture; results are shown for $p_{\text{causal}} = 100\%$ and 1%. **Right:** Each boxplot represents the distribution of the relative bias across all 14 LD-score-dependent architectures. White diamonds mark the average of each distribution. All boxplot whiskers mark the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

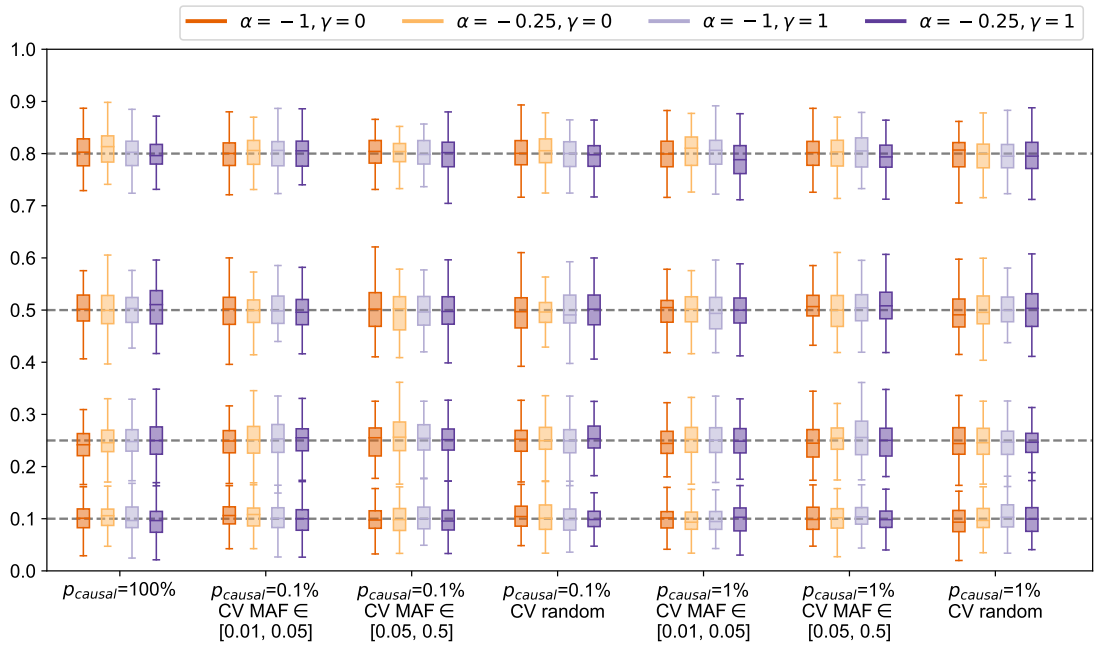


Figure S10: Distribution of h_g^2 estimates from GRE in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 8430$ individuals, $M = 14821$ array SNPs). Each boxplot represents the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

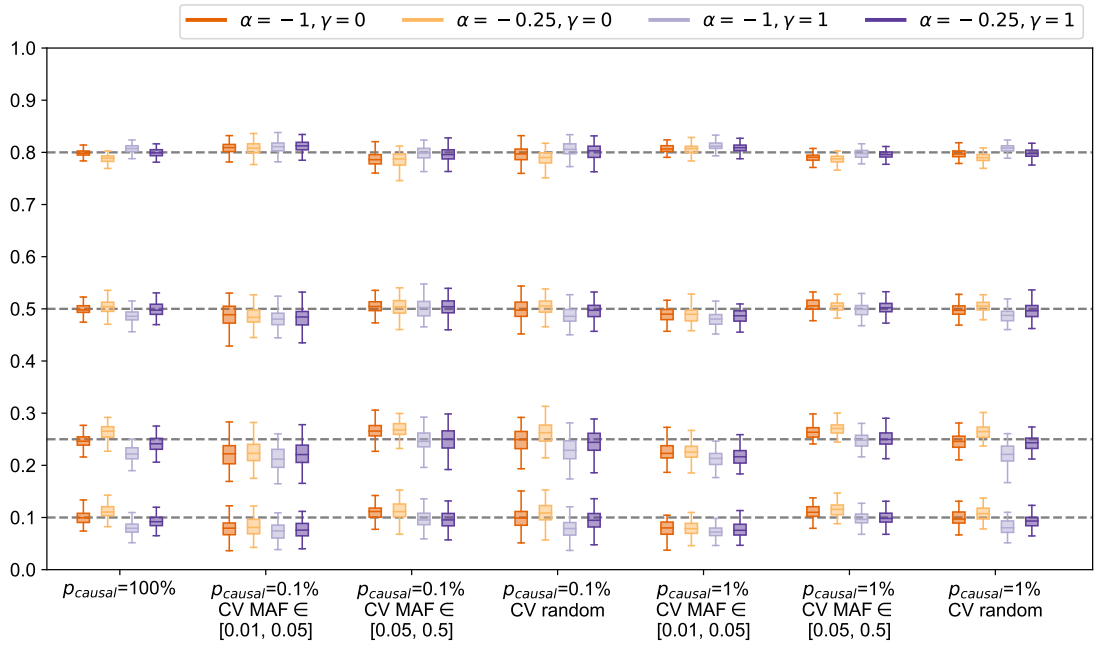


Figure S11: Distribution of h_g^2 estimates from single-component GREML in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 8430$ individuals, $M = 14821$ array SNPs). Each boxplot shows the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

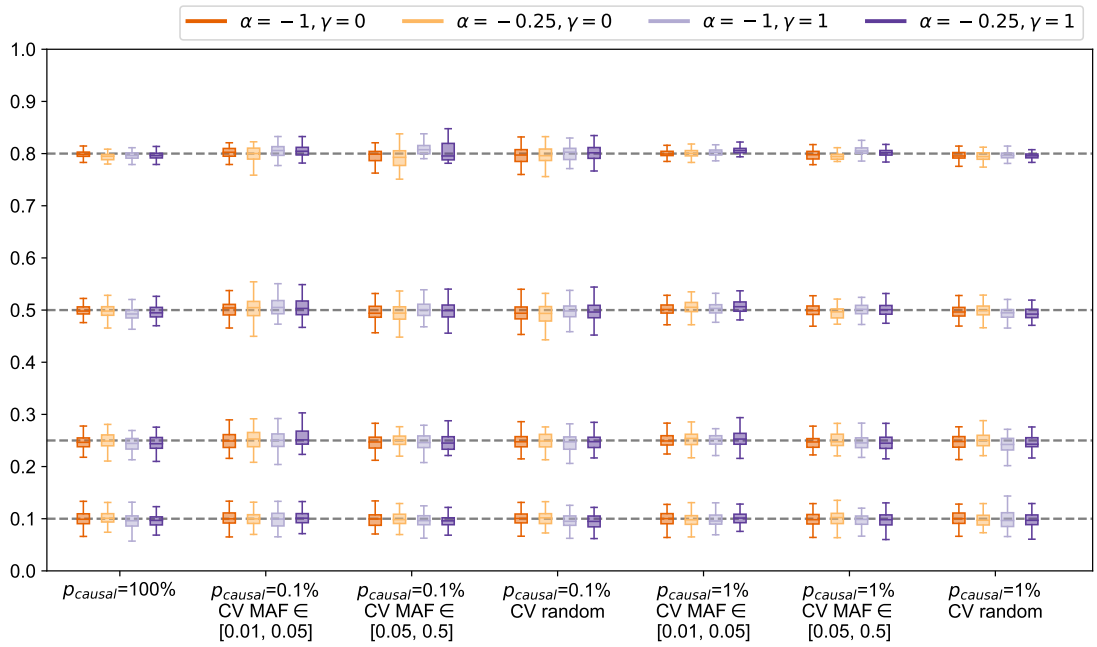


Figure S12: Distribution of h_g^2 estimates from GREML-LDMS-I in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 8430$ individuals, $M = 14281$ array SNPs). Each boxplot shows the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

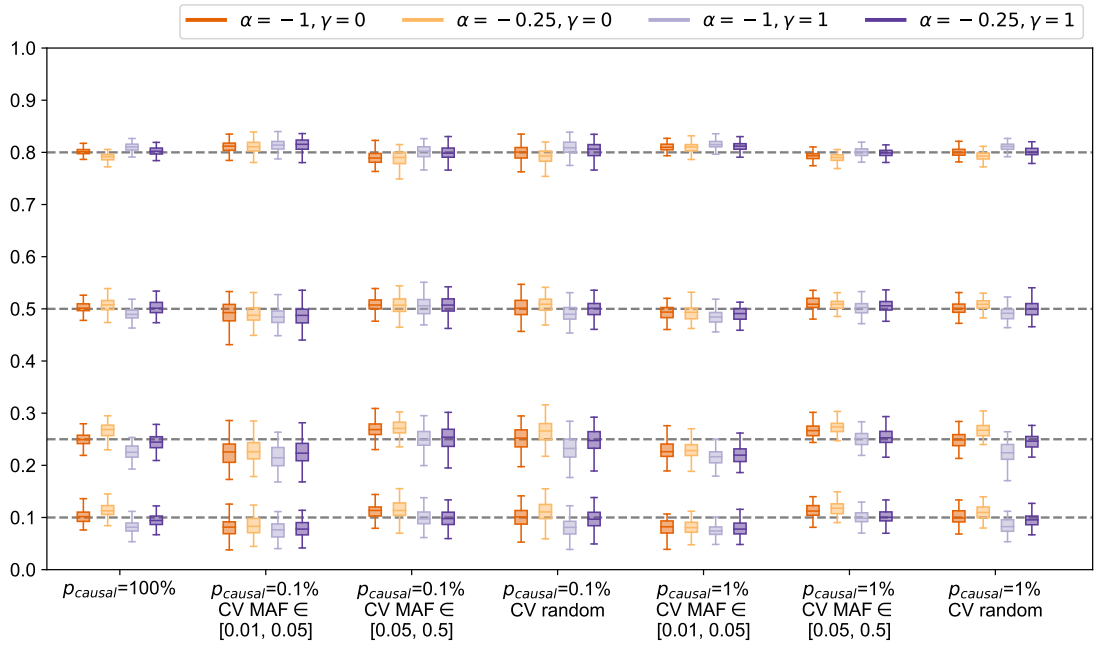


Figure S13: Distribution of h_g^2 estimates from BOLT-REML in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 8430$ individuals, $M = 14281$ array SNPs). Each boxplot represents the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

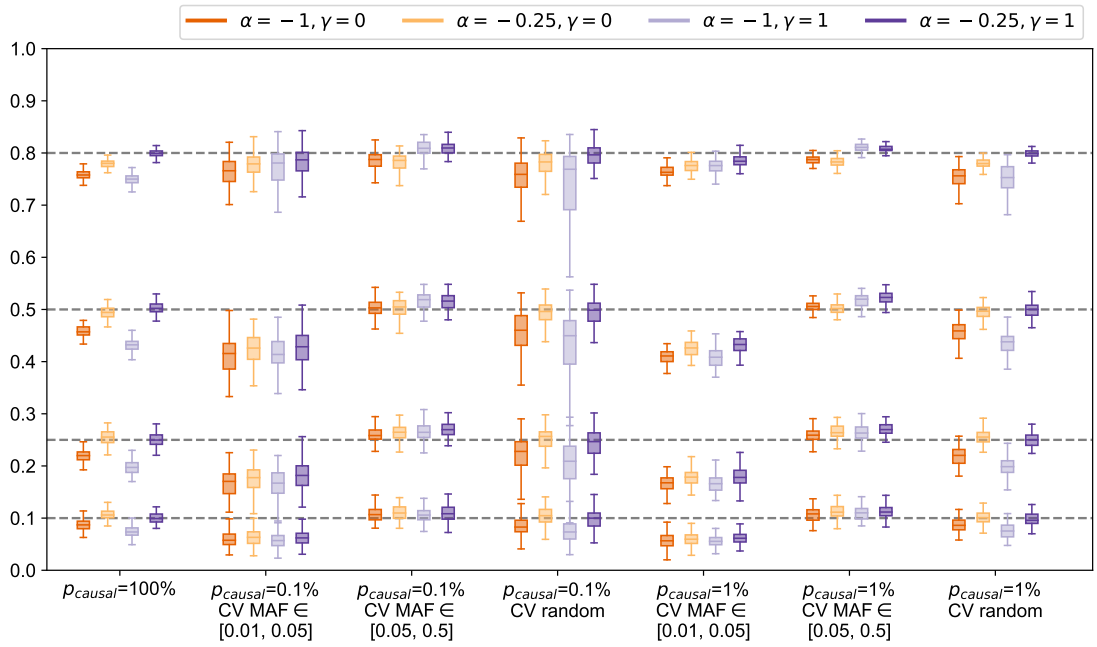


Figure S14: Distribution of h_g^2 estimates from LDAK in simulations across 112 LDAK-LD- and/or MAF-dependent architectures ($N = 8430$ individuals, $M = 14281$ array SNPs). Each boxplot represents the distribution of 100 estimates under a single architecture. Boxplot whiskers extend to the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

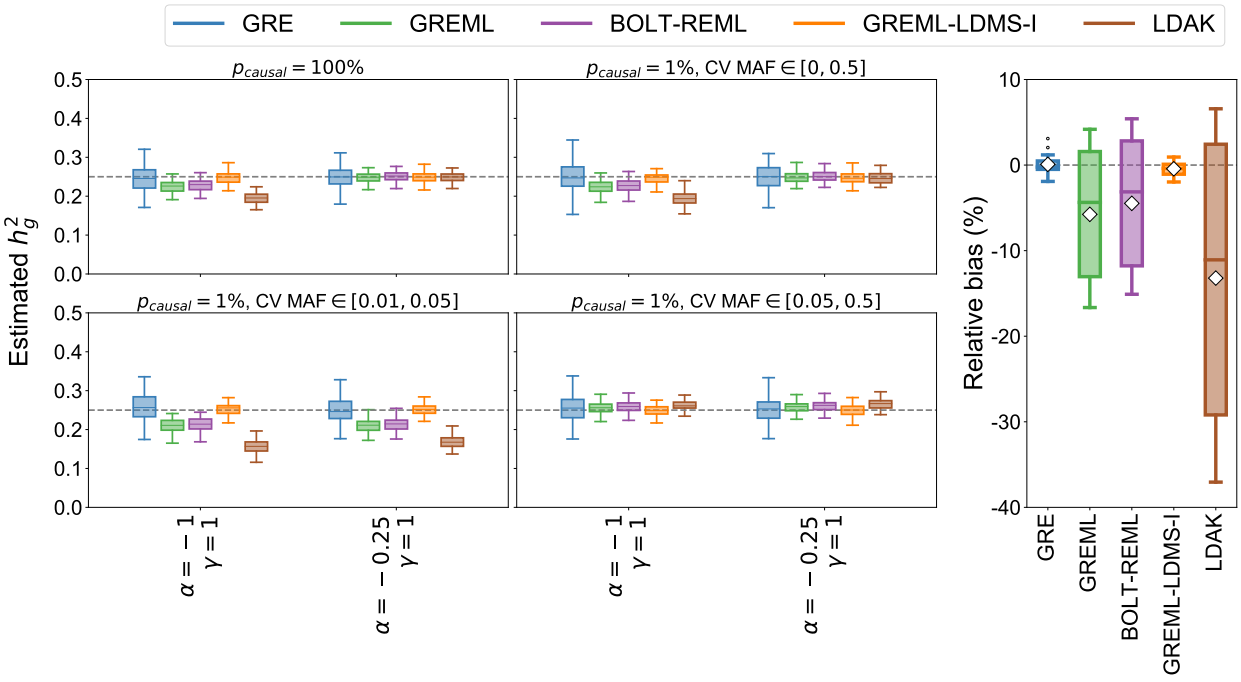


Figure S15: Comparison of methods across 14 MAF- and LD-score-dependent architectures ($N = 8430$ individuals, $M = 14281$ array SNPs). LD-score-dependent architectures are simulated by coupling the variance of each SNP to the inverse of its LD score (see Methods). **Left:** Each boxplot represents 100 estimates under a single architecture; results are shown for $p_{\text{causal}} = 100\%$ and 1%. **Right:** Each boxplot represents the distribution of the relative bias across all 14 LD-score-dependent architectures. White diamonds mark the average of each distribution. All boxplot whiskers mark the minimum and maximum estimates located within 1.5 × IQR from the first and third quartiles, respectively.

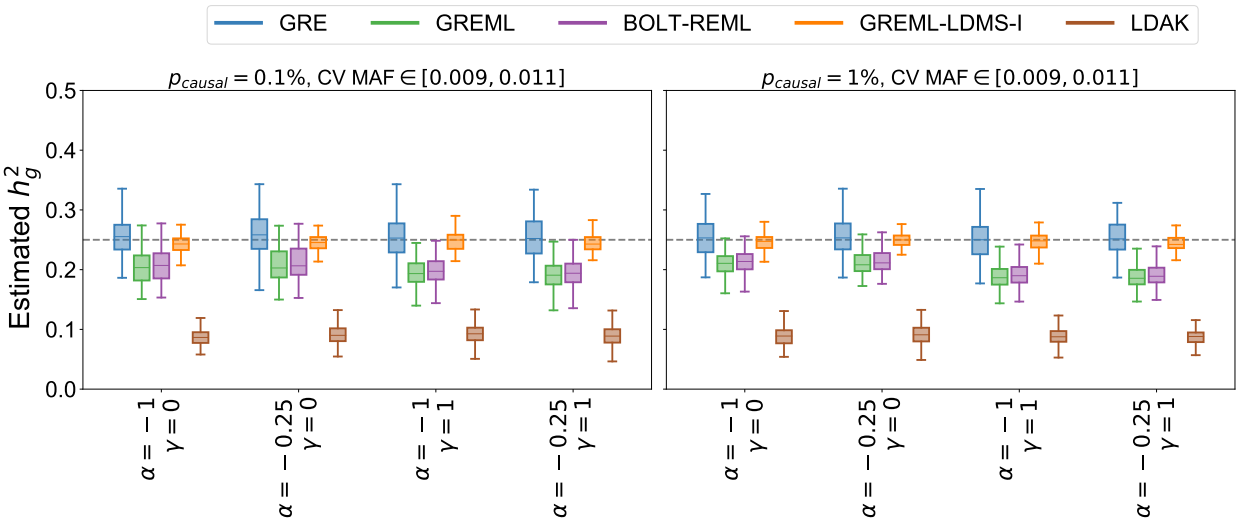


Figure S16: Comparison of GRE, GREML, BOLT-REML, GREML-LDMS-I, and LDAK in small-scale simulations ($N = 8430$ individuals, $M = 14821$ array SNPs) under MAF- and/or LDAK-LD-dependent architectures where all causal variants were drawn from the MAF range $[0.009, 0.011]$. Each boxplot contains estimates of h_g^2 from 100 simulations. The GRE estimator was computed with 22 chromosome-wide LD blocks. For GREML-LDMS-I, 8 GRMs were used (2 MAF bins $\times$ 4 LD quartiles). Boxplot whiskers mark the minimum and maximum estimates located within 1.5$\times$ IQR units from the first and third quartiles, respectively.

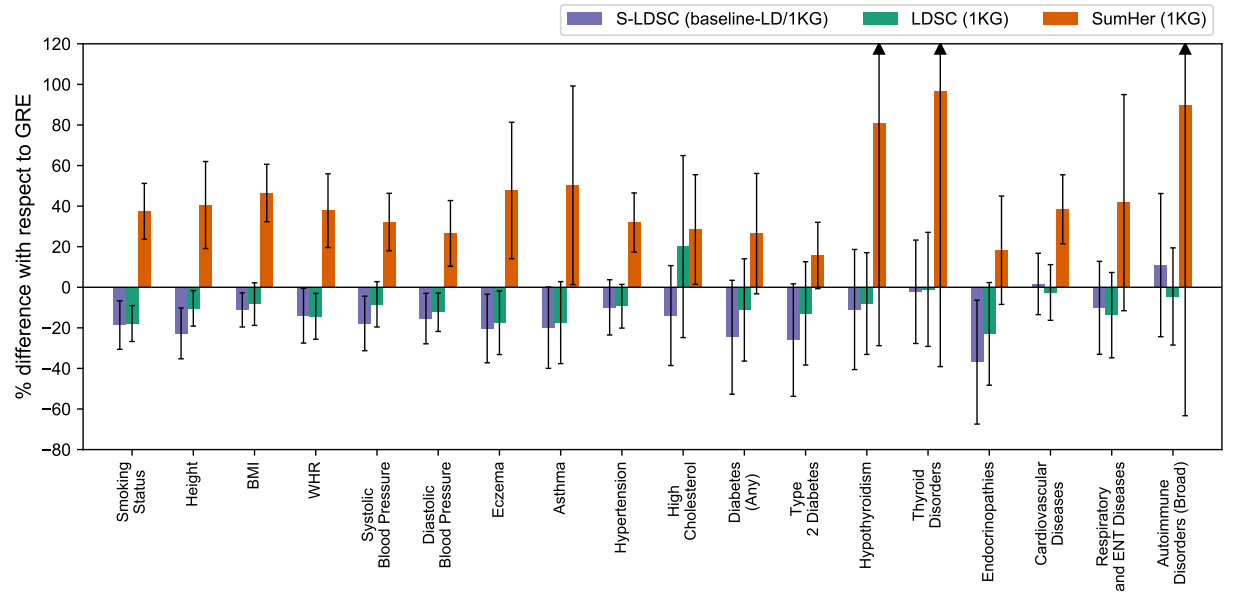


Figure S17: Percent difference of SNP-heritability estimates from LDSC (1KG), S-LDSC (baseline-LD/1KG), and SumHer (1KG) with respect to $\widehat{h}^2_{\text{GRE}}$ for 18 complex traits and diseases in the UK Biobank for which $\widehat{h}^2_{\text{GRE}} > 0.05$ ($N = 290$K unrelated British individuals and $M = 460$K typed SNPs; see Methods). Each bar represents the difference between the estimated SNP-heritability and $\widehat{h}^2_{\text{GRE}}$ as a percentage of $\widehat{h}^2_{\text{GRE}}$. Black bars mark ± 2 standard errors.

Supplementary Tables (see Excel file)

Table S1: Summary of $\widehat{h}^2_{\text{GRE}}$ in chromosome 22 simulations ($N = 337205$ individuals, $M = 9654$ array SNPs, 1 chromosome-wide LD block).

Table S2: Summary of $\widehat{h}^2_{\text{GRE}}$ in chromosome 22 simulations ($N = 337205$ individuals, $M = 9654$ array SNPs) where the number of blocks K was varied. Phenotypes were simulated for $h^2_g = 0.1$, $p_{causal} = 0.01$, $\alpha = -1$, and no LD weights.

Table S3: Summary of $\widehat{h}^2_{\text{GRE}}$ in genome-wide simulations ($N = 337205$ individuals, $M = 593300$ array SNPs, 22 chromosome-wide LD blocks).

Table S4: Comparison of analytical standard error of $\widehat{h}^2_{\text{GRE}}$ and empirical standard deviation of 100 $\widehat{h}^2_{\text{GRE}}$ estimates in simulations on chromosome 22.

Table S5: Comparison of analytical standard error of $\widehat{h}^2_{\text{GRE}}$ and empirical standard deviation of 100 $\widehat{h}^2_{\text{GRE}}$ estimates in genome-wide simulations.

Table S6: Summary of LDSC results in genome-wide simulations.

Table S7: Summary of SumHer results in genome-wide simulations.

Table S8: Boundaries of 10 MAF bins for S-LDSC (MAF).

Table S9: Summary of S-LDSC (MAF) results in genome-wide simulations.

Table S10: Summary of S-LDSC (MAF+LLD) results in genome-wide simulations.

Table S11: GRE, LDSC, SumHer, S-LDSC (MAF), S-LDSC (MAF+LLD) in genome-wide simulations with LD-score-dependent architectures.

Table S12: Summary of GREML in small-scale simulations.

Table S13: Summary of BOLT-REML in small-scale simulations.

Table S14: Summary of LDAK in small-scale simulations.

Table S15: Summary of GRE in small-scale simulations.

Table S16: Summary of GREML-LDMS-I in small-scale simulations.

Table S17: GRE, GREML, BOLT-REML, GREML-LDMS-I, LDAK in small-scale simulations with LD-score-dependent architectures.

Table S18: Estimates of h^2_g from the GRE approach, LDSC (in-sample), S-LDSC (baseline-LD/in-sample), SumHer (in-sample), LDSC (1KG), S-LDSC (baseline-LD/1KG) and SumHer (1KG) for 22 complex traits and diseases in the UK Biobank (N = 290K unrelated British individuals, M = 460K typed SNPs).