LDAK-KVIK performs fast and powerful mixed-model association analysis of quantitative and binary phenotypes

Jasper P. Hof[1] and Doug Speed[2,*]

1 Radboud University Medical Center, IQ Health Science Department, Nijmegen, The Netherlands

2 Aarhus University, Center for Quantitative Genetics and Genomics, Aarhus, Denmark

* Corresponding author: doug@qgg.au.dk

ABSTRACT

Mixed-model association analysis (MMAA) is the preferred tool for performing a genome-wide association study, because it enables robust control of type 1 error and increased statistical power to detect trait-associated loci. However, existing MMAA tools often suffer from long runtimes and high memory requirements. We present LDAK-KVIK, a novel MMAA tool for analyzing quantitative and binary phenotypes. Using simulated phenotypes, we show that LDAK-KVIK produces well-calibrated test statistics, both for homogeneous and heterogeneous datasets. LDAK-KVIK is computationally-efficient, requiring less than 20 CPU hours and 8Gb memory to analyse genome-wide data for 350k individuals. These demands are similar to those of REGENIE, one of the most efficient existing MMAA tools, and up to 30 times less than those of BOLT-LMM, currently the most powerful MMAA tool. When applied to real phenotypes, LDAK-KVIK has the highest power of all tools considered. For example, across 40 quantitative phenotypes from the UK Biobank (average sample size 349k), LDAK-KVIK finds 16% more significant loci than classical linear regression, whereas BOLT-LMM and REGENIE find 15% and 11% more, respectively. LDAK-KVIK can also perform gene-based tests; across the 40 quantitative UK Biobank phenotypes, LDAK-KVIK finds 18% more significant genes than the leading existing tool.

INTRODUCTION

Genome-wide association studies (GWAS) have greatly advanced our understanding of the genetic basis underlying complex diseases, offering valuable insights into disease mechanisms and potential therapeutic targets.[1,2] Since the first GWAS was carried out in 2005, sample sizes have regularly increased, such that studies involving over 100,000 individuals are now common.[3–7] Initially, GWAS relied on classical linear or logistic regression to test for association between single nucleotide polymorphisms (SNPs) and phenotypes. However, in recent years, mixed-model association analysis (MMAA) has become the method of choice.[8,9] MMAA can reduce false positives by accounting for cryptic relatedness, and increase true positives by factoring in the contributions of SNPs other than the one being tested.[10]

The most effective MMAA methods are two-step.[11–13] In Step 1, they construct leave-one-chromosome-out (LOCO) polygenic scores (PRS) and estimate λ, a test statistic scaling factor. In Step 2, they regress the phenotype on the SNPs, including the LOCO PRS as an offset, then scale the resulting test statistics. The computational demand of a two-step MMAA tool depends mainly on its algorithmic design, whereas its power depends primarily on the accuracy of its LOCO PRS.[14,15] For example, REGENIE[13] tends to be faster than BOLT-LMM[12] because its algorithm for estimating SNP effect sizes is block-based instead of genome-wide. However, BOLT-LMM tends to detect more significant associations than REGENIE because it usually constructs more accurate PRS.

In this paper, we introduce LDAK-KVIK, a novel two-step MMAA tool for analyzing quantitative and binary phenotypes. We first use simulated data to show that LDAK-KVIK controls type 1 error for both homogeneous and heterogeneous data sets, and also when analyzing highly imbalanced phenotypes (e.g., diseases with very few cases). We then apply LDAK-KVIK to large-scale data from the UK Biobank.[16,17] When used for single-SNP association analysis, LDAK-KVIK finds more significant associations than BOLT-LMM,[12] REGENIE,[13] fastGWA[18] and GCTA-LOCO,[10] four of the leading existing MMAA tools. Meanwhile, when used for gene-based association analysis, LDAK-KVIK finds more significant associations than LDAK-GBAT,[19] the leading existing tool for gene-based association testing. LDAK-KVIK is available in our software package LDAK (www.dougspeed.com).

RESULTS

Overview of LDAK-KVIK.

A detailed description of LDAK-KVIK is provided in the **Online Methods** and **Supplementary Notes 1-5**. Here we highlight its key features.

LDAK-KVIK is a computationally efficient MMAA tool. Firstly, it never needs to store genotypes for more than 512 SNPs, and thus has very low memory demands. Secondly, we have developed a chunk-based variational Bayes solver (illustrated in **Supplementary Figure 1**) that requires 5-20 times fewer updates than conventional variational Bayes solvers (**Supplementary Figure 2**).[12,20] Our solver not only estimates SNP effect sizes (required when constructing the Step 1 PRS), but also efficiently computes terms of the form $V^{-1}A$, where V is phenotypic variance matrix and A is a vector of genotypes or phenotypes (required when estimating heritability). Thirdly, we have developed a fast empirical algorithm for computing the saddlepoint approximation (SPA).[21]

LDAK-KVIK increases detection power, relative to existing MMAA tools, by using more realistic models for how SNP effect sizes vary across the genome.[22,23] Firstly, LDAK-KVIK models how per-SNP heritability

depends on minor allele frequency (MAF), whereas existing MMAA tools typically assume that per-SNP heritability is constant.[24,25] Secondly, LDAK-KVIK uses an elastic net prior distribution for SNP effect sizes (i.e., a mixture of a normal and a Laplace distribution), whereas existing MMAA tools generally restrict to mixtures of normal distributions.[26]

We have developed a novel test for structure, based on the average pairwise correlation between 512 SNPs randomly picked from the genome (the correlation is calculated after regressing out covariates). The outcome of this test determines how LDAK-KVIK calculates the test statistic scaling factor λ: if the test finds weak structure (specifically, estimates that the maximum average inflation of $\chi^2(1)$ test statistics is below 0.1), LDAK-KVIK sets $\lambda=1$, which is the correct value when analyzing homogeneous data; if the test finds strong structure, LDAK-KVIK replaces the elastic net prior distribution for SNP effect sizes with an infinitesimal prior, and estimates λ using the Grammar-Gamma Formula.[11]

In addition to testing SNPs individually for association with the phenotype, LDAK-KVIK can also perform gene-based tests. This is achieved by providing the results of the single-SNP analysis to our existing software LDAK-GBAT.[19] To avoid confusion, we refer to our new tool as LDAK-KVIK when testing SNPs for association, and as LDAK-KVIK-GBAT when testing genes for association.

Data.

We use genotype and phenotype data from the UK Biobank (obtained via application 21432).[16,17] In total, we construct four datasets: the "white dataset" contains 367,981 white British individuals, the "homogeneous dataset" contains 63,000 unrelated, white British individuals, the "twins dataset" contains 63,000 twins (generated by duplicating the genotypes of 31,500 individuals from the homogeneous dataset), while the "multi-ancestry dataset" contains 60,019 individuals of various ethnic backgrounds (including approximately 35k white, 5k Indian, 4k Caribbean and 3k African). After reducing to autosomal, biallelic SNPs with MAF>0.001, the white, homogeneous and twins datasets each contain 690,264 SNPs, while the multi-ancestry dataset contains 471,760 SNPs.

We first analyze simulated phenotypes. When generating these, we randomly select causal SNPs from the start of each chromosome then use the chromosome ends as null SNPs for evaluating type 1 error (illustrated in **Supplementary Figure 3**). We subsequently analyze 40 quantitative and 20 binary phenotypes from the UK Biobank. The quantitative phenotypes are height, body mass index, 20 biochemistry measurements (e.g., cholesterol, c-reactive protein, and urate), 16 blood measurements (e.g., haemoglobin concentration, lymphocyte percentage and platelet count) and two urine measurements (levels of creatinine and sodium). The binary

phenotypes are defined based on ICD-10 codes[27] (e.g., hypertension, obesity, and asthma) and have prevalences ranging from 0.02% to 29%. Additional details of the data are provided in **Supplementary Note 6**.

For the analyses below, we always include ten principal components as covariates; when analyzing real phenotypes, we additionally include age, sex, age^2 and age*sex. As explained above, LDAK-KVIK includes a test for structure. For the white, homogeneous and multi-ancestry datasets, LDAK-KVIK determines the structure is weak (the estimates of the maximum average inflation of test statistics are approximately 0.04, 0.002 and 0.03, respectively). By contrast, LDAK-KVIK determines the twins dataset has strong structure (the estimated maximum average inflation is 1.0).

Existing tools.

For single-SNP association testing, we compare LDAK-KVIK with classical linear and logistic regression, and with four existing MMAA tools: BOLT-LMM,[12] REGENIE,[13] fastGWA[18] and GCTA-LOCO[10] (a summary of each tool is provided in **Supplementary Note 7**). Note that REGENIE and fastGWA are designed for both quantitative and binary phenotypes, whereas BOLT-LMM and GCTA-LOCO are designed for quantitative phenotypes. For gene-based association testing, we compare LDAK-KVIK-GBAT with our existing tool LDAK-GBAT,[19] which we previously found to be consistently more powerful than five alternative tools (including MAGMA[28] and FastBAT[29]).

Heritability models.

We consider heritability models of the form $E\left[h_j^2\right] \propto \left[f_j\left(1-f_j\right)\right]^{1+\alpha}$, where $E\left[h_j^2\right]$ is the expected heritability contributed by SNP j, and f_j is its MAF. When $\alpha=-1$ all SNPs are expected to contribute equal heritability (this is the most commonly-used heritability model in human statistical genetics[24,30]). When $\alpha=-0.25$ SNPs with higher MAF are expected to contribute more heritability than SNPs with lower MAF (this model better reflects what is observed for human complex traits [22,31]). All four existing MMAA tools assume $\alpha=-1$ throughout their operations (e.g., REGENIE assumes $\alpha=-1$ when constructing the Step 1 PRS, while BOLT-LMM assumes $\alpha=-1$ both when constructing the Step 1 PRS and when estimating λ). By contrast, LDAK-KVIK estimates α from the data.

LDAK-KVIK controls type 1 error.

We simulate quantitative and binary phenotypes for the homogeneous, twins and mixed-ancestry datasets. For each dataset, we consider 12 different scenarios, obtained by varying the heritability (0.2 or 0.5), the number of

causal SNPs (5k or 20k), and for binary phenotypes, also the prevalence (10% or 1%). When generating causal SNP effect sizes, we assume $\alpha=-1$. We perform single-SNP analysis using LDAK-KVIK, then measure type 1 error based on the average $\chi^2(1)$ test statistic of null SNPs, and the proportions of null SNPs with p-values below 0.05, 0.001 and $5x10^{-5}$. **Supplementary Figure 4** shows that LDAK-KVIK has well-controlled type 1 error for all 12 scenarios across all three datasets. This remains the case when we instead perform a gene-based analysis (**Supplementary Figure 5**).

For comparison, **Supplementary Figures 6-9** provide results from analyzing the simulated phenotypes using BOLT-LMM, REGENIE, fastGWA and GCTA-LOCO. In general, each of the four tools controls type 1 error for all scenarios considered, except that REGENIE tends to produce inflated test statistics for common binary phenotypes (e.g., when analyzing the 40 binary phenotypes with prevalence 10% for the homogeneous dataset, the average $\chi^2(1)$ test statistic is 1.03).

LDAK-KVIK is computationally efficient.

Table 1 compares the runtime and memory requirements of LDAK-KVIK with existing MMAA tools when testing for association 690k directly-genotyped SNPs, using either the homogeneous or white datasets (63k and 368k individuals, respectively). For quantitative phenotypes, LDAK-KVIK and REGENIE have similar demands, typically requiring between 10 and 20 CPU hours and less than 16 Gb memory to analyze 368k individuals. fastGWA is the most computationally-efficient tool, but only if we ignore the one-off cost for constructing the genomic relationship matrix (which for 368k individuals, took 763 CPU hours). By contrast, BOLT-LMM and GCTA-LOCO are substantially more demanding. For example, BOLT-LMM takes approximately 450 CPU hours to analyse 368k individuals, while GCTA-LOCO takes over 400 CPU hours just to analyse 63k individuals (and thus it is not feasible for us to apply GCTA-LOCO to 368k individuals). The computational demands of LDAK-KVIK are similar when we switch from quantitative to binary phenotypes (which is also the case for REGENIE and fastGWA).

Supplementary Figure 10 and **Supplementary Table 1** provide further details of the LDAK-KVIK runtimes and results from additional analyses. For example, we see that for quantitative (binary) phenotypes, approximately 95% (85%) of the total time is spent in Step 1. This means that if we increase the number of association analysis SNPs from 690k to 10M, mimicking the situation where we perform a GWAS using imputed data, the impact is relatively modest (e.g., for quantitative phenotypes, the average time to analyse 368k individuals increases from 16 CPU hours to 26 CPU hours). Meanwhile, we see that the runtime of LDAK-KVIK-GBAT is approximately 15% longer runtime than that of LDAK-GBAT, due to the extra time required to perform the gene-based association analysis.

Note that the runtimes in **Table 1** correspond to analyzing phenotypes individually. However, both LDAK-KVIK and REGENIE are able to analyze multiple phenotypes simultaneously, which generally leads to a lower per-phenotype runtime. For example, **Supplementary Table 2** shows that when analyzing five and ten quantitative phenotypes at once, the per-phenotype runtimes of LDAK-KVIK are reduced by 62% and 72%, respectively.

		63k Individuals		368k Individuals	
	MMAA Tool	**CPU Hours**	**Memory (Gb)**	**CPU Hours**	**Memory (Gb)**
Quantitative Phenotypes	BOLT-LMM	18	12	452	61
	REGENIE	2.4	3	11	16
	fastGWA	0.2 (24)	1 (1)	0.3 (763)	3 (3)
	GCTA-LOCO	488	167	Not Feasible	
	LDAK-KVIK	2.5	2	16	6
Binary Phenotypes	REGENIE	3.3	3	15	16
	fastGWA	0.3 (24)	1 (46)	1.8 (764)	3 (3)
	LDAK-KVIK	2.7	2	19	6

Table 1. Computational requirements of MMAA tools. *We perform GWAS for five quantitative phenotypes (glucose, glycated haemoglobin, haemoglobin concentration, height and high-density lipoprotein) and five binary phenotypes (asthma, atrial fibrillation, chronic ischaemic heart disease, dental caries and residual haemorrhoidal skin tags). Each GWAS analyzes 690k SNPs using either the homogeneous or white dataset (63k and 368k individuals, respectively). All analyses were performed on AMD EPYC Genoa 9654 CPU processors, using either 4 CPUs (LDAK-KVIK, REGENIE and fastGWA) or 12 CPUs (BOLT-LMM and GCTA-LOCO). Values report the mean CPU hours and memory usage across either the five quantitative or five binary phenotypes. We report two sets of values for fastGWA, depending on whether we exclude or include the computation of the genomic relatedness matrix (which only needs to be done once per dataset). It was not feasible for us to apply GCTA-LOCO to 368k samples (we estimate it would require over 500Gb memory and over 10,000 CPU hours).*

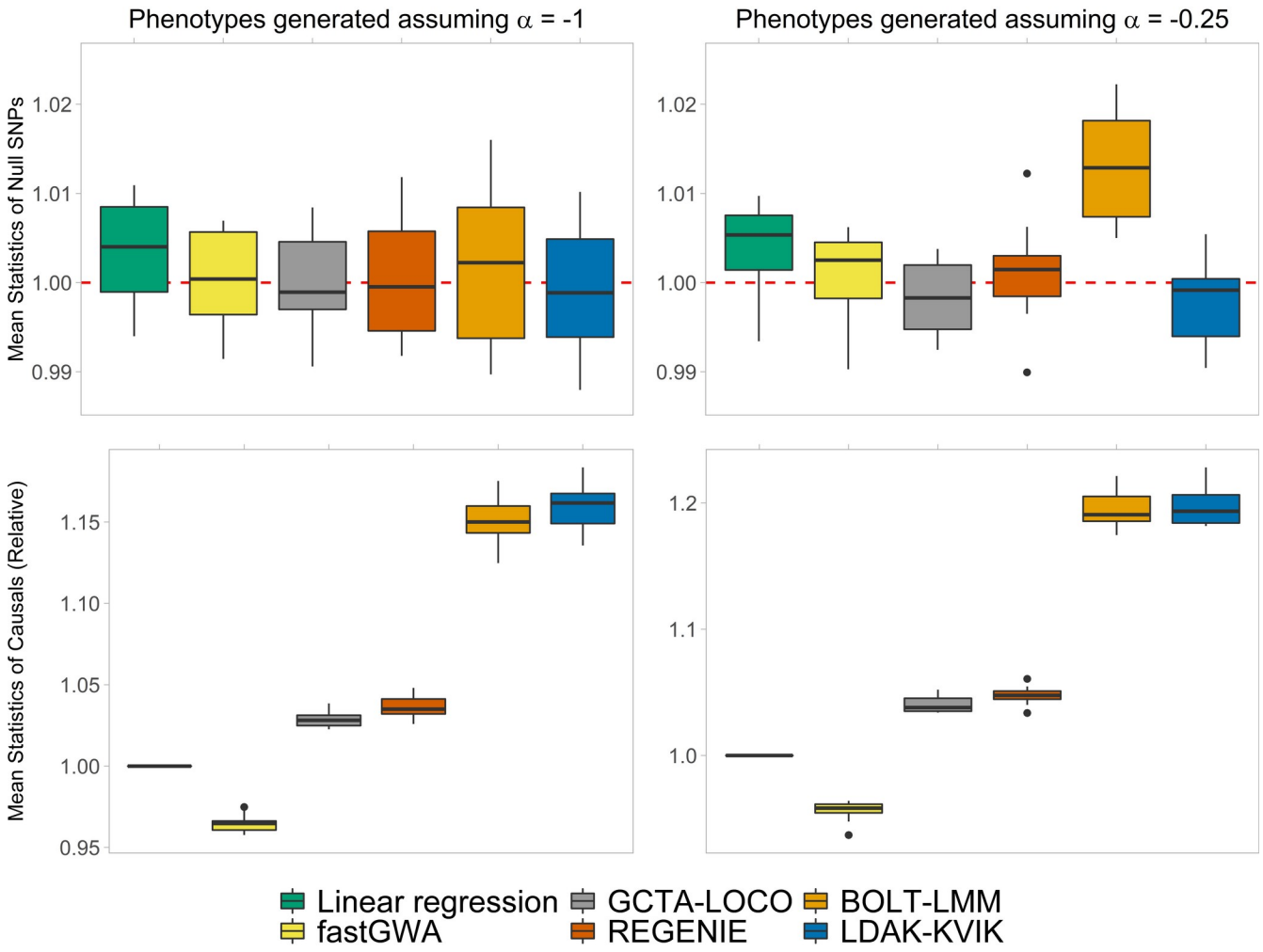

Figure 1. Type 1 error and power of MMAA tools when analyzing homogeneous data. *We generate quantitative phenotypes for 63k homogeneous individuals. Each phenotype has heritability 0.5 and 5k causal SNPs, with causal SNP effect sizes sampled assuming either* $\alpha=-1$ *(left) or* $\alpha=-0.25$ *(right). We analyze the phenotypes using linear regression, BOLT-LMM, REGENIE, fastGWA, GCTA-LOCO and LDAK-KVIK. Panels report the mean* $\chi^2(1)$ *test statistic of null SNPs (top) and causal SNPs (bottom) for each of ten replicates. The three horizontal lines in each box mark the medians and inter-quartile ranges. Note that in the top panels, the horizontal dashed lines mark the expected value if a tool is well-calibrated, while in the bottom panels, values are reported relative to the results from linear regression.*

LDAK-KVIK is powerful

Figure 1 and **Supplementary Figure 11** compare the power of MMAA tools when performing single-SNP analysis of the simulated phenotypes for the homogeneous dataset. For the quantitative phenotypes, LDAK-KVIK and BOLT-LMM are consistently the two most powerful MMAA tools, substantially ahead of REGENIE and GCTA-LOCO, while fastGWA generally has lowest power. For the binary phenotypes, the three MMAA tools generally have very similar power, reflecting that it is challenging to construct accurate PRS for phenotypes with low (observed-scale) heritability. **Supplementary Figure 12** shows that for gene-based analysis, LDAK-KVIK-GBAT tends to be substantially more powerful than LDAK-GBAT for quantitative phenotypes, while the two tools have similar power for binary phenotypes.

The importance of modeling the relationship between per-SNP heritability and MAF.

So far, we have only considered phenotypes generated assuming $\alpha=-1$ (all causal SNPs are expected to contribute equal heritability). We now also analyze phenotypes generated assuming $\alpha=-0.25$ (i.e., the per-SNP heritability of causal SNPs tends to increase with MAF). Changing from $\alpha=-1$ to $\alpha=-0.25$ has two main consequences. Firstly, we note that BOLT-LMM has inflated type 1 error, which is a consequence of it assuming $\alpha=-1$ when computing the test statistic scaling factor λ (**Figure 1**). Secondly, we find that the power of LDAK-KVIK generally increases relative to the other tools, reflecting that it can reliably infer the true value of α, and use this to construct more accurate Step 1 PRS.

Given the inflation observed for BOLT-LMM, we now perform additional simulations, where we switch from the homogeneous to the white dataset (i.e., increase the sample size from 63k to 368k). **Figure 2** shows that BOLT-LMM can have substantially-inflated type 1 error when applied to large datasets. For example, across ten quantitative phenotypes with heritability 0.5 and 5k causal SNPs, generated assuming $\alpha=-0.25$, the mean $\chi^2(1)$ test statistic of null SNPs is 1.09 (i.e., 9% higher than the expected value if BOLT-LMM was well-calibrated). This inflation is due to BOLT-LMM overestimating λ. For example, for the ten phenotypes just mentioned, its estimates of λ range from 1.08 to 1.11, despite the true value being close to one (because the dataset is approximately homogeneous). To avoid this inflation, we create BOLT-LMM-Unscaled, whose results match those from BOLT-LMM, except that we force $\lambda=1$. **Figure 2** shows that BOLT-LMM-Unscaled has well-controlled type 1 error when applied to the white dataset.

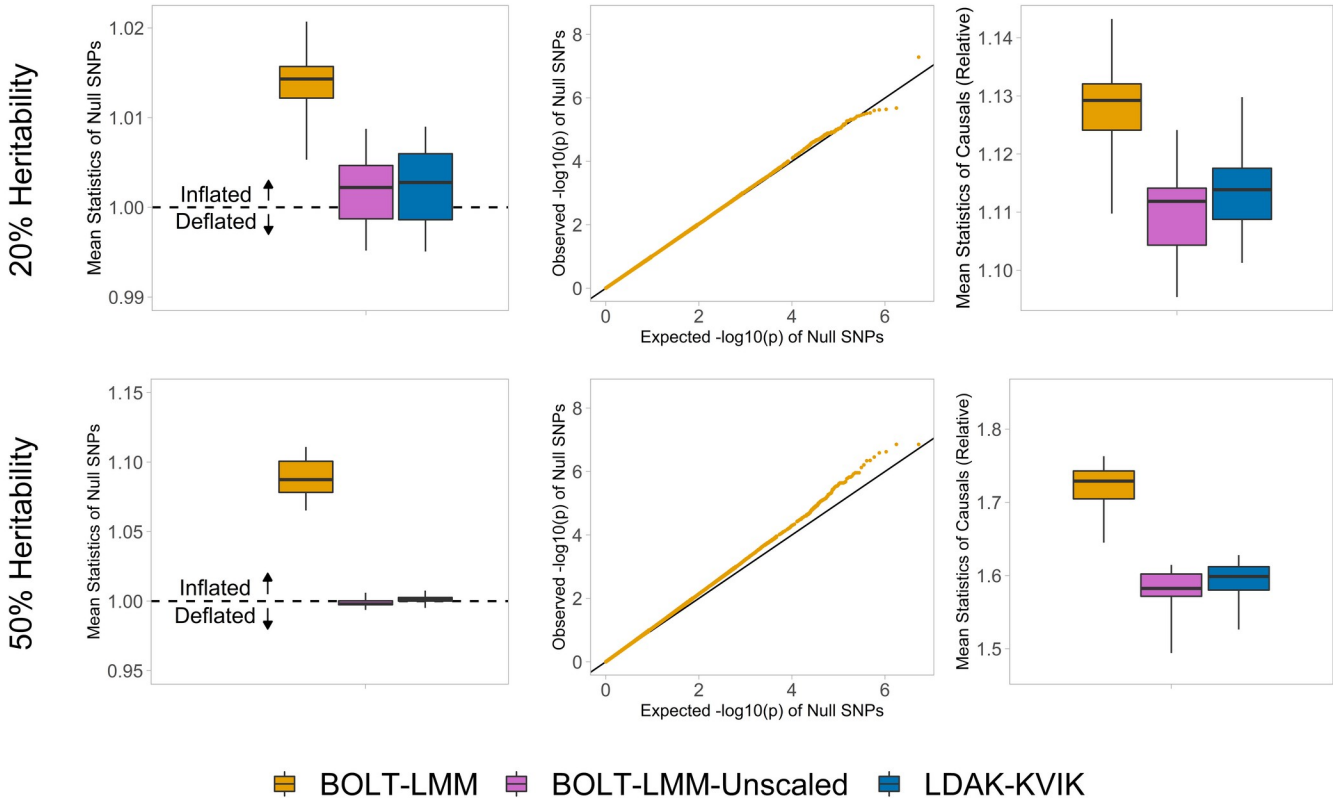

Figure 2. BOLT-LMM can produce inflated test statistics. *We generate quantitative phenotypes for 368k white individuals. Each phenotype has heritability 0.2 or 0.5 and 5k causal SNPs, with causal SNP effect sizes sampled assuming* $\alpha = -0.25$ *. We analyze the phenotypes using BOLT-LMM, BOLT-LMM-Unscaled and LDAK-KVIK. The left (right) panels reports the mean test statistic of null (causal) SNPs for each of ten replicates. The middle panels provide QQ-plots for BOLT-LMM (constructed using only null SNPs and combined across all ten replicates). In the left and right panels, the three horizontal lines in each box report the medians and inter-quartile ranges. Note that in the left panels, the horizontal dashed lines mark the expected value if a tool is well-calibrated, while in the right panels, values are reported relative to the results from linear regression.*

Single-SNP analysis of UK Biobank phenotypes.

Supplementary Figure 13 reports estimates of α for the 40 quantitative UK Biobank phenotypes. The mean estimate is -0.23 (s.d. 0.005), with α significantly greater than -1 for all phenotypes (P<0.01). These results, combined with those in **Figure 2**, indicate that BOLT-LMM will tend to produce inflated test statistics when applied to the quantitative UK Biobank phenotypes, and so for the following analyses, we replace BOLT-LMM with BOLT-LMM-Unscaled.

Figure 3 compares the number of associations found by Bolt-LMM-Unscaled, REGENIE, fastGWA and LDAK-KVIK when analyzing the 40 quantitative phenotypes. We find that LDAK-KVIK identifies 15.6% more independent, genome-wide significant ($P<5x10^{-8}$) SNPs than linear regression, which is slightly higher than BOLT-LMM-Unscaled (15.1%), and substantially higher than REGENIE (11.3%) and fastGWA (-1.9%). **Supplementary Figure 14** shows that ranking of MMAA tools is the same if we instead measure power based on the mean $\chi^2(1)$ test statistic of SNPs that are genome-wide significant from linear regression.

Supplementary Figure 15 shows that when analyzing the 20 binary phenotypes, all MMAA tools find similar numbers of associations. For example, REGENIE and LDAK-KVIK in total find 676 and 686 independent, genome-wide significant SNPs, respectively, which is only slightly higher than logistic regression (667 SNPs). These results are consistent with those when analyzing the simulated phenotypes, and again reflect the difficulty of constructing accurate PRS for binary phenotypes.

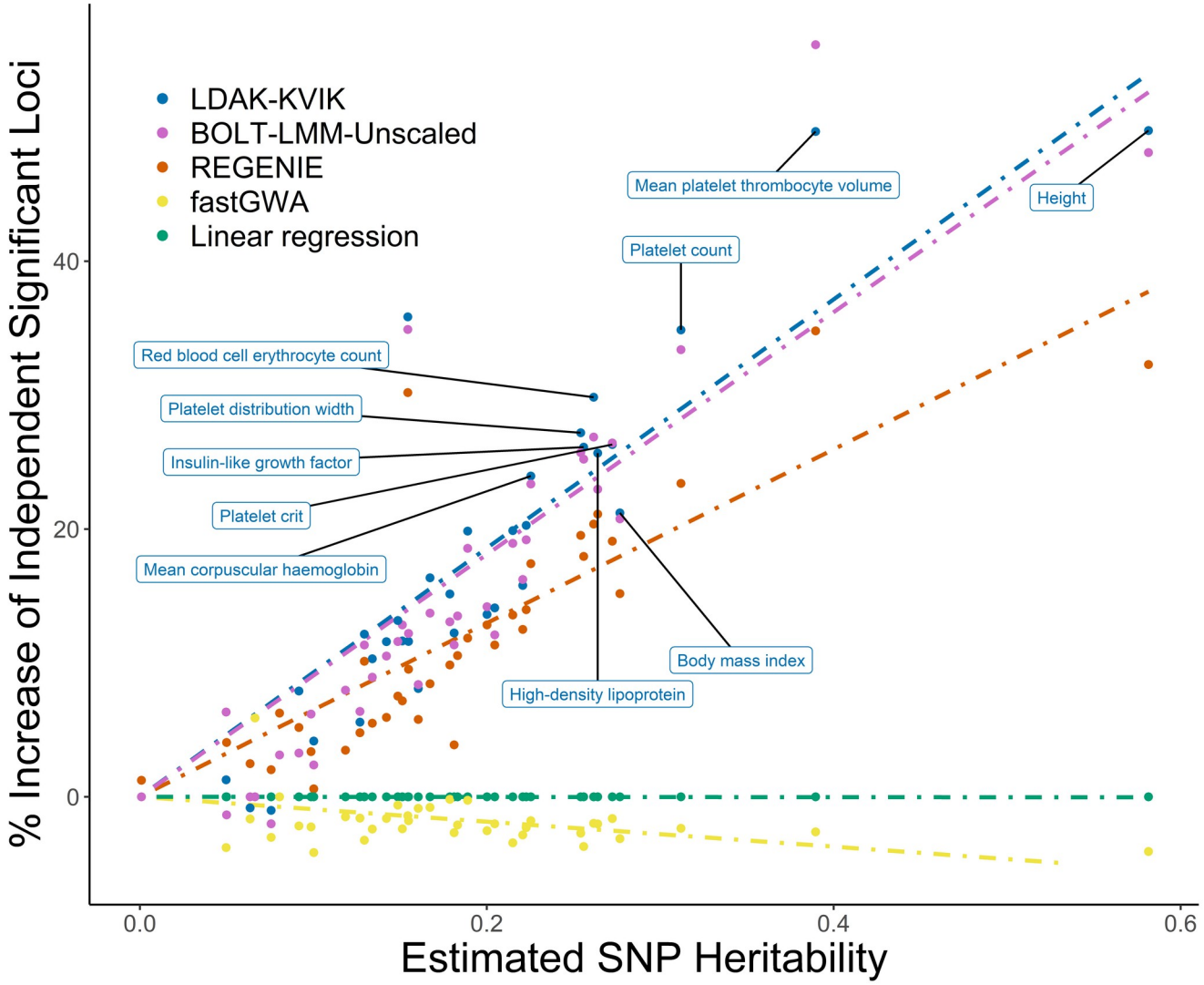

Figure 3. Performance of MMAA tools when analyzing 40 quantitative phenotypes from UK Biobank. *We analyze each phenotype using BOLT-LMM-Unscaled, REGENIE, fastGWA and LDAK-KVIK, then count the number of independent, genome-wide significant loci (SNPs with $P<5x10^{-8}$, filtered so that no pair within 1Mb has squared correlation above 0.1). Points compare the number of loci found by each tool, relative to the results from linear regression, with the estimated SNP heritability (obtained using our software SumHer[25]). The ten phenotypes with highest SNP heritability are named, while the dashed lines are obtained by regressing the relative numbers of loci found by each tool on the estimates of SNP heritability.*

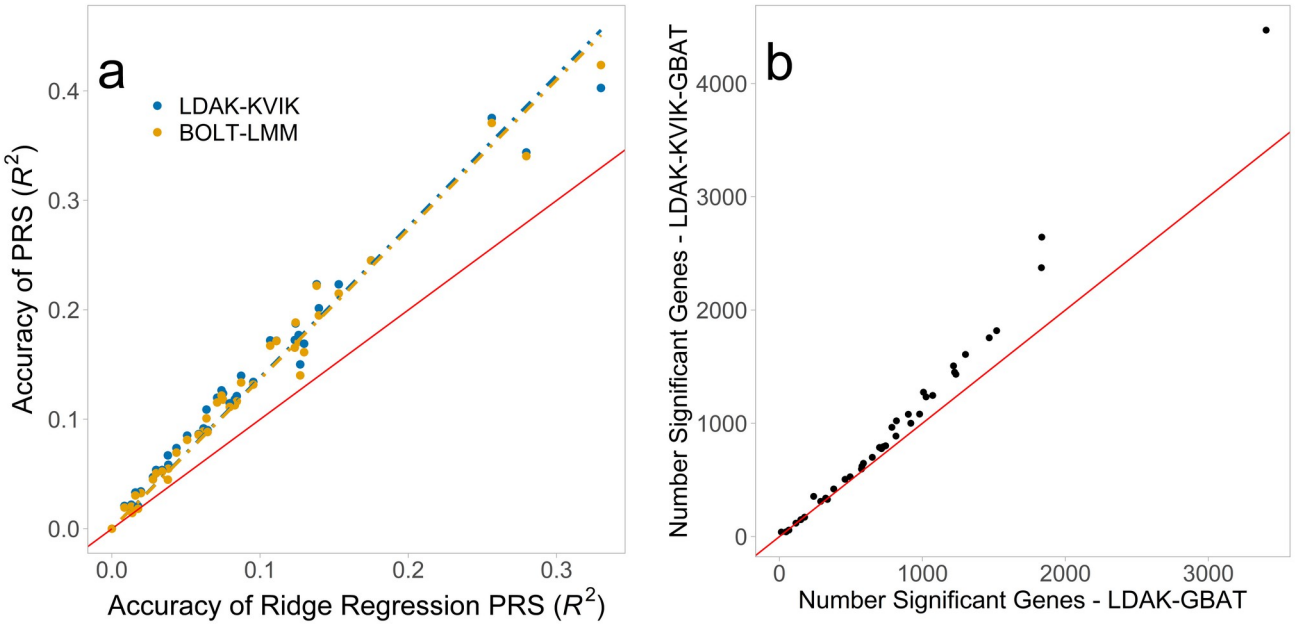

Figure 4. Accuracy of Step 1 PRS and power of LDAK-KVIK-GBAT. a, *We applied LDAK-KVIK and BOLT-LMM to the 40 quantitative phenotypes from the UK Biobank. Points report the accuracy of the Step 1 PRS from each tool, measured by the squared correlation between predicted and observed phenotypes across 41k samples (distinct from those used for the association analysis). For comparison, we also report the accuracy of Ridge Regression PRS, which are similar to the PRS constructed by REGENIE.* ***b,*** *Points compare the numbers of significant genes ($P<0.05/17,322=2.9x10^{-6}$) from LDAK-GBAT and LDAK-KVIK-GBAT, for each of the 40 quantitative phenotype. The diagonal lines mark y=x.*

Accuracy of Step 1 PRS.

Figure 4a compares the accuracy of the Step 1 PRS constructed by LDAK-KVIK and BOLT-LMM, for each of the 40 quantitative UK Biobank phenotypes. For comparison, we also consider Ridge Regression PRS, which are constructed using similar assumptions to the PRS constructed by REGENIE (the latter does not report effect sizes, so we can not measure the accuracy of its PRS directly). We see that the PRS accuracies mirror the detection powers of the three MMAA tools (**Figure 3**). For example, the PRS from LDAK-KVIK tend to be slightly more accurate than the PRS from BOLT-LMM and substantially more accurate than the PRS from Ridge Regression, explaining why LDAK-KVIK found slightly more significant loci than BOLT-LMM-Unscaled and substantially more significant loci than REGENIE. **Supplementary Figure 16** shows that for the 20 binary phenotypes, the Step 1 PRS tend to have very low accuracy, explaining why the power of the MMAA tools was similar to that of logistic regression.

Gene-based analysis of UK Biobank phenotypes.

We test 17,332 genes for association, defined based on RefSeq annotations.[32] **Figure 4b** and **Supplementary Figure 17** report the number of significant genes ($P<0.05/17{,}322=2.9\times10^{-6}$) from LDAK-GBAT and LDAK-KVIK-GBAT. Across the 40 quantitative phenotypes, LDAK-KVIK-GBAT finds on average 18.4% more significant genes than LDAK-GBAT. By contrast, across the 20 binary phenotypes, there is no advantage using LDAK-KVIK-GBAT instead of LDAK-GBAT (in total, the tools find 1508 and 1515 significant genes, respectively).

DISCUSSION

We have presented LDAK-KVIK, a novel tool for performing single-SNP and gene-based mixed-model association analysis. We have shown that LDAK-KVIK can be validly applied to homogeneous and heterogeneous datasets, and to both quantitative and binary phenotypes. LDAK-KVIK is computationally efficient; with access to parallel computing, it can analyse data for 100,000s of individuals within a few hours, and has low memory requirements. Furthermore, LDAK-KVIK is powerful; for example, when used for single-SNP analysis of quantitative phenotypes, LDAK-KVIK consistently finds more significant associations than the existing MMAA tools BOLT-LMM, REGENIE, fastGWA and GCTA-LOCO.

Compared to BOLT-LMM, the main advantage of LDAK-KVIK is its computational efficiency (e.g., when analyzing the UK Biobank phenotypes, LDAK-KVIK was 30 times faster and required five times less memory). This is mainly due to the development of a chunk-based variational Bayes solver. **Supplementary Figure 2** shows that our variational Bayes solver can construct PRS 5-10 times faster than the standard (genome-wide)

variational Bayes solver, and can compute terms of the form $V^{-1}A$ ten times faster than conjugate gradient descent. Further, our variational Bayes solver has a small memory footprint (less than 1Gb), because it never needs to store data for more than 256 SNPs at a time.

Compared to REGENIE, the main advantage of LDAK-KVIK is its power (e.g., across the 40 quantitative UK Biobank phenotypes, LDAK-KVIK found 3.7% (s.d. 0.6%) more independent, genome-wide significant loci, and when restricted to SNPs significant from linear regression, the test statistics from LDAK-KVIK were on average 2.8% (s.d. 0.3%) higher). This power increase mainly reflects that LDAK-KVIK uses more realistic models for the genetic architecture of complex traits. Specifically, LDAK-KVIK models the relationship between per-SNP heritabilities and MAF, and uses an elastic net prior distribution for SNP effect sizes. As shown in **Figure 4a**, these features result in LDAK-KVIK constructing more accurate PRS in Step 1, which then increases the chance of discovering associations in Step 2.

We recognize that LDAK-KVIK can be considered much slower than fastGWA. However, there are four points to note. Firstly, LDAK-KVIK is only slower if we ignore the time fastGWA takes to compute the genomic relatedness matrix. Although a one-off cost, this can be non-trivial (e.g., when analyzing 368k individuals, this step took 763 CPU hours, which is longer than the total time LDAK-KVIK required to analyse all 40 quantitative phenoptypes). Secondly, LDAK-KVIK is usually substantially more powerful than fastGWA, reflecting that fastGWA focuses on controlling type 1 error, instead of improving power relative to linear regression. Thirdly, fastGWA is only designed for homogeneous data, or data containing related individuals, whereas LDAK-KVIK can also be applied to datasets including individuals from multiple ancestries. Fourthly, LDAK-KVIK includes an option to perform an approximate version of fastGWA, that uses a much faster algorithm for identifying related pairs of individuals. **Supplementary Figure 18** shows that when applied to the 60 quantitative and binary UK Biobank phenotypes, our approximate version of fastGWA gives results very similar to those from the original version of fastGWA, yet in total takes under two CPU hours to analyze each trait (instead of 763 CPU hours).

We realise that LDAK-KVIK has some limitations. Firstly, we were unable to devise a reliable method for estimating λ when assuming a mixture prior for SNP effect sizes. Therefore, as a practical solution, LDAK-KVIK begins by testing for structure. If this test finds weak structure, then LDAK-KVIK uses a mixture prior and sets $\lambda=1$. However, if this test finds strong structure, LDAK-KVIK switches to an infinitesimal prior, and estimates λ using the Grammar-Gamma formula.[11] We recognise that our solution is imperfect, because switching to an infinitesimal prior typically results in less accurate Step 1 PRS, and therefore lower power to detect associations in Step 2 (**Supplementary Figures 19 & 20**). In **Supplementary Note 8** we summarize seven different approaches we tried when searching for a general method for estimating λ, in the hope that these might inspire others to succeed where we failed.

A second limitation is that LDAK-KVIK estimates α using a grid search, with five pre-defined values (-1, -0.75, -0.5, -0.25 and 0). While the use of pre-defined values enables high computational efficiency (because LDAK-KVIK can evaluate model fit for all five values simultaneously), we appreciate that it limits the accuracy of the algorithm. For example, when analyzing the 20 phenotypes underlying **Figure 2**, each of which was generated assuming $\alpha=-0.25$, LDAK-KVIK only infers the true value six times (for the remaining 14 phenotypes, its estimate of alpha is either -0.5 or 0). However, while imperfect, we believe that our algorithm is a marked improvement on the status quo, which is to simply assume $\alpha=-1$.

Thirdly, we found that when analyzing binary phenotypes, LDAK-KVIK was often only slightly more powerful than logistic regression. We note that this was the same for other MMAA tools. Furthermore, it partially reflects that we have focused on non-ascertained phenotypes (e.g., the UK Biobank is a population-based sample, so the proportion of cases for each ICD10 disease will be close to the disease's prevalence). For **Supplementary Figure 21**, we simulate ascertained binary phenotypes. We find that LDAK-KVIK continues to have good control of type 1 error, however its power advantage over logistic regression depends on the scenario considered (e.g., its advantage tends to increase when analyzing more common phenotypes, but reduce when analyzing rarer phenotypes).

We finish by pointing out that although designed for association testing, LDAK-KVIK also produces state-of-the-art PRS. In particular, previous works have shown that BOLT-LMM is one of the leading tools for constructing PRS.[14,23] Here we have found that LDAK-KVIK tends to produce more accurate PRS than BOLT-LMM, and has substantially lower computational demands.

DATA AVAILABILITY

Our study used data from UK Biobank, which we applied for and downloaded from https://www.ukbiobank.ac.uk. The UK Biobank has ethics approval from the North West Multi-centre Research Ethics Committee (MREC).

CODE AVAILABILITY

LDAK-KVIK is part of the software package LDAK, which can be downloaded from www.dougspeed.com.

ACKNOWLEDGMENTS

We thank Anders Halager and Dan Søndergaard (both Aarhus University) for programming suggestions, David Balding (University of Melbourne) for helpful comments on the manuscript, Hamed Heydari (University of Toronto) for advice on heritability estimation, and Soumeen Jin (Karolinska Institute) for testing LDAK-KVIK. DS is supported by the Aarhus University Research Foundation (AUFF), by the Independent Research Fund Denmark (project no. 7025-00094B) and by a European Research Council Consolidator Grant (ID 101088901, acronym ClassifyDiseases). The computing for this project was performed on the GenomeDK cluster (Aarhus University).

AUTHOR CONTRIBUTIONS

JH and DS jointly developed the software, performed the analysis and wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ONLINE METHODS

Here we provide a concise description of LDAK-KVIK; please see **Supplementary Notes 1-7** for an exhaustive version, as well as details of existing MMAA tools and the UK Biobank data. Note that when describing LDAK-KVIK, we first assume the phenotype is quantitative and that the data are approximately homogeneous, then later explain the changes required when the phenotype is binary, or when we detect heterogeneity.

Notation.

Suppose there are n individuals, each genotyped for m SNPs, recorded for q covariates and measured for a phenotype. Let the (n x m) matrix X' contain the genotypes, let the length-n vector Y' contain the phenotypes, and let the (n x q) matrix Z contain covariates. We use C to denote the total number of chromosomes, use X and Y to denote, respectively, the genotypes and phenotypes after regressing out the covariates, and use λ to denote the test statistic scaling factor. Without loss of generality, we assume X_j and Y are standardized to have mean zero and variance one.

LDAK-KVIK Step 1.

This step produces C LOCO Elastic Net PRS, each taking the form $P_c = \sum X_j \hat{\gamma}_j$, where $\hat{\gamma}_j$ is the estimated effect size for SNP j, and the sum is across all SNPs not on Chromosome c. This step also estimates λ.

LDAK-KVIK first performs our novel test for structure (described below); for now, we assume this test finds only weak structure. LDAK-KVIK next computes $\hat{h}_j^2$, estimates of the heritability contributed by each SNP. For this, it considers heritability models of the form

$$E[h_j^2] = w_j h^2 / W, \quad \text{with} \quad w_j = [f_j(1-f_j)]^{1+\alpha} \quad \text{and} \quad W = \sum w_j$$

where f_j is the MAF of SNP j, h^2 is the proportion of phenotypic variance explained by all SNPs, while α determines how per-SNP heritability depends on MAF. LDAK estimates α and h^2 using Randomized Haseman-Elston Regression[33] and Monte Carlo restricted maximum likelihood [12] (REML), then sets $\hat{h}_j^2$ to its expected value given these estimates. Note that Monte Carlo REML must calculate terms of the form $V^{-1}Y$, $KV^{-1}Y$, $V^{-1}r$ and $KV^{-1}r$, where V is an (n x n) variance matrix, K is an (n x n) genomic relatedness matrix (GRM), and r is a length-n vector whose elements are drawn from a standard normal distribution; we calculate these terms using our novel Variational Bayes solver (described below).

LDAK-KVIK then constructs C LOCO Elastic Net PRS, for which it assumes[26]

$$\gamma_j \sim p\,DE(a_j) + (1-p)N(0, b_j), \quad \text{with} \quad a_j = \sqrt{\frac{2p}{(1-F)\hat{h}_j^2}} \quad \text{and} \quad b_j = \frac{F\hat{h}_j^2}{1-p} \qquad (1)$$

where $DE(a_j)$ denotes a double exponential distribution with rate a_j. The parameters p and F determine the relative contributions of the double exponential and normal distributions. LDAK-KVIK finds suitable values for these using cross-validation (by default, LDAK-KVIK uses 90% of samples to construct genome-wide PRS corresponding to ten different pairs of values for p and F, then picks the pair that has lowest mean-squared error measured using the remaining 10% of samples). Note that all PRS are constructed using our novel variational Bayes solver (described below). Finally, LDAK-KVIK calculates λ. Because we are, for now, assuming there is only weak structure, LDAK-KVIK sets $\lambda = 1$.

LDAK-KVIK Step 2.

LDAK-KVIK tests SNP j for association using least-squares linear regression with the model $E[Y - P_c] = X_j \beta_j$, then scales the resulting $\chi^2(1)$ test statistic by λ.

Binary phenotypes.

Many of the operations in LDAK-KVIK are based on a linear model of the form $Y_i = \sum X_{i,j} \gamma_j + e_i$. In particular, this model is assumed (either explicitly or implicitly) when estimating h^2, when constructing PRS and when testing SNPs for association. When the phenotype is quantitative, X_j and Y contain standardized residuals from linearly regressing X_j' and Y', respectively, on Z, and LDAK-KVIK assumes $e_i \sim N(0, I(1-\hat{h}^2))$, where I is an (n x n) identity matrix and $\hat{h}^2$ is the estimate of h^2. When the phenotype is binary, LDAK-KVIK first computes the length-n vector μ', which contains estimates of the probabilities that each individual is a case given the covariates, and constructs the (n x n) diagonal matrix D, with $D_{i,i} \propto \mu'_i(1-\mu'_i)$ and trace(D^{-1})=n. LDAK-KVIK then sets X_j to the residual from regressing X_j' on Z using weighted linear regression with weight matrix D, sets $Y = D^{-1}(Y' - \mu')$, and assumes $e_i \sim N(0, D^{-1}(1-\hat{h}^2))$. These changes are motivated by the observation that the estimated SNP effect sizes from regressing Y' on Z and X_j' using logistic regression are approximately equal to those from regressing $D^{-1}(Y' - \mu')$ on Z and X_j' using weighted linear regression with weight matrix D (see **Supplementary Note 5** for a proof). Note that ensuring trace(D^{-1})=n allows us to continue to treat h^2 as a heritability (i.e., h^2 continues to represent the proportion of variance of Y explained by all SNPs).

When analyzing quantitative phenotypes, LDAK-KVIK obtains the (unscaled) Step 2 test statistics via a Wald Test (specifically, $U_j = \hat{\beta}_j^2 / Var(\hat{\beta}_j)$, the square of the estimated effect size for SNP j divided by its estimated variance). When analyzing binary phenotypes, LDAK-KVIK first obtains U_j via a Wald Test, but if the corresponding p-value (after scaling by λ) is below 0.1, recomputes U_j using our novel SPA solver (described below).

Novel test for structure.

LDAK-KVIK picks 512 SNPs semi-randomly from across the genome (specifically, LDAK-KVIK first randomly picks 10,000 SNPs, then retains the 512 SNPs with highest variance). It then computes $\bar{\rho}^2$, the average squared correlation between the r pairs of SNPs on different chromosomes. To test $\bar{\rho}^2$ for significance, we use the fact that under the null hypothesis (i.e., if the data are homogeneous and so SNPs on different chromosomes are independent), the squared correlations are beta distributed with parameters 1/2 and (n-2)/2, and so $\bar{\rho}^2$ will have expectation 1/(n-1) and variance 2(n-2)/[r(n-1)^2(n+1)].

As well as using $\bar{\rho}^2$ to test for structure, we consider $n\bar{\rho}^2$ an estimate of the maximum average inflation of test statistics due to structure, which is based on the following logic. Suppose that we could partition the genetic contribution to a phenotype as $G = L_j + D_j$ where L_j and D_j are the components local and distal to SNP j, respectively, whose heritability contributions are h^2_{Lj} and h^2_{Dj}, respectively. Then when testing SNP j for

association with the phenotype, we could write its expected $\chi^2(1)$ test statistic from classical linear regression as $E[S_j] \approx 1 + n(l_j^2 h_{Lj}^2 + d_j^2 h_{Dj}^2)$, where $l_j^2 = Cor(X_j L_j)^2$ and $d_j^2 = Cor(X_j D_j)^2$ are the proportions of local and distal genetic variation tagged by SNP j, respectively. Under this partitioning, $n d_j^2 h_{Dj}^2$ can be viewed as the expected inflation of S_j due to structure. Finally, if we assume that $\bar{\rho}^2$ is a reasonable estimator of $\bar{d}^2$, the average value of d^2_j across all SNPs, and recognize that $h_{Dj}^2 \leq 1$, then it follows that $n\bar{\rho}^2$ is an upper bound for the average inflation of test statistics due to structure.

LDAK-KVIK determines there is strong structure when $n\bar{\rho}^2 > 0.1$ and the corresponding p-value is below 0.001 (otherwise, it determines the structure is weak). Our test for structure is very fast (e.g., when analyzing 368k individuals, it takes less than one minute) and has low memory demands (because it is only necessary to store genotypes for 512 SNPs). When LDAK-KVIK determines there is strong structure, it makes the following changes. When analyzing a quantitative phenotype, it switches from the elastic net prior to the infinitesimal prior $\gamma_j \sim N(0, \hat{h}_j^2)$, then estimates λ using the Grammar-Gamma formula[11]. When analyzing a binary phenotype, LDAK-KVIK uses the approximate version of fastGWA described below.

Approximate version of fastGWA.

The LDAK-KVIK version of fastGWA differs from the original version in three ways. Firstly, whereas fastGWA constructs a GRM using all SNPs, our version uses only 512 SNPs (those used in the test for structure, described above). Secondly, when constructing the sparse GRM, fastGWA recommends truncating values below 0.05, whereas we truncate non-significant values (specifically, those with P>0.1/n, which typically corresponds to values below 0.2). Thirdly, when analyzing a binary phenotype, we assume the null model $Y \sim N(0, \sigma^2 K + E^{-1})$, where K is the sparse GRM and E is a diagonal matrix with $E_{i,i} = \mu'_i(1-\mu'_i)$. Note that this represents a simplified version of the quasi-likelihood used by fastGWA, because we set μ', and therefore also E, based only on the covariates, whereas fastGWA updates μ' to allow for the contribution of the sparse GRM. **Supplementary Figure 18** shows that despite these simplifications, our version of fastGWA performs similarly to the original version, both for quantitative and binary phenotypes.

LDAK-KVIK-GBAT.

Previously, we developed LDAK-GBAT, a tool for gene-based association analysis.[19] This tests each gene using the mixed model $Y \sim N(0, K_S \sigma_S^2 + I \sigma_g^2)$, where K_S is a local GRM computed using only SNPs within the gene. LDAK-GBAT finds the maximum likelihood estimate of σ_S^2, then obtains a p-value by testing if $\sigma_S^2 > 0$.

Importantly, LDAK-GBAT requires only GWAS summary statistics and a reference panel. Therefore, LDAK-KVIK-GBAT applies LDAK-GBAT using the association results from Step 2, and using 5000 randomly-picked individuals from the data as an in-sample reference panel.

Overview of Variational Bayes.

This section explains the general idea of using variational Bayes to construct PRS, whereas the next section provides specific details of our novel variational Bayes solver. Please note that a detailed description of variational Bayes is provided in the supplement of the BOLT-LMM paper.[12]

When constructing PRS, LDAK-KVIK estimates $P(\gamma|Y)$, the posterior distribution of SNP effect sizes given the data. To do this, we first construct a model likelihood $L(Y|\gamma)$ by assuming that Y has the multivariate normal distribution $Y \sim N(X\gamma, I(1-\hat{h}^2))$, where I is an (n x n) identity matrix. Then we specify $\pi(\gamma)$, a prior distribution for SNP effect sizes, and use variational Bayes to approximate $P(\gamma|Y)$ as the product of independent, single-parameter posterior distributions (one for each SNP):

$$P(\gamma|Y) \propto L(Y|\gamma) \times \pi(\gamma) \approx Q(\gamma) = \prod Q_j(\gamma_j)$$

LDAK-KVIK updates $Q(\gamma)$ one SNP at a time, by replacing the current $Q_j(\gamma_j)$ with the distribution that minimizes the difference between the approximate and true log likelihoods (measured by the Kullback-Leibler divergence). Once convergence has been achieved, the corresponding PRS is constructed by setting $\gamma_j = \eta_j$, where η_j is the expectation of $Q_j(\gamma_j)$.

As explained above, Step 1 of LDAK-KVIK uses variational Bayes to construct Elastic Net PRS, assuming the prior distribution defined in Equation 1. In this application, each $Q_j(\gamma_j)$ is a mixture of a left truncated normal distribution, a right truncated normal distribution, and a (non-truncated) normal distribution:

$$Q_j(\gamma) = p_{j-} N^-(\eta_{j-}, \sigma^2_{j-}) + p_{j+} N^+(\eta_{j+}, \sigma^2_{j+}) + (1 - p_{j-} - p_{j+}) N(\nu_{jn}, \sigma^2_{jn})$$

If $P_j = \sum X_{j'} \hat{\gamma}_{j'} - X_j \hat{\gamma}_j$ denotes a partial PRS where the effect sizes are based on the current estimate of $Q(\gamma)$, then $Q_j(\gamma_j)$ is updated by setting

$$\eta_{j-} = \frac{X_j^T(Y - P_j) + (1-\hat{h}^2) a_j}{X_j^T X_j}, \quad \eta_{j+} = \frac{X_j^T(Y - P_j) - (1-\hat{h}^2) a_j}{X_j^T X_j}, \quad \sigma^2_{j-} = \sigma^2_{j+} = \frac{1-\hat{h}^2}{X_j^T X_j}$$

$$\eta_{jn} = \frac{X_j^T(Y - P_j)}{X_j^T X_j + (1-\hat{h}^2)/b_j}, \quad \sigma^2_{jn} = \frac{1-\hat{h}^2}{X_j^T X_j + (1-\hat{h}^2)/b_j}, \quad p_{j-} = \frac{\nu_{j-}}{\nu_{j-} + \nu_{j+} + \nu_{jn}}, \quad p_{j+} = \frac{\nu_{j+}}{\nu_{j-} + \nu_{j+} + \nu_{jn}}$$

where $\nu_{j-} = \frac{p}{2} a_j \sqrt{2\pi\sigma^2_{j-}} \Phi\left(\frac{-\eta_{j-}}{\sqrt{\sigma^2_{j-}}}\right) \exp\left(\frac{\eta^2_{j-}}{2\sigma^2_{j-}}\right)$, $\nu_{j+} = \frac{p}{2} a_j \sqrt{2\pi\sigma^2_{j+}} \Phi\left(\frac{-\eta_{j+}}{\sqrt{\sigma^2_{j+}}}\right) \exp\left(\frac{\eta^2_{j+}}{2\sigma^2_{j+}}\right)$

and $v_{jn} = (1-p)\sqrt{\frac{\sigma^2_{jn}}{b_j}}\exp\left(\frac{\eta^2_{jn}}{2\sigma^2_{jn}}\right)$

Novel variational Bayes solver.

The variational Bayes solver implemented in BOLT-LMM uses sequential genome-wide scans.[12] This means that on each scan, it updates $Q_j(\gamma_j)$ once for each SNP in the genome (i.e., it first updates $Q_1(\gamma_1)$, then $Q_2(\gamma_2)$, and so on until $Q_m(\gamma_m)$). BOLT-LMM performs multiple scans, continuing until LL, the approximate log likelihood, changes by less than a specified tolerance (by default, 0.0005). Typically, BOLT-LMM requires 50-150 scans (convergence is slowest when analyzing highly-heritable traits with very large sample sizes).

By contrast, our novel variational Bayes solver partitions the genome into chunks (by default, each chunk contains 256 SNPs), then uses chunk-based scans (see **Supplementary Figure 1** for an illustration). On Scan 1, our solver first updates $Q_j(\gamma_j)$ for SNPs in Chunk 1, continuing until LL has converged (the default tolerance is $n/10^6$). Our solver then updates $Q_j(\gamma_j)$ for SNPs in Chunk 2, then for SNPs in Chunk 3, continuing until it reaches the final chunk in the genome. On subsequent scans, our solver repeats the same process, except that it only considers chunks that had a sizeable impact on LL (specifically, it only revisits a chunk if on the previous scan, the updates for that chunk caused LL to change by more than $n/10^6$). Our solver typically requires no more than ten scans to converge (because at this point, no chunks remain that have a sizeable impact on LL).

Our chunk-based solver has three main advantages over the genome-wide solver. Firstly, it prioritizes SNPs in regions that have a larger influence on LL (which is more efficient than simply treating all SNPs the same). Secondly, it enables on-the-fly reading of the genotypes (while this is, in theory, possible with the genome-wide solver, it would be necessary to read the data 50-150 times). Thirdly, it is more cache-friendly. **Supplementary Figure 2** compares our chunk-based solver with a genome-wide version when analyzing either 50k or 100k individuals. We see that both solvers have very similar accuracy (in that both produce models with very similar LL). However, we find that the chunk-based solver is substantially faster, reflecting that it performs fewer scans, and in turn fewer updates of $Q_j(\gamma_j)$. For example, when analyzing 100k individuals, the genome-wide solver on average requires 56 scans (so updates each $Q_j(\gamma_j)$ 56 times). By contrast, our chunk-based solver requires on average 4 scans, and on average updates each $Q_j(\gamma_j)$ 11 times. **Supplementary Figure 22** shows that the default convergence criterion suffices, in the sense that results are almost unchanged if the tolerance is made five times smaller.

Computing terms of the form $V^{-1}A$ and $KV^{-1}A$.

Our variational Bayes solver can be used to construct Ridge Regression PRS by assuming the prior distribution $\gamma_j \sim N(0, \hat{h}^2_j)$. In this application, each $Q_j(\gamma_j)$ has the form $N(\eta_j, \sigma^2_j)$, and is updated by setting

$$\eta_j = \frac{X_j^T(Y - P_j)}{X_j^T X_j + (1-\hat{h}^2)/\hat{h}^2_j} \quad \text{and} \quad \sigma^2_j = \frac{1-\hat{h}^2}{X_j^T X_j}$$

However, for this choice of prior, the posterior mean has an explicit form (derived in **Supplementary Note 4**). In particular, it can be shown that if P is a (genome-wide or LOCO) Ridge Regression PRS, then $P = \hat{h}^2 K V^{-1} Y$ and $Y - P = (1-\hat{h}^2) V^{-1} Y$ where V is a (genome-wide or LOCO) variance matrix. It follows that we can estimate terms of the form $V^{-1}Y$ and $KV^{-1}Y$ by using our variational Bayes solver to construct Ridge Regression PRS for Y, then dividing the estimated PRS by $\hat{h}^2$ or dividing the corresponding residuals by $(1-\hat{h}^2)$. More generally, we can estimate terms of the form $V^{-1}A$ and $KV^{-1}A$, where A is an arbitrary length-n vector, by performing the same steps but replacing Y with A (i.e., instead of constructing a PRS for Y, we construct a PRS for A). Therefore, our novel variational Bayes solver can not only be used to construct PRS, but also to compute the terms required when performing either Monte Carlo REML or using the Grammar-Gamma formula.[11]

Overview of SPA.

Here we summarize the SPA; for a fuller description, we recommend reading the supplement of the REGENIE paper.[13] If A is a random variable, then its cumulant-generating function (CGF) is $K_A(t) = \log(E[\exp(tA)])$. The first, second and third derivatives of $K_A(t)$ are

$$K^1_A(t) = \frac{m_1(t)}{m_0(t)}, \quad K^2_A(t) = \frac{m_2(t)}{m_0(t)} - \frac{m_1(t)^2}{m_0(t)^2} \quad \text{and} \quad K^3_A(t) = \frac{m_3(t)}{m_0(t)} - \frac{3m_1(t)m_2(t)}{m_0(t)^2} + \frac{2m_1(t)^2}{m_0(t)^3}$$

where $m_k(t) = E[A^k \exp(tA)]$. The CGF is additive, such that if A and B are independent random variables, then $K_{aA+bB}(t) = K_A(at) + K_B(bt)$. Similar relationships hold for the first, second and third derivatives:

$$K^1_{aA+bB}(t) = aK^1_A(at) + bK^1_B(bt), \quad K^2_{aA+bB}(t) = a^2 K^2_A(at) + b^2 K^2_B(bt) \quad \text{and} \quad K^3_{aA+bB}(t) = a^3 K^3_A(at) + b^3 K^3_B(bt)$$

Now suppose A arises as a test statistic from regressing a phenotype on X_j. The most common way to calculate a p-value is to compute $U_j = (A - E[A])^2 / Var(A)$, where E[A] and Var(A) are, respectively, the expectation and estimated variance of A under the null hypothesis, then to compare U_j to a $\chi^2(1)$ distribution (or equivalently, compare its square root to a standard normal distribution). However, this approach assumes the null distribution of A is approximately normal, which can be inappropriate (e.g., if the phenotype is binary and very imbalanced, or when testing rare variants).

The SPA provides an alternative way to compute a p-value. It involves computing

$$U'_j=\left(w+\log\left(\frac{v}{w}\right)/w\right)^2 \quad \text{with} \quad w=sign(t')\sqrt{2(t'A-K_A(t'))} \quad \text{and} \quad v=t'\sqrt{K_A^2(t')} \qquad (2)$$

where t' is the solution to $K^1(t')=A$. The SPA p-value is then obtained by comparing U'$_j$ to a $\chi^2(1)$ distribution.

Novel SPA solver.

Suppose the test statistic from regressing a phenotype B on X_j takes the form $A=a_1B_1 + \ldots + a_nB_n$. Our solver starts by calculating $K_B(at)=\log\left(\sum \exp(atB_i)/n\right)$ for 41 predetermined values of a (evenly spaced between -2 and 2), and for 256 predetermined values of t (ranging from t_{min} to t_{max}, as described below). Note that these calculations correspond to assuming that the B_i are independent and identically distributed, and that their true distribution matches the observed distribution. Our solver similarly computes 41 x 256 realizations for each of $K^1_B(at)$, $K^2_B(at)$ and $K^3_B(at)$. We refer to the predetermined values of a and t as "bin centres" and "knots", respectively.

When LDAK-KVIK analyzes a quantitative phenotype, the (uncalibrated) Step 2 test statistic for SNP j is $A=X_j^T(Y-P_c)$, so we set $a_i=X_{ij}$ and $B_i=Y_i-P_{ci}$ in the above equations; when LDAK-KVIK analyzes a binary phenotype, the test statistic is $A=X_j^T D(Y-P_c)$, so we set $a_i=X_{ij}$ and $B_i=D_{ii}(Y_i-P_{ci})$. Note that when calculating the CDF of A and its derivatives, it is necessary to calculate terms of the form $K_B(a_it)$, $K^1_B(a_it)$ and $K^2_B(a_it)$, and in general, a_i will not match one of the 41 bin centres. Therefore, we approximate these terms using first, second and third order Taylor Series. Specifically, if a' denotes the bin centre closest to a_i, then we use the three approximations

$$K_B(a_it)\approx K_B(a't)+K_B^1(a't)(a_it-a't)+K_B^2(a't)(a_it-a't)^2/2+K_B^3(a't)(a_it-a't)^3/6$$

$$K_B^1(a_it)\approx K_B^1(a't)+K_B^2(a't)(a_it-a't)+K_B^3(a't)(a_it-a')^2/2$$

and $$K_B^2(a_it)\approx K_B^2(a't)+K_B^3(a't)(a_it-a't)$$

Our solver uses the second approximation to estimate $K^1_A(t) = a_1K^1_B(a_1t) + \ldots + a_nK^1_B(a_nt)$ for each of the 256 knots. It then identifies t_L and t_R, the knots immediately left and right of the solution to $K^1_A(t)=A$, and uses the first and last approximations to calculate $K_A(t_L)$, $K_A(t_R)$, $K^2_A(t_L)$ and $K^2_A(t_R)$. Lastly, our solver uses linear interpolation to estimate t', $K_A(t')$ and $K^2_A(t')$. Specifically, it computes $\epsilon=(K_A^1(t_L)-A)/(K_A^1(t_L)-K_A^1(t_R))$, which estimates the location of the solution, relative to the closest two knots (e.g., if $\varepsilon=0.5$, then t' is approximately halfway between t_L and t_R), then sets

$$t'=(1-\epsilon)t_L+\epsilon t_R, \quad K_A(t')=(1-\epsilon)K_A(t_L)+\epsilon K_A(t_R) \quad \text{and} \quad K_A^2(t')=(1-\epsilon)K_A^2(t_L)+\epsilon K_A^2(t_R)$$

Our SPA solver now has all terms required to compute U'$_j$, the $\chi^2(1)$ test statistic defined in Equation 2. If our solver performed the above calculations naively, it would likely be no faster than existing solvers, because computing each of the CGF and its derivatives still requires n operations (e.g., to compute $K_A(t)$ for a particular knot, it is necessary compute $K_B(a_i t)$ for each individual). However, the number of operations can be reduced dramatically by first summarizing X_j with respect to the bin centres. Specifically, if b_j denotes the jth bin centre, and the function I(i,j) indicates whether a_i is closest to b_j, then our solver computes C_1, …, C_9, nine length-41 count vectors whose elements are

$$C_{1,j}=\sum I(i,j), \quad C_{2,j}=\sum I(i,j)a_i, \quad C_{3,j}=\sum I(i,j)a_i^2$$

$$C_{4,j}=\sum I(i,j)(a_i-b_j), \quad C_{5,j}=\sum I(i,j)a_i(a_i-b_j), \quad C_{6,j}=\sum I(i,j)a_i^2(a_i-b_j)$$

$$C_{7,j}=\sum I(i,j)(a_i-b_j)^2, \quad C_{8,j}=\sum I(i,j)a_i(a_i-b_j)^2, \quad C_{9,j}=\sum I(i,j)(a_i-b_j)^3$$

Given these nine vectors, we can rewrite the CDF of A and its derivatives as

$$K_A(t)=\sum\left(C_{1,j}K_A(b_j)+tC_{4,j}K_A^1(b_j)+t^2C_{7,j}K_A^2(b_j)+t^3C_{9,j}K_A^3(b_j)\right)$$

$$K_A^1(t)=\sum\left(C_{2,j}K_A^1(b_j)+tC_{5,j}K_A^2(b_j)+t^2C_{8,j}K_A^3(b_j)\right)$$

$$K_A^2(t)=\sum\left(C_{3,j}K_A^2(b_j)+tC_{6,j}K_A^3(b_j)\right)$$

which can be computed using 164, 123 and 82 operations, respectively (numbers that, for large datasets, are substantially smaller than n).

When deciding the predetermined knot values, we initially use 256 quantiles from a Cauchy distribution, scaled such that $t_{min}=-2000\sqrt{n}$ and $t_{max}=2000\sqrt{n}$. However, if our solver encounter a SNP whose test statistic is either below $K^1_A(t_{min})$ or above $K^1_A(t_{max})$, then it resets the knot values so that t_{min} and t_{max} are five times larger than their original values (this was not necessary for any of the analyses in this paper).

As explained above, our SPA assumes the true distribution of B, the adjusted phenotype, matches its empirical distribution. This has the advantage that our SPA solver can be (validly) applied to any phenotype, whereas existing SPA solvers generally assume the phenotype has a Bernouilli distribution, so can only be applied to binary phenotypes. Furthermore, our SPA solver makes no requirements on the form of X_j (although the predetermined bin centres assume that all elements of X_j are between -2 and 2, if this is not the case, X_j can simply be rescaled). This means that our solver can automatically be applied to dosage data produced by genotype imputation (or in fact non-SNP data).

Supplementary Figure 23 shows that the default values for the numbers of bin centres and knots suffice, in the sense that results are almost unchanged if we increase either of these values (e.g., increase the number of bin centres from 41 to 100, or the number of knots from 256 to 512).

1. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 59 (2021).
2. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5–22 (2017).
3. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* **40**, 1652–1666 (2011).
4. Keaton, J. M. *et al.* Genome-wide analysis in over 1 million individuals of European ancestry yields improved polygenic risk scores for blood pressure traits. *Nature Genetics* **56**, 778–791 (2024).
5. Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).
6. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
7. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology* **44**, 1137–1147 (2014).
8. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).
9. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* **11**, 407–409 (2014).
10. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100–106 (2014).
11. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components–based method for whole-genome association analysis. *Nature Genetics* **44**, 1166–1170 (2012).
12. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).
13. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53**, 1097–1103 (2021).
14. Loh, P. Mixed-model association for biobank-scale datasets. **50**, 906–908 (2018).
15. Campos, A. I. *et al.* Boosting the power of genome-wide association studies within and across ancestries by using polygenic scores. *Nature Genetics* **55**, 1769–1776 (2023).

16. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. (2018) doi:10.1038/s41586-018-0579-z.
17. Sudlow, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *PLoS Med.* **12**, e1001779 (2015).
18. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51**, 1749–1755 (2019).
19. Berrandou, T., Balding, D. & Speed, D. LDAK-GBAT: Fast and powerful gene-based association testing using summary statistics. *Am. J. Hum. Genet.* **110**, 23–29 (2023).
20. MacKay, D. J. *Information Theory, Inference and Learning Algorithms*. (Cambridge university press, 2003).
21. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* **101**, 37–49 (2017).
22. Speed, D., Holmes, J. & Balding, D. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).
23. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications* **12**, 1–9 (2021).
24. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
25. Speed, D. & Balding, D. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
26. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301–320 (2005).
27. Organization, W. H. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. vol. 1 (World Health Organization, 1992).
28. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).
29. Bakshi, A. *et al.* Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Scientific Reports* **6**, 32894 (2016).
30. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *American journal of human genetics* **91**, 1011–21 (2012).
31. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).

32. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65 (2007).
33. Pazokitoroudi, A. *et al.* Efficient variance components analysis across millions of genomes. *Nat Commun* **11**, 4020 (2020).