nature genetics

SUPPLEMENTARY INFORMATION

In the format provided by the authors and unedited.

Supplementary Information for “Mixed model association for biobank-scale data sets”

Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, Alkes L Price

Contents

Supplementary Note

1 UK Biobank data

We analyzed genetic data from the UK Biobank full release consisting of 487,409 samples typed at $\sim$800,000 markers and imputed to $\sim$93 million variants [5]. We restricted the sample set to 459,327 individuals of European ancestry (based on self-reported white ethnicity), and for linear regression analyses, we further restricted the sample set to 337,539 British-ancestry individuals passing principal component analysis filters and containing no third-degree or closer relationships [5]. (The ancestry filter eliminated $\sim$50,000 samples and the relatedness filter eliminated an additional $\sim$70,000 samples.) We restricted the genotyped marker set to autosomal markers with missingness $<$10% and minor allele frequency (MAF) $>$0.1%, leaving 672,292 markers. We analyzed $\sim$20 million imputed variants with MAF $>$0.1% (applying this filter within BOLT-LMM).

In our running time benchmarks, we also analyzed genetic data from the UK Biobank interim release of 152,249 samples imputed to $\sim$72 million variants. Applying analogous exclusions produced a sample set of 145,613 European-ancestry individuals typed at 651,011 autosomal markers with missingness $<$10% and MAF$>$0.1%. We used QCTOOL v2 to convert imputed data between the BGEN v1.1 and v1.2 formats (to benchmark previous versions of BOLT-LMM, which only support the BGEN v1.1 format).

We analyzed 23 phenotypes selected based on phenotyping rate $>$80% (Supplementary Table 1), SNP-heritability h_g^2>0.08 (Supplementary Table 2), and low correlation between traits. We performed basic QC on each trait, removing outliers outside the reasonable range for each quantitative trait and quantile normalizing within sex strata after correcting for covariates as described in previous GWAS [11–16].

In all association analyses, we included assessment center, genotyping array, sex, age, and age squared as covariates. In linear regression analyses (implemented in the BOLT-LMM software), we also included 20 principal components to correct for ancestry (provided with the UK Biobank data release [5]). In our primary BOLT-LMM analyses, we included 20 principal components computed on our filtered marker set using the FastPCA [17] algorithm (as implemented in PLINK 2.0 [18] `--pca approx`). In auxiliary BOLT-LMM analyses, we varied the number of principal components included as covariates (Supplementary Table 7).

2 BOLT-LMM version 2.3

Our new release of the BOLT-LMM software (version 2.3) performs much faster processing of imputed genotypes, which we discovered was the bottleneck for analyses of extremely large imputed data sets (e.g., $\sim$93 million variants in the UK Biobank N=500K release). This step of the BOLT-LMM computation, which occurs after the model-fitting steps and scales only linearly in sample size and variant count, nonetheless accounted for the large majority of running time for previous versions of BOLT-LMM on UK Biobank data. To overcome this bottleneck, we implemented fast multi-threaded support for analysis of imputed genotypes in the new BGEN v1.2 file format (used to encode genotype probabilities in the UK Biobank full release). Streaming BGEN data and decompressing it in a thread-safe manner (and also applying MAF and INFO filters to allow early-exit when possible) requires careful implementation given that BGEN is a compressed format that is not guaranteed to be indexed. BOLT-LMM v2.3 analyzes these imputed genotypes by reading compressed probability data for blocks of 400 variants at a time from disk and then analyzing these data in parallel compute threads. Analysis of a single variant involves decompressing the genotype probabilities, computing the variant's allele frequency and INFO score (and exiting early if MAF or INFO thresholds are not met), projecting out covariates, and computing the BOLT-LMM test statistic (via a dot product with the residualized phenotype and rescaling term [3]).

For analyses of very large data sets with BOLT-LMM v2.3, we additionally now recommend including principal components as covariates for the purpose of increasing the rate of convergence of the iterative computations performed during BOLT-LMM's model-fitting steps [3]. For phenotypes with high heritability, these steps of the computation account for the majority of run time (after the improvements to processing of imputed data described above) but can be sped up by including PC covariates. Projecting out top PCs (which can be computed rapidly using FastPCA [17]) improves the conditioning of the matrix computations that BOLT-LMM implicitly performs, improving convergence; details are provided in Section 5 below.

3 Power analyses

We benchmarked statistical power of three association analysis approaches: linear regression (using 20 principal component covariates), BOLT-LMM using a Gaussian SNP effect prior (BOLT-LMM-inf, equivalent to the standard "infinitesimal" mixed model [1, 2, 19–24]), and BOLT-LMM using its default mixture-of-Gaussians prior on SNP effect sizes, which accounts for larger-effect SNPs [3]. We tested all three methods on the subset of N=337,539 unrelated British samples, and we additionally tested BOLT-LMM-inf and BOLT-LMM (which are robust to sample structure) on the full set of N=459,327 European-ancestry samples.

We performed two types of benchmarks to assess statistical power afforded by each analysis. First, we counted independent genome-wide significant associations identified by each analysis. To obtain counts that were robust to linkage disequilibrium among associated variants and avoided double-counting of loci, we used PLINK's LD clumping algorithm [18] using LD computed in N=113,851 unrelated British individuals [25] at 9.6 million imputed SNPs with MAF>0.1% and

INFO>0.6 (corresponding to a conservative genome-wide significance threshold of $p<5\times10^{-9}$). We used a stringent 5Mb window and R^2 threshold of 0.01 for LD clumping, and we further collapsed associated SNPs within 100kb of each other. These analyses demonstrated that BOLT-LMM-inf and BOLT-LMM achieved considerable power gains over linear regression when run on the same set of N=337,539 unrelated British samples (21% and 28% increases in locus discovery, respectively; Supplementary Table 2). Expanding the sample set to include all European individuals (allowing relatives) achieved even larger boosts in power (76% and 84%, respectively; Fig. 1a and Supplementary Table 2).

Second, to provide additional insight into the power gain achieved by BOLT-LMM, we examined the amounts of phenotypic variance explained by BOLT-LMM's internal linear predictors and the increases in χ^2 test statistics (for BOLT-LMM-inf and BOLT-LMM vs. linear regression) at associated SNPs. We previously showed that these two quantities are tightly coupled [3]; the intuition is that independent of its ability to analyze data sets containing sample structure (and thereby gain power by analyzing more samples), BOLT-LMM also achieves increased power by implicitly conditioning on polygenic predictions using genome-wide SNPs [3,6]. Conditioning on polygenic predictions effectively reduces noise in an association test, producing a multiplicative boost in χ^2 statistics at associated loci in a manner similar to increasing sample size.

We compared the variance explained by BOLT-LMM's linear predictor—using either the default mixture-of-Gaussians prior on SNP effect sizes or the single-Gaussian BOLT-LMM-inf model, equivalent to best linear unbiased prediction (BLUP)—to the variance theoretically explained by an optimal linear predictor, i.e., SNP-heritability $h_g{}^2$. Variance explained by the linear predictors within BOLT-LMM and BOLT-LMM-inf were estimated internally by the BOLT-LMM software via out-of-sample benchmarks (training on 80% of samples and testing on the remaining 20%). We observed that for several traits, BOLT-LMM successfully predicted more than half of SNP-heritability; for height and hair color, BOLT-LMM predicted >40% of phenotypic variance (Fig. 1b, Supplementary Fig. 1, and Supplementary Table 3).

Estimation of effective sample size. To estimate the effective sample size achieved by BOLT-LMM and BOLT-LMM-inf, we then measured the boosts in χ^2 association statistics of BOLT-LMM and BOLT-LMM-inf (on either N=337,539 or 459,327 samples) versus linear regression on N=337,539 unrelated British samples. Specifically, for BOLT-LMM (resp. BOLT-LMM-inf) on N=459,327 samples, we computed the median ratio of BOLT-LMM (resp. BOLT-LMM-inf) χ^2 statistics on N=459,327 samples to linear regression χ^2 statistics (on N=337,539 samples) across genotyped SNPs with $\chi^2>30$ (roughly corresponding to the usual 5×10^{-8} genome-wide significance threshold) in BOLT-LMM N=337,539 analyses. We ascertained associated SNPs in this way—using an association test different from the two methods being compared—to avoid biasing our benchmarks in favor of either BOLT-LMM (resp. BOLT-LMM-inf) (N=459,327) or linear regression (*N*=337,539). We conducted the other benchmarks (for BOLT-LMM and BOLT-LMM-inf on *N*=337,539 samples) in an analogous manner, swapping the roles of the *N*=337,539 and *N*=459,327 data sets. We observed boosts in χ^2 test statistics at associated SNPs (equivalently, boosts in effective sample size) that tracked closely with the proportions of variance predicted by BOLT-LMM and BOLT-LMM-inf (Fig. 1b, Supplementary Fig. 1, and Supplementary Table 3). (We verified that these boosts were robust to the ascertainment threshold χ^2>30 for selecting associated benchmark SNPs. Replacing this threshold with χ^2>20 or χ^2>40 had a negligible effect on the results: the estimated boosts in effective sample size (Supplementary Table 3b) change by a mean multiplicative factor of 0.98 (standard deviation 0.03) when using a threshold of χ^2>20 instead of χ^2>30. Similarly, using χ^2>40 instead of χ^2>30 changes these numbers by a mean multiplicative factor of 1.01 (standard deviation 0.02).)

For traits in which BOLT-LMM predicted only a very small fraction of phenotypic variance (e.g., hypothyroidism and smoking status), we observed that BOLT-LMM *N*=459,327 analyses still achieved moderate gains in association power over linear regression on *N*=337,539 unrelated British samples; here, BOLT-LMM still benefited from the increased sample size (achieving power equivalent to ∼430K unrelated individuals; Supplementary Table 3). For traits in which BOLT-LMM predicted large fractions of phenotypic variance (e.g., height and hair color), we observed that BOLT-LMM *N*=459,327 analyses achieved power equivalent to linear regression on up to ∼700K unrelated samples (Fig. 1b, Supplementary Fig. 1, and Supplementary Table 3). As expected, BOLT-LMM achieved substantial additional gains over BOLT-LMM-inf for traits with larger-effect SNPs (e.g., hair color, tanning ability, and blood cell traits; Supplementary Table 3).

We also considered estimating effective sample size based on polygenic prediction accuracy, but this approach is subject to the concern that prediction accuracy is a function not only of sample size and SNP-heritability, but also of the effective number of chromosome segments (M_e)—which may be different in different sample sets [26, 27]. Here, we found that M_e differed very slightly between the *N*=337,539 and *N*=459,327 sample sets: using the inverse variance of off-diagonal genetic relationship matrix (GRM) entries to estimate M_e [26,27], we obtained estimates of M_e=106K for the *N*=337K data and M_e=104K for the *N*=459K data. The observation of near-identical M_e between the *N*=337,539 and *N*=459,327 sample sets—the latter of whcih contains related individuals—suggests that the the level of relatedness present in UK Biobank is low enough not to substantially affect the overall genetic structure of the data set, which is consistent with our observation that linear regression on the full *N*=459K data—without accounting for relatedness—only exhibits slight confounding (Supplementary Table 4).

Comparison to other UK Biobank LMM analyses. LMM analyis of at "atlas" of UK Biobank traits has also been undertaken in [9]. We compared BOLT-LMM p-values to GeneATLAS p-values [9] by regressing BOLT-LMM –log$_{10}p$-values on GeneATLAS –log$_{10}p$-values for height.

We observed a high correlation of 0.96 and a regression slope of 1.31 (driven by greater significance at top associated SNPs), indicating that BOLT-LMM was achieving substantially greater power. This observation was expected given that (a) the BOLT-LMM analysis included an additional ∼50K non-white-British individuals, (b) the BOLT-LMM analysis used a LOCO (leave-one-chromosome-out) approach when conditioning on polygenic predictions to increase statistical power, whereas the GeneATLAS analysis used an odd/even-chromosome leave-out approach that conditioned on only half the genome, and (c) the BOLT-LMM analysis modeled non-infinitesimal genetic architectures and was thus better able to condition on larger-effect SNPs (whereas the "atlas" analysis conditioned on polygenic predictions constructed assuming an infinitesimal architecture, as in standard mixed model analysis).

4 Calibration analyses

To assess the calibration of BOLT-LMM (i.e., control of false positives) when used to analyze all *N*=459,327 European samples (keeping related individuals) we performed benchmarks using LD score regression [7]. For each phenotype, we considered BOLT-LMM *N*=459,327 association statistics, linear regression *N*=337,539 association statistics computed using 20 principal component covariates (as a negative control robust to confounding), and linear regression *N*=337,539 association statistics computed without PC covariates (as a positive control susceptible to slight confounding from population stratification among British individuals). We used the LDSC software to run LD score regression on each set of association statistics using the baselineLD model [8] (which applies stratified LD score regression, S-LDSC [28]). (In brief, the baselineLD model is comprised of 59 "baseline" annotations (based on coding, UTR, promoter, and intronic regions, histone marks, DNAse hypersensitivity sites, ENCODE annotations, conserved regions, and enhancers), 10 MAF bin annotations, and 6 LD-related annotations.) We used LD scores from 1000 Genomes EUR samples [29]; LD scores need to be estimated using sequence data, so estimating LD scores within the UK Biobank sample was not an option. (Computing LD scores using imputed data is not recommended [7].)

We previously proposed using the LD score regression intercept as a way of distinguishing polygenicity from confounding as possible sources of increased association test statistics [7]. In theory, SNPs with larger numbers of LD partners have more opportunities to tag causal variants, such that regressing observed χ^2 statistics (for a properly calibrated association test) against the LD score of a SNP should produce a regression line with a y-intercept of 1 (even if the mean χ^2 statistic across all SNPs is larger than 1 due to polygenicity); in contrast, the y-intercept will be larger than 1 if the association test is confounded by ancestry or relatedness. In practice, we previously observed that LD score intercepts were typically close to 1 but slightly larger than 1 due to deviations from the theoretical model (e.g., attenuation bias) [7].

Here, we observe that in highly-powered analyses of traits with substantial heritability, these deviations push the LDSC intercept well above 1 for uninflated association tests, e.g., PC-corrected linear regression on unrelated British samples (Supplementary Fig. 2a and Supplementary Table 4). The reason is that in such analyses, the mean χ^2 test statistic is much larger than 1 (e.g., $\sim$4 for linear regression N=337,539 and $\sim$7 for BOLT-LMM N=459,327 analysis of height, after excluding SNPs explaining $>$0.1% of variance), such that even a slight deviation from theory results in large intercepts (here, as high as 1.5). In general, we observe that LD score regression intercepts tend to rise with SNP-heritability and sample size (Supplementary Fig. 2a and Supplementary Table 4). This behavior of the LDSC intercept makes test statistic inflation difficult to discern based on the value of the LDSC intercept alone: for example, for the height phenotype, linear regression on N=337,539 unrelated British samples *without* principal component covariates—which is susceptible to inflation—and BOLT-LMM on N=459,327 European samples both have LDSC intercepts of nearly 1.5.

Fortunately, accounting for differences in mean χ^2 statistic for different phenotypes and association methods improves the interpretability of the LDSC intercept. The *attenuation ratio*, (LDSC intercept – 1) / (mean χ^2 – 1), calibrates the intercept against the overall shift in χ^2 statistics (due to polygenicity for uninflated association tests). Here we observe that for each trait, PC-corrected linear regression and BOLT-LMM have near-identical attenuation ratios, typically around 0.08 (Fig. 1c), whereas uncorrected linear regression typically has larger attenuation ratios, indicating confounding (Supplementary Fig. 2b and Supplementary Table 4). Across 23 traits, we observe similar mean attenuation ratios of 0.078 (s.e.m. 0.006) for PC-corrected linear regression (N=337,539) and 0.082 (0.005) for BOLT-LMM (N=459,327), a statistically insignificant difference (BOLT-LMM had the higher attenuation ratio for 12 of 23 traits); in contrast, we observe a much higher mean attenuation ratio of 0.104 (0.012) for uncorrected linear regression (N=337,539). We also observe a slightly higher mean attenuation ratio of 0.085 (0.006) for PC-corrected linear regression on all N=459,327 European samples (higher than N=337,539 PC-corrected linear regression for 17 of 23 traits; binomial p=0.01), indicating slight confounding from relatedness, as expected. These observations provide confidence that BOLT-LMM is successfully controlling for sample structure (as expected for mixed model methods) [1, 2]. We note that attenuation ratios are broadly smaller under the baselineLD model, which incorporates genome annotations [8], than under the original LDSC model (Supplementary Table 5), consistent with better model fit upon incorporating genome annotations.

Intuition for attenuation ratios. Attenuation bias in LDSC analyses is essentially a measure of model misspecification: the basic assumption of LD score regression is that association chi-square statistics should (on average) increase linearly with the extent to which a SNP tags other potentially causal SNPs—i.e., the LD score of the SNP [7]. If this model holds perfectly, then the regression has an intercept of 1 and an attenuation ratio of 0; on the other extreme, if LD scores are completely non-informative, the regression becomes flat with an intercept equal to the mean chi-square statistic and an attenuation ratio of 1. In general, most LDSC analyses exhibit modest nonzero attenuation ratios, as previously noted in refs. [3, 7].

As we show in Supplementary Fig. 2 and Supplementary Tables 4 and 5, attenuation ratios vary among traits (as expected, given that different traits have genetic architectures with different levels of agreement to the LDSC model); however, for a given trait, attenuation ratios are largely consistent between analyses of N=337K unrelated individuals vs. N=459K related individuals (also as expected, given that increasing sample size or relatedness does not affect the underlying genetic architecture of a trait). We also observe that attenuation ratios under the original LDSC model (Supplementary Table 5) are generally larger than attenuation ratios under the baselineLD model (Supplementary Table 4), consistent with improved model fit upon augmenting the LDSC model with information about genomic annotations.

Estimation of heritability parameters. While estimation of SNP-heritability for UK Biobank traits was not a primary goal of this manuscript, we do report SNP-heritability estimates to help with interpretation of our results on power and calibration. These estimates were obtained from BOLT-LMM N=337K analyses, which estimated h_g^2during model-fitting.

LDSC also reports estimates of heritability parameters during execution. However, unlike BOLT-LMM, LDSC does not estimate the quantity traditionally termed “SNP-heritability” and denoted h_g^2. Whereas BOLT-LMM (and other restricted maximum likelihood approaches) estimate the proportion of population variance explained by genotyped SNPs (h_g^2), LDSC estimates the *causal heritability contributed by common SNPs (excluding those of large effect)*; this quantity is usually smaller than h_g^2 [28]. As an illustrative example to help intuition, consider a MAF=1% SNP and a MAF=5% SNP in linkage disequilibrium. If the MAF=1% SNP is causal and untyped while the MAF=5% SNP is not causal but is typed, then BOLT-LMM will partially include the causal effect in its h_g^2 estimate, while LDSC will include zero effect (because the MAF=1% causal SNP is not common). If the MAF=1% SNP is causal and both SNPs are typed, then BOLT-LMM will include the full causal effect in its h_g^2 estimate, while LDSC will still include zero effect (because the MAF=1% causal SNP is not common).

In Supplementary Table 9, we compare BOLT-LMM h_g^2 estimates to the heritability parameters estimated by LDSC (under either the baselineLD model [8] or the original LDSC model [7]). Across 23 analyzed traits, the average ratio of the LDSC heritability parameter estimate to the BOLT-LMM h_g^2 estimate was 0.68 (s.e.m. 0.03) under the baselineLD model and 0.54 (s.e.m. 0.01) under the original LDSC model. The ratio showed no correlation to the LDSC intercept ($R = –0.03$, $p = 0.9$).

5 Running time analyses

We benchmarked the running time of BOLT-LMM v2.3 (with 20 principal component covariates to increase convergence rate; see below), the previous version of BOLT-LMM [3], and linear regression using 20 principal component covariates (implemented efficiently within the BOLT-LMM software; cf. Bycroft et al. Table S9 [5]) in example analyses of the years-of-education phenotype. We ran each method on all European-ancestry individuals in the UK Biobank interim and full data releases, analyzing $\sim$72M and $\sim$93M imputed SNPs, respectively, and imposing a MAF>0.1% filter on minor allele frequency. (We ran linear regression on all European-ancestry individuals for the sake of run time comparison even though this analysis would not be performed in practice due to potential confounding from sample structure. Also, for BOLT-LMM v1 analysis of the full data release, we analyzed imputed data from only chromosome 22 and extrapolated the computational cost to the full genome.) We performed all analyses using 8 threads of a 2.10 GHz Intel Xeon E5-2683 v4 processor and reported the median of 5 runs (Fig. 1d and Supplementary Table 6), observing a $\sim$4x speedup of BOLT-LMM v2.3 over the previous version, achieving speed comparable to linear regression.

We further explored the effect of including varying numbers of principal components as covariates in BOLT-LMM analyses to improve convergence speed. During its model-fitting steps, BOLT-LMM applies iterative methods (specifically, conjugate gradient iteration and variational Bayes) to eliminate computationally expensive matrix operations that scale quadratically or cubically with sample size [3]. The cost of a single iteration scales only linearly with *N*; however, we previously observed that the number of iterations required to achieve convergence increases slowly with *N* [3]. Our analyses here (Supplementary Table 7) demonstrate that convergence can be sped up by including principal component covariates (which effectively reduce the condition number of the underlying matrix computations), thus achieving close-to-linear scaling of run time with sample size. (Intuitively, projecting out PC covariates produces faster BOLT-LMM convergence by reducing the amount of genetic structure in the GRM; this structure captured by top PCs is generally unrelated to genetic effects on phenotype.) We note that to achieve increased convergence, principal components need to be computed on the set of SNPs used in the mixed model; PCs that do not match the implicit genetic relationship matrix (GRM) will not improve conditioning. We also note that after model-fitting, BOLT-LMM performs a linear-time association test on imputed SNPs (which we sped up separately using multi-threading; see Section 2); the speedup described here only applies to the model-fitting step.

Comparison to distributed computing approaches. We note that an alternatives exist to BOLT-LMM's algorithmic approach to efficient mixed model analysis at large sample sizes. In particular, the DISSECT software [30], applied in the ref. [9], can utilize distributed computing across a large compute cluster. This method enables $O(N^3)$ analyses of very large numbers of individuals and traits by efficiently distributing (and redistributing) data and computation across a suitably large compute farm (e.g., 5,040 processor cores working together using 5TB of memory [9]). For the use case of analyzing thousands of traits, the DISSECT approach can be more efficient than BOLT-LMM as it requires only one expensive eigendecomposition (assuming missing phenotype values are imputed to allow analysis of a single set of individuals), after which GWAS analyses take only linear time (in SNPs and samples) per trait. In contrast, BOLT-LMM's running time scales superlinearly but subquadratically with sample size.

BOLT-LMM is designed for a different use case. In contrast to phenome-wide, "atlas"-stype GWAS analyses (e.g., ref. [9]), we have designed BOLT-LMM to be useful to research groups focused on specific phenotypes, who often wish to run carefully tailored and/or iterated analyses: e.g., QC-ing and normalizing phenotypes, restricting to various subsets of individuals, applying different covariate adjustments, etc. A good example is ref. [31], which applied BOLT-LMM to blood cell traits that had been carefully adjusted for technical covariates (e.g., time between blood collection and analysis, instrument drift, calibration events and episodes of malfunction), resulting in de-noising of phenotypes by up to 40%. In contrast, most "atlas"-style endeavors apply minimal phenotype QC given the infeasibility of performing detailed QC on hundreds of traits.

6 Unbalanced case-control analyses

As pointed out in a recent preprint introducing the SAIGE method [10], association tests (such as BOLT-LMM) that estimate p-values based on a chi-square distribution can incur inflated type I error rates when used to analyze highly unbalanced case-control traits (due to deviation from asymptotic normality). The extent to which chi-square-based p-values suffer miscalibration for binary traits is a function of three variables: sample size, minor allele frequency, and case-control ratio. Specifically, miscalibration occurs when the minor allele count (MAC) multiplied by the case fraction is small (corresponding to the conventional wisdom that chi-square test statistics break down when expected counts are small). Importantly, genomic control and LD score-based assessments of inflation do not detect failure to control type I error due to this phenomenon.

To assess the effect of unbalanced case-control ratios on biobank-scale BOLT-LMM analyses, we performed a suite of simulations that vary the three key parameters (N, MAF, and case fraction), exploring type I error control across significance thresholds from 1×10^{-4} to 5×10^{-8} (Supplementary Table 8). Specifically, we considered N = 450K, 150K, 50K; case fraction = 30%, 10%, 3%, 1%; and MAF tranches with boundaries 0.0001, 0.001, 0.01, 0.1, 1. We simulated case-control traits using a liability threshold model with in which heritable variance ($h_g{}^2$=0.5) was distributed across odd-numbered chromosomes under an infinitesimal genetic architecture (all 672K genotyped SNPs causal with normally distributed effect sizes with variance proportional to $(p(1-p))^{-0.3}$, where p=MAF [32]). We supplied an LD-pruned set of 495K genome-wide SNPs to BOLT-LMM for model-fitting, and we evaluated association test statistics on imputed SNPs on even chromosomes (which had zero effect) to assess p-values under the null distribution. We performed 20 replicates of each simulation and aggregated type I error across replicates. The results of these analyses, presented in Supplementary Table 8, demonstrate that at UK Biobank sample size, BOLT-LMM p-values are well-calibrated for case fractions $>10\%$ and MAF$>0.1\%$. At lower combinations of sample size, case fraction, and MAF, p-values displayed inflated significance, as expected.

The 23 real phenotypes we analyzed in this manuscript include 7 binary traits with minimum case fraction 4.2% (Supplementary Table 1). We are confident that BOLT-LMM p-values are well-calibrated for these analyses based on three lines of reasoning:

1. Based on our simulations (Supplementary Table 8), BOLT-LMM p-values for N=459K and MAF$>0.1\%$ (the lower limit of SNPs we considered) are well-calibrated for case fractions $>10\%$, and for a case fraction of 3%, BOLT-LMM only overestimates significance for rare SNPs (MAF$<1\%$). The lowest case fractions of the binary traits we analyzed were 4–5% for T2D and hypothyroidism and 14% for respiratory disease (Supplementary Table 1), implying no cause for concern for most of the traits we analyzed.

2. We verified that all genome-wide significant loci identified in our analyses for T2D (62 loci), hypothyroidism (111 loci), and respiratory disease (76 loci) have a genome-wide significant hit with MAF$>1\%$, ruling out the possibility of false-positive loci driven solely by rare SNPs. As noted above, our simulations show that BOLT-LMM p-values are valid at this MAF threshold for case fractions $>3\%$ (Supplementary Table 8).

3. We did not observe a systematic increase in power gain for binary vs. quantitative traits, which would be expected if BOLT-LMM's increases in power were driven by false discoveries specific to binary traits. Instead, binary and quantitative traits exhibit consistent gains in power (Fig 1a and Supplementary Table 2).

While we believe for the reasons above that the analyses of binary traits presented here are robust, the saddlepoint approximation of SAIGE [10] generally solves the problem of p-value miscalibration in unbalanced case-control scenarios, and as such, we recommend using SAIGE rather than BOLT-LMM for analyses of highly unbalanced binary traits (especially if rare variants are to be analyzed). We do still recommend BOLT-LMM over SAIGE for analyses of (reasonably) balanced case-control traits, as (i) BOLT-LMM performs leave-one-chromosome-out (LOCO) analysis to guard against power loss due to proximal contamination; (ii) BOLT-LMM models non-infinitesimal genetic architectures, thereby achieving gains in power for traits with larger-effect loci (e.g., tanning ability; compare BOLT-LMM to BOLT-LMM-inf in Supplementary Fig. 1 and Supplementary Tables 2 and 3); and (iii) BOLT-LMM is slightly faster than SAIGE in the benchmarks of ref. [10].

References

1. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208 (2006).

2. 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).

3. Loh, P.-R. *et al.* Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).

4. 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 Medicine* **12**, 1–10 (2015).

5. Bycroft, C. *et al.* Genome-wide genetic data on $\sim$500,000 UK Biobank participants. *bioRxiv* (2017).

6. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525–526 (2012).

7. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295 (2015).

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

9. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *bioRxiv* 176834 (2017).

10. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *bioRxiv* (2017).

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

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

13. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187 (2015).

14. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature Genetics* **44**, 491–501 (2012).

15. Loth, D. W. *et al.* Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nature Genetics* **46**, 669–677 (2014).

16. Ehret, G. B. *et al.* Genetic variants in novel pathways influence blood pressure and cardio-vascular disease risk. *Nature* **478**, 103–109 (2011).

17. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent evolution of *ADH1B* in Europe and East Asia. *American Journal of Human Genetics* **98**, 456–472 (2016).

18. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 1–16 (2015).

19. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

20. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354 (2010).

21. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355–360 (2010).

22. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).

23. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012).

24. 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).

25. Galinsky, K. J., Loh, P.-R., Mallick, S., Patterson, N. J. & Price, A. L. Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. *American Journal of Human Genetics* **99**, 1130–1139 (2016).

26. Goddard, M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245–257 (2009).

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

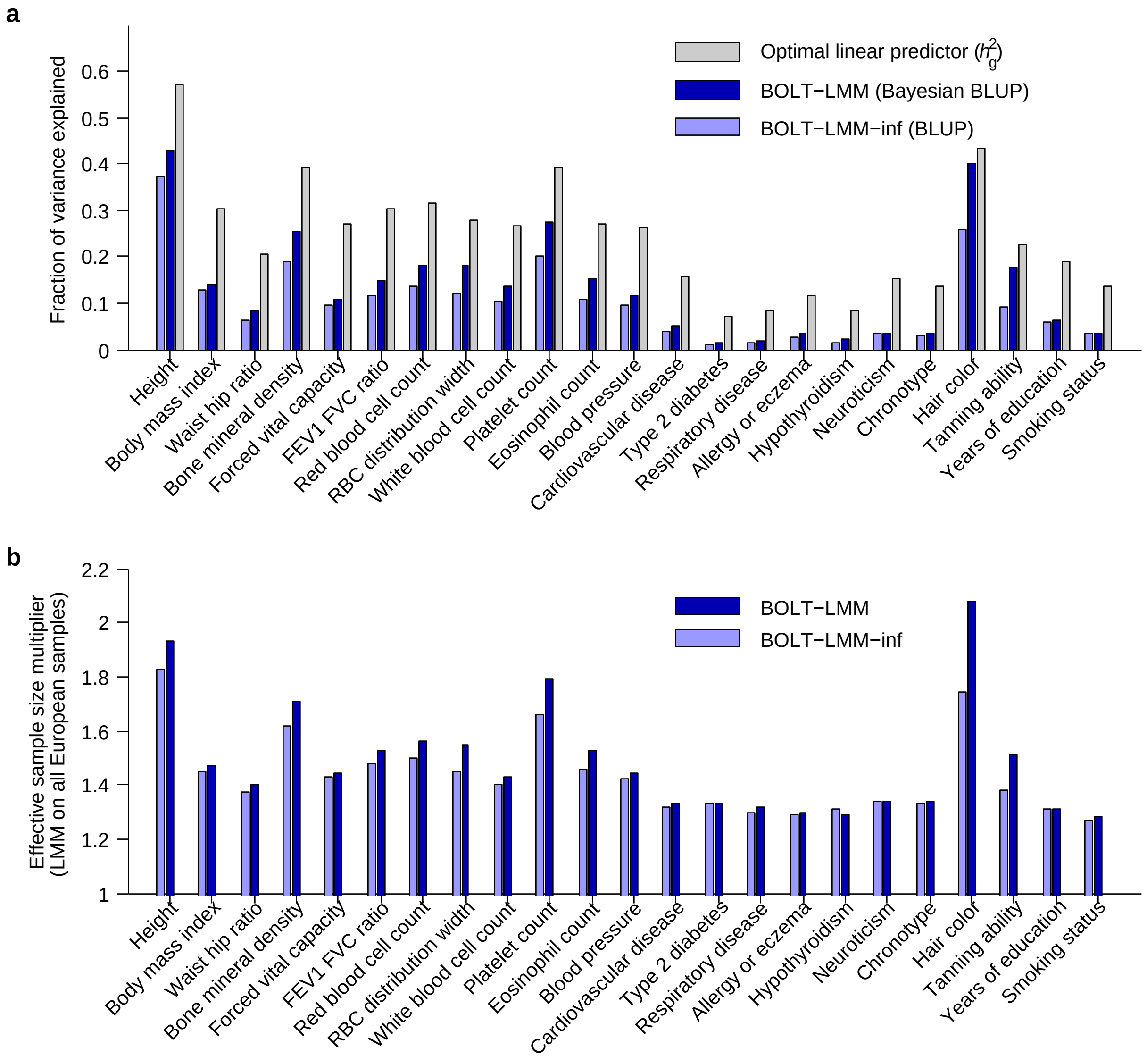
28. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

29. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

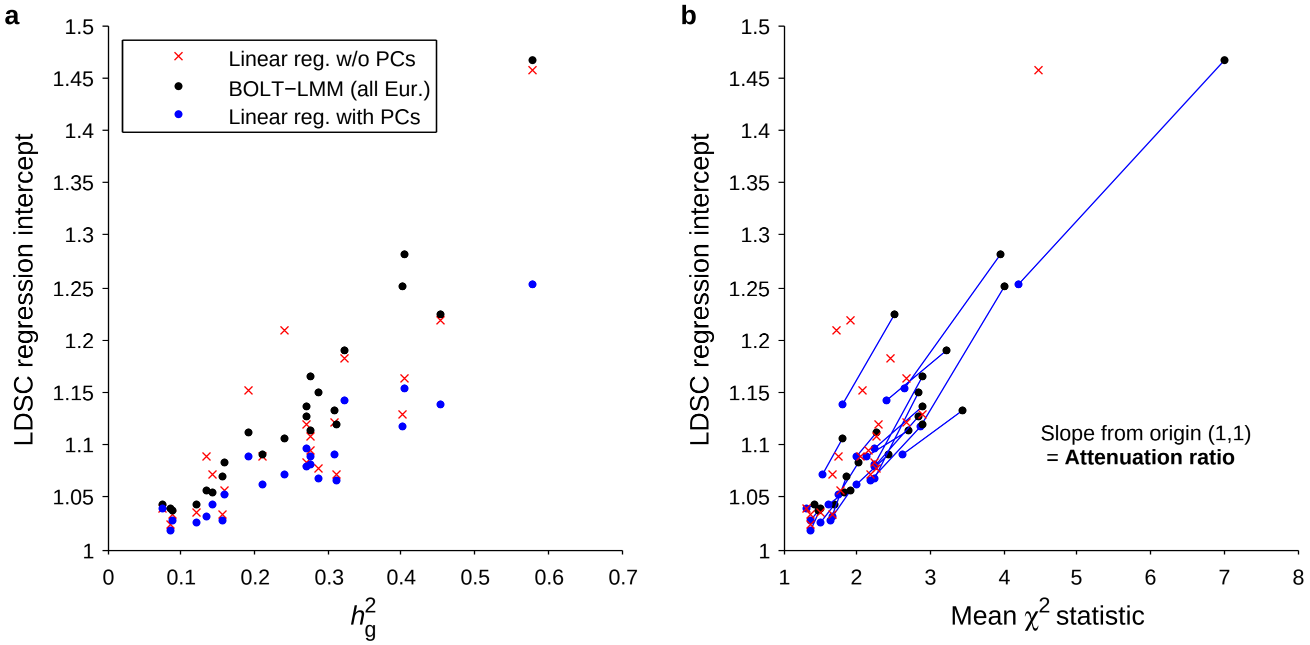
30. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J. A. & Tenesa, A. A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. *Nature Communications* **6** (2015).

31. Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).

32. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nature Genetics* **47**, 1385–1392 (2015).



Supplementary Figure 1. Conditioning on polygenic predictions from genome-wide SNPs boosts association power. (**a**) Comparison of variance explained by BOLT-LMM's linear predictor—using either the default mixture-of-Gaussians prior on SNP effect sizes, which accounts for larger-effect SNPs [3], or the single-Gaussian "infinitesimal" model (BOLT-LMM-inf, equivalent to best linear unbiased prediction, BLUP)—and variance theoretically explained by an optimal linear predictor, i.e., SNP-heritability h_g^2. BOLT-LMM and BOLT-LMM-inf results (on *N*=459,327 European-ancestry samples) are from out-of-sample prediction performed internally by the BOLT-LMM software (holding out 20% of samples for testing). (**b**) Boost in effective sample size using BOLT-LMM or BOLT-LMM-inf on *N*=459,327 European samples vs. linear regression on *N*=337,539 unrelated British samples, as assessed by multiplicative increase in χ^2 statistics at associated SNPs (Supplementary Note). Numerical data are provided in Supplementary Table 3.



Supplementary Figure 2. LD score regression intercepts. Plotted points correspond to analyses of 23 phenotypes using 3 association methods (under the baselineLD model [8] of LDSC). (**a**) LD score regression intercepts [7] tend to rise with SNP-heritability and sample size, even for association tests robust to confounding (e.g., linear regression on N=337,539 unrelated British samples using 20 principal component covariates and BOLT-LMM on N=459,327 European samples). This behavior of the LDSC intercept makes test statistic inflation difficult to discern based on the value of the LDSC intercept alone: for example, for the height phenotype, linear regression on N=337,539 unrelated British samples *without* principal component covariates—which is susceptible to inflation—and BOLT-LMM on N=459,327 European samples both have LDSC intercepts of nearly 1.5. (**b**) Accounting for differences in mean χ^2 statistic for different phenotypes and association methods improves the interpretability of the LDSC intercept. Deviations from the theoretical model assumed by LD score regression (e.g., attenuation bias [7]) push the LDSC intercept above 1—even for uninflated association tests—toward the mean χ^2 test statistic (which can be much larger than 1 for highly-powered analyses of traits with substantial heritability, e.g., $\sim$7 for BOLT-LMM analysis of height, after excluding SNPs explaining $>$0.1% of variance). The attenuation ratio, (intercept – 1) / (mean χ^2 – 1), calibrates the intercept against the overall shift in χ^2 statistics (due to polygenicity for uninflated association tests). In these data, we observe that for each trait, PC-corrected linear regression and BOLT-LMM (connected by a line segment) have near-identical attenuation ratios, typically around 0.08, whereas uncorrected linear regression typically has larger attenuation ratios, indicating confounding. Numerical data are provided in Supplementary Table 4.

Supplementary Table 1. Number of phenotyped individuals analyzed per UK Biobank trait.

Phenotype	*N*	Fraction phenotyped	Case fraction
Height	458303	1.00	–
Body mass index	457824	1.00	–
Waist hip ratio	458417	1.00	–
Bone mineral density	445921	0.97	–
Forced vital capacity	371949	0.81	–
FEV1 FVC ratio	371949	0.81	–
Red blood cell count	445174	0.97	–
RBC distribution width	442700	0.96	–
White blood cell count	444502	0.97	–
Platelet count	444382	0.97	–
Eosinophil count	439938	0.96	–
Blood pressure (systolic)	422771	0.92	–
Cardiovascular disease	459324	1.00	0.319
Type 2 diabetes	459324	1.00	0.042
Respiratory disease	459324	1.00	0.140
Allergy or eczema	458699	1.00	0.230
Hypothyroidism	459324	1.00	0.049
Neuroticism	372066	0.81	–
Chronotype (morning person)	410520	0.89	0.625
Hair color	452720	0.99	–
Tanning ability	449984	0.98	0.610
Years of education	454813	0.99	–
Smoking status	457683	1.00	–
Self-reported white, QC pass	459327	–	–

Phenotypes we analyzed were available for large majorities of the 459,327 UK Biobank participants we analyzed who self-reported white ancestry and passed genotyping QC (Supplementary Note). Throughout this manuscript, when we refer to analyses of 459K European-ancestry individuals, we take it to be understood that the actual number of individuals analyzed per phenotype is slightly smaller than 459K and varies depending on phenotyping rate.

Supplementary Table 2. Number of independent GWAS loci identified by different association analysis methods.

		N=337K unrelated British			N=459K all European		
		Linear	BOLT-LMM		Linear	BOLT-LMM	
Phenotype	h_g^2	regression	inf.	non-inf.	regression*	inf.	non-inf.
Height	0.579	1086	1479	1540	1488	1992	2098
Body mass index	0.308	300	379	387	540	645	665
Waist hip ratio	0.210	217	241	255	326	365	384
Bone mineral density	0.401	537	681	713	758	947	978
Forced vital capacity	0.277	203	244	251	347	406	412
FEV1 FVC ratio	0.313	308	368	391	472	552	566
Red blood cell count	0.324	406	485	505	597	697	714
RBC distribution width	0.288	354	387	418	467	544	570
White blood cell count	0.272	347	387	404	499	555	584
Platelet count	0.404	558	694	751	766	955	1007
Eosinophil count	0.277	342	403	414	506	576	625
Blood pressure	0.271	282	332	346	461	516	522
Cardiovascular disease	0.160	126	131	135	206	210	213
Type 2 diabetes	0.074	38	39	41	59	62	62
Respiratory disease	0.086	46	46	50	69	75	76
Allergy or eczema	0.120	99	99	99	145	149	153
Hypothyroidism	0.088	69	67	70	109	112	111
Neuroticism	0.156	36	41	43	74	75	78
Chronotype	0.143	53	54	57	95	100	101
Hair color	0.454	210	273	326	263	352	436
Tanning ability	0.242	95	105	113	121	129	136
Years of education	0.193	95	89	91	207	165	172
Smoking status	0.134	32	37	36	88	93	96
All phenotypes		5839	7061	7436	8663	10272	10759

Counts of independent genome-wide significant associations ($p<5\times10^{-9}$) are reported for three types of association tests: linear regression using 20 principal component covariates, BOLT-LMM using a Gaussian SNP effect prior (the standard "infinitesimal" mixed model, BOLT-LMM-inf), and BOLT-LMM using its default mixture-of-Gaussians prior on SNP effect sizes, which accounts for larger-effect SNPs [3]. We tested all three methods on N=337,539 unrelated British samples and on all individuals who reported white ethnicity (N=459,327 European-ancestry samples). Linear regression on the full N=459K European-ancestry set is denoted with an asterisk, as these analyses are expected to be slightly confounded by relatedness. For reference, we also report SNP-heritability (h_g^2) estimated by BOLT-LMM on the N=337K unrelated British samples during model-fitting.

For each analysis, we counted independent associations by performing stringent LD clumping (requiring $R^2<0.01$ in 5Mb windows) and further collapsing associated SNPs within 100kb of each other (Supplementary Note).

Supplementary Table 3. Conditioning on polygenic predictions boosts association power.

(a) Proportion of variance explained (BOLT-LMM prediction R^2 in cross-validation and h_g^2)

	N=337K unrelated British			N=459K all European		
Phenotype	BOLT-LMM-inf	BOLT-LMM	h_g^2	BOLT-LMM-inf	BOLT-LMM	h_g^2
Height	0.332	0.397	0.579	0.373	0.429	0.570
Body mass index	0.111	0.122	0.308	0.129	0.139	0.303
Waist hip ratio	0.055	0.075	0.210	0.063	0.084	0.205
Bone mineral density	0.168	0.236	0.401	0.191	0.255	0.391
Forced vital capacity	0.079	0.089	0.277	0.095	0.108	0.272
FEV1 FVC ratio	0.102	0.136	0.313	0.115	0.147	0.303
Red blood cell count	0.119	0.166	0.324	0.137	0.181	0.314
RBC distribution width	0.101	0.166	0.288	0.119	0.180	0.279
White blood cell count	0.091	0.124	0.272	0.105	0.137	0.266
Platelet count	0.170	0.255	0.404	0.201	0.275	0.394
Eosinophil count	0.091	0.139	0.277	0.108	0.154	0.272
Blood pressure	0.079	0.101	0.271	0.096	0.116	0.264
Cardiovascular disease	0.034	0.042	0.160	0.039	0.049	0.155
Type 2 diabetes	0.009	0.014	0.074	0.010	0.015	0.073
Respiratory disease	0.013	0.019	0.086	0.014	0.020	0.083
Allergy or eczema	0.022	0.033	0.120	0.025	0.035	0.115
Hypothyroidism	0.012	0.023	0.088	0.014	0.024	0.085
Neuroticism	0.027	0.028	0.156	0.033	0.036	0.151
Chronotype	0.024	0.027	0.143	0.030	0.034	0.138
Hair color	0.234	0.397	0.454	0.257	0.401	0.434
Tanning ability	0.080	0.173	0.242	0.092	0.177	0.226
Years of education	0.050	0.052	0.193	0.061	0.064	0.188
Smoking status	0.025	0.026	0.134	0.033	0.035	0.136

(b) Boost in effective sample size (vs. linear regression on N=337K unrelated British samples)

	N=337K unrelated British		N=459K all European		
Phenotype	BOLT-LMM-inf	BOLT-LMM	BOLT-LMM-inf	BOLT-LMM	BOLT-LMM N_{eff}
Height	1.37x	1.45x	1.83x	1.93x	650K
Body mass index	1.14x	1.15x	1.45x	1.47x	500K
Waist hip ratio	1.06x	1.08x	1.37x	1.40x	470K
Bone mineral density	1.21x	1.29x	1.62x	1.71x	580K
Forced vital capacity	1.10x	1.11x	1.43x	1.44x	490K
FEV1 FVC ratio	1.13x	1.17x	1.48x	1.53x	520K
Red blood cell count	1.14x	1.19x	1.50x	1.56x	530K
RBC distribution width	1.12x	1.19x	1.45x	1.55x	520K
White blood cell count	1.09x	1.11x	1.40x	1.43x	480K
Platelet count	1.23x	1.33x	1.66x	1.79x	600K
Eosinophil count	1.11x	1.17x	1.46x	1.53x	520K
Blood pressure	1.13x	1.15x	1.42x	1.44x	480K
Cardiovascular disease	1.04x	1.04x	1.32x	1.33x	450K
Type 2 diabetes	1.01x	1.01x	1.33x	1.33x	450K
Respiratory disease	1.02x	1.02x	1.30x	1.32x	440K
Allergy or eczema	1.05x	1.05x	1.29x	1.30x	440K
Hypothyroidism	1.03x	1.05x	1.31x	1.29x	430K
Neuroticism	1.02x	1.02x	1.34x	1.34x	450K
Chronotype	1.04x	1.05x	1.33x	1.34x	450K
Hair color	1.30x	1.60x	1.74x	2.08x	700K
Tanning ability	1.06x	1.17x	1.38x	1.51x	510K
Years of education	1.02x	1.03x	1.31x	1.31x	440K
Smoking status	1.04x	1.05x	1.27x	1.28x	430K

See caption of Supplementary Fig. 1.

Supplementary Table 4. LDSC intercepts increase with mean χ^2 statistics while attenuation ratios are consistently low for BOLT-LMM and linear regression with PC covariates.

(a) LDSC intercept (jackknife s.e.) and mean χ^2 statistic for different association methods

		LDSC intercept (baselineLD model)			Mean χ^2 statistic		
Phenotype	h_g^2	Lin. reg. w/o PCs	LR, N=337K	BOLT, 459K	LR w/o PCs	LR, 337K	BOLT, 459K
Height	0.579	1.456 (0.035)	1.252 (0.033)	1.468 (0.057)	4.48	4.19	7.00
Body mass index	0.308	1.121 (0.016)	1.090 (0.015)	1.132 (0.019)	2.68	2.63	3.45
Waist hip ratio	0.210	1.088 (0.016)	1.062 (0.015)	1.091 (0.019)	2.05	2.00	2.44
Bone mineral density	0.401	1.129 (0.034)	1.117 (0.034)	1.250 (0.048)	2.88	2.86	4.01
Forced vital capacity	0.277	1.095 (0.013)	1.089 (0.013)	1.113 (0.015)	2.14	2.13	2.70
FEV1 FVC ratio	0.313	1.071 (0.015)	1.065 (0.016)	1.119 (0.021)	2.20	2.18	2.89
Red blood cell count	0.324	1.183 (0.025)	1.142 (0.026)	1.189 (0.036)	2.46	2.39	3.23
RBC distribution width	0.288	1.078 (0.025)	1.067 (0.025)	1.151 (0.032)	2.26	2.24	2.83
White blood cell count	0.272	1.120 (0.018)	1.097 (0.018)	1.137 (0.021)	2.28	2.24	2.89
Platelet count	0.404	1.163 (0.029)	1.154 (0.029)	1.281 (0.044)	2.68	2.66	3.96
Eosinophil count	0.277	1.108 (0.021)	1.082 (0.021)	1.165 (0.030)	2.26	2.23	2.89
Blood pressure	0.271	1.084 (0.015)	1.079 (0.015)	1.128 (0.018)	2.24	2.23	2.84
Cardiovascular disease	0.160	1.056 (0.013)	1.053 (0.013)	1.083 (0.016)	1.77	1.76	2.03
Type 2 diabetes	0.074	1.040 (0.012)	1.040 (0.012)	1.043 (0.015)	1.31	1.31	1.41
Respiratory disease	0.086	1.024 (0.011)	1.019 (0.011)	1.040 (0.013)	1.37	1.36	1.49
Allergy or eczema	0.120	1.034 (0.014)	1.026 (0.014)	1.043 (0.016)	1.51	1.50	1.70
Hypothyroidism	0.088	1.033 (0.013)	1.029 (0.013)	1.036 (0.013)	1.37	1.36	1.48
Neuroticism	0.156	1.034 (0.015)	1.028 (0.015)	1.069 (0.011)	1.66	1.65	1.85
Chronotype	0.143	1.072 (0.012)	1.043 (0.012)	1.055 (0.013)	1.66	1.62	1.84
Hair color	0.454	1.218 (0.057)	1.139 (0.054)	1.224 (0.078)	1.92	1.82	2.52
Tanning ability	0.242	1.209 (0.032)	1.071 (0.029)	1.105 (0.042)	1.73	1.53	1.81
Years of education	0.193	1.152 (0.013)	1.089 (0.012)	1.112 (0.013)	2.09	1.98	2.26
Smoking status	0.134	1.088 (0.011)	1.032 (0.010)	1.057 (0.011)	1.76	1.67	1.92

(b) LDSC attenuation ratio: (intercept – 1) / (mean χ^2 – 1); jackknife s.e.

Phenotype	LR w/o PCs, 337K	LR, 337K	BOLT, 337K	LR, 459K*	BOLT, 459K
Height	0.131 (0.010)	0.079 (0.010)	0.079 (0.010)	0.084 (0.009)	0.078 (0.009)
Body mass index	0.072 (0.009)	0.055 (0.009)	0.056 (0.009)	0.061 (0.008)	0.054 (0.008)
Waist hip ratio	0.084 (0.015)	0.061 (0.015)	0.067 (0.015)	0.060 (0.013)	0.063 (0.013)
Bone mineral density	0.069 (0.018)	0.063 (0.018)	0.080 (0.017)	0.072 (0.017)	0.083 (0.016)
Forced vital capacity	0.083 (0.012)	0.079 (0.012)	0.076 (0.011)	0.080 (0.010)	0.066 (0.009)
FEV1 FVC ratio	0.060 (0.013)	0.055 (0.013)	0.058 (0.013)	0.063 (0.011)	0.063 (0.011)
Red blood cell count	0.125 (0.018)	0.102 (0.018)	0.094 (0.018)	0.096 (0.016)	0.085 (0.016)
RBC distribution width	0.062 (0.020)	0.054 (0.020)	0.076 (0.020)	0.068 (0.018)	0.082 (0.017)
White blood cell count	0.093 (0.014)	0.078 (0.014)	0.079 (0.013)	0.078 (0.012)	0.072 (0.011)
Platelet count	0.097 (0.017)	0.093 (0.017)	0.092 (0.017)	0.096 (0.016)	0.095 (0.015)
Eosinophil count	0.085 (0.017)	0.067 (0.017)	0.079 (0.017)	0.087 (0.015)	0.087 (0.016)
Blood pressure	0.067 (0.012)	0.065 (0.012)	0.066 (0.011)	0.075 (0.010)	0.069 (0.010)
Cardiovascular disease	0.073 (0.017)	0.070 (0.017)	0.073 (0.017)	0.080 (0.015)	0.081 (0.016)
Type 2 diabetes	0.130 (0.040)	0.129 (0.041)	0.129 (0.041)	0.114 (0.036)	0.105 (0.036)
Respiratory disease	0.064 (0.030)	0.052 (0.031)	0.055 (0.030)	0.085 (0.026)	0.080 (0.026)
Allergy or eczema	0.067 (0.028)	0.051 (0.029)	0.066 (0.027)	0.063 (0.024)	0.061 (0.023)
Hypothyroidism	0.089 (0.035)	0.079 (0.035)	0.121 (0.029)	0.067 (0.029)	0.076 (0.026)
Neuroticism	0.051 (0.024)	0.043 (0.024)	0.068 (0.017)	0.065 (0.017)	0.082 (0.013)
Chronotype	0.109 (0.018)	0.070 (0.019)	0.075 (0.019)	0.067 (0.015)	0.066 (0.015)
Hair color	0.236 (0.062)	0.170 (0.066)	0.141 (0.049)	0.185 (0.065)	0.147 (0.051)
Tanning ability	0.288 (0.045)	0.134 (0.054)	0.120 (0.053)	0.146 (0.053)	0.129 (0.051)
Years of education	0.140 (0.012)	0.091 (0.012)	0.090 (0.013)	0.100 (0.010)	0.089 (0.010)
Smoking status	0.117 (0.015)	0.047 (0.015)	0.062 (0.015)	0.064 (0.012)	0.062 (0.012)

See caption of Supplementary Fig. 2.

Supplementary Table 5. LDSC intercepts and attenuation ratios are higher under the original LDSC model (vs. baselineLD model).

(a) LDSC intercept (jackknife s.e.) and mean χ^2 statistic for different association methods

		LDSC intercept (original LDSC model)			Mean χ^2 statistic		
Phenotype	$h_g{}^2$	Lin. reg. w/o PCs	LR, *N*=337K	BOLT, 459K	LR w/o PCs	LR, 337K	BOLT, 459K
Height	0.579	1.706 (0.031)	1.493 (0.030)	1.870 (0.043)	4.70	4.40	7.62
Body mass index	0.308	1.235 (0.017)	1.202 (0.016)	1.318 (0.019)	2.69	2.64	3.48
Waist hip ratio	0.210	1.217 (0.015)	1.185 (0.015)	1.267 (0.018)	2.05	2.01	2.44
Bone mineral density	0.401	1.315 (0.022)	1.305 (0.022)	1.497 (0.028)	3.12	3.10	4.51
Forced vital capacity	0.277	1.194 (0.014)	1.190 (0.014)	1.267 (0.017)	2.14	2.13	2.70
FEV1 FVC ratio	0.313	1.218 (0.015)	1.212 (0.014)	1.326 (0.018)	2.24	2.22	2.96
Red blood cell count	0.324	1.323 (0.020)	1.272 (0.019)	1.393 (0.026)	2.58	2.51	3.44
RBC distribution width	0.288	1.181 (0.017)	1.171 (0.017)	1.274 (0.022)	2.42	2.40	3.15
White blood cell count	0.272	1.264 (0.018)	1.238 (0.018)	1.349 (0.023)	2.34	2.30	2.97
Platelet count	0.404	1.283 (0.020)	1.272 (0.020)	1.461 (0.026)	2.85	2.83	4.35
Eosinophil count	0.277	1.232 (0.019)	1.207 (0.019)	1.324 (0.024)	2.38	2.36	3.15
Blood pressure	0.271	1.200 (0.014)	1.195 (0.013)	1.303 (0.017)	2.24	2.23	2.85
Cardiovascular disease	0.160	1.127 (0.012)	1.123 (0.012)	1.181 (0.015)	1.77	1.76	2.03
Type 2 diabetes	0.074	1.066 (0.009)	1.065 (0.009)	1.084 (0.010)	1.31	1.31	1.42
Respiratory disease	0.086	1.076 (0.010)	1.071 (0.010)	1.103 (0.011)	1.37	1.36	1.49
Allergy or eczema	0.120	1.114 (0.011)	1.107 (0.011)	1.148 (0.012)	1.51	1.50	1.70
Hypothyroidism	0.088	1.081 (0.011)	1.078 (0.011)	1.103 (0.012)	1.37	1.36	1.49
Neuroticism	0.156	1.079 (0.010)	1.074 (0.010)	1.113 (0.010)	1.66	1.65	1.85
Chronotype	0.143	1.114 (0.011)	1.082 (0.010)	1.103 (0.011)	1.66	1.62	1.84
Hair color	0.454	1.212 (0.017)	1.133 (0.015)	1.238 (0.021)	3.02	2.89	4.80
Tanning ability	0.242	1.219 (0.013)	1.075 (0.011)	1.109 (0.013)	2.35	2.12	2.70
Years of education	0.193	1.216 (0.012)	1.147 (0.011)	1.187 (0.012)	2.09	1.98	2.26
Smoking status	0.134	1.149 (0.010)	1.085 (0.010)	1.125 (0.011)	1.76	1.67	1.92

(b) LDSC attenuation ratio: (intercept – 1) / (mean χ^2 – 1); jackknife s.e.

Phenotype	LR w/o PCs, 337K	LR, 337K	BOLT, 459K
Height	0.191 (0.008)	0.145 (0.009)	0.132 (0.006)
Body mass index	0.139 (0.010)	0.123 (0.010)	0.128 (0.008)
Waist hip ratio	0.206 (0.015)	0.184 (0.015)	0.185 (0.013)
Bone mineral density	0.148 (0.010)	0.145 (0.010)	0.142 (0.008)
Forced vital capacity	0.170 (0.013)	0.168 (0.013)	0.157 (0.010)
FEV1 FVC ratio	0.176 (0.012)	0.173 (0.012)	0.166 (0.009)
Red blood cell count	0.204 (0.013)	0.180 (0.013)	0.161 (0.011)
RBC distribution width	0.127 (0.012)	0.122 (0.012)	0.128 (0.010)
White blood cell count	0.197 (0.014)	0.184 (0.014)	0.177 (0.012)
Platelet count	0.153 (0.011)	0.149 (0.011)	0.138 (0.008)
Eosinophil count	0.168 (0.014)	0.153 (0.014)	0.150 (0.011)
Blood pressure	0.162 (0.011)	0.159 (0.011)	0.164 (0.009)
Cardiovascular disease	0.165 (0.016)	0.162 (0.016)	0.177 (0.014)
Type 2 diabetes	0.212 (0.029)	0.211 (0.030)	0.202 (0.025)
Respiratory disease	0.207 (0.028)	0.198 (0.028)	0.209 (0.022)
Allergy or eczema	0.223 (0.022)	0.211 (0.022)	0.211 (0.017)
Hypothyroidism	0.220 (0.029)	0.214 (0.029)	0.212 (0.025)
Neuroticism	0.120 (0.015)	0.113 (0.015)	0.134 (0.012)
Chronotype	0.172 (0.016)	0.134 (0.016)	0.123 (0.013)
Hair color	0.104 (0.008)	0.070 (0.008)	0.063 (0.005)
Tanning ability	0.162 (0.009)	0.067 (0.010)	0.064 (0.008)
Years of education	0.199 (0.011)	0.149 (0.011)	0.148 (0.010)
Smoking status	0.197 (0.014)	0.127 (0.015)	0.137 (0.012)

Compare to Supplementary Table 4.

Supplementary Table 6. Running time of association methods on UK Biobank data.

Data set	Linear regression	BOLT-LMM v1	BOLT-LMM v2.3
N=150K	0.49 days	2.43 days	0.62 days
N=500K	1.62 days	9.34 days	2.54 days

Run time benchmarks for association analyses using BOLT-LMM v2.3 (with 20 principal component covariates to increase convergence rate; Supplementary Table 7), the previous version of BOLT-LMM [3], and linear regression using 20 principal component covariates (implemented efficiently within the BOLT-LMM software; cf. Bycroft et al. Table S9 [5]). We analyzed the years-of-education phenotype as a representative trait, and we ran all methods on the same set of all European-ancestry individuals in the UK Biobank N=150K and N=500K data releases (Supplementary Note), analyzing $\sim$72M and $\sim$93M imputed SNPs, respectively, and imposing a MAF$>$0.1% filter on minor allele frequency. Analyses used 8 threads on a 2.10 GHz Intel Xeon E5-2683 v4 processor.

Supplementary Table 7. Faster convergence of BOLT-LMM iterative computations using principal component covariates.

Principal component covariates	Conjugate gradient iterations	Variational Bayes iterations
0	85	178
10	75	82
20	63	79
30	55	80
40	49	73

During its model-fitting steps, BOLT-LMM applies iterative methods (specifically, conjugate gradient iteration and variational Bayes) to eliminate computationally expensive matrix operations that scale quadratically or cubically with sample size [3]. The cost of a single iteration scales only linearly with N; however, we previously observed that the number of iterations required to achieve convergence increases slowly with N [3]. Our analyses here demonstrate that convergence can be sped up by including principal component covariates (which effectively reduce the condition number of the underlying matrix computations), thus achieving close-to-linear scaling of run time with sample size. The iteration counts reported in this table are for total numbers of iterations performed in BOLT-LMM's conjugate gradient steps (for estimating parameters and fitting the infinitesimal mixed model) and variational Bayes steps (for estimating parameters and fitting the mixture-of-Gaussians model) [3]. We note that to achieve increased convergence, principal components need to be computed on the set of SNPs used in the mixed model; PCs that do not match the implicit genetic relationship matrix (GRM) will not improve conditioning. We also note that after model-fitting, BOLT-LMM performs a linear-time association test on imputed SNPs; the speedup described here only applies to the model-fitting step.

Supplementary Table 8. Type I error inflation of BOLT-LMM when testing rare variants in unbalanced case-control settings.

Case		*N*=450K			*N*=150K			*N*=50K		
frac.	MAF	α=1e-4	α=1e-6	α=5e-8	α=1e-4	α=1e-6	α=5e-8	α=1e-4	α=1e-6	α=5e-8
0.3	0.1	0.00012	1.2e-6	1.7e-8	0.00009	4.6e-7	1.0e-7	0.00010	2.9e-7	0
	0.01	0.00009	5.4e-7	0	0.00011	4.3e-7	2.2e-8	0.00011	1.5e-6	4.3e-8
	0.001	0.00010	8.6e-7	1.5e-8	0.00010	1.1e-6	8.9e-8*	0.00010	1.1e-6	1.0e-7*
	0.0001	0.00010	1.3e-6*	7.1e-8**	0.00010*	1.5e-6***	1.5e-7***	0.00011***	1.4e-6**	8.6e-8**
0.1	0.1	0.00012	1.7e-6	0	0.00010	8.7e-7	0	0.00010	9.2e-7	1.7e-8
	0.01	0.00010	1.4e-6	0	0.00012	1.5e-6	4.3e-8	0.00012*	1.4e-6	8.7e-8
	0.001	0.00010*	1.4e-6*	3.0e-8	0.00012***	2.7e-6***	4.0e-7***	0.00018***	5.5e-6***	5.3e-7***
	0.0001	0.00019***	6.0e-6***	7.7e-7***	0.00035***	1.9e-5***	3.0e-6***	0.00074***	5.8e-5***	1.3e-5***
0.03	0.1	0.00012	2.1e-6	3.4e-8	0.00012	1.8e-6	1.2e-7	0.00010	7.9e-7	0
	0.01	0.00010	8.9e-7	0	0.00011	1.6e-6	2.2e-8	0.00014**	2.3e-6**	8.7e-8
	0.001	0.00015***	4.1e-6***	5.5e-7*	0.00022***	9.6e-6***	1.2e-6***	0.00045***	3.2e-5***	7.6e-6***
	0.0001	0.00050***	3.8e-5***	8.1e-6***	0.00115***	1.4e-4***	4.1e-5***	0.00277***	5.4e-4***	2.1e-4***
0.01	0.1	0.00009	8.0e-7	1.7e-8	0.00010	4.1e-7	1.2e-7	0.00011	9.4e-7	1.4e-7
	0.01	0.00011	4.8e-7	4.3e-8	0.00014***	3.7e-6*	1.7e-7	0.00019***	6.5e-6***	5.0e-7**
	0.001	0.00024***	1.0e-5***	1.3e-6***	0.00050***	3.4e-5***	6.8e-6***	0.00112***	1.4e-4***	4.0e-5***
	0.0001	0.00127***	1.7e-4***	5.3e-5***	0.00303***	6.4e-4***	2.6e-4***	0.00672***	2.2e-3***	1.2e-3***
0.003	0.1	0.00009	6.3e-7	6.8e-8	0.00011	4.8e-6	0	0.00012	2.1e-6	1.7e-8
	0.01	0.00013***	2.1e-6**	1.5e-7*	0.00017***	4.8e-6***	8.3e-7***	0.00039***	2.3e-5***	3.1e-6***
	0.001	0.00056***	5.1e-5***	1.1e-5***	0.00128***	1.7e-4***	5.1e-5***	0.00302***	6.3e-4***	2.6e-4***
	0.0001	0.00345***	8.0e-4***	3.4e-4***	0.00737***	2.5e-3***	1.4e-3***	0.01764***	8.3e-3***	5.1e-3***
0.001	0.1	0.00011	1.3e-6	1.0e-7	0.00010	9.9e-7	3.4e-8	0.00014**	2.6e-6	6.8e-8
	0.01	0.00019***	5.9e-6***	5.9e-7**	0.00037***	2.1e-5***	3.0e-6***	0.00086***	1.0e-4***	2.6e-5***
	0.001	0.00128***	1.7e-4***	5.1e-5***	0.00301***	6.5e-4***	2.6e-4***	0.00640***	2.3e-3***	1.2e-3***
	0.0001	0.00743***	2.5e-3***	1.4e-3***	0.01755***	8.3e-3***	5.2e-3***	0.01954***	1.3e-2***	1.0e-2***

This table presents results from a suite of simulations that vary the three key parameters affecting type I error control of BOLT-LMM (and in general, chi-square-based regression tests) on case-control traits: sample size (*N*), minor allele frequency (MAF), and case fraction. For each combination of *N* and case fraction, we simulated 20 binary traits with heritable variance distributed across odd-numbered chromosomes, and we assessed MAF-stratified type I error rates for association tests on SNPs on even chromosomes (Supplementary Note Section 6). Type I error rates with statistically significant inflation are indicated with asterisks (* = p<0.05, ** = p<0.01, *** = p<0.001; z-test across 20 simulation replicates). All tests were two-sided.

Supplementary Table 9. Comparison of heritability parameter estimates from REML and LDSC.

	REML h_g^2 (s.e.)	LDSC heritability parameter (jackknife s.e.)	
Phenotype		baselineLD model	original model
Height	0.578 (0.002)	0.476 (0.019)	0.377 (0.019)
Body mass index	0.309 (0.003)	0.247 (0.007)	0.194 (0.007)
Waist hip ratio	0.212 (0.002)	0.156 (0.007)	0.107 (0.007)
Bone mineral density	0.404 (0.003)	0.284 (0.017)	0.246 (0.024)
Forced vital capacity	0.277 (0.003)	0.206 (0.007)	0.151 (0.006)
FEV1 FVC ratio	0.313 (0.003)	0.237 (0.011)	0.164 (0.010)
Red blood cell count	0.323 (0.003)	0.214 (0.011)	0.170 (0.016)
RBC distribution width	0.289 (0.003)	0.195 (0.013)	0.168 (0.015)
White blood cell count	0.273 (0.003)	0.193 (0.007)	0.139 (0.010)
Platelet count	0.404 (0.003)	0.260 (0.013)	0.227 (0.021)
Eosinophil count	0.278 (0.003)	0.198 (0.011)	0.146 (0.013)
Blood pressure	0.272 (0.003)	0.206 (0.008)	0.147 (0.007)
Cardiovascular disease	0.159 (0.002)	0.117 (0.005)	0.081 (0.004)
Type 2 diabetes	0.073 (0.002)	0.045 (0.003)	0.031 (0.002)
Respiratory disease	0.084 (0.002)	0.057 (0.004)	0.035 (0.003)
Allergy or eczema	0.119 (0.002)	0.081 (0.006)	0.051 (0.004)
Hypothyroidism	0.086 (0.002)	0.053 (0.005)	0.034 (0.004)
Neuroticism	0.156 (0.003)	0.119 (0.008)	0.091 (0.005)
Chronotype	0.142 (0.002)	0.104 (0.005)	0.079 (0.004)
Hair color	0.453 (0.002)	0.115 (0.017)	0.271 (0.104)
Tanning ability	0.241 (0.002)	0.072 (0.012)	0.138 (0.056)
Years of education	0.194 (0.002)	0.148 (0.005)	0.113 (0.004)
Smoking status	0.134 (0.002)	0.104 (0.004)	0.077 (0.003)

This table compares SNP-heritability h_g^2 (estimated by BOLT-LMM using restricted maximum likelihood) with the heritability parameter estimated by LDSC (using either the baselineLD model [8] or the original LDSC model [7]) in analyses of N=337K unrelated British individuals. We note that LDSC does not estimate the quantity traditional called "SNP-heritability" and denoted h_g^2; instead LDSC estimates the "causal heritability contributed by common SNPs, excluding those of large effect" (Supplementary Note Section 4). We also note that the values in the REML h_g^2 column are very slightly different from those reported in preceding Supplementary Tables, which were computed in BOLT-LMM association analyses during model-fitting; we computed the estimates in this table using the BOLT-REML [32] algorithm in order to also obtain standard errors.