Efficient Bayesian mixed model analysis increases association power in large cohorts

Published in final edited form as:
Nat Genet. 2015 March ; 47(3): 284–290. doi:10.1038/ng.3190.

Efficient Bayesian mixed model analysis increases association power in large cohorts

Po-Ru Loh[1,2], **George Tucker**[1,3,4], **Brendan K Bulik-Sullivan**[2,5], **Bjarni J Vilhjálmsson**[1,2], **Hilary K Finucane**[3], **Rany M Salem**[2,6], **Daniel I Chasman**[7], **Paul M Ridker**[7], **Benjamin M Neale**[2,5], **Bonnie Berger**[3,4], **Nick Patterson**[2], and **Alkes L Price**[1,2,8]

[1]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA.

[2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.

[3]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[4]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA.

[5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA.

[6]Department of Endocrinology, Children's Hospital Boston, Boston, Massachusetts, USA.

[7]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA.

[8]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA.

Abstract

Linear mixed models are a powerful statistical tool for identifying genetic associations and avoiding confounding. However, existing methods are computationally intractable in large cohorts, and may not optimize power. All existing methods require time cost $O(MN^2)$ (where N = #samples and M = #SNPs) and implicitly assume an infinitesimal genetic architecture in which effect sizes are normally distributed, which can limit power. Here, we present a far more efficient mixed model association method, BOLT-LMM, which requires only a small number of O(MN)-time iterations and increases power by modeling more realistic, non-infinitesimal genetic architectures via a Bayesian mixture prior on marker effect sizes. We applied BOLT-LMM to nine quantitative traits in 23,294 samples from the Women's Genome Health Study (WGHS) and observed significant increases in power, consistent with simulations. Theory and simulations show

Correspondence should be addressed to P.-R.L. (loh@hsph.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu).

URLs.

BOLT-LMM software and source code, http://www.hsph.harvard.edu/alkes-price/software/.
LTMLM method, http://biorxiv.org/content/early/2014/09/04/008755.

Author Contributions

P. L., N. P., and A. L. P. designed experiments. P. L. performed experiments. P. L., G. T., B. K. B., B. J. V, H. K. F., and A. L. P. analyzed data. D. I. C. and P. M. R. provided data. All authors wrote the paper.

Competing Financial Interests

The authors declare no competing financial interests.

NIH-PA Author Manuscript
NIH-PA Author Manuscript

that the boost in power increases with cohort size, making BOLT-LMM appealing for GWAS in large cohorts.

Linear mixed models are emerging as the method of choice for association testing in genome-wide association studies (GWAS) because they account for both population stratification and cryptic relatedness and achieve increased statistical power by jointly modeling all genotyped markers[1–12]. However, existing mixed model methods still have limitations. First, mixed model analysis is computationally expensive. Despite a series of recent algorithmic advances, current algorithms require either $O(MN^2)$ or $O(M^2N)$ total running time, where M is the number of markers and N is the sample size. This cost is becoming prohibitive for large cohorts, forcing existing methods to subsample the markers so that $M<N$ (ref.[5]). Second, current mixed model methods fall short of achieving maximal statistical power owing to suboptimal modeling assumptions regarding the genetic architectures underlying phenotypes. The standard linear mixed model implicitly assumes that all variants are causal with small effect sizes drawn from independent Gaussian distributions—the “infinitesimal model”—whereas in reality, complex traits are estimated to have roughly a few thousand causal loci[13,14].

Methodologically, efforts to more accurately model non-infinitesimal genetic architectures have followed two general thrusts. One approach is to apply the standard infinitesimal mixed model but adapt the input data. For example, large-effect loci can be explicitly identified and conditioned out as fixed effects[7], or the mixed model can be applied to only a selected subset of markers[9,11,15,16]. A more flexible alternative approach is to adapt the mixed model itself by taking a Bayesian perspective and modeling SNP effects with non-Gaussian prior distributions that better accommodate both small- and large-effect loci. Such methods were pioneered in livestock genetics to improve prediction of genetic values[17] and have been extensively developed in the plant and animal breeding literature for the purpose of genomic selection[18]. These techniques are of interest in the association testing setting because models that improve prediction should in theory enable corresponding improvements in association power (via conditioning on other associated loci when testing a candidate marker[9,12]). Here, we present an algorithm that performs mixed model analysis in a small number of $O(MN)$-time iterations and increases power by modeling non-infinitesimal genetic architectures. Our algorithm fits a Gaussian mixture model of SNP effects[19], using a fast variational approximation[20–22] to compute approximate phenotypic residuals, and tests the residuals for association with candidate markers via a retrospective score statistic[23] that provides a bridge between Bayesian modeling for phenotype prediction and the frequentist association testing framework. We calibrate our statistic using an approach based on the recently developed LD Score regression technique[24]. The entire procedure operates directly on raw genotypes stored compactly in memory and does not require computing or storing a genetic relationship matrix. In the special case of the infinitesimal model, we achieve results equivalent to existing methods at dramatically reduced time and memory cost.

We provide an efficient software implementation of our algorithm, BOLT-LMM, and demonstrate its computational efficiency on simulated data sets of up to 480,000 individuals.

Our simulations also show that BOLT-LMM achieves increased association power over standard infinitesimal mixed model analysis of traits driven by a few thousand causal SNPs. We applied BOLT-LMM to perform mixed model analysis of nine quantitative traits in 23,294 samples from the Women's Genome Health Study (WGHS)[25] and observed increased association power equivalent to up to 10% increase in effective sample size. We demonstrate through theory and simulations that the power boost increases with cohort size, making BOLT-LMM a promising approach for large-scale GWAS.

Results

Overview of Methods

The BOLT-LMM algorithm consists of four main steps, each of which require a small number of $O(MN)$-time iterations. These steps are: (1a) Estimate variance parameters; (1b) Compute infinitesimal mixed model association statistics (denoted BOLT-LMM-inf); (2a) Estimate Gaussian mixture parameters; (2b) Compute Gaussian mixture model association statistics (BOLT-LMM). Step 1a computes results nearly identical to standard variance components analysis but applies a stochastic approximation algorithm[26,27] that reduces time and memory cost by circumventing spectral decomposition, which is expensive for large sample sizes. Instead, the approximation algorithm only requires solving linear systems of mixed model equations, which can be accomplished efficiently using conjugate gradient iteration[28,29]. Step 1b likewise circumvents spectral decomposition by introducing a new retrospective mixed model association statistic similar to GRAMMAR-Gamma[10] and MASTOR[23], which we compute—up to a calibration constant—using only solutions to linear systems of equations. We estimate the calibration constant by computing and comparing the new statistic and the standard prospective mixed model statistic at a random subset of SNPs, which can likewise be accomplished efficiently using conjugate gradient iteration. This procedure is similar in spirit to GRAMMAR-Gamma calibration but requires only $O(MN)$-time iterations.

Steps 2a and 2b are Gaussian mixture parallels of steps 1a and 1b. BOLT-LMM's non-infinitesimal model amounts to a generalization of the standard mixed model, which from a Bayesian perspective imposes a Gaussian prior distribution on SNP effect sizes. BOLT-LMM relaxes this assumption by using a mixture of two Gaussians as the prior, giving the model greater flexibility to accommodate large-effect SNPs while maintaining effective modeling of genome-wide effects (e.g., ancestry). Exact posterior inference is no longer tractable under the generalized model, so BOLT-LMM instead computes a variational approximation[20–22] that converges after a small number of $O(MN)$-time iterations. Step 2a applies this method within 5-fold cross-validation to estimate best-fit parameters for the prior distribution (taking into account variance parameters estimated in Step 1a) based on out-of-sample prediction accuracy. If the prediction accuracy of the best-fit Gaussian mixture model exceeds that of the infinitesimal model by at least a specified amount, Step 2b is then run to compute association statistics by testing each SNP against the residual phenotype obtained from the Gaussian mixture model and calibrating the test statistics against the results of Step 1b using LD Score regression[24]. Otherwise, the BOLT-LMM association statistic is the same as BOLT-LMM-inf. Both Step 1b and Step 2b are performed using a leave-one-chromosome-out (LOCO) scheme to avoid proximal contamination[5,9,12]. (The software also supports subdividing chromosomes into more segments; see Online Methods.) The key properties of BOLT-LMM regarding speed and modeling assumptions are compared to existing methods in Table 1.

Computational cost of BOLT-LMM versus existing methods

To analyze the computational performance of BOLT-LMM, we simulated data sets of sizes ranging from N=3,750 to 480,000 individuals and M=300,000 SNPs. We used genotypes from the WTCCC2 data set[30] analyzed in ref.[12], which contains 15,633 individuals of European ancestry, to form mosaic chromosomes, and we used a phenotype model in which 5,000 SNPs explained 20% of phenotypic variance (Supplementary Note).

We benchmarked BOLT-LMM against existing mixed model association methods, running each method for up to 10 days on machines with 96GB of memory. BOLT-LMM completed all analyses through N=480,000 individuals within these constraints, whereas previous methods could only analyze a maximum of N=7,500–30,000 individuals (Fig. 1 and Supplementary Table 1). All previous methods require $O(MN^2)$ running time (for $M>N$), whereas the running time of BOLT-LMM scales roughly with $MN^{1.5}$ (Fig. 1a and Supplementary Fig. 1a). We also observed substantial savings in memory use with BOLT-LMM (Fig. 1b and Supplementary Fig. 1b), which requires little more than the $MN/4$ bytes of memory needed to store raw genotypes (as in GenABEL software[31]).

The running time of BOLT-LMM depends not only on the cost of matrix arithmetic, which scales linearly with M and N, but also the number of $O(MN)$-time iterations required for convergence, which empirically scales roughly as $N^{0.5}$ (Supplementary Fig. 1) and also varies with heritability, relatedness, and population structure (Supplementary Note and Supplementary Fig. 2). These observations apply both to the full Gaussian mixture modeling performed by BOLT-LMM and to the subset of the computation (Steps 1a and 1b) needed to compute BOLT-LMM-inf infinitesimal mixed model association statistics, which in our benchmarks required ≈40% of the full BOLT-LMM run time (Fig. 1a and Supplementary Fig. 1a). Our results show that even on very large data sets, BOLT-LMM is efficient enough to enable mixed model analysis using a Gaussian mixture prior, which we recommend because of its potential to increase power.

Power and false positive control of BOLT-LMM in simulations

To assess the power of BOLT-LMM to detect associated loci, we performed additional simulations using real genotypes from the WTCCC2 data set, which is an ancestry-stratified sample containing both Northern and Southern European samples. We simulated phenotypes with 1,250–10,000 causal SNPs[13,14] explaining 50% of phenotypic variance and an additional 60 standardized effect SNPs explaining 2% of variance. We included the latter category of SNPs to allow direct power comparisons across different simulation setups, as the 60 standardized effect SNPs always explain the same total amount of variance regardless of other simulation parameters. We further introduced environmental differences in ancestry by including a phenotypic component aligned with the top principal component that explained an additional 1% of variance. (We note that principal component analysis is not part of BOLT-LMM; it is unnecessary to perform PCA when running mixed model association methods[12].) We chose causal SNPs randomly from the first halves of chromosomes, leaving the second halves of chromosomes to contain only non-causal SNPs (Supplementary Note).

We computed χ^2 association statistics using linear regression with 10 principal components (PCA)[32], GCTA-LOCO[12], BOLT-LMM-inf, and BOLT-LMM. We were unable to test FaST-LMM-Select[15] on this data set because of its memory requirements (Fig. 1). For each method, we computed means of its χ^2 statistics over standardized effect SNPs and compared these means across simulations involving different numbers of causal SNPs (Fig. 2a and Supplementary Table 2). We observed that BOLT-LMM achieved power gains by modeling non-infinitesimal architectures. For the sparsest genetic architecture (1,250 causal SNPs plus 60 standardized effect SNPs), we observed a 25% increase in mean BOLT-LMM χ^2 statistics at standardized effect SNPs compared to GCTA-LOCO and BOLT-LMM-inf infinitesimal mixed model χ^2 statistics. This metric is readily interpretable as corresponding to a 25% increase in effective sample size; for completeness, we also computed traditional power curves at two significance thresholds (Supplementary Fig. 3). The power gain of the Gaussian mixture model decreased with increasing numbers of causal SNPs (Fig. 2a). This behavior is expected because the advantage of the Gaussian mixture lies in its ability to more accurately model a small fraction of SNPs with larger effects amid a majority of SNPs with near-zero effects. Larger numbers of causal SNPs explaining a fixed proportion of variance result in smaller effect sizes per causal SNP, giving BOLT-LMM less opportunity for power gain. In contrast, all methods other than BOLT-LMM had performance independent of the number of causal SNPs, consistent with the fact that none of these methods model non-infinitesimal genetic architectures. GCTA-LOCO and BOLT-LMM-inf mean χ^2 statistics at standardized effect SNPs were essentially identical and slightly exceeded PCA, consistent with theory[12]. We also tested EMMAX[3] and GEMMA[6], which are vulnerable to proximal contamination[5,9,12]; these methods suffered loss of power relative to PCA (Supplementary Fig. 4a), consistent with theory[12].

To further explore the relationship between the magnitude of Gaussian mixture model power gain and other parameters of the data set, we also varied the proportion of variance explained by causal SNPs (Fig. 2b) and the number of individuals (Fig. 2c). We observed that the power boost of BOLT-LMM over infinitesimal mixed model analysis (GCTA-LOCO, BOLT-LMM-inf) increased with each of these parameters. In further simulations using data sets of size N=30,000 and N=60,000 (Supplementary Note) and simulated phenotypes with M_{causal}=250–15,000 causal SNPs explaining 15–35% of the variance, we observed that the effectiveness of the Gaussian mixture model is closely tied to $h_g^2 N/M_{\text{causal}}$ (where h_g^2 is the heritability parameter estimated by BOLT-LMM; see Online Methods for interpretation); intuitively, this quantity measures the effective number of samples per causal SNP (Supplementary Fig. 5). These results are consistent with theory (Supplementary Note and Supplementary Table 2 of ref.[12]), which explains that even in the absence of confounding, mixed model analysis provides a power gain over marginal regression by conditioning on the estimated effects of other SNPs when testing a candidate SNP[9,12]. As sample size increases, the power gain of both methods approaches an asymptote corresponding to an increase in effective sample size of $1/(1-h_g^2)$, but for sparse genetic architectures, the Gaussian mixture model approaches this asymptote much faster.

To verify that BOLT-LMM is correctly calibrated and robust to confounding, we also computed mean χ^2 statistics across SNPs on the second halves of chromosomes, simulated to all have zero effect ("null SNPs"). Because our simulated phenotypes included an ancestry effect, linear regression without correcting for population stratification suffered 35% inflation. In contrast, the BOLT-LMM and BOLT-LMM-inf statistics were both well-calibrated (Supplementary Fig. 4b, Supplementary Table 3, and Supplementary Table 4). We further verified that Type I error was properly controlled (Online Methods and Supplementary Table 5) and that the distribution of statistics at null SNPs did not deviate noticeably from a 1 d.o.f. chi-squared distribution (Supplementary Fig. 6a,b). Genomic inflation factors[33] for BOLT-LMM and BOLT-LMM-inf exceeded 1 in these simulations (Supplementary Fig. 6c,d), consistent with polygenicity of the simulated phenotype and use of a mixed model statistic that successfully avoids proximal contamination[12,13]. In contrast, EMMAX and GEMMA had deflated test statistics (Supplementary Fig. 4b).

To examine the tightness of the variational approximation used by BOLT-LMM for Bayesian model fitting and to enable comparison with FaST-LMM-Select, we ran a small-scale simulation using the same setup as above but only one-third of the samples (N=5,211). We simulated genetic architectures with 1,250 causal SNPs explaining 70% of phenotypic variance (and 60 additional standardized effect SNPs explaining 2% of variance and ancestry explaining 1%, as before). We ran PCA, BOLT-LMM-inf, BOLT-LMM, FaST-LMM-Select, and a modified version of BOLT-LMM in which we replaced the variational iteration of Step 2b with a Markov chain Monte Carlo (MCMC) Gibbs sampler. In the limit of infinite sampling iterations, MCMC would produce exact versions of the posterior approximations computed by BOLT-LMM. In these simulations, the variational iteration (i.e., standard BOLT-LMM) achieved statistically identical results to MCMC (Supplementary Table 6a), supporting the choice of variational Bayes for BOLT-LMM. We also observed that while BOLT-LMM-inf achieved a power gain over PCA and BOLT-LMM achieved a further power gain over BOLT-LMM-inf (consistent with previous simulations), FaST-LMM-Select achieved lower power than BOLT-LMM-inf and BOLT-LMM (Supplementary Table 6a). Upon repeating this experiment with the number of causal SNPs reduced to 500, we observed that FaST-LMM-Select achieved a power gain in between BOLT-LMM-inf and BOLT-LMM (Supplementary Table 6b). Finally, we observed that the LD Score calibration approach used by BOLT-LMM also worked well when applied to FaST-LMM-Select, validating this calibration approach (Supplementary Table 6).

Lastly, we investigated the similarity between the BOLT-LMM-inf mixed model statistic and existing methods at the individual SNP level. Despite its use of an infinitesimal model, the BOLT-LMM-inf statistic is not identical to any existing mixed model statistic because it is an approximate test statistic and avoids proximal contamination (Online Methods and Table 1). Nonetheless, we observed that BOLT-LMM-inf statistics very nearly match GCTA-LOCO statistics (which use the standard prospective model), with R^2>0.999 (Supplementary Table 7 and Supplementary Fig. 7).

Application of BOLT-LMM to WGHS phenotypes

To assess the efficacy of Gaussian mixture model analysis for increasing power on real phenotypes, we analyzed nine phenotypes in the Women's Genome Health Study (N=23,294 samples, M=324,488 SNPs after QC) (Online Methods). These phenotypes consisted of five lipid phenotypes, height, body mass index, and two blood pressure phenotypes; we chose to analyze these phenotypes because of the availability of large-scale GWAS results.

We compared the power of three association tests: linear regression with 10 principal components (PCA)[32], infinitesimal mixed model analysis with BOLT-LMM-inf, and Gaussian mixture modeling with BOLT-LMM. Because of memory constraints (Fig. 1), we were unable to run GCTA-LOCO[12], FaST-LMM[5], or FaST-LMM-Select[15], which are the only previous methods that avoid proximal contamination (Table 1); however, GCTA-LOCO and BOLT-LMM-inf statistics are near-identical (Supplementary Table 7 and Supplementary Fig. 7). To compare power among these methods, we computed two roughly equivalent metrics: mean χ^2 statistics at known associated loci, a direct but somewhat noisy approach due to having only 19–180 loci for each trait (Supplementary Table 8), and out-of-sample prediction R^2 (measured in cross-validation) using all SNPs for the mixed model methods and using only PCs for linear regression. For mixed model analysis, the latter metric estimates the ability of the mixed model to condition on effects of other SNPs when testing a candidate SNP, which drives its power (Online Methods)[12,34].

BOLT-LMM achieved higher power than PCA for all traits studied (Fig. 3 and Supplementary Table 9). Most of the increase was due to gains over infinitesimal mixed model analysis, with the magnitude of this power gain increasing with inferred concentration of genetic effects at few loci (Supplementary Table 10). Standard errors of the direct method of assessing improvement (mean χ^2 at known loci) were somewhat high (0.6–2.2%; Fig. 3a and Supplementary Table 9), so the improvement was statistically significant ($p<0.05$) for only 6 of 9 traits. According to the prediction R^2 metric, improvements were statistically significant for all traits ($p<0.0002$) (Fig. 3b and Supplementary Table 9). The largest gains were achieved for lipid traits; for ApoB, a lipoprotein closely related to LDL cholesterol, BOLT-LMM analysis achieved a 10% increase in mean χ^2 statistics versus PCA and a 9% increase versus infinitesimal mixed model analysis at known loci. To verify that these increases were not merely driven by a few loci with the largest effects, we also computed flat averages across loci of improvements in χ^2 statistics (restricting to loci replicating in WGHS with at least nominal $p<0.05$ significance to reduce statistical noise) and obtained consistent results (Supplementary Table 8). Simulations show that these improvements will increase with sample size (Fig. 2c and Supplementary Fig. 5).

We also observed that infinitesimal mixed model analysis achieved statistically significant power gains over PCA, with the magnitude of the power gains increasing with the heritability parameter h_g^2 (Fig. 3 and Supplementary Table 9). For height (h_g^2=0.47 in WGHS), the moderately large sample size of WGHS (N=23,294) was enough to obtain a 6% increase in BOLT-LMM-inf χ^2 statistics versus PCA, consistent with theory[12,34]. Again, larger sample sizes will enable further gains[12,34].

To verify that BOLT-LMM successfully corrected for confounding from population structure, we computed mean χ^2 statistics across all typed SNPs and genomic inflation factors for the three methods compared above as well as uncorrected marginal linear regression. We observed that PCA, BOLT-LMM-inf, and BOLT-LMM statistics were consistently calibrated, while uncorrected linear regression statistics were inflated, especially for height (Supplementary Table 11). We further verified that genetic variation at the lactase gene had a false-positive genome-wide significant association with height using uncorrected marginal regression[35] which disappeared when using PCA, BOLT-LMM-inf, and BOLT-LMM (Supplementary Table 12).

Discussion

We have described a new algorithm for fast Bayesian mixed model association, BOLT-LMM, and demonstrated that its running time scales only with $\approx MN^{1.5}$ and its memory usage is only $\approx MN/4$ bytes, resulting in orders-of-magnitude improvements in computational efficiency over existing methods for large data sets. We have further shown in simulations and analyses of WGHS phenotypes that the Gaussian mixture modeling capability of BOLT-LMM enables increased association power over standard mixed model analysis while controlling false positives. Among WGHS lipid traits, we observed power increases equivalent to increases in effective sample size of up to 10% over PCA and 9% over standard mixed model analysis.

BOLT-LMM is an advance for two main reasons. First, as sample sizes continue to increase, mixed model analysis is simultaneously becoming more important—in order to correct for population structure and cryptic relatedness in very large data sets—yet less practical with existing methods, all of which have $\geq O(MN^2)$ time complexity (for $M>N$) and high memory requirements. The algorithmic innovations of BOLT-LMM overcome this computational barrier (Fig. 1). (Our implementation uses $\approx MN/4$ bytes of memory, which is already much less in practice than existing methods. In theory, existing algorithms have a memory complexity of $O(N^2)$, while BOLT-LMM's memory complexity could be reduced to $O(M+N)$ by iteration on data.) Second, the ability of BOLT-LMM to better model non-infinitesimal genetic architectures enables a power gain relative to standard mixed model analysis. Recent methodological progress in this direction includes the multi-locus mixed model (MLMM)[7], which identifies and conditions out large-effect loci as fixed effects, and FaST-LMM-Select and related methods[9,11,15,16,36], which adopt a sparse regression framework that restricts the mixed model to a subset of markers. However, these methods all face the same $O(MN^2)$ computational hurdle as standard mixed model analysis.

Bayesian methods have previously been developed that apply non-infinitesimal models to improve the accuracy of genetic risk prediction. These methods extend in principle to association testing, although the Bayes factors and posterior inclusion probabilities that are naturally produced by Bayesian analysis do not directly translate to customary GWAS frequentist test statistics[37]. The variational Bayes spike regression (vBsr) method[38] is a recent step toward addressing this issue, proposing a z-statistic heuristically calibrated by assuming that the vast majority of variants are unassociated (as in genomic control[33]), but such a technique is prone to deflation when large sample sizes cause inflation due to polygenicity[13,24]. BOLT-LMM sidesteps this difficulty via its hybrid approach of leaving each chromosome out in turn, fitting a Bayesian model on the remaining SNPs, and then applying a retrospective hypothesis test for association of left-out SNPs with the residual phenotype. In contrast to modeling all SNPs simultaneously and assessing evidence for association using Bayesian posterior inference[37], our approach generalizes existing mixed model methods that are widely used, and we believe its ability to harness the power of Bayesian analysis while still computing frequentist statistics will be useful to GWAS practitioners. Additionally, such a hybrid approach lends itself readily to efficiently testing millions of imputed SNP dosages for association while including only typed SNPs in the mixed model, which we recommend to limit computational costs.

While BOLT-LMM improves upon existing mixed model association methods in both speed and power, BOLT-LMM still has limitations. First, the power gain that BOLT-LMM offers over existing methods via its more flexible prior on SNP effect sizes is contingent on the true genetic architecture being sufficiently non-infinitesimal and the sample size being sufficiently large (Supplementary Fig. 5). Second, BOLT-LMM, like existing mixed model methods, is susceptible to loss of power when used to analyze large ascertained case-control data sets in diseases of low prevalence[12]. We recommend BOLT-LMM for randomly ascertained quantitative traits, ascertained case-control studies of diseases with prevalence ≥5% (Supplementary Table 13)—e.g., type 2 diabetes, heart disease, common cancers, hypertension, asthma—and studies of rarer diseases in large, non-ascertained population cohorts[39,40]. For large ascertained case-control studies of rarer diseases, we are developing a method of modeling ascertainment using posterior mean liabilities (LTMLM); applying the techniques of BOLT-LMM to these posterior mean liabilities is an avenue for future research. Third, while mixed model analysis is effective in correcting for many forms of confounding, performing careful data quality control remains critical to avoiding false positives. Fourth, our work does not attempt to estimate the extent to which the heritability parameter estimated by BOLT-LMM (denoted h_g^2) may be influenced by population structure or relatedness, nor does it conduct or evaluate genetic prediction in external validation samples from an independent cohort[34]. Fifth, we have not studied the performance of mixed model methods in data sets dominated by family structure[23]. Sixth, the running time of BOLT-LMM scales with the number of phenotypes analyzed; for data sets with a very large number of phenotypes (P), the GRAMMAR-Gamma method[10], which has running time $O(MN^2+MNP)$ (reviewed in ref.[12]) may be faster. Seventh, we have only tested BOLT-LMM in human data sets, which have very different patterns of linkage disequilibrium and genetic architectures from plant and animal data. In particular, given that some approximations we make may be violated in non-human data sets (e.g., treating the denominator of the prospective test statistic as near-constant[10]), we are unsure whether the BOLT-LMM statistic is valid in these scenarios. Similarly, these assumptions should be viewed with caution when testing very rare variants. Finally, we have developed fast mixed model analysis for a mixed model with one random genetic effect; extending the algorithm to model multiple variance components[41] is a direction for future work.

Online Methods

Standard mixed model association methods

Standard methods employ a model

$$y = x_{\text{test}}\beta_{\text{test}} + g + e, \quad (1)$$

where y is the phenotype, x_{test} is the candidate SNP being tested, g is the genetic effect, and e is the environmental effect. We assume for now that all have been mean-centered and there are no covariates; we treat covariates by projecting them out from both genotypes and phenotypes, which is equivalent to including them as fixed effects (Supplementary Note). The genetic and environmental effects are modeled as random effects, while the candidate SNP is modeled as a fixed effect with coefficient β_{test}, and the goal is to test the null hypothesis $\beta_{\text{test}}=0$. Under the standard infinitesimal model, the genetic effect is modeled as

$$g = X_{\text{GRM}}\beta_{\text{GRM}}, \quad (2)$$

where X_{GRM} is an $N \times M_{\text{GRM}}$ matrix, each column of which contains normalized genotypes corresponding to a SNP included in the model, and β_{GRM} is an M_{GRM}-vector of random SNP effect sizes all drawn from the same normal distribution, so that g has a multivariate normal distribution with covariance $\text{Cov}(g) \propto X_{\text{GRM}}X_{\text{GRM}}'$. Note that in order to avoid proximal contamination[5,9,12], the M_{GRM} SNPs used in X_{GRM} should vary depending on which SNP x_{test} is being tested: the candidate SNP x_{test} (and SNPs in linkage disequilibrium with it) should be excluded from X_{GRM} to avoid modeling its effect twice. BOLT-LMM adopts a leave-one-chromosome-out (LOCO) scheme[5,12] in which X_{GRM} leaves out SNPs on the same chromosome as x_{test}.

The matrix $X_{\text{GRM}}X_{\text{GRM}}'/M_{\text{GRM}}$ is conventionally called the genetic relationship matrix (GRM) or empirical kinship matrix K, and we write

$$\text{Cov}(g) = \sigma_g^2 X_{\text{GRM}} X_{\text{GRM}}'/M_{\text{GRM}} = \sigma_g^2 K, \quad (3)$$

where σ_g^2 is a variance parameter. Environmental effects are assumed i.i.d. normal, so e is also multivariate normal with

$$\text{Cov}(e) = \sigma_e^2 I, \quad (4)$$

where I denotes the $N \times N$ identity matrix and σ_e^2 is another variance parameter.

In practice, the variance parameters σ_g^2 and σ_e^2 are unknown. Several existing methods[3,10,12] therefore take a two-step approach to computing association statistics: first estimate the variance parameters (with the SNP x_{test} removed from the model) using restricted maximum likelihood (REML), and then compute the prospective chi-squared (1 d.o.f.) test statistic (as previously proposed in family-based tests[42])

$$\chi^2_{\mathrm{LMM}} = \frac{\left(x'_{\mathrm{test}} V^{-1} y\right)^2}{x'_{\mathrm{test}} V^{-1} x_{\mathrm{test}}}, \quad (5)$$

where

$$V = \mathrm{Cov}(y) = \sigma_g{}^2 K + \sigma_e{}^2 I, \quad (6)$$

setting the variance parameters $\sigma_g{}^2$ and $\sigma_e{}^2$ to their estimates under the null hypothesis β_{test}=0. Within a LOCO scheme, the test statistic becomes

$$\chi^2_{\mathrm{LMM-LOCO}} = \frac{\left(x'_{\mathrm{test}} V^{-1}_{\mathrm{LOCO}} y\right)^2}{x'_{\mathrm{test}} V^{-1}_{\mathrm{LOCO}} x_{\mathrm{test}}}, \quad (7)$$

where we have written V_{LOCO} for V to explicitly indicate that the chromosome containing x_{test} is left out of the GRM.

Recent computational advances have also enabled computation of exact likelihood ratio test statistics that model the variance parameters while testing the candidate SNP[5,6]. While exact statistics are more accurate in situations with very large-effect SNPs, approximate methods produce near-identical results in typical human genetics scenarios[3,10,12].

BOLT-LMM-inf mixed model statistic

The BOLT-LMM-inf infinitesimal mixed model statistic is slightly different:

$$\chi^2_{\mathrm{BOLT-LMM-inf}} = \frac{\left(x'_{\mathrm{test}} V^{-1}_{\mathrm{LOCO}} y\right)^2}{c_{\mathrm{inf}}}, \quad (8)$$

where c_{inf} is a constant calibration factor estimated as

$$c_{\mathrm{inf}} = \frac{\mathrm{Mean}\left(x'_{\mathrm{test}} V^{-1}_{\mathrm{LOCO}} y\right)^2}{\mathrm{Mean}\chi^2_{\mathrm{LMM-LOCO}}} \quad (9)$$

so that

$$\mathrm{Mean}\chi^2_{\mathrm{BOLT-LMM-inf}} = \mathrm{Mean}\chi^2_{\mathrm{LMM-LOCO}}. \quad (10)$$

In practice, for computational efficiency, we take means over 30 pseudorandom SNPs not significantly associated with the phenotype (χ^2<5 estimated with the GRAMMAR statistic[43]). We have observed empirically that 30 random SNPs are enough to estimate the calibration factor to within 1% (Supplementary Table 14).

We can view the BOLT-LMM-inf statistic either as an approximation of the standard prospective statistic (which treats phenotypes as random) or as a retrospective statistic (which treats genotypes as random and builds a null model on SNPs). The first perspective is motivated by the observation that in human genetics applications, the denominator of the prospective statistic in equation (5), $x_{test}'V^{-1}x_{test}$, is nearly independent of the SNP x_{test}

being tested[10]. From this perspective, BOLT-LMM-inf is similar to GRAMMAR-Gamma[10], with two key differences: (1) BOLT-LMM-inf is computed via much faster algorithms (described below) for performing initial variance parameter estimation and estimating the calibration constant, and (2) BOLT-LMM-inf avoids proximal contamination via LOCO analysis. Alternatively, we can also view BOLT-LMM-inf as a retrospective quasi-likelihood score test similar to $T^{SCORE\text{–}R}$ (ref.[44]) and MASTOR[23] (Supplementary Note).

BOLT-LMM Gaussian mixture model association statistic

We now generalize BOLT-LMM-inf by observing that the vector $V_{\text{LOCO}}^{-1}y$ appearing in equation (8) is a scalar multiple of the residual phenotype vector $\sigma_e^2 V_{\text{LOCO}}^{-1}y$ from best linear unbiased prediction (BLUP). Thus, the $\chi^2_{\text{BOLT-LMM-inf}}$ statistic is equivalent to computing (and then calibrating) squared correlations between SNPs x_{test} and BLUP residuals. The power of mixed model association is driven by the fact that SNPs x_{test} are tested against these "de-noised" residual phenotypes from which other SNP effects estimated by the mixed model have been conditioned out[9,12].

We may generalize this approach by defining

$$\chi^2_{\text{BOLT-LMM}}=\frac{(x'_{\text{test}}y_{\text{resid-LOCO}})^2}{c}, \quad (11)$$

where $y_{\text{resid-LOCO}}$ denotes a generalized residual phenotype vector obtained after fitting a Gaussian mixture extension of the standard LMM (using SNPs not on the same chromosome as x_{test}) and c denotes a calibration factor, estimated so that the LD Score regression intercept[24] of $\chi^2_{\text{BOLT-LMM}}$ matches that of the (properly calibrated) $\chi^2_{\text{BOLT-LMM-inf}}$ statistic. Under the infinitesimal model, $y_{\text{resid-LOCO}}$ is proportional to $V_{\text{LOCO}}^{-1}y$, so $\chi^2_{\text{BOLT-LMM}}$ reduces to $\chi^2_{\text{BOLT-LMM-inf}}$. The general $\chi^2_{\text{BOLT-LMM}}$ statistic can still be interpreted as a retrospective quasi-likelihood score test and is thus asymptotically chi-squared distributed.

To define the Gaussian mixture LMM extension, it is helpful to first frame the standard LMM in a Bayesian formulation. The null model of BOLT-LMM-inf is

$$y=X_{\text{LOCO}}\beta_{\text{LOCO}}+e, \quad (12)$$

where SNP effects β_m (m indexing SNPs not on the left-out chromosome) are independently drawn from the Gaussian prior distribution

$$\beta_m \sim N(0,{\sigma_g}^2)/M_{\text{LOCO}} \quad (13)$$

and environmental effects e_n (n indexing samples) are independently drawn from $e_n \sim N(0, \sigma_e^2)$. Performing best linear unbiased prediction amounts to computing the posterior mean of the genetic effect $X_{\text{LOCO}}\beta_{\text{LOCO}}$.

To generalize this model to non-infinitesimal genetic architectures, we replace the Gaussian prior on SNP effect sizes with a more general distribution; this approach has been extensively applied by the "Bayesian alphabet" of genomic prediction methods in the animal breeding literature[17–19]. In BOLT-LMM, we use a spike-and-slab mixture of two Gaussians[19] as the prior:

$$\beta_m \sim N(0, {\sigma_{\beta,1}}^2) \text{with probability} p. \ \beta_m \sim N(0, {\sigma_{\beta,2}}^2) \text{with probability} 1-\mathrm{p}. \quad (14)$$

This mixture more flexibly models the heavier-tailed distributions of genetic effects of typical (non-infinitesimal) phenotypes. Explicitly, if $p \ll 1$ and ${\sigma_{\beta,1}}^2 \gg {\sigma_{\beta,2}}^2$, the first component of the mixture is a "slab" that models the existence of a small number of relatively large-effect loci, while the second component is a "spike" that models the assumption that most SNPs have near-zero—but not exactly zero—effect on the phenotype. (Note, however, that all SNPs are assigned the same mixture prior; i.e., SNPs are not individually allocated to one or the other component.) It is important that the spike component have nonzero variance so as to capture genome-wide effects on phenotype such as ancestry or relatedness; then, when testing SNPs for association, these genome-wide effects are conditioned out from residual phenotypes, protecting against confounding. The prior could in principle be further generalized; we chose to use a mixture of two Gaussians to keep the model fairly simple and because Gaussian distributions produce convenient analytical formulas during model-fitting.

Under this generalized model, posterior means no longer correspond to BLUP, but we can still approximately fit the Bayesian model (once per left-out chromosome) and obtain residuals

$$y_{\text{resid}-\text{LOCO}} = y - X_{\text{LOCO}} \beta_{\text{LOCO}}, \quad (15)$$

where β_{LOCO} are estimated posterior mean effect sizes. Plugging these residuals into equation (11) gives the BOLT-LMM Gaussian mixture model association test statistic.

Fast iterative algorithm

The BOLT-LMM software performs a four-step computation for mixed model association analysis, stopping after the first two steps when specialized to the infinitesimal model. We outline the algorithm here and provide full details and pseudocode in the Supplementary Note.

Step 1a: Estimate variance parameters

A key feature of BOLT-LMM is estimation of variance parameters ${\sigma_g}^2$ and ${\sigma_e}^2$ using only linear-time iterations without building or decomposing any covariance matrices. We use a Monte Carlo REML approach[26,27] that eliminates all $O(MN^2)$ and $O(N^3)$-time matrix computations, requiring only the solution of linear systems of mixed model equations. We solve the mixed model equations using conjugate gradient iteration, which requires only $O(MN)$-time matrix-vector products[28,29] (Supplementary Note).

Step 1b: Compute and calibrate BOLT-LMM-inf statistics

Having variance parameter estimates from Step 1a, it is straightforward to compute (for each LOCO rep) the quantity ${V_{\text{LOCO}}}^{-1}y$ in the numerator of the BOLT-LMM-inf statistic, equation (8), using conjugate gradient iteration as above. Completing the computation of the numerator of $\chi^2_{\text{BOLT-LMM-inf}}$ then just amounts to calculating one dot product per SNP x_{test}, which requires only $O(MN)$ additional cost across all SNPs. Moreover, this computation can easily be performed for additional SNPs not included in the mixed model but at which association statistics are desired; BOLT-LMM handles imputed "dosage" data in this way. To compute the calibration constant c_{inf} in equation (9), BOLT-LMM rapidly computes the prospective statistic $\chi^2_{\text{LMM-LOCO}}$ from equation (7) at 30 random SNPs by applying conjugate gradient iteration to compute $V_{\text{LOCO}}^{-1}x_{\text{test}}$ for each of the 30 selected SNPs x_{test}. Finally, in addition to computing χ^2 association statistics, BOLT-LMM also computes effect size estimates for all SNPs tested (Supplementary Note).

There is a slight mismatch between the variance parameters estimated in Step 1a, which BOLT-LMM computes once using all SNPs—not leaving any chromosomes out—and the theoretically optimal parameter estimates that would be obtained by refitting once per left-out chromosome. However, we have observed in simulations that slight mis-specification of the variance parameters has a negligible impact (<0.5%) on the calibration of the BOLT-LMM-inf and BOLT-LMM statistics (Supplementary Table 4). Because very slight miscalibration is not a concern for confounding from population stratification at highly differentiated markers (Supplementary Table 12) and has little impact on Type I error (Supplementary Table 5), the BOLT-LMM software does not by default refit variance parameters for each LOCO rep. If extremely precise calibration is desired, we provide a runtime option to refit variance parameters for each LOCO rep, at the cost of a factor of 2–3 in running time. We believe that LOCO strikes a good balance in terms of achieving ≈95% of the potential power gain (by jointly fitting ≈95% of markers that are not in LD with the candidate marker) while keeping run time down[12], but we also provide a runtime option to partition the genome more finely (e.g., into 100 segments rather than 22), again at the cost of a factor of 2–3 in running time.

Step 2a: Estimate Gaussian mixture prior parameters

The first step of BOLT-LMM Gaussian mixture model association analyis is to estimate parameters of the generalized prior on SNP effect sizes. As written in equation (14), this mixture has three parameters: $\sigma_{\beta,1}^2$ and $\sigma_{\beta,2}^2$, the variances of the two Gaussians, and p, the probability of drawing from the first Gaussian. To reduce the complexity of parameter estimation, we constrain the total variance of the mixture to equal the variance σ_g^2/M estimated under the infinitesimal model in Step 1a:

$$p{\sigma_{\beta,1}}^2+(1-\mathrm{p}){\sigma_{\beta,2}}^2={\sigma_g}^2/M. \quad (16)$$

We reparameterize the remaining two degrees of freedom using the parameters p and f_2, where f_2 denotes the proportion of the total mixture variance within the second Gaussian (the "spike" component that models small genome-wide effects):

$$f_2=\frac{(1-p)\sigma^2_{\beta,2}}{p\sigma^2_{\beta,1}+(1-p)\sigma^2_{\beta,2}}. \quad (17)$$

Because the model fit is insensitive to the precise values of the mixture parameters, we test a discrete set of model parameter combinations: $f_2 \in \{0.5,0.3,0.1\}$, $p\in$

{0.5,0.2.0.1,0.05,0.02,0.01}. Note that f_2=0.5, p=0.5 corresponds to the infinitesimal model: when $f_2=1-p$, the two Gaussians are identical and the mixture is degenerate. We bound f_2 from below to ensure that at least a small amount (10%) of the mixture variance is assigned to the spike component, protecting against confounding from genome-wide effects. We bound p from below to prevent the model from trying to fit too strongly to a few SNPs, which makes model-fitting computationally difficult and also increases susceptibility to confounding. BOLT-LMM performs model selection among the 18 possible parameter pairs (f_2, p) by performing cross-validation to optimize mean-squared prediction R^2.

BOLT-LMM uses a variational approximation to fit Bayesian linear regressions with Gaussian mixture priors. Approximation methods are necessary for Bayesian inference in this setting because exact posterior means involve intractable integrals. We apply a fully factored variational approximation[21,22,38] that repeatedly loops through the SNPs, updating the estimated effect size of each SNP with its posterior mean conditional on current estimates of all other SNP effects. This iteration has also previously been termed "iterative conditional expectation (ICE)"[20]. The variational Bayes framework puts this iteration on a sound theoretical footing as an optimization of an approximate log likelihood function; the iteration monotonically increases this function and is guaranteed to converge[45]. In fact, we show that the optimization can be reformulated as cyclic coordinate descent applied to a penalized regression problem arising from Bayesian linear regression using a transformed prior (Supplementary Note). The approximate log likelihood also serves as a convenient convergence criterion: BOLT-LMM stops the iteration when the increase in approximate log likelihood over one full update cycle drops below 0.01.

While the core variational iteration that BOLT-LMM uses is identical to previous methods[20–22,38] up to the choice of SNP effect size prior, BOLT-LMM uses cross-validation to estimate hyperparameters[15] rather than doing so within the variational iteration[22,38] or based on variational approximate log likelihoods[21]. We found this approach to be more robust to slackness of the variational approximation caused by linkage disequilibrium.

Step 2b: Compute and calibrate BOLT-LMM Gaussian mixture model statistics

After inferring parameters of the mixture prior in Step 2a, BOLT-LMM uses the same variational iteration to estimate posterior mean residuals $y_{\text{resid-LOCO}}$ (independently for each left-out chromosome). The numerators of the BOLT-LMM Gaussian mixture model statistic from equation (11) are then easily obtained as dot products with test SNPs, leaving only the constant calibration factor c in the denominator to be calculated. Unlike the case of the infinitesimal model, here we do not have a prospective statistic to calibrate against, so we instead apply LD Score regression[24] (Supplementary Note). In practice, the calibration factor is usually quite close to 1 (e.g., 1.00 to two decimal places for all WGHS traits; see Supplementary Table 15).

WGHS data set

The Women's Genome Health Study (WGHS) is a prospective cohort of initially healthy, female North American health care professionals. We analyzed 23,294 individuals with selfreported European ancestry with genotyping at 324,488 SNPs after QC (Supplementary Note).

Interpretation of heritability parameter

The heritability parameter (denoted h_g^2) estimated by BOLT-LMM may in general include some contribution from cryptic relatedness or population structure[46], and thus may not strictly correspond to the heritability explained by genotyped SNPs[47]. Ref.[3] refers to this parameter as "pseudo-heritability" for this reason. Because the WGHS samples that we primarily analyze here do not contain substantial relatedness or population structure, we have simply used the notation h_g^2 to avoid complicating the discussion.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to M. Lipson, S. Simmons, A. Gusev, K. Galinsky, J. Yang, P. Visscher, Z. Zhu, and D. Gudbjartsson for helpful discussions. This research was supported by NIH grant R01 HG006399 and NIH fellowship F32 HG007805. H. K. F. was supported by the Fannie and John Hertz Foundation. The WGHS is supported by HL043851 and HL080467 from the National Heart, Lung, and Blood Institute and CA047988 from the National Cancer Institute, the Donald W. Reynolds Foundation and the Fondation Leducq, with collaborative scientific support and funding for genotyping provided by Amgen.

References

1. Yu J, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics. 2006; 38:203–208. [PubMed: 16380716]
2. Kang HM, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178:1709–1723. [PubMed: 18385116]
3. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010; 42:348–354. [PubMed: 20208533]
4. Zhang Z, et al. Mixed linear model approach adapted for genome-wide association studies. Nature Genetics. 2010; 42:355–360. [PubMed: 20208535]
5. Lippert C, et al. FaST linear mixed models for genome-wide association studies. Nature Methods. 2011; 8:833–835. [PubMed: 21892150]
6. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012; 44:821–824. [PubMed: 22706312]
7. Segura V, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature Genetics. 2012; 44:825–830. [PubMed: 22706313]
8. Korte A, et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nature Genetics. 2012; 44:1066–1071. [PubMed: 22902788]
9. Listgarten J, et al. Improved linear mixed models for genome-wide association studies. Nature Methods. 2012; 9:525–526. [PubMed: 22669648]
10. Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. Nature Genetics. 2012
11. Listgarten J, Lippert C, Heckerman D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nature Genetics. 2013; 45:470–471. [PubMed: 23619783]
12. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics. 2014; 46:100–106. [PubMed: 24473328]

NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

13. Yang J, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011; 19:807–812. [PubMed: 21407268]
14. Stahl EA, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nature Genetics. 2012; 44:483–489. [PubMed: 22446960]
15. Lippert C, et al. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. Scientific Reports. 2013; 3
16. Rakitsch B, Lippert C, Stegle O, Borgwardt K. A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics. 2013; 29:206–214. [PubMed: 23175758]
17. Meuwissen T, Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157:1819–1829. [PubMed: 11290733]
18. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013; 193:327–345. [PubMed: 22745228]
19. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genetics. 2013; 9:e1003264. [PubMed: 23408905]
20. Meuwissen T, Solberg TR, Shepherd R, Woolliams JA. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol. 2009; 41
21. Carbonetto P, Stephens M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis. 2012; 7:73–108.
22. Logsdon BA, Hoffman GE, Mezey JG. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinformatics. 2010; 11:58. [PubMed: 20105321]
23. Jakobsdottir J, McPeek MS. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. American Journal of Human Genetics. 2013; 92:652–666. [PubMed: 23643379]
24. Bulik-Sullivan B, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. (in press).
25. Ridker PM, et al. Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000 initially healthy American women. Clinical Chemistry. 2008; 54:249–255. [PubMed: 18070814]
26. García-Cortés LA, Moreno C, Varona L, Altarriba J. Variance component estimation by resampling. Journal of Animal Breeding and Genetics. 1992; 109:358–363.
27. Matilainen K, Mäntysaari EA, Lidauer MH, Strandén I, Thompson R. Employing a Monte Carlo Algorithm in Newton-Type Methods for Restricted Maximum Likelihood Estimation of Genetic Parameters. PLoS ONE. 2013; 8:e80821. [PubMed: 24339886]
28. Legarra A, Misztal I. Computing strategies in genome-wide selection. Journal of Dairy Science. 2008; 91:360–366. [PubMed: 18096959]
29. VanRaden P. Efficient methods to compute genomic predictions. Journal of Dairy Science. 2008; 91:4414–4423. [PubMed: 18946147]
30. Sawcer S, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476:214. [PubMed: 21833088]
31. Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007; 23:1294–1296. [PubMed: 17384015]
32. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38:904–909. [PubMed: 16862161]
33. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]
34. Wray NR, et al. Pitfalls of predicting complex traits from SNPs. Nature Reviews Genetics. 2013; 14:507–515.
35. Campbell CD, et al. Demonstrating stratification in a European American population. Nature Genetics. 2005; 37:868–872. [PubMed: 16041375]

36. Tucker G, Price AL, Berger BA. Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. Genetics. 2014
37. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009; 10:681–690.
38. Logsdon BA, Carty CL, Reiner AP, Dai JY, Kooperberg C. A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. Bioinformatics. 2012; 28:1738–1744. [PubMed: 22563072]
39. Styrkarsdottir U, et al. Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature. 2013
40. Do CB, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genetics. 2011; 7:e1002141. [PubMed: 21738487]
41. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Research. 2014 gr–169375.

References (Online Methods)

42. Chen W-M, Abecasis GR. Family-based association tests for genomewide association scans. American Journal of Human Genetics. 2007; 81:913–926. [PubMed: 17924335]
43. Aulchenko YS, De Koning D-J, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007; 177:577–585. [PubMed: 17660554]
44. Chen W-M, Manichaikul A, Rich SS. A generalized family-based association test for dichotomous traits. American Journal of Human Genetics. 2009; 85:364–376. [PubMed: 19732865]
45. Boyd, SP.; Vandenberghe, L. Convex Optimization. Cambridge University Press; 2004.
46. Yang J, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics. 2011; 43:519–525. [PubMed: 21552263]
47. Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010; 42:565–569. [PubMed: 20562875]

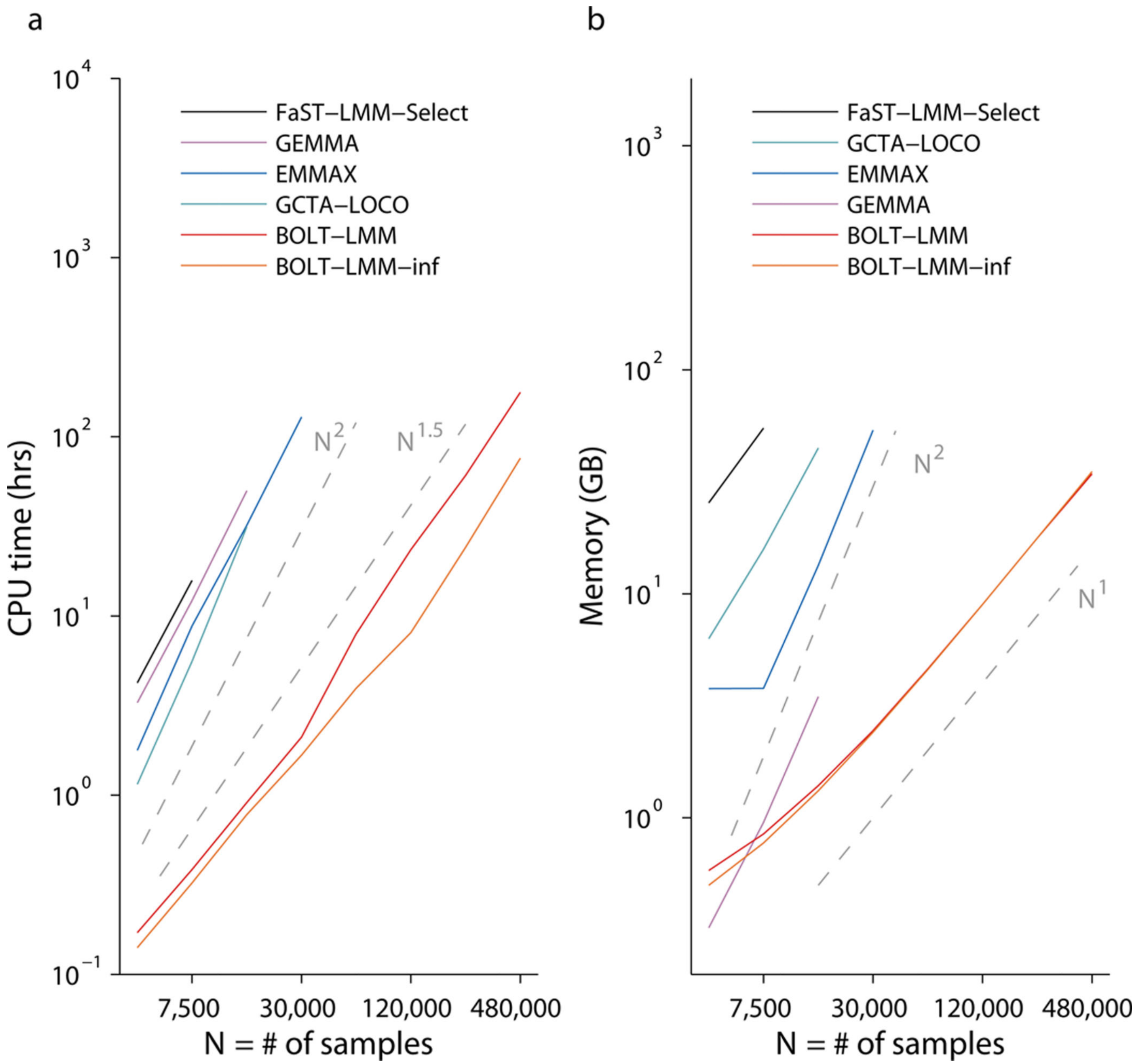

Figure 1. Computational performance of mixed model association methods

Log-log plots of (**a**) run time and (**b**) memory as a function of sample size (*N*). Slopes of the curves correspond to exponents of power-law scaling with *N*. Benchmarking was performed on simulated data sets in which each sample was generated as a mosaic of genotype data from 2 random "parents" from the WTCCC2 data set (*N*=15,633, *M*=360K) and phenotypes were simulated with M_{causal}=5,000 SNPs explaining h^2_{causal}=0.2 of phenotypic variance. Reported run times are medians of five identical runs using one core of a 2.27 GHz Intel Xeon L5640 processor. We caution that running time comparisons may vary by a small constant factor as a function of computing environment. FaST-LMM-Select (resp. GCTA-LOCO, EMMAX) memory usage exceeded the 96GB available at *N*=15K (resp. 30K, 60K).

GEMMA encountered a runtime error (segmentation fault) at *N*=30K. Software versions: FaST-LMM-Select, v2.07; GCTA-LOCO, v1.24; EMMAX, v20120210; GEMMA, v0.94. Numerical data are provided in Supplementary Table 1.

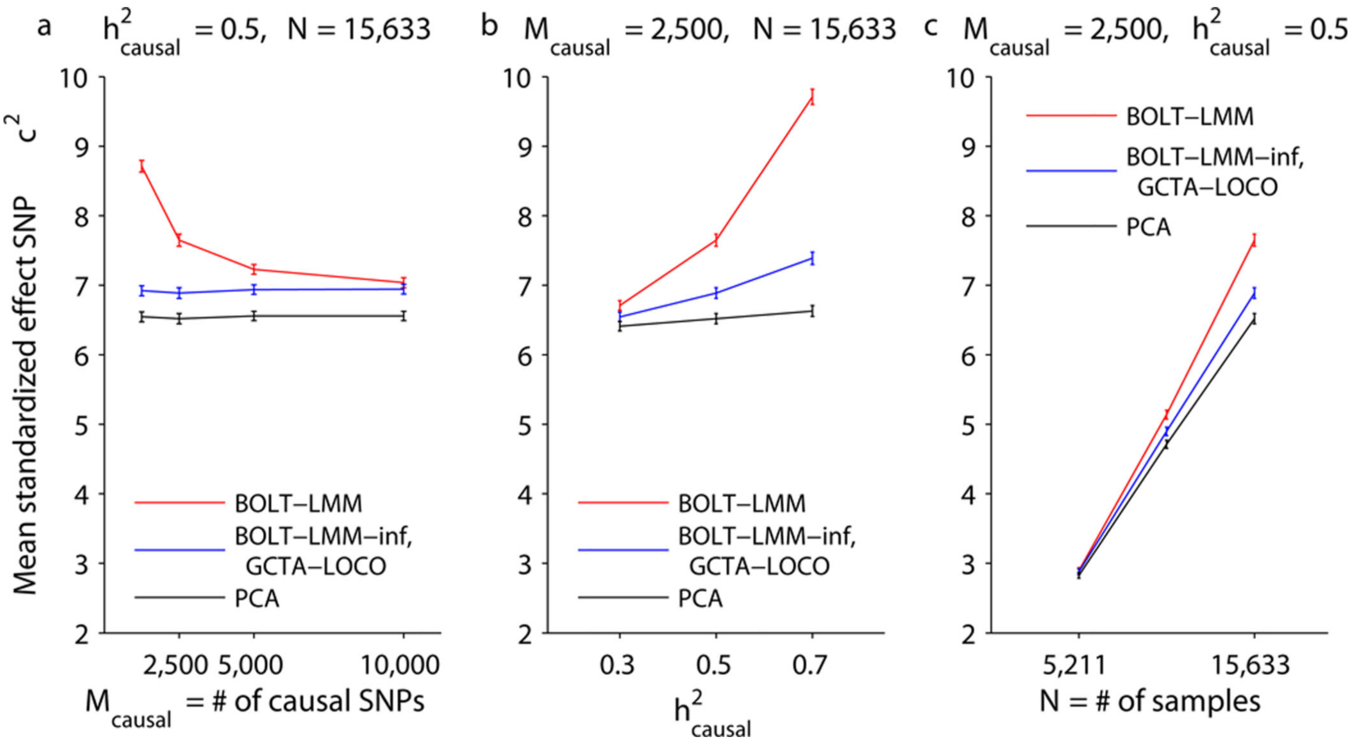

Figure 2. BOLT-LMM increases power to detect associations in simulations

Mean χ^2 at standardized effect SNPs as a function of (**a**) number of causal SNPs, (**b**) proportion of variance explained by causal SNPs, (**c**) number of samples. Simulations used real genotypes from the WTCCC2 data set (N=15,633, M=360K) and simulated phenotypes with the specified number of causal SNPs explaining the specified proportion of phenotypic variance and 60 more standardized effect SNPs explaining an additional 2% of the variance. Error bars, s.e.m., 100 simulations. We verified on the first 5 simulations that the BOLT-LMM-inf and GCTA-LOCO statistics are nearly identical (Supplementary Table 7). Numerical data are provided in Supplementary Table 2.

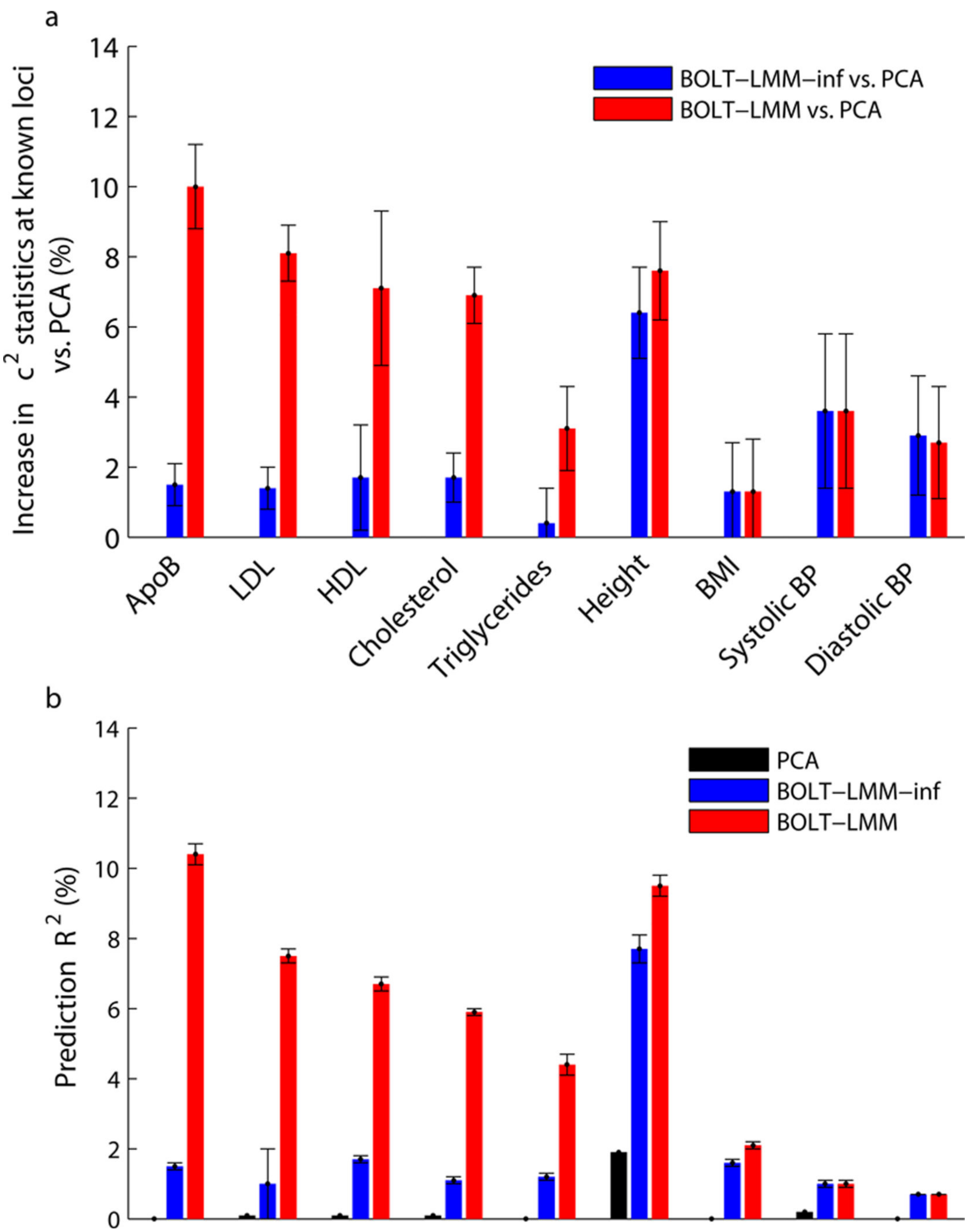

Figure 3. BOLT-LMM increases power to detect associations for WGHS phenotypes

We compare power (measured using two roughly equivalent metrics) of linear regression using 10 principal components, standard (infinitesimal) mixed model analysis, and BOLT-LMM Gaussian mixture model analysis. **(a)** Percent increases in χ^2 statistics across known loci using mixed model methods vs. PCA: ratios of sums of χ^2 statistics over typed SNPs in highest LD with published associated SNPs. **(b)** Prediction R^2 values from 5-fold cross-validation: each fold was left out in turn and predictions were computed by fitting all SNP effects simultaneously (for mixed model methods) or estimating covariate effects (for PCA) using the training folds. (Note that BOLT-LMM-inf is equivalent to BLUP prediction here.) We show PCA in **(b)** because the small amount of variance that the PCs explain (due to population stratification) provides a baseline that allows translating prediction R^2 to the power gain of mixed model association vs. regression with PC covariates. That is, the correspondence between association power and prediction accuracy is such that the red bars in **(a)** roughly correspond to differences between red and black bars in **(b)**, and analogously for blue bars (Online Methods). Error bars, jackknife s.e. over **(a)** known loci (Supplementary Table 8); **(b)** 5 cross-validation folds. Numerical data are provided in Supplementary Table 9.

Table 1

Comparison of fast mixed model association methods that model all SNPs.

Method[a]	Requires $O(MN^2)$ time	Avoids proximal contamination	Models non-infinitesimal genetic architecture
EMMAX [3]	X		
FaST-LMM [5]	X[b]	X	
FaST-LMM-Select [9, 11, 15]	X[b]	X	X[c]
GEMMA [6]	X		
GRAMMAR-Gamma [10]	X[d]		
GCTA-LOCO [12]	X	X	
BOLT-LMM		X	X

[a] For methods that have been updated over multiple publications, we cite and list characteristics of the latest published version.

[b] If $M<N$, FaST-LMM and FaST-LMM-Select can complete in $O(M^2N)$ time.

[c] FaST-LMM-Select models non-infinitesimal genetic architectures by restricting the mixed model to a subset of SNPs; a caveat of this approach is that it may incur susceptibility to confounding from stratification[12].

[d] GRAMMAR-Gamma requires $O(MN^2)$ time for only the initial computation of the genetic relationship matrix but not for computing association test statistics. For a detailed breakdown of computational complexity per algorithmic step, see Table 1 of ref.[12].