Supplementary Information for "Efficient Bayesian mixed model analysis increases association power in large cohorts"

Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, Alkes L Price

Contents

Nature Genetics: doi:10.1038/ng.3190

Nature Genetics: doi:10.1038/ng.3190

1 BOLT-LMM algorithm details

The BOLT-LMM algorithm is overviewed in Online Methods. Here, we provide additional details for components of the computation not fully described earlier.

1.1 Model setup

1.1.1 Initialization: Missing data, normalization, and covariates

Given a data set of genotypes and phenotypes, we apply the following procedure to create a normalized genotype matrix X. We begin by dealing with missing data as follows. First, we perform QC by filtering out SNPs and individuals with missing rates exceeding thresholds (default 10% for each). Second, we filter out individuals with missing phenotypes. Third, if the analysis includes covariates, we filter out individuals with any missing covariates. (As an alternative to this approach, known as "complete case analysis," we also implement the option to use the "missing indicator method," which adds missing status indicator variables as additional covariates.) Finally, we replace missing genotypes in the remaining data with per-SNP averages. We denote by N and M the numbers of samples and SNPs remaining post-QC.

Next, we mean-center each SNP and normalize the SNPs to have equal sample variance. We also mean-center the phenotypes. We model covariates by projecting them out from both genotypes and phenotypes, which is equivalent to including them as fixed effects. Explicitly, we compute a basis spanning the covariate vectors and subtract out the components of the SNP and phenotype vectors along the basis vectors. This procedure is mathematically equivalent to multiplying by an orthogonal projection matrix $\mathbf{P}_{\text{fixed}}$ that projects vectors to an $(N - C)$-dimensional subspace $\mathcal{S} \subset \mathbb{R}^N$ orthogonal to the covariates, where C is the number of independent covariates (including the all-1s vector, which is implicitly included as a covariate upon mean-centering). For comparison, GCTA-LOCO [12] version 1.24 provides two options for treating covariates: (1) projecting out covariates from the phenotype vector only (the default, which is an approximate approach); or (2) fitting covariates as fixed effects along with the candidate SNP in each association test (the `--mlma-no-adj-covar` option, which is equivalent to projecting covariates out from both genotypes and phenotypes). The GCTA-LOCO documentation notes that the latter option significantly reduces the computational efficiency of GCTA-LOCO, but the BOLT-LMM implementation is not

Nature Genetics: doi:10.1038/ng.3190

subject to this loss of computational efficiency.

We denote by X the final $N \times M$ matrix of genotypes and y the final N-vector of phenotypes after QC, normalization, and projection. While y and all columns of X (i.e., SNP vectors) are N-dimensional vectors, it is important to keep in mind that they all actually belong to the $(N - C)$-dimensional subspace $\mathcal{S}$ left after projecting by $\mathbf{P}_{\text{fixed}}$. For example, when estimating variance parameters, accounting for the loss of degrees of freedom distinguishes restricted maximum likelihood (REML) analysis—the preferred approach, which we implement—from maximum likelihood (ML), and when computing χ^2 test statistics, we need to use $N - C$ instead of N as the sample size.

1.1.2 Models

We therefore use the following general model to estimate hyperparameters (i.e., variance parameters under the infinitesimal model and Gaussian mixture parameters under the non-infinitesimal model):

$$y = X\beta + e_{\text{proj}}, \tag{18}$$

where

$$\begin{aligned}
y &\in \mathcal{S} \subset \mathbb{R}^N \text{ : projected phenotypes} \\
X &= N \times M \text{ matrix}, x_m \in \mathcal{S} \subset \mathbb{R}^N \text{ for } m = 1, \ldots, M \text{ : projected genotypes} \\
\beta &\in \mathbb{R}^M \text{ : iid random effects with } E[\beta_m] = 0, \text{Var}(\beta_m) = \sigma_\beta^2 \\
e_{\text{proj}} &\in \mathcal{S} \subset \mathbb{R}^N \text{ : iid random normal noise with } e_{\text{proj}} \sim N(0, \sigma_e^2 \mathbf{P}_{\text{fixed}}).
\end{aligned}$$

Reiterating the point above, this "mixed" model does not contain any fixed effects; the fixed effects have been projected out, so that the entire model—including the noise term e_{proj}—lives in the subspace $\mathcal{S}$.

Model for estimating infinitesimal model parameters. When estimating variance parameters under the infinitesimal model in Step 1a of the BOLT-LMM algorithm, we assume the SNP effect

Nature Genetics: doi:10.1038/ng.3190

prior is the normal distribution

$$\beta_m \sim N(0, \sigma_\beta^2)$$

and we estimate σ_β^2 and σ_e^2 using the stochastic REML approximation algorithm described in Section 1.2. Note that the following notation is often used for the total genetic effect $X\beta$ and its covariance:

$$\begin{aligned} g &= X\beta \\ K &= \frac{XX'}{M} \quad \text{(the "empirical kinship" or GRM)} \\ \sigma_g^2 K &= \text{Cov}(g) \end{aligned}$$

in which case

$$\sigma_\beta^2 = \sigma_g^2 / M.$$

Model for estimating Gaussian mixture model parameters. When estimating Gaussian mixture parameters in Step 2a of the BOLT-LMM algorithm (after having obtained estimates of σ_g^2 and σ_e^2 in Step 1a), we generalize the SNP effect prior to the three-parameter mixture of normals

$$\beta_m \sim \begin{cases} N(0, \sigma_{\beta,1}^2) & \text{with probability } p \\ N(0, \sigma_{\beta,2}^2) & \text{with probability } 1-p, \end{cases} \tag{19}$$

where we require

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

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

$$f_2 = \frac{(1-p)\sigma_{\beta,2}^2}{p\sigma_{\beta,1}^2 + (1-p)\sigma_{\beta,2}^2}.$$

Thus, Step 2a consists of estimating f_2 and p.

Nature Genetics: doi:10.1038/ng.3190

Prospective model for computing association statistics. When performing association tests (under the infinitesimal model) in Step 1b of the BOLT-LMM algorithm, we modify the above models by (1) including the SNP being tested as a fixed effect and (2) leaving out all SNPs on its chromosome from the random effect:

$$y = x_{\text{test}}\beta_{\text{test}} + X_{\text{LOCO}}\beta + e_{\text{proj}}. \tag{20}$$

As discussed in Online Methods, we do not by default re-estimate the hyperparameters σ_g^2, σ_e^2 (and f_2, p in the Gaussian mixture case; see below) when performing association tests; instead, we simply reuse the estimates obtained above from the slightly different models above, which include all SNPs in the random effect (and no fixed effect). However, we also offer the option of re-estimating σ_g^2 and σ_e^2 with each chromosome left out in turn.

Retrospective formulation of association tests. In Online Methods, we defined the BOLT-LMM-inf χ^2 test statistic as an approximation of the standard prospective mixed model χ^2 statistic using the model of equation (20). We then briefly noted that the BOLT-LMM-inf association test can also be viewed as a retrospective quasi-likelihood score test similar to $T^{SCORE-R}$ [44] and MASTOR [23]. Following ref. [23], this viewpoint models the phenotype y as known and the SNP x_{test} as random, being drawn from a distribution with covariance $\text{Cov}(x_{\text{test}}) = \Phi$ and mean

$$E[x_{\text{test}} \mid y] = (\Phi V^{-1} y)\alpha, \tag{21}$$

where we wish to test the null hypothesis $\alpha = 0$. Note that x_{test} is a discrete random variable, so we cannot formally model it as a multivariate normal and perform likelihood-based analysis, but the retrospective mean model is enough to obtain a quasilikelihood score test statistic that is asymptotically χ^2-distributed:

$$\chi^2_{\text{LMM-R}} = \frac{(x_{\text{test}}\Phi^{-1}(\Phi V^{-1}y))^2}{(\Phi V^{-1}y)'\Phi^{-1}(\Phi V^{-1}y)} = \frac{(x_{\text{test}}V^{-1}y)^2}{(V^{-1}y)'\Phi(V^{-1}y)}. \tag{22}$$

Previous formulations of this statistic [23, 44] have been aimed at family studies in which the pedigree matrix may be used as the covariance matrix Φ. For the general case of individuals with

Nature Genetics: doi:10.1038/ng.3190

unknown pedigree, a seemingly natural choice is to substitute the GRM, but unfortunately, the GRM does not accurately reflect the covariance matrix of the distribution of unassociated (i.e., null) SNPs, which form the null model under the retrospective framework. In particular, for large sample sizes, this approach incurs deflation because it incorporates SNPs associated with y into the covariance model (for x_{test}, treated as random), causing overestimation of the variance of $x_{\text{test}}V^{-1}y$. This phenomenon is not overcome by LOCO analysis because unlike with prospective modeling, the problem is not proximal contamination; rather, it is polygenicity: to obtain unbiased variance estimates based on genotype data, we would need to eliminate associated SNPs, which runs into the same difficulty faced by genomic control [13, 24].

While the denominator of ${\chi^2}_{\text{LMM-R}}$ in equation (22), $(V^{-1}y)'\Phi(V^{-1}y)$, cannot be computed without Φ (which is in general inaccessible as discussed above), it does not involve the candidate SNP x_{test} and can therefore be treated as a constant calibration factor, so that in practice, equation (22) is equivalent to the BOLT-LMM-inf formulation in equation (8). (More precisely, within a LOCO scheme, the matrix V depends on which chromosome x_{test} belongs to, so there should formally be a different denominator for each chromosome, but all are approximately equal.) The retrospective model thus provides an alternative justification for the BOLT-LMM-inf statistic.

Importantly, the retrospective viewpoint also enables a natural generalization of BOLT-LMM-inf to the more powerful BOLT-LMM association test for non-infinitesimal phenotypes. The basic observation motivating this extension is that in the retrospective mixed model statistic ${\chi^2}_{\text{LMM-R}}$, the vector $V^{-1}y$ is a scalar multiple of the residual after best linear unbiased prediction (BLUP)—i.e., the component of the phenotype remaining after conditioning out SNP effects estimated by the mixed model, which drives the power of mixed model association—but this vector may be replaced by any other vector not involving x_{test}. More precisely, we generalize the infinitesimal retrospective mean model, equation (21), by replacing the BLUP residual $\sigma_e^2 V^{-1}y$ with the residual phenotype $y_{\text{resid-LOCO}}$ upon conditioning out the effects of SNPs not on the same chromosome as x_{test} (where $y_{\text{resid-LOCO}}$ is estimated by fitting a Gaussian mixture extension of the standard LMM):

$$E[x_{\text{test}} \mid y] = (\Phi y_{\text{resid-LOCO}})\alpha. \tag{23}$$

This model produces ${\chi^2}_{\text{BOLT-LMM}}$ as a quasi-likelihood score test for $\alpha \stackrel{?}{=} 0$.

Nature Genetics: doi:10.1038/ng.3190

1.1.3 Estimated SNP effect sizes

When performing a genome-wide association study, it is of interest not only to analyze χ^2 statistics for association but also estimated SNP effect sizes; to this end, BOLT-LMM outputs such effect sizes. We explain below how BOLT-LMM estimates effect sizes, but first, we discuss an important distinction between two uses of the term "estimated effect size" that can be a source of confusion.

Distinction between estimated effect sizes in association tests vs. prediction. When performing mixed model GWAS, each SNP x_m is in turn tested for association with the phenotype y, and the "estimated effect size" of x_m is defined to be the fitted value $\hat{\beta}_{m,\text{assoc}}$ of the fixed effect x_m in the mixed model given by equation (20), taking $x_{\text{test}} = x_m$. Importantly, the model in equation (20) is *different for each SNP*, with the fixed effect x_m changing from test to test (and the SNPs included in the random effect also changing from chromosome to chromosome, assuming the LOCO framework is being used). If we denote the (estimated) phenotypic covariance by $V_{-\text{chr}(m)}$, the mixed model association effect size is given by

$$\hat{\beta}_{m,\text{assoc}} = \frac{x_m' V_{-\text{chr}(m)}^{-1} y}{x_m' V_{-\text{chr}(m)}^{-1} x_m}. \tag{24}$$

Note that equation (24) is very similar to but not the same as the formula for the prospective mixed model test statistic:

$$\chi_m^2 = \frac{(x_m' V_{-\text{chr}(m)}^{-1} y)^2}{x_m' V_{-\text{chr}(m)}^{-1} x_m}. \tag{25}$$

(We also note in passing that the BLUP residual is given by $\sigma_e^2 V^{-1} y$, which is why the standard mixed model association test is similar to testing the SNP for correlation with the BLUP residual.)

In contrast, when performing mixed model prediction (i.e., BLUP), a *single* model is fit, equation (18), in which all SNPs are included as random effects. Each SNP receives an "estimated effect size" when the model is fit, but these effect sizes,

$$\hat{\beta}_{m,\text{BLUP}} = \frac{\sigma_g^2}{M} x_m' V^{-1} y, \tag{26}$$

are conceptually completely different from the association effect size estimates $\hat{\beta}_{m,\text{assoc}}$. The effect sizes $\hat{\beta}_{m,\text{BLUP}}$ are meaningful precisely for the purpose of prediction; that is, they have the

Nature Genetics: doi:10.1038/ng.3190

property that building a linear combination of the SNPs using the coefficients $\hat{\beta}_{m,\text{BLUP}}$ produces the BLUP prediction. In particular, these coefficients take into account linkage disequilibrium between nearby SNPs to correctly weight their contributions. These same statements apply to the effect sizes that BOLT-LMM iterates on internally when using variational Bayes to fit the Bayesian mixed model.

For a concrete example that may aid intuition, consider two SNPs in perfect LD. When testing association, each SNP should be ignored when testing the other; the fact that we happen to have two redundant tags should not change the strength of association. However, when creating a risk score (i.e., prediction), it would be incorrect to assign each SNP the effect size assuming only one copy of the SNP: doing so would overweight the contribution of these SNPs. Instead, the LD needs to be accounted for by acknowledging the presence of both SNPs and splitting the total effect between the two SNPs.

To summarize, we emphasize that association effect sizes must not be confused with prediction effect sizes; in particular, it is inappropriate to use the former as coefficients for polygenic prediction.

Estimated effect sizes reported by BOLT-LMM. Estimating association effect sizes—i.e., $\hat{\beta}_{m,\text{assoc}}$ given in equation (24)—is computationally challenging for the same reasons that computing χ^2 association statistics in equation (25) is challenging: building and inverting the covariance matrix is slow. Fortunately, it turns out that we can harness that fact that we already have a way to estimate χ^2 statistics: namely, BOLT-LMM-inf. Combining equations (24) and (25), we have

$$\hat{\beta}_{m,\text{assoc}} = \frac{x_m' V_{-\text{chr}(m)}^{-1} y}{x_m' V_{-\text{chr}(m)}^{-1} x_m} = \frac{\chi_m^2}{x_m' V_{-\text{chr}(m)}^{-1} y} \approx \frac{\chi_{m,\text{BOLT-LMM-inf}}^2}{x_m' V_{-\text{chr}(m)}^{-1} y}. \tag{27}$$

BOLT-LMM already computes both the statistics $\chi_{m,\text{BOLT-LMM-inf}}^2$ and the quantities $V_{-\text{chr}(m)}^{-1} y$, so the final expression in equation (27) is simple to compute. This quantity is the effect size estimate that the BOLT-LMM software reports. (Note that the BOLT-LMM software reports this effect size regardless of whether the infinitesimal or Gaussian mixture prior is used because there is no natural analog of the above for the Gaussian mixture model.)

Nature Genetics: doi:10.1038/ng.3190

1.2 Variance component estimation: Monte Carlo REML approximation

Here we provide details of our stochastic algorithm for estimating REML variance parameters in Step 1a of BOLT-LMM. The crux of the method is to employ the observation that the REML first-order conditions on σ_g^2 and σ_e^2 are equivalent [48, equation (14.8)] to the system of equations

$$E[\sum \hat{e}^2_{\text{rand}}] = \sum \hat{e}^2_{\text{data}}, \quad E[\sum \hat{\beta}^2_{\text{rand}}] = \sum \hat{\beta}^2_{\text{data}}, \tag{28}$$

where both the left and right sides of each equation, defined below, are functions of the assumed variance parameters σ_g^2 and σ_e^2. (The summations are over components of each vector, which is a slight abuse of notation for $||\hat{e}||^2$ and $||\hat{\beta}||^2$.)

- On the right sides, $\hat{e}_{\text{data}}$ and $\hat{\beta}_{\text{data}}$ are best linear unbiased predictions (BLUP) on the observed phenotype data assuming the basic linear mixed model $y = X\beta + e_{\text{proj}}$, equation (18), with

$$\text{Cov}(y) = \sigma_g^2 \frac{XX'}{M} + \sigma_e^2 \mathbf{P}_{\text{fixed}}.$$

- On the left sides, expectations are taken over the same quantities, with the phenotype data replaced by *random* y_{rand} *generated according to the model* $y = X\beta + e_{\text{proj}}$ with variance parameters set to the assumed σ_g^2 and σ_e^2.

We can therefore estimate the left sides (for a particular choice of σ_g^2 and σ_e^2) by Monte Carlo sampling, i.e., by generating random $y_{\text{rand}} = X\beta_{\text{rand}} + e_{\text{rand}}$ from $\beta_{\text{rand},m} \sim N(0, \sigma_g^2/M), e_{\text{rand},n} \sim N(0, \sigma_e^2)$ and then performing BLUP on y_{rand}—i.e., solving the mixed model equations (Section 1.2.1)—using the assumed σ_g^2 and σ_e^2 [26, 27]. Only a few Monte Carlo trials are required; BOLT-LMM computes 3–15 trials, with the number of trials decreasing as N increases, reflecting the way in which the standard error depends on sample size and the number of trials. Explicitly, BOLT-LMM by default uses $4 \times 10^9/N^2$ trials, with a minimum of 3 and a maximum of 15. The number of trials can also be varied as a user-specified parameter.

Explicitly, defining

$$\delta = \sigma_e^2/\sigma_g^2 \tag{29}$$

$$H = \frac{XX'}{M} + \delta \mathbf{I}_N, \tag{30}$$

Nature Genetics: doi:10.1038/ng.3190

the BLUP estimates $\hat{\beta}$ and $\hat{e}$ for a phenotype vector y are:

$$\hat{\beta} = \frac{1}{M} X' H^{-1} y \tag{31}$$

$$\hat{e} = \delta H^{-1} y. \tag{32}$$

(Note that technically, the scaled covariance matrix is actually $H_{\mathcal{S}} = \sigma_g^2 \mathrm{Cov}(y) = \frac{XX'}{M} + \delta \mathbf{P}_{\text{fixed}}$; however, because all of our vectors belong to the subspace $\mathcal{S}$, using $\mathbf{I}_N$ instead of $\mathbf{P}_{\text{fixed}}$ produces the correct result and has the convenience of making the operator H invertible in the full space $\mathbb{R}^N$.)

Equations (29)–(32) show that for a fixed value of δ, the BLUP predictions $\hat{e}$ and $\hat{\beta}$ are constant. Thus, the right sides of the REML first order conditions (28), involving BLUP on the observed phenotypes, depend only on the variance ratio δ and are independent of the variance scale, which we may parameterize by σ_g^2 (in which case $\sigma_e^2 = \sigma_g^2 \delta$ becomes a dependent variable). The left sides scale proportionately with σ_g^2 because scaling up the variances scales up randomly generated phenotypes y_{rand}. Therefore, finding σ_g^2 and σ_e^2 that solve the pair of equations (28) is equivalent to: (1) finding the value of the single parameter δ such that

$$\frac{E[\sum \hat{\beta}^2]}{E[\sum \hat{e}^2]} = \frac{\sum \hat{\beta}_{\text{data}}^2}{\sum \hat{e}_{\text{data}}^2}, \tag{33}$$

and (2) choosing σ_g^2 to scale the expectations to match the values observed on the data. (This procedure is analogous to the usual REML trick of optimizing over δ and then setting $\hat{\sigma}_g^2 = y' H^{-1} y / (N - C)$ [49].)

We propose an algorithm that rapidly estimates the ratio of expectations in equation (33) and uses this estimate within a one-parameter search over δ. Define

$$f_{\text{REML}}(\log \delta) = \log \left(\frac{\sum \hat{\beta}_{\text{data}}^2}{\sum \hat{e}_{\text{data}}^2} \Big/ \frac{E[\sum \hat{\beta}^2]}{E[\sum \hat{e}^2]} \right).$$

For a fixed value of δ, we produce Monte Carlo estimates of the expectations by generating random phenotypes

$$y_{\text{rand}} = X \beta_{\text{rand}} + e_{\text{rand, proj}},$$

Nature Genetics: doi:10.1038/ng.3190

where

$$\begin{aligned} \beta_{\text{rand}} &\sim \text{iid } N\left(0, \sqrt{1/M}\right) \\ e_{\text{rand}} &\sim \text{iid } N(0, \sqrt{\delta}) \\ e_{\text{rand, proj}} &= \mathbf{P}_{\text{fixed}} e_{\text{rand}}. \end{aligned}$$

(Note that as the variance scale parameter σ_g^2 is irrelevant for this part of the computation, we set it to 1 for convenience. Note also that the use of the projection matrix $\mathbf{P}_{\text{fixed}}$ is what makes this procedure compute REML rather than ML estimates.) We then run BLUP (using the chosen value of δ) using conjugate gradient iteration [28, 29] (Section 1.2.1) to obtain Monte Carlo estimates of $E[\sum \hat{\beta}^2]/E[\sum \hat{e}^2]$, and we likewise run BLUP to compute $\hat{\beta}_{\text{data}}$ and $\hat{e}_{\text{data}}$, which together give an estimate $\hat{f}_{\text{REML}}(\delta)$ of $f_{\text{REML}}(\delta)$.

We wish to find δ such that $f_{\text{REML}}(\log \delta) = 0$. We have observed empirically in simulations and real data sets that $f_{\text{REML}}(\log \delta)$ increases monotonically with δ except possibly at extremely small or large values of δ outside the range of reasonable parameter values (corresponding to, say, $0.01 < h_{\text{g}}{}^2 < 0.99$). Thus, one approach to finding the unique zero of f_{REML} within the reasonable parameter range is binary search, with the caveat that some care is needed because we can only compute noisy Monte Carlo estimates $\hat{f}_{\text{REML}}$ of f_{REML}.

We implement the following more robust approach. Instead of independently re-randomizing phenotypes y_{rand} used to compute estimates $\hat{f}_{\text{REML}}$ at different values of δ, we a generate a single set of random phenotypic component pairs

$$\begin{aligned} \beta_{\text{rand}} &\sim N\left(0, \sqrt{1/M}\right) \\ e_{\text{rand, unscaled}} &\sim N(0, 1) \\ e_{\text{rand, unscaled, proj}} &= \mathbf{P}_{\text{fixed}} e_{\text{rand, unscaled}} \end{aligned}$$

and then use these component pairs to generate random phenotypes

$$y_{\text{rand}} = X\beta_{\text{rand}} + \sqrt{\delta} \cdot e_{\text{rand, unscaled, proj}}.$$

for any given δ. Using this approach (taken in ref. [26]), the Monte Carlo estimate $\hat{f}_{\text{REML}}(\log \delta)$ becomes a smooth function of $\log \delta$, allowing us to use the secant method, a finite difference approximation of Newton's method, to perform zero-finding on $\hat{f}_{\text{REML}}$.

We note that satisfying the first-order conditions is not in general a sufficient condition to have found the REML global optimum, but for the case of only one non-identity variance component, the likelihood surface appears to be well-behaved, and we have never observed an instance in which this estimation procedure failed.

1.2.1 Conjugate gradient iteration to solve mixed model equations

The core computational subroutine needed for Monte Carlo REML (and also for calibration of the BOLT-LMM-inf statistic) is solution of the mixed model equations, i.e., computation of expressions of the form $V^{-1}x$, where

$$V = \sigma_g^2 \frac{XX'}{M} + \sigma_e^2 \mathbf{I}_N.$$

We solve the mixed model equations using conjugate gradient iteration as in refs. [28, 29], which requires only $O(MN)$-time matrix-vector products. More precisely, computing $V^{-1}x$ using conjugate gradient iteration only requires computing products of V with vectors, which we can compute in $O(MN)$-time without forming V by leaving V in factored form $\sigma_g^2 XX'/M + \sigma_e^2 \mathbf{I}_N$, distributing the product, and multiplying from right to left. Note that V is positive definite because XX' is positive semidefinite and $\sigma_g^2, \sigma_e^2 > 0$. Moreover, because most real phenotypes have only a small fraction of heritable variance, the identity component $\sigma_e^2 \mathbf{I}_N$ typically accounts for much of the covariance V, so that V is a well-conditioned matrix and conjugate gradient iteration converges rapidly (as determined by the residual norm dropping below a specified relative tolerance, which by default we set to 0.0005).

1.3 Gaussian mixture model fitting: Variational Bayes iteration

As outlined in Online Methods, the core component of the BOLT-LMM Gaussian mixture modeling algorithm is a variational iteration [20–22, 38] that computes approximate posterior mean effect sizes β_m for the Gaussian mixture model version of equation (18). We sketch the iteration in slightly more detail here, leaving a full description for Section 2.

Nature Genetics: doi:10.1038/ng.3190

The idea of the iteration is to obtain successively better approximations of the SNP effect sizes by cyclically updating each estimated SNP effect with its posterior mean conditional on the current estimates of all other SNP effects. Explicitly, we begin by initializing each SNP effect β_m to 0 and initializing the residual phenotype $y_{\text{resid}} = y - X\beta$ to y. Then, each iteration performs the following loop over SNPs $m = 1, \ldots, M$:

1. Remove effect of SNP m from residual. $y_{\text{resid},-m} \leftarrow y_{\text{resid}} + x_m\beta_m$
2. Re-estimate effect of SNP m. $\beta_m \leftarrow$ posterior mean given $y_{\text{resid},-m}$
3. Replace effect of SNP m in residual. $y_{\text{resid}} \leftarrow y_{\text{resid},-m} - x_m\beta_m$

Step 2 amounts to computing the posterior mean of β_m given its assumed mixture prior in equation (19), the environmental noise distribution $e_{\text{proj}} \sim N(0, \sigma_e^2 \mathbf{P}_{\text{fixed}})$, and the residual phenotype $y_{\text{resid},-m}$ given current estimates of all other SNP effects. Steps 1 and 3 are a computational technique allowing the residual in Step 2 not to be computed from scratch for each SNP, so that an entire iteration (updating each SNP effect once) takes $O(MN)$ time.

The basic structure of this iteration is the shared with previous variational methods [20–22, 38], with the update Step 2 varying among methods depending on the assumed SNP effect prior. As noted in Online Methods, an additional slight difference is that BOLT-LMM estimates the hyperparameters σ_g^2, σ_β^2 using REML and f_2, p using cross-validation [15], whereas previous approaches have performed hyperparameter estimation within the variational iteration [22, 38] or using the variational approximate log likelihood [21]. We used cross-validation instead because we found cross-validation estimates to be more robust to decreases in the quality of the variational approximation caused by linkage disequilibrium.

1.4 Cross-validation for estimating Gaussian mixture parameters

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. More precisely, for each parameter pair, for each cross-validation fold, BOLT-LMM estimates posterior mean SNP effect sizes β_m on the training data and uses these effect size estimates to make predictions on the left-out fold, which are then compared to the actual left-out phenotype values. (Note that effect size estimates used as coefficients in genomic prediction are conceptually completely different from effect sizes

Nature Genetics: doi:10.1038/ng.3190

output by an association test, as discussed in Section 1.1.3.) The parameter pair (f_2, p) producing the highest prediction R^2 (mean across folds) is then selected. However, if the best model only reduces prediction error by less than 1% (in absolute prediction R^2) compared to the infinitesimal model, BOLT-LMM terminates the Gaussin mixture analysis and does not proceed to Step 2b to compute association statistics. We implement this feature as the default option to save unnecessary computation in situations where the Gaussian mixture model provides little or no improvement in association power.

We perform 5-fold cross-validation to select best-fit Gaussian mixture model parameters according to maximal prediction R^2. For large sample sizes $N >$ 10,000, computing a subset of the cross-validation folds is already sufficient to obtain parameter estimates that achieve near-optimal association power, so by default, we compute only enough folds to make predictions on 10,000 test fold samples. This observation holds because absolute prediction R^2 corresponds directly to the power of the corresponding retrospective association test (as measured by increase in χ^2 statistics at truly associated loci, or equivalently, increase in effective sample size; see Online Methods), and the standard error of prediction R^2 estimated via cross-validation scales as the inverse square root of the number of test samples.

1.5 LD Score regression

LD Score regression [24] observes that for complex traits, properly calibrated χ^2 association statistics approximately obey the following linear trend (on average across the genome):

$$\chi^2 \sim 1 + (\text{constant}) \cdot (\text{LD Score}), \tag{34}$$

where loosely speaking, the LD Score of the test SNP is its effective number of LD partners. Because the intercept of this regression relation is 1, it follows that if we have computed a set of χ^2 statistics up to an unknown constant factor c, then one way to estimate c is to compute the intercept of the regression of χ^2 statistics on LD Scores. (Note that this approach only works on test statistics that are not vulnerable to proximal contamination; equation (34) does not hold for χ^2 statistics that are deflated due to LD between the candidate SNP and SNPs included in the model.)

Nature Genetics: doi:10.1038/ng.3190

To achieve very precise calibration, BOLT-LMM adopts a slightly more complex procedure that matches the intercept of LD Score regression on $\chi^2_{\text{BOLT-LMM}}$ (after calibration) to that of the properly-calibrated infinitesimal statistic $\chi^2_{\text{BOLT-LMM-inf}}$. We do so to increase robustness of the calibration to violations of the modeling assumptions underlying LD Score regression, which may result in attenuation bias (i.e., inflation of the intercept toward the mean χ^2) that causes the straightforward LD Score correction to association statistics to be conservative, though still much less conservative than genomic control. Because LD Score regression on either $\chi^2_{\text{BOLT-LMM}}$ or $\chi^2_{\text{BOLT-LMM-inf}}$ should be affected roughly equally by attenuation bias, matching LD intercepts guards against this potential problem (Supplementary Table 6).

1.6 Performance optimization

Beyond the broad algorithmic techniques we have described, the BOLT-LMM software employs a variety of performance optimizations that provide large constant-factor savings (>10x) in memory and running time over a straightforward implementation.

1.6.1 Memory

Computation on raw genotypes. The key optimization that we implement for memory efficiency is direct computation on raw genotypes. Raw genotypes can be stored in only 2 bits per base, versus 64 bits (8 bytes) for standard double-precision floating point values. However, all previous methods necessarily work with floating point matrices (either the $N \times N$ GRM or a floating point representation of the $N \times M$ normalized genotype matrix) because they perform spectral decomposition. In contrast, because BOLT-LMM applies iterative methods that use the genotype matrix only in matrix multiplications, it is enough for us to store the raw genotypes plus lookup tables containing normalization information that we apply on-the-fly. Explicitly, for each SNP, rather than storing a normalized genotype vector in memory, we simply store its raw allele count (0, 1, 2, or missing) for each individual and additionally record its mean allele count and normalization constant. Then, when performing computations involving the SNP, we build a properly normalized SNP vector using the above information. Importantly, this vector can be thrown away after the computation, thus keeping BOLT-LMM's memory footprint small.

Nature Genetics: doi:10.1038/ng.3190

An additional subtlety arises when working with covariates because normalized genotype vectors no longer contain only 4 values after projecting out covariates. In this case, we store the components of each normalized genotype vector along a basis that spans the covariates, and we do the same for the phenotype vector. We treat these "covariate component" values as additional coordinates that we carry along in all computations, so that whenever we need to compute a dot product (the basic computation all of our iterations use), we can do so by taking the usual dot product and subtracting the dot product of the additional covariate component vector.

Streaming computation of association tests. In analyses of data sets containing very large numbers of SNPs (e.g., millions of imputed SNPs) that may be stored as real-valued dosages, it is often desirable to compute association statistics for all SNPs but use only a subset of genotyped SNPs in the mixed model. In this situation, we retain memory efficiency by reading and analyzing SNPs *not* used in the mixed model only when performing final computation of retrospective association test statistics. That is, we first compute a set of residual phenotypes $y_{\text{resid-LOCO}}$ using only the subset of typed SNPs in the model; then, we successively read each test SNP and compute and output its association test, throwing away the SNP after completing the computation. This streaming computation allows us never to store the full set of SNPs in memory.

1.6.2 Computational speed

Batch computation using optimized matrix subroutines. The iterative methods we have described reduce the association computation to basic building blocks of vector operations (for variational iteration) and matrix-vector operations (for conjugate gradient iteration). These operations can be performed using optimized implementations of the Basic Linear Algebra Subprograms (BLAS), but all BLAS libraries achieve maximal speed when performing matrix-matrix "Level 3 BLAS" operations. We therefore batch our computations into matrix-matrix multiplications by performing simultaneous updates across SNPs and parameter values.

Conjugate gradient iteration uses Level 2 BLAS operations as written. Variational iteration only uses Level 1 BLAS as written (Section 1.3), so to step up to Level 2 BLAS, we perform block updates of SNP effect sizes. Explicitly, the basic variational iteration consists of re-estimating each SNP effect in turn conditioned on current estimates of other SNP effects, which requires computing

Nature Genetics: doi:10.1038/ng.3190

dot products of each SNP with the residual phenotype vector. Instead of computing dot products one at a time (a Level 1 BLAS operation), we compute dot products of a block of 64 SNPs at once (a matrix-vector multiplication, which is Level 2 BLAS). When we subsequently update the effect size of each SNP in turn, we just need to be careful to update its dot product to reflect changes that have been made to the residual vector due to previous updates of previous SNPs within the block. We do so by precomputing a 64×64 correlation matrix for each SNP block.

To step up from Level 2 BLAS to Level 3 BLAS, we make use of the fact that all computationally intensive steps of the BOLT-LMM algorithm require multiple replicates of almost the same computation: Step 1a computes BLUP on different random phenotypes; Step 1b solves a set of similar linear systems for different LOCO reps and calibration SNPs; Step 2a computes variational Bayes assuming different hyperparameter values; Step 2b computes variational Bayes for different LOCO reps. With some care, we can in each case simultaneously perform iterations across the different replicates. The upshot is that by using batch computations, BOLT-LMM performs all $O(MN)$-time operations using BLAS 3 matrix-matrix multiplications (DGEMM). Our software distribution uses the well-tuned Intel Math Kernel Library (MKL) BLAS implementation.

Multithreading. Conveniently, most BLAS implementations also support optimized multithreaded computation on multi-core processors, which are now commonplace. In practice, multithreading rarely decreases computation time by a factor equal to the number of cores used because of various overhead costs, but Level 3 BLAS operations often come close to achieving this theoretically maximal speedup. The BOLT-LMM software supports multithreading through BLAS, and we recommend using this option whenever multiple cores are available. We performed our benchmarking analyses using single-core computation simply to give a fair comparison against methods that do not support multithreading.

Low-level optimization. We reduce the overhead of loading raw genotypes and building normalized genotype vectors using a few additional low-level tricks. Instead of looking up normalized genotype values one at a time when loading a SNP, we build an intermediate lookup table that maps all 256 possible values of 4 consecutive genotypes (stored in 1 byte) to a 4-vector of normalized values. When loading these 4-vectors from the lookup table into the normalized SNP vector being

Nature Genetics: doi:10.1038/ng.3190

built, we use streaming SIMD extensions (SSE instructions) to perform multiple loads at once.

2 Theory

2.1 Variational Bayes

In Bayesian analysis, one specifies a probability model over observations and model parameters, often wishing to obtain posterior mean estimates of parameters of interest. Unfortunately, the posterior distribution is typically infeasible to integrate over. The idea of the variational framework (termed "variational Bayes") is to approximate the posterior distribution with a factorized form that is much easier to work with computationally. The factorized form can be integrated over, allowing calculation of approximate posterior mean estimates. For a full discussion of variational Bayes, we refer to ref. [50]. Previous variational methods for Bayesian linear regression in genetics have been presented in refs. [21,22,38]. Here, we briefly summarize key aspects of variational Bayes for quick reference and to establish notation. We then derive formulas giving the variational iteration used by BOLT-LMM, and we further derive theory establishing equivalence between this family of variational methods and penalized linear regression.

2.1.1 Terminology and notation

We begin by considering a general probability model $p(y, \beta)$ over observations y and parameters β. We assume that $p(y, \beta)$ takes into account the prior distributions of β, so that $p(y, \beta)$ is the joint probability of sampling parameters β and observing y.

- The true *log likelihood* (LL) of observing y is the log of the integral of $p(y, \beta)$ over all possible values of the parameters β:

$$\text{true LL} = \log \int p(y, \beta) d\beta.$$

  This integral is typically intractable to compute.

- The true *posterior distribution* of the parameters β conditional on observing y is given by

Nature Genetics: doi:10.1038/ng.3190

normalizing the joint probability $p(y, \beta)$ by the likelihood:

$$\text{true posterior distribution} = p(\beta \mid y) = \frac{p(y, \beta)}{\int p(y, \beta) d\beta}.$$

- The *variational approximation to the posterior distribution* is the best approximation of the posterior distribution $p(\beta \mid y)$ with a distribution $q(\beta)$ that factors:

$$\text{approx posterior distribution} = q(\beta) = \prod_i q_i \approx p(\beta \mid y).$$

  The factors q_i are usually constrained to have simple forms. For instance, if β represents a set of individual parameters β_i, each factor q_i may be required to be a function of only one parameter β_i, in which case q_i is an approximate marginal posterior for β_i. This *fully factorized* variational approximation is the approach we consider here.

- The *approximate log likelihood* is a variational lower bound on the true log likelihood, given by:

$$\begin{aligned} \text{approx LL} = \mathcal{L}(q) &= \int q(\beta) \log \frac{p(y, \beta)}{q(\beta)} d\beta \\ &= \text{true LL} - D_{KL}(q(\beta) \parallel p(\beta \mid y)), \end{aligned} \tag{35}$$

  where D_{KL} denotes the Kullback-Liebler (KL) divergence between probability distributions. Note that the second line says that the gap between the variational lower bound and the true LL is given by the KL divergence between the approximating distribution q and the true posterior distribution $p(\beta \mid y)$.

Variational iteration, which we describe in the next section, successively refines the approximation of $p(\beta \mid y)$ with $q(\beta)$ by iteratively updating the factors q_i, reducing the KL divergence $D_{KL}(q(\beta) \parallel p(\beta \mid y))$ and therefore monotonically improving the lower bound $\mathcal{L}(q)$. The point of the iteration is that at convergence, the distribution $q(\beta)$ will ideally be a good approximation of the true posterior distribution $p(\beta \mid y)$ that can easily be integrated over (because of its factored form) to perform approximate posterior inference.

Nature Genetics: doi:10.1038/ng.3190

Before moving on, we note that it follows from the above that the faithfulness of the approximating distribution $q(\beta)$ to the true posterior $p(\beta \mid y)$ determines the accuracy of the resulting approximate posterior inference. Because we only consider approximating distributions $q(\beta)$ that factor as $\prod_i q_i$, the optimal quality of the approximation depends on the extent to which the posterior distribution $p(\beta \mid y)$ can be approximately factored in the prescribed manner. For Bayesian linear regression, which is our focus here, correlation among regressors (i.e., linkage disequilibrium among markers) tends to pose a challenge for some types of inference using variational methods. For example, when regressors are correlated, the approximate posterior is often too tightly concentrated (i.e., overconfident of parameter localization). However, aggregate inference can still be robust. We refer to ref. [21] for an in-depth exploration of these issues.

2.1.2 General variational iteration

The variational Bayes algorithm iteratively updates each approximate marginal distribution q_i using the update step

$$\log(q_i(\beta_i)) \leftarrow E_{q_{i'},\, i' \neq i}[\log p(y, \beta)] + C.$$

As written, the above update step updates the entire distribution q_i—a *function* of β_i—to its new optimum (as a distribution, optimizing in the sense of calculus of variations). Indeed, the only assumption made by the variational approximation is that the approximate posterior distribution $q(\beta)$ factors (i.e., the parameters are assumed to be conditionally independent, conditional on the observed outputs): no assumption is made about the functional form of the factors $q_i(\beta_i)$. However, the factor distributions q_i are typically characterized by sets of sufficient statistics, so that in each iteration, only the sufficient statistics for factor i need to be updated based on the current values of the sufficient statistics for each of the other factors. Moreover, we will show in Section 2.1.3 that in the case of Bayesian linear regression, the factor distributions take the form of conditional posterior distributions. For the Gaussian mixture priors we use in BOLT-LMM, these conditional posteriors retain the Gaussian mixture form, just with different parameters.

Convex optimization theory guarantees the convergence of variational iteration [45]. Moreover, convergence of the approximate log likelihood, which monotonically increases during the iteration, serves as a convergence criterion. At convergence, approximate posterior means of the parameters

β_i can be read off from the approximate marginal distributions q_i: for the special case of fully factorized Bayesian linear regression, these values are among the sufficient statistics updated at each iteration.

2.1.3 Variational iteration for Bayesian linear regression

We now specialize to the Bayesian linear regression model

$$\begin{aligned} y &= X\beta + e \\ e &\sim N(0, \sigma_e^2 \mathbf{I}_N) \\ \text{prior on } \beta_m &= \pi_m(\beta_m) \end{aligned}$$

That is, we assume each regression coefficient (i.e., SNP effect) has a prior distribution $\pi_m(\beta_m)$ and the response y is the sum of regressor effects plus iid Gaussian noise with variance σ_e^2. (Note that this noise model does not take into account projecting out covariates as described in Section 1.1; projecting out C independent covariates just reduces N to $N - C$ in the equations below.) The joint probability $p(y, \beta)$ of parameters β and observations y satisfies

$$p(y, \beta) = p_{\text{noise}}(y - X\beta) \prod \pi_m(\beta_m) = \left(\frac{1}{\sqrt{2\pi\sigma_e^2}} \right)^N \exp\left(-\frac{||y - X\beta||^2}{2\sigma_e^2} \right) \prod_{m=1}^{M} \pi_m(\beta_m). \quad (36)$$

Below we derive the fact that the fully factorized variational approximation matches the iterative conditional update algorithm discussed in Section 1.3. The fully factorized variational approximation takes the form

$$q(\beta) = \prod_{m=1}^{M} q_m(\beta_m) \approx p(\beta \mid y).$$

The variational update step for the approximate marginal distribution q_m optimizes the KL divergence $D_{KL}(q||p)$ over all distributions q_m, fixing the other marginals $q_{m'}$ to their current distributions, thereby monotonically increasing the approximate (lower bound) log likelihood $\mathcal{L}(q)$. The

Nature Genetics: doi:10.1038/ng.3190

update amounts to:

$$\begin{aligned}
\log(q_m(\beta_m)) &\leftarrow E_{q_{m'}(\beta_{m'}),\, m' \neq m}[\log p(y, \beta)] + C \\
&= E_{q_{m'}(\beta_{m'}),\, m' \neq m}\left[-\frac{||y - X\beta||^2}{2\sigma_e^2} + \sum_{m'=1}^{M} \log \pi_{m'}(\beta_{m'})\right] + C \\
&= -\frac{||y - X_{-m}\overline{\beta}_{-m} - x_m\beta_m||^2}{2\sigma_e^2} + \log \pi_m(\beta_m) + C \\
&= -\frac{||y_{\text{resid},-m} - x_m\beta_m||^2}{2\sigma_e^2} + \log \pi_m(\beta_m) + C,
\end{aligned}$$

where $\overline{\beta}_{-m}$ denotes the vector of estimated posterior mean effect sizes at SNPs other than m according to the current approximate marginal posterior distributions $q_{m'}(\beta_{m'})$. (In the above sequence of equations, we absorb all terms independent of β_m into the constant C; the use of "$+C$" in all lines does not imply that the constant is the same from line to line.) To see how the posterior mean estimates $\overline{\beta}_{-m}$ enter the equation in the third line above, consider expanding the quadratic term $||y - X\beta||^2$ inside the expectation. (Note that taking an expectation over $q_{m'}(\beta_{m'}),\ m' \neq m$ corresponds to integrating over $q_{m'}(\beta_{m'})d\beta_{m'}$.)

- For terms that are linear in $\beta_{m'}$ (for $m' \neq m$), $E[\beta_{m'}]$ can be replaced with $\overline{\beta}_{m'}$ by linearity of expectation over $q_{m'}(\beta_{m'})$.

- Terms that are quadratic in $\beta_{m'}$ are independent of β_m, so $E[\beta_{m'}^2]$ can be replaced by $\overline{\beta}_{m'}^2 + \text{Var}_{q_{m'}}(\beta_{m'})$ to re-complete the square $||y - X_{-m}\overline{\beta}_{-m} - x_m\beta_m||^2$. The leftover terms, proportional to $\text{Var}_{q_{m'}}(\beta_{m'}) = E_{q_{m'}}[\beta_{m'}^2] - \overline{\beta}_{m'}^2$, are constant (with respect to β_m) and can be absorbed into the constant term.

Finally, the contributions of the prior terms $\pi_{m'}(\beta_{m'})$ for $m' \neq m$ also become additive constants upon taking expectations and can be absorbed into the constant term as well.

Exponentiating the result above gives

$$q_m(\beta_m) \propto \exp\left(-\frac{||y_{\text{resid},-m} - x_m\beta_m||^2}{2\sigma_e^2}\right)\pi_m(\beta_m), \tag{37}$$

which is simply a conditional posterior distribution for β_m (given the prior $\pi_m(\beta_m)$ and conditional on setting all other $\beta_{m'}$ to their variational expected means). That is, for Bayesian linear regression,

Nature Genetics: doi:10.1038/ng.3190

the optimal approximating marginal distribution $q(\beta_m)$ (making no assumptions about the form of each marginal, only requiring that the approximating posterior fully factorizes) turns out to be the conditional posterior. In particular, as we derive in Section 2.1.4, if the prior has the nice form of a mixture of Gaussians, then the approximating marginal keeps that form (with different means, standard deviations, and weights).

Moreover, the approximate log likelihood $\mathcal{L}(q)$ can be obtained by substituting the joint probability $p(y, \beta) = p_{\text{noise}}(y - X\beta) \prod \pi_m$, given in equation (36), into equation (35):

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\beta) \log \frac{p(y,\beta)}{q(\beta)} d\beta \\
&= E_q[\log p_{\text{noise}}(y - X\beta)] + \sum E_q[\log \pi_m / q_m] \\
&= E_q[\log p_{\text{noise}}(y - X\beta)] - \sum D_{KL}(q_m || \pi_m) \\
&= -\frac{N}{2} \log 2\pi\sigma_e^2 - \frac{||y - X\overline{\beta}||^2}{2\sigma_e^2} - \frac{\sum ||x_m||^2 \text{Var}_{q_m}(\beta_m)}{2\sigma_e^2} - \sum D_{KL}(q_m || \pi_m) \\
&= -\frac{N}{2} \log 2\pi\sigma_e^2 - \frac{||y - X\overline{\beta}||^2}{2\sigma_e^2} - \frac{\sum ||x_m||^2 \text{Var}_{q_m}(\beta_m)}{2\sigma_e^2} + \sum (E_{q_m}[\log \pi_m] + H(q_m)),
\end{aligned} \tag{38}$$

where $\overline{\beta}_m = E_{q_m}[\beta_m]$, using the same trick of completing the square as above, only now we have to be careful to account for the leftover constant terms involving $\text{Var}_{q_m}(\beta_m)$. In the last line, $H(q_m)$ denotes information entropy. Once again, we note that if C independent covariates are projected out from our model as described in Section 1.1, then we simply reduce N to $N - C$ in equation (38) to account for the lost degrees of freedom.

2.1.4 Update equations for Gaussian mixture prior

We now specialize to Bayesian linear regression with the specific Gaussian mixture prior

$$\beta_m \sim \begin{cases} N(0, \sigma_{\beta,1}^2) & \text{with probability } p \\ N(0, \sigma_{\beta,2}^2) & \text{with probability } 1 - p, \end{cases} \tag{39}$$

Nature Genetics: doi:10.1038/ng.3190

i.e.,

$$\pi_m(\beta_m) = p \cdot \frac{1}{\sqrt{2\pi\sigma^2_{\beta,1}}} \exp\left(-\frac{\beta_m^2}{2\sigma^2_{\beta,1}}\right) + (1-p) \cdot \frac{1}{\sqrt{2\pi\sigma^2_{\beta,2}}} \exp\left(-\frac{\beta_m^2}{2\sigma^2_{\beta,2}}\right).$$

To obtain the the conditional marginal distribution $q_m(\beta_m)$, we substitute the above into equation (37), obtaining

$$q_m(\beta_m) = p_m \cdot \frac{1}{\sqrt{2\pi\tau^2_{m,1}}} \exp\left(-\frac{(\beta_m - \overline{\beta}_{m,1})^2}{2\tau^2_{m,1}}\right) + (1-p_m) \cdot \frac{1}{\sqrt{2\pi\tau^2_{m,2}}} \exp\left(-\frac{(\beta_m - \overline{\beta}_{m,2})^2}{2\tau^2_{m,2}}\right),$$

i.e.,

$$\beta_m \sim_{q_m} \begin{cases} N\left(\overline{\beta}_{m,1}, \tau^2_{m,1}\right) & \text{with probability } p_m \\ N\left(\overline{\beta}_{m,2}, \tau^2_{m,2}\right) & \text{with probability } 1-p_m, \end{cases}$$

where

$$\begin{aligned} \hat{\beta}_m &:= \frac{y^T_{\text{resid},-m} x_m}{||x_m||^2} \\ \overline{\beta}_{m,k} &:= \frac{\sigma^2_{\beta,k}}{\sigma^2_{\beta,k} + (\sigma^2_e/||x_m||^2)} \hat{\beta}_m \\ \tau^2_{m,k} &:= \frac{1}{\frac{1}{\sigma^2_e/||x_m||^2} + \frac{1}{\sigma^2_{\beta,k}}} = \frac{\sigma^2_{\beta,k}\sigma^2_e/||x_m||^2}{\sigma^2_{\beta,k} + (\sigma^2_e/||x_m||^2)} \\ p_m &:= \frac{\frac{p}{s_1} e^{-\hat{\beta}^2_m/2s^2_1}}{\frac{p}{s_1} e^{-\hat{\beta}^2_m/2s^2_1} + \frac{1-p}{s_2} e^{-\hat{\beta}^2_m/2s^2_2}} \end{aligned}$$

where s_1^2 and s_2^2 are the variances of the marginal distribution of $\hat{\beta}_m$ in each case (assuming $y_{\text{resid},-m} = x_m\beta_m + e$ and marginalizing over random β_m and e):

$$s^2_k := \sigma^2_{\beta,k} + \frac{\sigma^2_e}{||x_m||^2}$$

for $k = 1, 2$. These equations explicitly give the update formulas for the BOLT-LMM variational iteration (Step 2 in Section 1.3), in which each estimated SNP effect is set to its conditional posterior mean.

To compute the approximate log likelihood for use in testing convergence of the iteration, it is convenient to consider a reparameterization of the model (following ref. [21]) in which instead of

Nature Genetics: doi:10.1038/ng.3190

having one parameter β_m for each SNP (with $\pi_m(\beta_m)$ a mixture of two Gaussians), we introduce a state parameter $s_m \in \{0, 1\}$ and consider a prior $\pi_m(\beta_m, s_m)$ with:

$$\begin{aligned} \pi_m(\beta_m \mid s_m = 0) &= \text{Gaussian 1: } N(0, \sigma^2_{\beta,1}) \\ \pi_m(\beta_m \mid s_m = 1) &= \text{Gaussian 2: } N(0, \sigma^2_{\beta,2}) \\ \pi_m(s_m = 0) &= \text{mixture fraction: } p. \end{aligned}$$

This parameterization gives exactly the same model as the original Bayesian linear regression, with the only difference being that the mixture prior on β_m in equation (39) is replaced by the joint prior distribution $\pi_m(\beta_m, s_m)$ above with the extra hidden state s_m.

We can now apply the variational approach to factorize the approximate posterior distribution over all SNPs—in (β_m, s_m) pairs—as a product of joint distributions $q_m(\beta_m, s_m)$. Each optimal factor distribution q_m has the property that when integrated over s_m, it matches the variational approximation $q_m(\beta_m)$ from the original parameterization, implying that the estimated posterior means are identical. This parameterization lends itself more easily to computation of KL divergences, however. We have the formula:

$$\begin{aligned} D_{KL}(q_m(\beta_m) \parallel \pi_m(\beta_m)) &= D_{KL}(q_m(\beta_m, s_m) \parallel \pi_m(\beta_m, s_m)) \\ &= p_m \log \frac{p_m}{p} + (1 - p_m) \log \frac{1 - p_m}{1 - p} \\ &\quad - \frac{p_m}{2} \left(1 + \log \frac{\tau^2_{m,1}}{\sigma^2_{\beta,1}} - \frac{\tau^2_{m,1} + \overline{\beta}^2_{m,1}}{\sigma^2_{\beta,1}} \right) \\ &\quad - \frac{1 - p_m}{2} \left(1 + \log \frac{\tau^2_{m,2}}{\sigma^2_{\beta,2}} - \frac{\tau^2_{m,2} + \overline{\beta}^2_{m,2}}{\sigma^2_{\beta,2}} \right). \end{aligned}$$

Writing

$$\text{Var}_{q_m}(\beta_m) = E_{q_m}[\beta^2_m] - \overline{\beta}^2_m = p_m(\tau^2_{m,1} + \overline{\beta}^2_{m,1}) + (1 - p_m)(\tau^2_{m,2} + \overline{\beta}^2_{m,2}) - \overline{\beta}^2_m,$$

we have all of the terms needed to compute $\mathcal{L}(q)$ according to equation (38).

We note that in the limit $\sigma_{\beta,2} \to 0$ of the point-normal prior used by ref. [21], the formulas

Nature Genetics: doi:10.1038/ng.3190

become

$$\begin{aligned} D_{KL}(q_m(\beta_m) \parallel \pi_m(\beta_m)) &= p_m \log \frac{p_m}{p} + (1-p_m) \log \frac{1-p_m}{1-p} \\ &\quad - \frac{p_m}{2}\left(1 + \log \frac{\tau_{m,1}^2}{\sigma_{\beta,1}^2} - \frac{\tau_{m,1}^2 + \overline{\beta}_{m,1}^2}{\sigma_{\beta,1}^2}\right) \\ \text{Var}_{q_m}(\beta_m) &= p_m(\tau_{m,1}^2 + \overline{\beta}_{m,1}^2) - \overline{\beta}_m^2 \\ &= p_m(\tau_{m,1}^2 + \overline{\beta}_{m,1}^2) - (p_m \overline{\beta}_{m,1})^2, \end{aligned}$$

matching ref. [21].

2.1.5 Equivalence with penalized linear regression

We now show that the variational iteration of Section 2.1.3 for Bayesian linear regression is equivalent to coordinate descent applied to a penalized linear regression problem. Equivalently, the fully factorized variational approximation to Bayesian linear regression (for an arbitrary choice of prior on regressor effects) can be recast as applying a transformation to the prior and then finding a posterior mode of the (new) Bayesian linear regression with transformed prior. More precisely, we have the following *equivalence of optimization problems*:

1. Maximize the variational approximate log likelihood (for the fully factored VB approximation to Bayesian linear regression).

2. Minimize the penalized linear regression objective function

$$||y - X\beta||^2 + \sum \text{penalty}_m(\beta_m),$$

   where the penalty is derived from the prior in the original Bayesian linear regression.

3. Find a posterior mode of the Bayesian linear regression with prior

$$\tilde{\pi}_m(\beta_m) \propto \exp(-\text{penalty}_m(\beta_m)),$$

   which we can view as a transformation of the original prior $\pi_m(\beta_m)$.

Nature Genetics: doi:10.1038/ng.3190

We also have the following *equivalence of algorithms*:

1. Variational iteration applied to the original Bayesian linear regression.

2. Coordinate descent applied to the corresponding penalized linear regression.

Implications for numerical optimization. The equivalence of variational Bayes with penalized linear regression (in the context of Bayesian linear regression) elucidates some numerical properties of the algorithm that are not immediately apparent. In particular, penalized linear regression is in general a numerically challenging non-convex optimization. We can therefore expect variational iteration to be susceptible to numerical issues such as convergence to local optima. Some methods have tried to address this problem by repeating the iteration for different random choices of update orders or different initialization points [21,22,38]. In our simulations, we found that using a single run of the iteration was typically already sufficient for BOLT-LMM to achieve most of the available power gain, however, so we opted to use just a single run to avoid increasing computational cost.

Mathematical derivation. As discussed in Section 2.1.1, variational theory gives the following lower bound to the log likelihood, which variational iteration tries to maximize:

$$\mathcal{L}(q) = -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{||y - \sum x_m E_{q_m}[\beta_m]||^2}{2\sigma_e^2} - \frac{\sum ||x_m||^2 \text{Var}_{q_m}(\beta_m)}{2\sigma_e^2} - \sum D_{KL}(q_m||\pi_m).$$

Note that $\mathcal{L}(q)$ is a *functional*, i.e., it evaluates factored distributions $q = q_1 \cdots q_M$ that attempt to approximate the posterior. We will show that in fact, this maximization is equivalent to optimizing the objective function (over effect size estimates $\hat{\beta}_m$) of a penalized linear regression:

$$L(\hat{\beta}) = -\frac{N}{2}\log 2\pi\sigma_e^2 - \frac{||y - X\hat{\beta}||^2}{2\sigma_e^2} - \sum \text{penalty}_m(\hat{\beta}_m).$$

The key is that we can define a mapping $\hat{\beta}_m \mapsto q_m(\,\cdot\,;\hat{\beta}_m)$ with $E_{q_m}[\beta_m] = \hat{\beta}_m$ and define

$$\text{penalty}_m(\hat{\beta}_m) := \frac{||x_m||^2 \text{Var}(q_m(\,\cdot\,;\hat{\beta}_m))}{2\sigma_e^2} + \sum D_{KL}(q_m(\,\cdot\,;\hat{\beta}_m)||\pi_m),$$

Nature Genetics: doi:10.1038/ng.3190

with the property that

$$\begin{aligned}
\text{(penalized LR objective)} \quad L(\hat{\beta}) &= \mathcal{L}\left(q = \prod q_m(\,\cdot\,;\hat{\beta}_m)\right) \quad \text{(approx LL with derived } q_m\text{)} \\
&\leq \max\left\{\mathcal{L}(q) : q = \prod q_m, E_{q_m}[\beta_m] = \hat{\beta}_m\right\} \qquad (40) \\
&\leq \max \mathcal{L}(q). \qquad (41)
\end{aligned}$$

(Note that if the SNPs are normalized such that $||x_m||$ are all equal, then the penalty is independent of m.)

The mapping $\hat{\beta}_m \mapsto q_m(\,\cdot\,;\hat{\beta}_m)$ is derived from the variational update step, which we know chooses the q_m that optimizes $\mathcal{L}(q)$ conditional on the current choices of the other marginal distributions. Thus, equality holds in (40) at any convergence "point" of the iteration. Here a convergence "point" really means a choice of approximating distributions $q_1, \ldots, q_M$; however, if the above mapping exists, these distributions may be parameterized by $\hat{\beta}_1, \ldots, \hat{\beta}_M$. Equality holds in (41) at the global maximizer of the variational approximate log likelihood. It follows that $L(\hat{\beta})$, which is upper-bounded by the global maximum of $\mathcal{L}(q)$, attains that global maximum at $\hat{\beta}_{m,\text{opt}} = E_{q_{m,\text{opt}}}[\beta_m]$. That is, the solution to the penalized linear regression optimization of $L(\hat{\beta})$ corresponds to the solution to the variational optimization of $\mathcal{L}(q)$: if we could solve the penalized linear regression optimally, we would have the optimal variational Bayes solution.

All that remains is to define the mapping $\hat{\beta}_m \mapsto q_m(\,\cdot\,;\hat{\beta}_m)$, which we obtain indirectly from the variational update step via the key observation that the optimal marginal q_m conditioned on all other marginals depends only on a single statistic: the correlation of x_m with the residual $y_{\text{resid},-m} = y - \sum_{m' \neq m} x_{m'} E_{q_{m'}}[\beta_{m'}]$. That is, we have:

$$x_m^T y_{\text{resid},-m} \quad \mapsto \quad q_m \quad \mapsto \quad \begin{cases} \text{penalty terms Var}(q_m), D_{KL}(q_m||\pi_m) \\ \text{conditional posterior mean } E_{q_m}[\beta_m] \end{cases}$$

Moreover, the conditional posterior mean $E_{q_m}[\beta_m]$ is an increasing function of the correlation statistic $x_m^T y_{\text{resid},-m}$, meaning that we can invert the final mapping to obtain $E_{q_m}[\beta_m] \mapsto q_m$, giving the desired mapping $\hat{\beta}_m \mapsto q_m(\cdot;\hat{\beta}_m)$. If we wanted to derive an explicit formula for $\text{penalty}_m(\beta_m)$, we could try to invert the mapping $x_m^T y_{\text{resid},-m} \mapsto E_{q_m}[\beta_m]$, but in general there is probably no closed form.

Nature Genetics: doi:10.1038/ng.3190

2.2 Convergence rate of BOLT-LMM iterative computations

Because BOLT-LMM applies iterative methods for numerical linear algebra, its running time depends not only on the cost of matrix operations, which scales linearly with M and N, but also the number of $O(MN)$-time iterations required for convergence, which is largely determined by the condition number of the phenotypic covariance matrix and in particular increases with sample size (N), heritability, relatedness, and population structure. Simulations show empirically that the number of iterations scales roughly with $\sqrt{N}$ and does not change dramatically within the range of typical values of the other parameters, hence our estimate of an overall running time that scales roughly with $MN^{1.5}$ (Supplementary Fig. 1 and Supplementary Fig. 2).

More precisely, we measured (for varying N) the numbers of iterations required by BOLT-LMM-inf and BOLT-LMM to reach convergence, as defined by: (1) the conjugate gradient residual norm dropping below a relative tolerance of 0.0005, in BOLT-LMM-inf; and (2) the variational Bayes change in approximate log likelihood dropping below 0.01, in BOLT-LMM. For both methods, we observed that the number of iterations increased roughly as $\sqrt{N}$ (Supplementary Fig. 1). For BOLT-LMM-inf, our intuition (based on random matrix theory) is that the conditioning of the covariance matrix becomes modestly worse as the sample size increases; we believe this phenomenon is responsible for the $\approx \sqrt{N}$ increase in conjugate gradient iterations. For BOLT-LMM, which faces a nonlinear problem due to its Gaussian mixture prior on effect sizes, we are not aware of theoretical bounds on the convergence rate of variational iteration. However, the nonlinearity appears not to be a major hurdle in practice, perhaps because heritability is typically low and also because we limit the nonlinearity by bounding the extent to which the mixture prior deviates from normality (by bounding the mixture parameters f_2 and p).

We note that the high observed-scale heritabilities of case-control ascertained traits and inflated heritabilities of phenotypes in family data, in combination with population structure or relatedness among samples, may reduce the computational efficiency of BOLT-LMM (Supplementary Fig. 2); an avenue for future investigation is to mitigate this effect by including PCs to improve the numerical conditioning of the computation.

Nature Genetics: doi:10.1038/ng.3190

2.3 Correspondence between association power and prediction accuracy

Intuitively, mixed model analysis to test a candidate marker for association gains power over linear regression (assuming no confounding) by conditioning on other markers modeled as random effects. This intuition is especially clear in the BOLT-LMM statistical formulation, where we retrospectively test candidate markers for association with residual phenotypes. Residualizing eliminates the component of phenotype that was successfully predicted by other markers, so that the subsequent association test is trying to detect candidate marker association signal amid phenotypic variance ("noise") that has been decreased by a factor of 1-R^2. (Here, R^2 denotes out-of-sample prediction R^2; even though in practice we residualize by genetic predictions computed in-sample, the extent to which these predictions capture true genetic effects corresponds to out-of-sample prediction R^2.) Note that residualizing leaves the amount of candidate marker signal unchanged, assuming no proximal contamination and assuming a randomly ascertained phenotype; violation of these assumptions results in loss of power [12].

Quantitatively, a reduction of association test noise by a factor of 1-R^2 results in an increase in χ^2 statistics by a factor of 1/(1-R^2) at truly associated loci. (More precisely, the quantity χ^2-1 increases by 1/(1-R^2) on average, but for $\chi^2 \gg 1$ at known loci, this distinction is minor.) This relationship can be seen explicitly in comparing equation (1) of ref. [34] for prediction R^2 using BLUP,

$$R^2 = \frac{h_g^2}{1 + \frac{M}{Nh_g^2}(1 - R^2)}, \tag{42}$$

with Supplementary Table 2 of ref. [12], which gives the following formulas for mean χ^2 statistics using linear regression vs. mixed model association:

$$\text{mean } \chi^2_{\text{LR}} \text{ at causal SNPs} \;=\; 1 + \frac{Mh_g^2}{M_{\text{causal}}} \tag{43}$$

$$\text{mean } \chi^2_{\text{LMM}} \text{ at causal SNPs} \;=\; 1 + \frac{Mh_g^2}{M_{\text{causal}}} \cdot \frac{1}{1 - R^2}, \tag{44}$$

where R^2 in equation (44) (denoted $r^2 h_g^2$ in ref. [12]) satisfies the same quadratic equation (42).

In Fig. 3a, we plot increases in mean χ^2 statistics at known loci for mixed model methods vs. PCA, and in Fig. 3b, we plot absolute prediction R^2. As explained above, the increase in mean χ^2 statistics at known loci for mixed model methods vs. linear regression—assuming no

Nature Genetics: doi:10.1038/ng.3190

confounding—is approximately $1/(1-R^2)-1 \approx R^2$. When including principal components in linear regression to avoid confounding, we get an increase in mean χ^2 statistics at known loci for mixed model methods vs. PCA of approximately ${R^2}_{\text{LMM}} - {R^2}_{\text{PCA}}$, where the latter is the prediction R^2 obtained by linear prediction using only the principal components. Thus, the red bars in Fig. 3a roughly correspond to differences between red and black bars in Fig. 3b, and analogously for blue bars. The correspondence is approximate because of the approximations mentioned above and two additional subtleties: (1) the LOCO scheme that BOLT-LMM uses to avoid proximal contamination renders the mixed model unable to condition out effects of markers on the same chromosome as the candidate marker; (2) our measurement of prediction R^2 uses 5-fold cross-validation, reducing the training sample size by 20%. However, these two effects have small magnitudes and act in opposite directions, so that ultimately the correspondence is very tight, as is visually apparent in Fig. 3.

We note two caveats to the correspondence between association power and prediction accuracy claimed above. First, in ascertained case-control data sets, mixed model methods that achieve positive prediction R^2 may actually reduce association power due to ascertainment-induced "linkage disequilibrium" between causal markers [12]. The result of this phenomenon is that residualizing by other SNPs has a deleterious, masking effect on the association signal at a candidate SNP, similar to proximal contimation [5, 9, 12]. Second, in data sets with pervasive family relatedness, the correspondence between association power and prediction accuracy is again less clear-cut. Our intuition here is similar: family relatedness introduces "long-range LD" that produces an effect similar to proximal contamination.

3 Parameters of mixed model software used in analyses

We ran BOLT-LMM with default options except in analyses investigating power of the Gaussian mixture model, in which we used the `--forceNonInf` option to fit the Gaussian mixture model even in scenarios with no expected power gain.

We ran the 2012-02-10 intel64 release of EMMAX with default options. We ran GEMMA version 0.94 with default options. We ran GCTA version 1.24 with the `--mlma-loco` option for leave-one-chromosome-out analysis. We ran FaST-LMM version 2.07 using all markers to

compute the similarity matrix. We ran FaST-LMM-Select version 2.07 with the autoSelect option

```
-autoSelectSearchValues
  "0,1,3,10,30,100,300,1000,3000,10000,30000,100000,300000"
```

to test increasing numbers of selected SNPs up to a maximum of 300,000 SNPs. We benchmarked the precompiled executable distribution of each software package.

4 Pseudocode outlining BOLT-LMM algorithm

Step 1a: Estimate variance parameters.
Algorithm: Use secant iteration to find approximate zero of derivative of log likelihood.

```
X := normalized genotypes   # N x M matrix
y := phenotype vector       # N-vector

MCtrials := max(min(4e9/N^2, 15), 3)  # set number of Monte Carlo trials
for t = 1 to MCtrials
  for j = 1 to M
    β_rand[j,t] := randn() * sqrt(1/M) # generate random SNP effects
  for i = 1 to N
    e_rand,unscaled[i,t] := randn() # generate random environmental effects
end for

h_1^2 := 0.25, logDelta_1 := log((1-h_1^2)/h_1^2)
f_1 := evalfREML(logDelta_1)  # perform first fREML evaluation

if f_1 < 0, h_2^2 := 0.125; else h_2^2 := 0.5
logDelta_2 := log((1-h_2^2)/h_2^2)
f_2 := evalfREML(logDelta_2)  # perform second fREML evaluation

for s = 3 to 7   # perform up to 5 steps of secant iteration
  logDelta_s := (logDelta_s-2*f_s-1 - logDelta_s-1*f_s-2) / (f_s-1 - f_s-2)
  if abs(logDelta_s - logDelta_s-1) < 0.01 # check convergence
    break
  f_s := evalfREML(logDelta_s)  # perform next fREML evaluation
end for
```


Nature Genetics: doi:10.1038/ng.3190

return σ_g^2 := $y'H^{-1}y$/N, σ_e^2 := δ * σ_g^2

###

function f = evalfREML(logDelta)
  for t = 1 to MCtrials
    # build random phenotypes using pre-generated components
    y_{rand}[:,t] := X * β_{rand}[:,t] + $\sqrt{\delta}$ * $e_{\text{rand,unscaled}}$[:,t]

    # compute $H^{-1}y_{\text{rand}}$[:,t], where H = XX'/M+δI
    $H^{-1}y_{\text{rand}}$[:,t] := conjugateGradientSolve(X, δ, y_{rand}[:,t])

    # compute BLUP estimated SNP effect sizes and residuals
    $\hat{\beta}_{\text{rand}}$[:,t] := 1/M * X' * $H^{-1}y_{\text{rand}}$[:,t]
    $\hat{e}_{\text{rand}}$[:,t] = δ * $H^{-1}y_{\text{rand}}$[:,t]
  end for

  # compute BLUP estimated SNP effect sizes and residuals for real phenotypes
  $H^{-1}y_{\text{data}}$:= conjugateGradientSolve(X, δ, y)
  $\hat{\beta}_{\text{data}}$[:,t] := 1/M * X' * $H^{-1}y_{\text{data}}$[:,t]
  $\hat{e}_{\text{data}}$[:,t] = δ * $H^{-1}y_{\text{data}}$[:,t]

  # evaluate fREML
  f := log((sum($\hat{\beta}^2_{\text{data}}$) / sum($\hat{e}^2_{\text{data}}$)) / (sum($\hat{\beta}^2_{\text{rand}}$) / sum($\hat{e}^2_{\text{rand}}$)))
end

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

σ_g^2,σ_e^2 := variance parameters calculated in Step 1a

precompute $V_{\text{-chr}}^{-1}y$, where $V_{\text{-chr}} = \sigma_g^2 \text{X}_{\text{-chr}}\text{X}_{\text{-chr}}'/\text{M}_{\text{-chr}} + \sigma_e^2\text{I}$
for chr = 1 to 22
  $V_{\text{-chr}}^{-1}y$:= conjugateGradientSolve($\text{X}_{\text{-chr}}$, σ_g^2, σ_e^2, y)
end for

compute calibration for BOLT-LMM-inf statistic using 30 random SNPs

Nature Genetics: doi:10.1038/ng.3190

```
for t = 1 to 30
  m := random SNP in {1..M}
  x := X[:,m]   # normalized genotype vector for chosen SNP m
  chr := chromosome containing chosen SNP m
  V_{-chr}^{-1}x := conjugateGradientSolve(X_{-chr}, σ_g^2, σ_e^2, x)
  prospectiveStat[t] := (x'V_{-chr}^{-1}y)^2/(x'V_{-chr}^{-1}x)
  uncalibratedRetrospectiveStat[t] := N * (x'V_{-chr}^{-1}y)^2/(norm(x)^2*norm(V_{-chr}^{-1}y)^2)
end for

infStatCalibration := sum(uncalibratedRetrospectiveStats) / sum(prospectiveStats)

# compute BOLT-LMM-inf mixed model statistics at all SNPs
for m = 1 to M
  x := X[:,m]
  chr := chromosome containing SNP m
  boltLMMinf[m] := N * (x'V_{-chr}^{-1}y)^2/(norm(x)^2*norm(V_{-chr}^{-1}y)^2) / infStatCalibration
end for

return boltLMMinf[:]
```


Nature Genetics: doi:10.1038/ng.3190

Step 2a: Estimate Gaussian mixture prior parameters.

Algorithm: Optimize prediction mean-squared error in cross-validation.

```
σ_g^2, σ_e^2 := variance parameters calculated in Step 1a

for f_2 in {0.5, 0.3, 0.1}
  for p in {0.5, 0.2, 0.1, 0.05, 0.02, 0.01}
    MSE[f_2, p] := 0
    for CVfold in 1 to 5
      foldIndivs := subset of samples {1..N} in fold CVfold
      X_train := X[-foldIndivs,:], y_train = y[-foldIndivs]
      X_test := X[foldIndivs,:], y_test = y[foldIndivs]
      [β_fit, y_resid] := fitVariationalBayes(X_train, y_train, σ_g^2, σ_e^2, f_2, p)
      y_pred := X_test * β_fit
      MSE[f_2, p] += norm(y_pred - y_test)^2
    end for
  end for
end for

return [f_2, p] = argmin(MSE)

###

function [β_fit, y_resid] = fitVariationalBayes(X, y, σ_g^2, σ_e^2, f_2, p)
  σ_β,1^2 := σ_g^2/M * (1-f_2) / p # set Gaussian variances
  σ_β,2^2 := σ_g^2/M * f_2 / (1-p)
  β_fit := zero M-vector # initialize SNP effect estimates to 0
  y_resid := y             # initialize residual phenotype to y
  approxLL := -inf
  for t = 1 to maxIters   # perform VB iterations until convergence or maxIters
    approxLLprev := approxLL
    approxLL := -N/2 * log(2πσ_e^2)
    # update SNP effect estimates in turn and accumulate contributions to approxLL
    for m = 1 to M
      x := X[:,m]
      y_resid += β_fit[m] * x # remove effect of SNP m from residual
      β̂_m := x' * y_resid / norm(x)^2 # see formulas in 2.1.4
```


Nature Genetics: doi:10.1038/ng.3190

```
      β_fit[m] := p_m * β̄_{m,1} + (1−p_m) * β̄_{m,2} # set effect size to conditional posterior mean
      approxLL −= norm(x)^2/(2σ_e^2) * Var_{q_m}(β_m) + D_KL(q_m||π_m) # update approxLL (see 2.1.4)
      y_resid −= β_fit[m] * x # update residual with new effect of SNP m
    end for
    approxLL −= norm(y_resid)^2 / (2σ_e^2)

    # test convergence
    if approxLL − approxLLprev < 0.01
      break
  end for
end
```

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

```
σ_g^2, σ_e^2, f_2, p := mixture parameter estimates from Steps 1b and 2a
boltLMMinf[:] := infinitesimal mixed model stats computed in Step 1b

# compute uncalibrated BOLT-LMM statistics
for chr in 1 to 22   # leave-one-chromosome-out (LOCO) to avoid proximal contamination
  # fit model using all SNPs not on chromosome; compute residuals
  [β_fit, y_resid] := fitVariationalBayes(X_−chr, y, σ_g^2, σ_e^2, f_2, p)
  for m in (SNPs on chr)
    uncalibratedBoltLMM[m] := N * (X[:,m]'y_resid)^2 / (norm(X[:,m])^2*norm(y_resid)^2)
  end for
end for

# calibrate BOLT-LMM statistics using LD Score
interceptBoltLMMinf := LDscoreIntercept(boltLMMinf)   # see LD Score ref. [24]
interceptUncalibratedBoltLMM := LDscoreIntercept(uncalibratedBoltLMM)
LDscoreCalibration := interceptUncalibratedBoltLMM / interceptBoltLMMinf

# apply calibration
for m in 1 to M
  boltLMM[m] := uncalibratedBoltLMM[m] / LDscoreCalibration
end for

return boltLMM[:]
```


Nature Genetics: doi:10.1038/ng.3190

5 Data sets

5.1 Simulations

We simulated data sets from the WTCCC2 data with the goal of creating realistic test scenarios including linkage disequilibrium and possible confounding from population stratification or relatedness. The WTCCC2 data set contains 15,633 samples genotyped at 360,557 SNPs after QC; full details are given in ref. [12]. In the simulations that we ran to assess power and Type I error, we used these genotypes directly (subsampling to N = 5,211 samples for some simulations).

5.1.1 Simulated genotypes

In the simulations we used to benchmark computational costs, we simulated genotype data with N = 3,750 to 480,000 individuals and M = 150,000 or 300,000 SNPs by subsampling the desired number of SNPs from the WTCCC2 data set and then independently building N individuals as mosaics of individuals from the original data set using the following procedure. First, specify a number of "ancestors" A for each individual to have. Then, for each simulated individual, select A ancestors at random from among the original individuals and create the simulated individual's genotype data by chopping the genome into segments of 1,000 SNPs and copying each segment from a randomly sampled (diploid) ancestor.

Note that this approach retains realistic LD among SNPs, and for small values of A, it also retains population structure and introduces relatedness among individuals that share ancestors. In our simulations, we used $A = 2$ (substantial structure and relatedness) or $A = 10$ (low levels of structure and relatedness).

5.1.2 Simulated phenotypes

We simulated phenotypes as sums of up to four components: genetic effects from a specified number of causal SNPs, genetic effects from a specified number of standardized effect SNPs, population stratification, and environmental effects. We simulated genetic effects by selecting SNPs at random from the first half of the list of typed SNPs in each chromosome and at least 2Mb and 2cM before the middle SNP. (Likewise, when assessing calibration at null SNPs, we only tested

Nature Genetics: doi:10.1038/ng.3190

SNPs in the second halves of chromosomes and at least 2Mb and 2cM beyond the middle SNP, to avoid contamination of null SNPs from LD with causal SNPs.) We sampled effect sizes for the selected normalized SNPs from a standard normal distribution. We simulated population stratification by taking the first principal component of the normalized genotype matrix. We sampled environmental effects for all individuals from a standard normal. To create phenotypes with specified proportions of variance explained by each of these effects, we normalized each component (an N-vector across samples) and created a linear combination of the components with weights equal to the square roots of the desired variance fractions. Note that standardized effect SNPs thus have effect sizes that vary from SNP to SNP; it is the average variance explained by these SNPs that is standardized.

5.1.3 LD Scores for simulated data

Because causal SNPs in our simulated phenotypes are selected only from among genotyped SNPs, we computed reference LD Scores (used to calibrate the BOLT-LMM statistic in our simulations) by summing LD only to typed SNPs—as opposed to all SNPs, which is appropriate for real phenotypes [24]—in 379 European-ancestry samples from the 1000 Genomes Project [51].

5.2 WGHS data

The Women's Genome Health Study (WGHS) is a prospective cohort of initially healthy, female North American health care professionals at least 45 years old at baseline representing participants in the Women's Health Study (WHS) who provided a blood sample at baseline and consent for blood-based analyses. The WHS was a 2x2 trial beginning in 1992-1994 of vitamin E and low dose aspirin in prevention of cancer and cardiovascular disease with about 10 years of follow-up. Since the end of the trial, follow-up has continued in observational mode. Genotyping in the WGHS sample was performed using the HumanHap300 Duo "+" chips or the combination of the HumanHap300 Duo and iSelect chips (Illumina, San Diego, CA) with the Infinium II protocol. For quality control, all samples were required to have successful genotyping using the BeadStudio v. 3.3 software (Illumina, San Diego, CA) for at least 98% of the SNPs. A subset of 23,294 individuals were identified with self-reported European ancestry that could be verified on the basis of

multidimensional scaling analysis of identity by state using 1443 ancestry informative markers in PLINK v. 1.06. In the final data set of these individuals, a total of 339,596 SNPs were retained with MAF $>1\%$, successful genotyping in 90% of the subjects, and deviations from Hardy-Weinberg equilibrium not exceeding $p = 10^{-6}$ in significance. In our analyses, we further eliminated non-autosomal SNPs, duplicate SNPs, and custom SNPs, leaving 324,488 SNPs.

Nature Genetics: doi:10.1038/ng.3190

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

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

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

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

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

7. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* **44**, 825–830 (2012).

8. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics* **44**, 1066–1071 (2012).

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

10. 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* (2012).

11. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics* **45**, 470–471 (2013).

12. 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).

13. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* **19**, 807–812 (2011).

14. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics* **44**, 483–489 (2012).

15. Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports* **3** (2013).

Nature Genetics: doi:10.1038/ng.3190

16. Rakitsch, B., Lippert, C., Stegle, O. & Borgwardt, K. A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics* **29**, 206–214 (2013).

17. Meuwissen, T., Hayes, B. & Goddard, M. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).

18. de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. & Calus, M. P. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345 (2013).

19. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* **9**, e1003264 (2013).

20. Meuwissen, T., Solberg, T. R., Shepherd, R. & Woolliams, J. A. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet Sel Evol* **41** (2009).

21. Carbonetto, P. & Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7**, 73–108 (2012).

22. Logsdon, B. A., Hoffman, G. E. & Mezey, J. G. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11**, 58 (2010).

23. Jakobsdottir, J. & McPeek, M. S. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *American Journal of Human Genetics* **92**, 652–666 (2013).

24. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* (in press).

25. Ridker, P. M. *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* **54**, 249–255 (2008).

26. García-Cortés, L. A., Moreno, C., Varona, L. & Altarriba, J. Variance component estimation by resampling. *Journal of Animal Breeding and Genetics* **109**, 358–363 (1992).

27. Matilainen, K., Mäntysaari, E. A., Lidauer, M. H., Strandén, I. & Thompson, R. Employing a Monte Carlo Algorithm in Newton-Type Methods for Restricted Maximum Likelihood Estimation of Genetic Parameters. *PLoS ONE* **8**, e80821 (2013).

28. Legarra, A. & Misztal, I. Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360–366 (2008).

29. VanRaden, P. Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414–4423 (2008).

30. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214 (2011).

31. Aulchenko, Y. S., Ripke, S., Isaacs, A. & Van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).

32. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006).

33. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

34. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* **14**, 507–515 (2013).

35. Campbell, C. D. *et al.* Demonstrating stratification in a European American population. *Nature Genetics* **37**, 868–872 (2005).

36. Tucker, G., Price, A. L. & Berger, B. A. Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics* (2014).

37. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* **10**, 681–690 (2009).

38. Logsdon, B. A., Carty, C. L., Reiner, A. P., Dai, J. Y. & Kooperberg, C. A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. *Bioinformatics* **28**, 1738–1744 (2012).

39. Styrkarsdottir, U. *et al.* Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* (2013).

40. Do, C. B. *et al.* Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genetics* **7**, e1002141 (2011).

41. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research* gr–169375 (2014).

42. Chen, W.-M. & Abecasis, G. R. Family-based association tests for genomewide association scans. *American Journal of Human Genetics* **81**, 913–926 (2007).

43. Aulchenko, Y. S., 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* **177**, 577–585 (2007).

44. Chen, W.-M., Manichaikul, A. & Rich, S. S. A generalized family-based association test for dichotomous traits. *American Journal of Human Genetics* **85**, 364–376 (2009).

45. Boyd, S. P. & Vandenberghe, L. *Convex Optimization* (Cambridge University Press, 2004).

Nature Genetics: doi:10.1038/ng.3190

46. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**, 519–525 (2011).

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

48. McCulloch, C., Searle, S. & Neuhaus, J. *Generalized, linear, and mixed models* (Wiley, 2008), 2nd edn.

49. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).

50. Bishop, C. M. *et al. Pattern recognition and machine learning*, vol. 1 (springer New York, 2006).

51. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

52. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463 (2010).

53. Sul, J. H. & Eskin, E. Mixed models can correct for population structure for genomic regions under selection. *Nature Reviews Genetics* **14**, 300–300 (2013).

54. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. Response to sul and eskin. *Nature Reviews Genetics* **14**, 300–300 (2013).

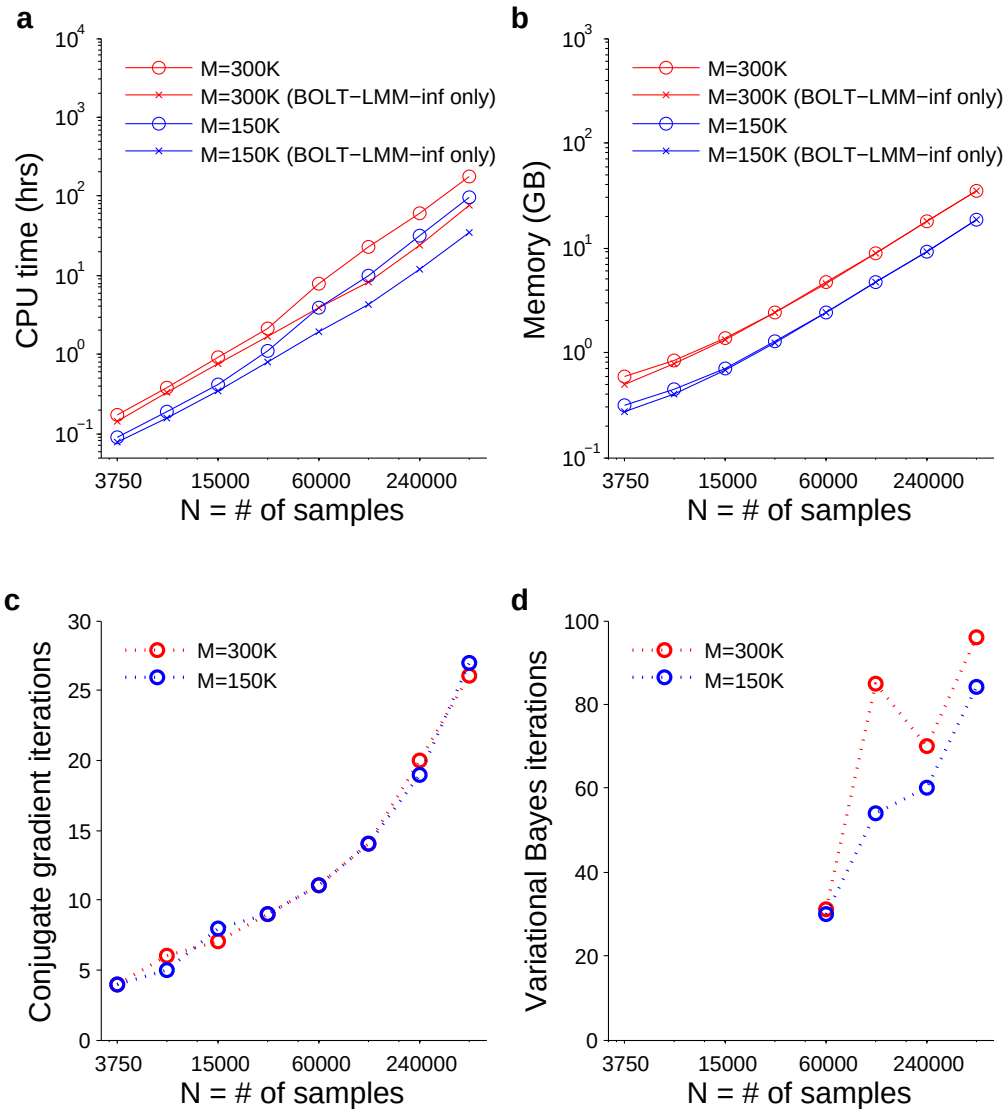
55. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics* (2013).

56. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

57. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**, 937–948 (2010).

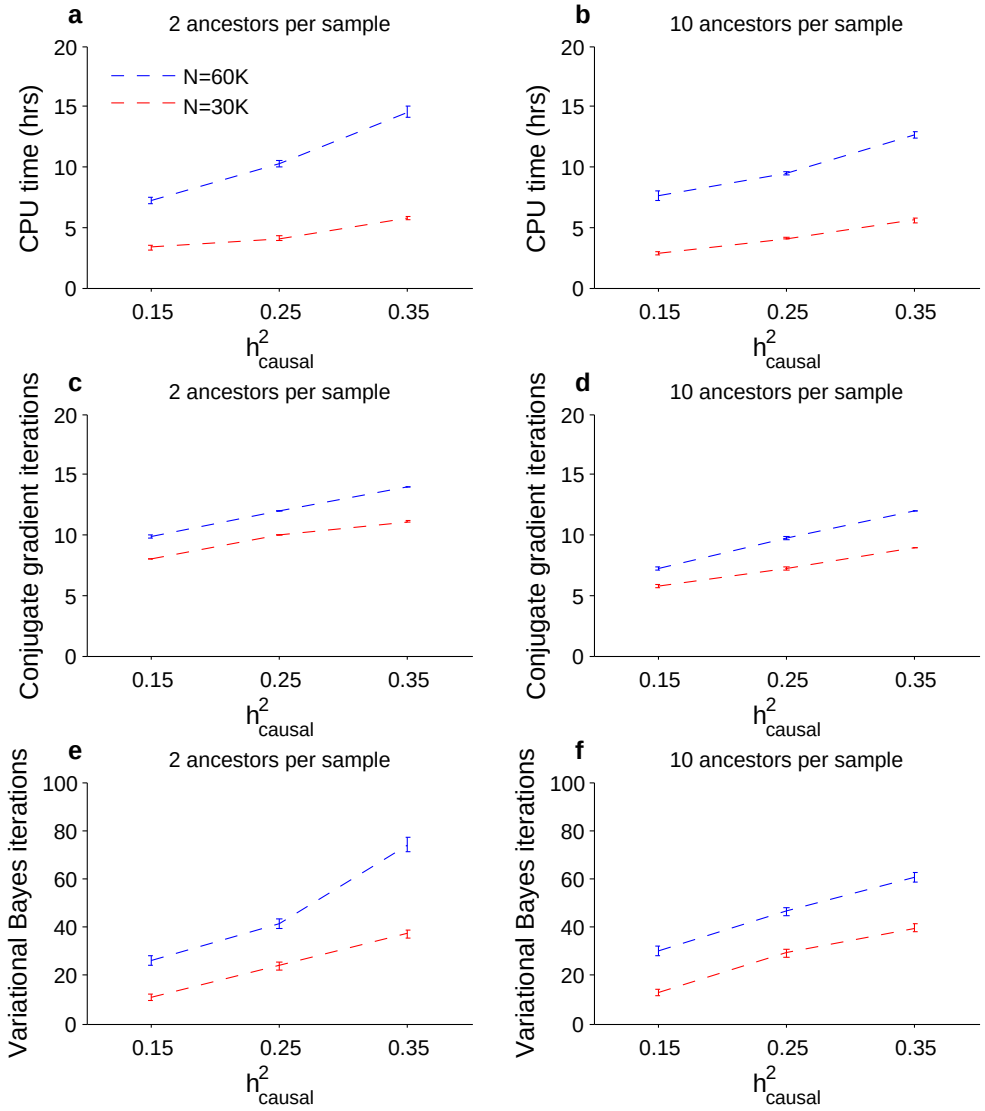
58. Ehret, G. B. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).

Nature Genetics: doi:10.1038/ng.3190



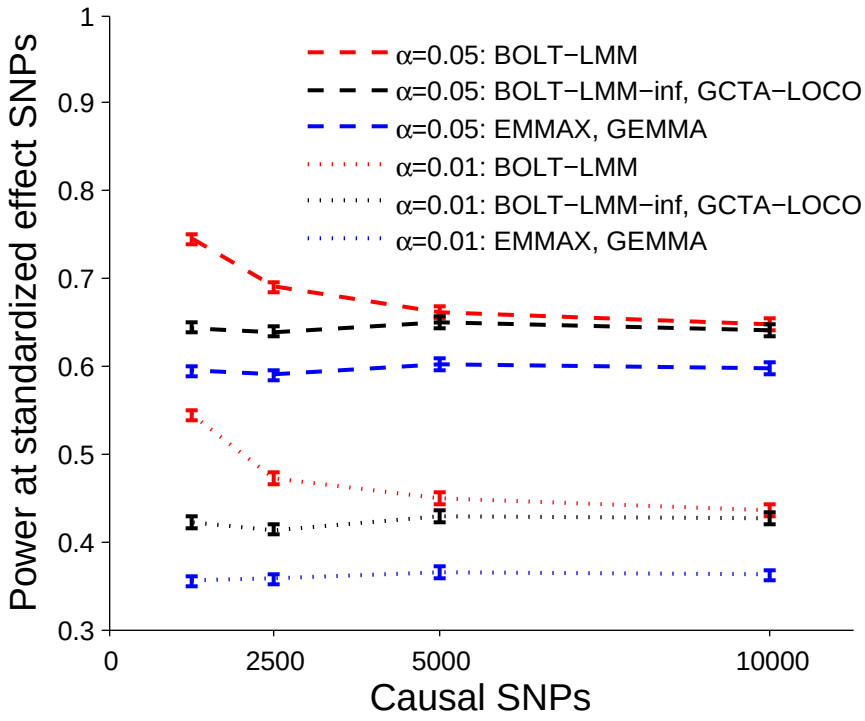
Supplementary Figure 1. Detailed computational cost metrics from running BOLT-LMM on simulated data sets with increasing sample size (Fig. 1): (**a**) running time, (**b**) memory usage, (**c**) conjugate gradient iterations used in LOCO analysis, (**d**) variational Bayes iterations used in LOCO analysis (max among 22 LOCO reps). Note that the conjugate gradient computation (**c**) is required to compute both the BOLT-LMM-inf and BOLT-LMM statistics, whereas the variational Bayes computation (**d**) is relevant only to the BOLT-LMM statistic. Additionally, BOLT-LMM skips the LOCO variational Bayes computation when estimated improvement in prediction R^2 using the Gaussian mixture model is small (<1%); in (**d**), this behavior occurred for N <60K. (Note that in such cases, some variational Bayes work is still needed in making this determination; see Supplementary Note Section 1.4 for details.) Each plotted point corresponds to one simulation with M_{causal} = 5,000 SNPs explaining $h^2{}_{\text{causal}}$ = 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.

Nature Genetics: doi:10.1038/ng.3190



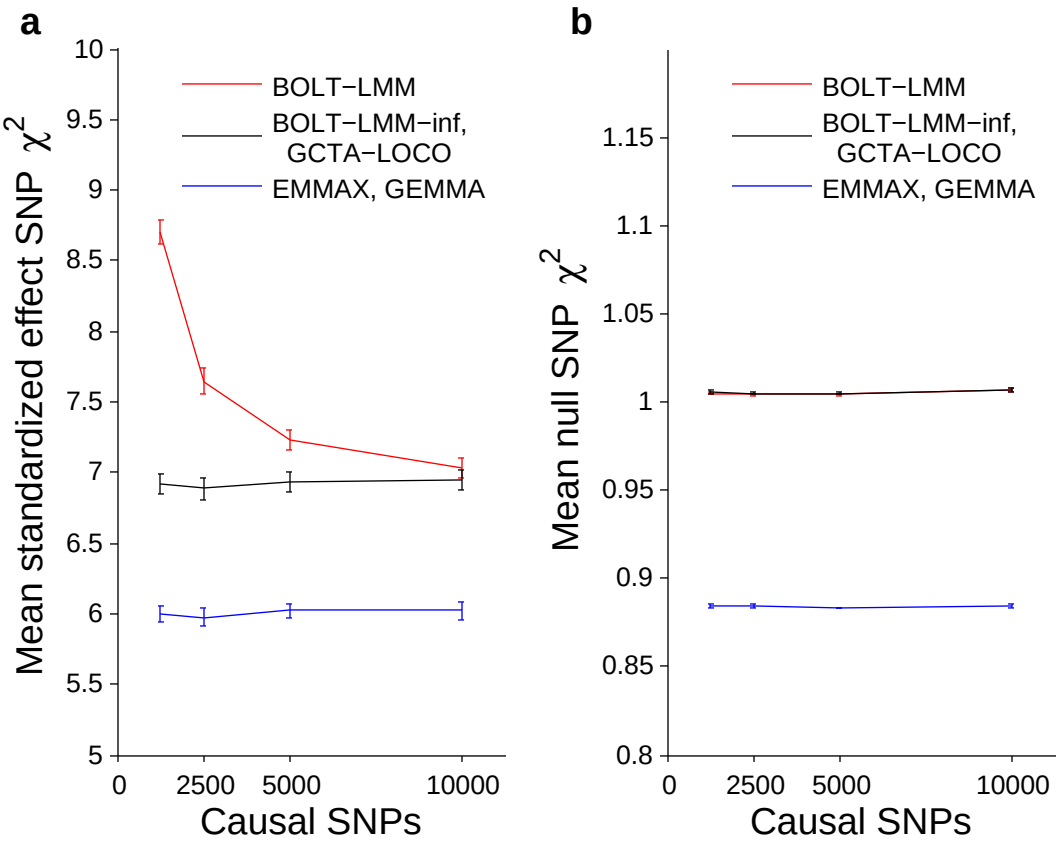
Supplementary Figure 2. Dependence of BOLT-LMM computational cost metrics (see Supplementary Fig. 1) on heritability explained by genotyped SNPs (h^2_g), sample size, and sample structure. BOLT-LMM was run on simulated data sets with N = 30,000 or 60,000 samples, each generated as a mosaic of genotype data from 2 (left panels) or 10 (right panels) random "ancestors" from the WTCCC2 data set (N = 15,633, M = 360K). Phenotypes were simulated with M_{causal} = 5,000 SNPs explaining $h^2{}_{\text{causal}}$ = 0.15–0.35 of phenotypic variance. In order to measure variational Bayes iterations (**e**, **f**) used in LOCO analysis for all parameter combinations, BOLT-LMM Gaussian mixture model analysis was run to completion even when estimated improvement in prediction R^2 using the Gaussian mixture model was small (i.e., the default behavior was overriden). Error bars, s.e.m., 10 simulations.

Nature Genetics: doi:10.1038/ng.3190



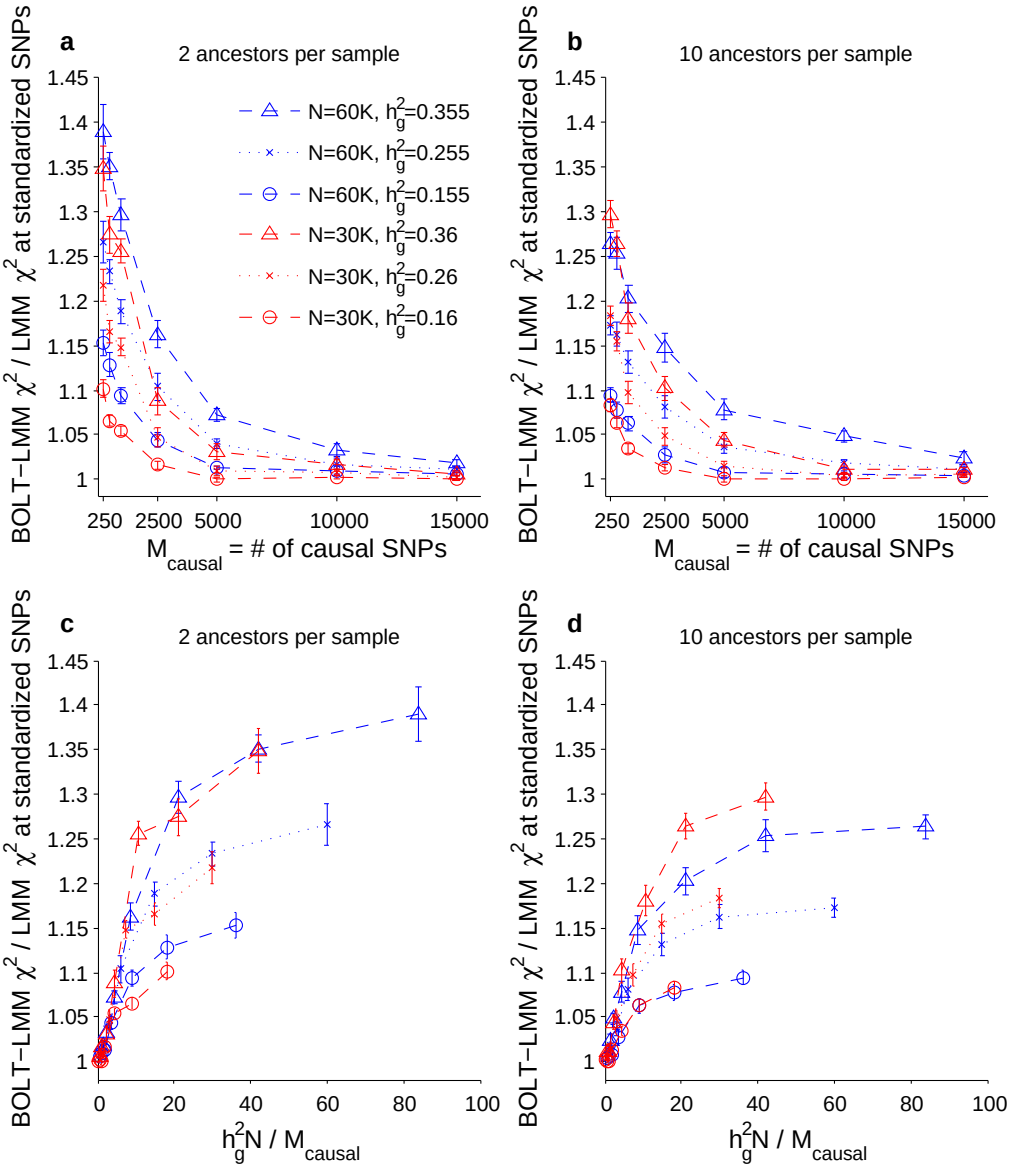
Supplementary Figure 3. Power of BOLT-LMM and existing mixed model association methods to detect standardized effect SNPs in the simulations of Fig. 2 (real genotypes from the WTCCC2 data set with N = 15,633 and M = 360K, simulated phenotypes with varying numbers of causal SNPs). Error bars, s.e.m., 100 simulations.

Nature Genetics: doi:10.1038/ng.3190



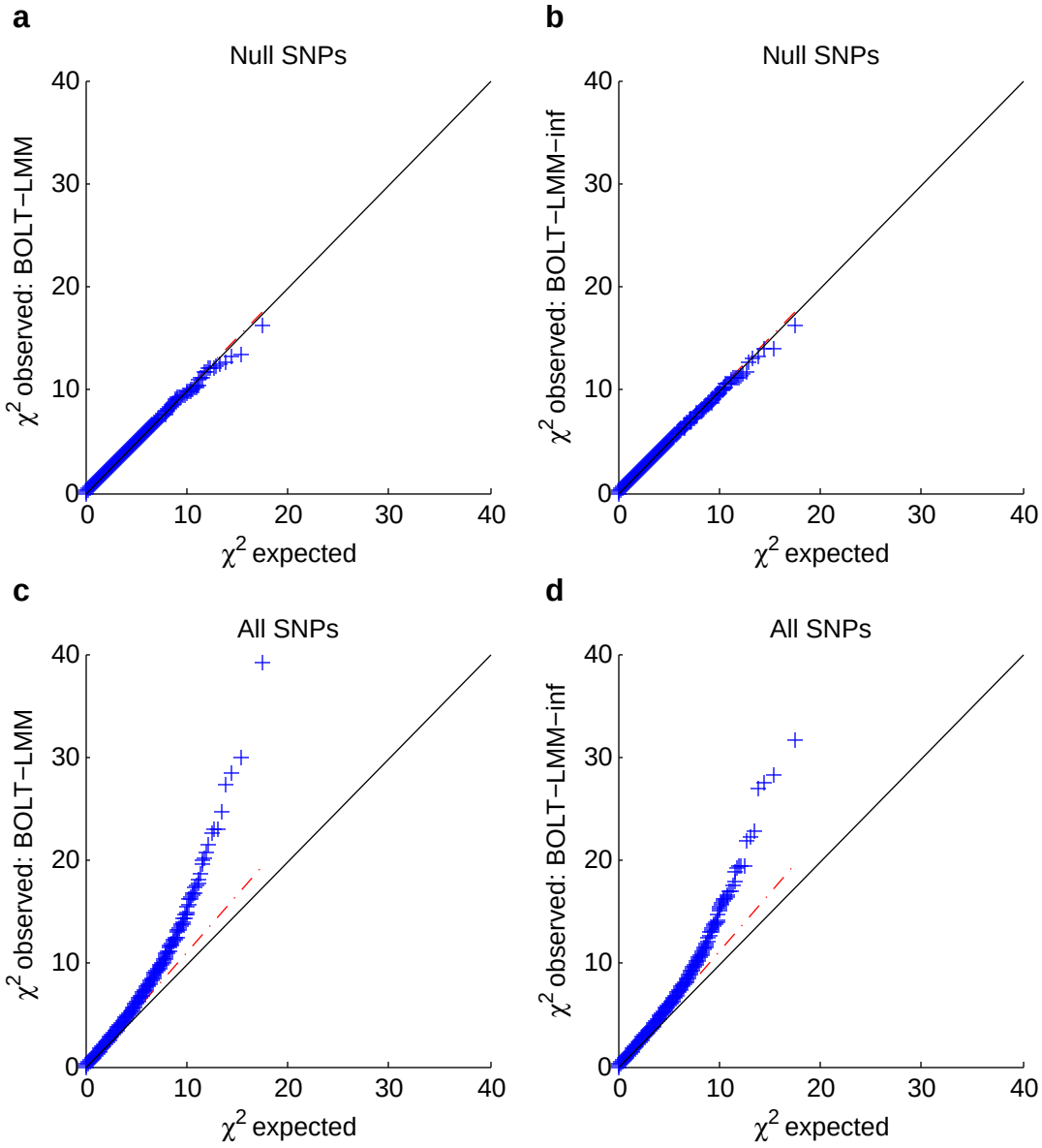
Supplementary Figure 4. BOLT-LMM increases power to detect associations in simulations while maintaining false positive control. (**a**) Mean χ^2 at standardized effect SNPs. (**b**) Mean χ^2 at null SNPs (i.e., SNPs not in LD with causal SNPs). Simulations used real genotypes from the WTCCC2 data set (N = 15,633, M = 360K) and simulated phenotypes with the specified number of causal SNPs explaining 50% of phenotypic variance and 60 more standardized effect SNPs explaining an additional 2% of the variance. Error bars, s.e.m., 100 simulations. Plotted data are for BOLT-LMM, BOLT-LMM-inf, and EMMAX statistics. We verified on the first 5 simulations that the BOLT-LMM-inf and GCTA-LOCO statistics were nearly identical and that the EMMAX and GEMMA statistics were nearly identical (Supplementary Table 7).

Nature Genetics: doi:10.1038/ng.3190



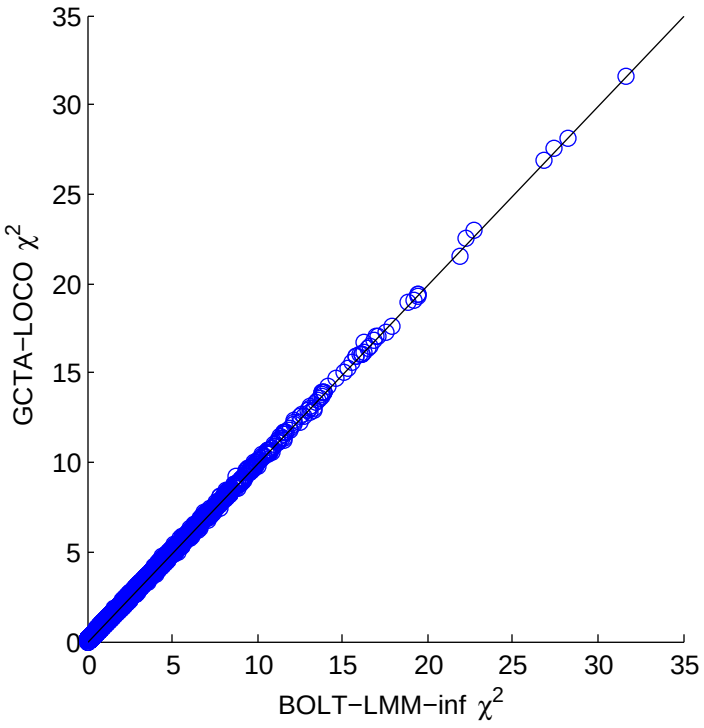
Supplementary Figure 5. Power increase of BOLT-LMM Gaussian mixture model association analysis over standard mixed model analysis as a function of sample size, number of causal SNPs, and heritability explained by genotyped SNPs (h_g^2). Plotted are ratios of mean χ^2 BOLT-LMM vs. BOLT-LMM-inf statistics at standardized effect SNPs as a function of (**a**) M_{causal} or (**b**) $h_g^2 N / M_{\text{causal}}$. This metric is equivalent to increase in effective sample size using the Gaussian mixture model. Data sets used N = 30,000 or 60,000 simulated individuals, each generated as a mosaic of genotype data from 2 (left panels) or 10 (right panels) random "ancestors" from the WTCCC2 data set (N = 15,633, M = 360K). Phenotypes were simulated with M_{causal} = 2,500–15,000 SNPs explaining h^2_{causal} = 0.15–0.35 of phenotypic variance and 60 more standardized effect SNPs explaining an additional 0.005 (N = 60K) or 0.01 (N = 30K) of the variance. In all simulations, both BOLT-LMM and BOLT-LMM-inf statistics were properly calibrated (mean χ^2 = 1.00–1.01 at null SNPs). Error bars, s.e.m., 10 simulations.

Nature Genetics: doi:10.1038/ng.3190



Supplementary Figure 6. Q-Q plots of BOLT-LMM χ^2 statistics (**a**, **c**) and BOLT-LMM-inf χ^2 statistics (**b**, **d**). The observed quantiles of both association statistics at null SNPs (**a**, **b**) match theoretical χ^2 quantiles. The observed test statistics over all SNPs (**c**, **d**) show lift-off consistent with polygenicity as simulated. Data shown are from one simulation using the setup of Fig. 2 (real genotypes from the WTCCC2 data set with N = 15,633 and M = 360K, simulated phenotypes with 5060 causal SNPs explaining 0.52 of phenotypic variance). To reduce clutter, 5% of the SNPs are plotted.

Nature Genetics: doi:10.1038/ng.3190



Supplementary Figure 7. Scatter plot of BOLT-LMM-inf vs. GCTA-LOCO χ^2 statistics. BOLT-LMM-inf and GCTA-LOCO are expected to differ slightly because GCTA-LOCO is the standard prospective statistic whereas BOLT-LMM-inf is a retrospective statistic. Data shown are from one simulation using the setup of Fig. 2 (real genotypes from the WTCCC2 data set, simulated phenotypes with 5060 causal SNPs explaining 0.52 of phenotypic variance). To reduce clutter, 5% of the SNPs are plotted.

Nature Genetics: doi:10.1038/ng.3190

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 1. Computational performance of mixed model association methods

N	BOLT-LMM	BOLT-LMM-inf	GCTA-LOCO	EMMAX	GEMMA	FaST-LMM	FaST-LMM-Select
3,750	0.171 hr / 0.582 GB	0.141 hr / 0.5 GB	1.15 hr / 6.31 GB	1.78 hr / 3.77 GB	3.28 hr / 0.323 GB	1.22 hr / 25.4 GB	4.24 hr / 25.5 GB
7,500	0.385 hr / 0.847 GB	0.324 hr / 0.771 GB	5.58 hr / 15.8 GB	8.79 hr / 3.78 GB	12.2 hr / 0.953 GB	3.34 hr / 50.9 GB	15.8 hr / 54.9 GB
15,000	0.91 hr / 1.39 GB	0.78 hr / 1.32 GB	31.4 hr / 44.8 GB	32.1 hr / 13.4 GB	50 hr / 3.47 GB	NA[a]	NA[a]
30,000	2.11 hr / 2.45 GB	1.67 hr / 2.41 GB	NA[a]	129 hr / 53.7 GB	NA[b]	NA[a]	NA[a]
60,000	7.93 hr / 4.62 GB	3.94 hr / 4.59 GB	NA[a]	NA[a]	NA[b]	NA[a]	NA[a]
120,000	23.4 hr / 8.96 GB	8.07 hr / 8.96 GB	NA[a]	NA[a]	NA[b]	NA[a]	NA[a]
240,000	60.8 hr / 17.7 GB	24.1 hr / 17.7 GB	NA[a]	NA[a]	NA[b]	NA[a]	NA[a]
480,000	177 hr / 34.3 GB	76 hr / 35.1 GB	NA[a]	NA[a]	NA[b]	NA[a]	NA[a]

This table provides the numerical data (running times and memory benchmarks for various mixed model methods) plotted in Fig. 1 as well as FaST-LMM v2.07 (using all markers). [a]Method did not complete due to exceeding the 96GB of memory available. [b]Method did not complete due to a runtime error (segmentation fault).

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

(a)

M_{causal}	PCA	BOLT-LMM-inf	BOLT-LMM
1,250	6.55 (0.07)	6.92 (0.07)	8.71 (0.08)
2,500	6.52 (0.07)	6.89 (0.08)	7.65 (0.09)
5,000	6.56 (0.07)	6.94 (0.07)	7.23 (0.07)
10,000	6.56 (0.07)	6.94 (0.07)	7.04 (0.07)

(b)

h^2_{causal}	PCA	BOLT-LMM-inf	BOLT-LMM
0.3	6.41 (0.07)	6.54 (0.07)	6.71 (0.07)
0.5	6.52 (0.07)	6.89 (0.08)	7.65 (0.09)
0.7	6.63 (0.08)	7.39 (0.09)	9.71 (0.11)

(c)

N	PCA	BOLT-LMM-inf	BOLT-LMM
5,211	2.83 (0.04)	2.88 (0.04)	2.90 (0.04)
10,422	4.71 (0.06)	4.90 (0.06)	5.14 (0.06)
15,633	6.52 (0.07)	6.89 (0.08)	7.65 (0.09)

These tables provide the numerical data (mean χ^2 statistics at standardized effect SNPs) plotted in Fig. 2.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 3. Calibration of BOLT-LMM statistics in simulations with varying genetic architectures and sample sizes

(a)

M_{causal}	PCA	BOLT-LMM-inf	BOLT-LMM
1,250	1.009 (0.001)	1.006 (0.001)	1.005 (0.001)
2,500	1.008 (0.001)	1.005 (0.001)	1.004 (0.001)
5,000	1.007 (0.001)	1.005 (0.001)	1.004 (0.001)
10,000	1.009 (0.001)	1.007 (0.001)	1.007 (0.001)

(b)

h^2_{causal}	PCA	BOLT-LMM-inf	BOLT-LMM
0.3	1.005 (0.001)	1.004 (0.001)	1.005 (0.001)
0.5	1.008 (0.001)	1.005 (0.001)	1.004 (0.001)
0.7	1.011 (0.001)	1.009 (0.001)	1.000 (0.001)

(c)

N	PCA	BOLT-LMM-inf	BOLT-LMM
5,211	1.002 (0.001)	1.004 (0.001)	1.004 (0.001)
10,422	1.005 (0.001)	1.005 (0.001)	1.005 (0.001)
15,633	1.008 (0.001)	1.005 (0.001)	1.004 (0.001)

We report mean χ^2 statistics at null SNPs in the simulations of Fig. 2. The very slight inflation of BOLT-LMM and BOLT-LMM-inf arises from the use of approximate variance parameter estimates and from the fact that standard mixed model methods eliminate the effects of stratification nearly perfectly but not completely (Supplementary Table 4).

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 4. Calibration of BOLT-LMM statistics using different variance component estimation procedures

Variance component estimation procedure	Leave-out segments	$h^2_{\text{strat}} = 0.01$			$h^2_{\text{strat}} = 0$		
		GCTA-LOCO	BOLT-inf	BOLT	GCTA-LOCO	BOLT-inf	BOLT
Once with no left-out SNPs	22	NA	1.005	1.004	NA	1.004	1.002
	100	NA	1.002	1.000	NA	1.001	0.999
Independently for each leave-out segment	22	1.002	1.001	0.999	1.000	0.999	0.997
	100	NA	1.000	0.998	NA	0.998	0.997

Mean GCTA-LOCO, BOLT-LMM-inf, and BOLT-LMM chi-squared statistics at null SNPs computed using different methods for variance component analysis and different numbers of leave-out segments. We used the same simulation setup as in Fig. 2 (real genotypes from the WTCCC2 data set, simulated phenotypes with 2560 causal SNPs explaining 0.52 of phenotypic variance). We performed two sets of simulations: one with environmental stratification h^2_{strat} explaining 0.01 of the variance, and the other with no environmental stratification. Thus, the top left entries of 1.005 for BOLT-LMM-inf and 1.004 for BOLT-LMM correspond to the calibration results in Supplementary Table 3 for M_{causal} = 2,500, h^2_{causal} = 0.5, N = 15,633 (appearing in each of Supplementary Table 3a,b,c). Data shown are from 100 simulations, which gave a standard error of 0.001 for all numbers above.

By default, the BOLT-LMM algorithm estimates variance parameters once using all SNPs and then reuses the variance estimates when computing association statistics, leaving each of the 22 chromosomes out in turn. This procedure is an approximation of the theoretically precise method of re-estimating variance parameters once per leave-out segment, which the BOLT-LMM software offers as an alternative option (Online Methods). The quality of the approximation can also be improved by subdividing chromosomes for leave-out analysis; we compare results with 22 vs. 100 leave-out segments, which the BOLT-LMM software allows as well (Online Methods). GCTA-LOCO re-estimates variance components once for each of 22 left-out chromosomes.

These simulations demonstrate that very slight (0.1–0.5%) inflation of χ^2 test statistics can stem from two causes: (1) reusing the same variance parameter estimates across all LOCO reps rather than re-estimating them for each LOCO rep; and (2) near-complete but imperfect correction for stratification by mixed model methods [52]. The first source of slight inflation is specific to the BOLT-LMM approximation procedure but can be reduced by partitioning the genome into finer leave-out segments or eliminated by refitting variance parameters for each LOCO rep at a small increase in computational cost (Online Methods). This inflation does not increase with sample size (Supplementary Table 3c). The second source of slight inflation is shared by all mixed model methods, as evidenced by the 0.2% inflation of GCTA-LOCO in the $h^2_{\text{strat}} = 0.01$ simulation, and scales with sample size and proportion of variance explained by ancestry in the same manner as inflation caused by genuine polygenic effects (see Supplementary Table 2 of ref. [12]). To completely eliminate this source of very slight inflation, it is necessary to either include principal components as fixed effects or include a second variance component that models ancestry [12, 52–54]. However, we believe that this very slight inflation is not a significant problem for standard mixed model methods or BOLT-LMM.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 5. Type I error of BOLT-LMM and EMMAX association tests in simulations

Method	$\alpha = 10^{-2}$		$\alpha = 10^{-4}$		$\alpha = 10^{-6}$		$\alpha = 5 \times 10^{-8}$	
BOLT-LMM	1.02×10^{-2}	705103	1.04×10^{-4}	7221	1.05×10^{-6}	73	4×10^{-8}	3
BOLT-LMM-inf	1.02×10^{-2}	707636	1.03×10^{-4}	7119	1.07×10^{-6}	74	6×10^{-8}	4
EMMAX	0.63×10^{-2}	434081	0.39×10^{-4}	2673	0.32×10^{-6}	22	0×10^{-8}	0

Type I error rates and counts for test statistics at null SNPs in the simulations of Fig. 2 (real genotypes from the WTCCC2 data set, simulated phenotypes with causal SNPs explaining 52% of phenotypic variance and environmental stratification explaining 1% of phenotypic variance). To increase statistical resolution, we combined data from all 400 simulations (with 1310, 2560, 5060, and 10060 causal SNPs, 100 simulations each) for a total of 69,293,600 hypothesis tests. Type I error was statistically indistinguishable between different subsets of 100 simulations. The very slight upward bias of actual vs. expected Type I error rates for the BOLT-LMM-inf and BOLT-LMM association tests is explained by very slight (0.1–0.5%) inflation of χ^2 test statistics (Supplementary Table 4), which can be mitigated by performing a more precise variance parameter calculation if desired (Online Methods).

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 6. Power and calibration of BOLT-LMM statistics, BOLT-LMM with MCMC, and FaST-LMM-Select in simulations

(a) M_{causal} = 1,250

Method	Mean χ^2 at std. effect SNPs	Mean χ^2 at null SNPs	LD Score intercept
PCA	2.843 (0.041)	1.003 (0.001)	1.008 (0.001)
BOLT-LMM-inf	2.923 (0.041)	1.005 (0.001)	1.008 (0.001)
BOLT-LMM	3.174 (0.045)	1.003 (0.001)	1.008 (0.001)
BOLT-LMM with MCMC	3.154 (0.045)	1.004 (0.001)	1.008 (0.001)
FaST-LMM-Select	2.852 (0.037)	1.021 (0.003)	1.026 (0.002)

(b) M_{causal} = 500

Method	Mean χ^2 at std. effect SNPs	Mean χ^2 at null SNPs	LD Score intercept
PCA	2.918 (0.045)	1.004 (0.001)	1.010 (0.001)
BOLT-LMM-inf	2.988 (0.047)	1.005 (0.001)	1.010 (0.001)
BOLT-LMM	4.181 (0.064)	1.002 (0.001)	1.010 (0.001)
FaST-LMM-Select	3.520 (0.044)	1.017 (0.002)	1.021 (0.002)

We report mean χ^2 statistics at standardized effect SNPs, mean χ^2 statistics at null SNPs, and LD Score intercepts [24] of association statistics computed by linear regression with PC covariates (PCA), variants of BOLT-LMM, and FaST-LMM-Select over 100 simulations. Each simulation used real genotypes from N = 5,211 WTCCC2 samples and simulated phenotypes with (a) 1,250 or (b) 500 causal SNPs explaining 70% of variance, 60 standardized effect SNPs explaining 2% of variance, and a 1% ancestry effect. In simulation group (a), we ran the MCMC modification of BOLT-LMM for 1,000 iterations of the Gibbs sampler through all SNPs. We do not report MCMC results for (b) because the Gibbs sampler had not converged after 1,000 iterations; we hypothesize the MCMC was slow to mix in this case because of the small number of causal SNPs resulting in a highly nonlinear model involving a few large effect sizes.

In simulation group (a) with M_{causal} = 1,250, FaST-LMM-Select typically selected 100 top SNPs to include in the mixed model; in (b) with M_{causal} = 500, FaST-LMM-Select typically selected 300 top SNPs. In (a), we observed that FaST-LMM achieved lower power than BOLT-LMM-inf (as measured by mean χ^2 at standardized effect SNPs), which we found surprising, especially given that FaST-LMM-Select reported that including 100 top SNPs in the mixed model achieved a 7% average reduction in cross-validation prediction MSE (which is in between that achieved by BOLT-LMM-inf and BOLT-LMM, consistent with our expectations). We are unsure why this improvement in out-of-sample prediction MSE did not translate into improved association power, although we also noticed that in (b), the association power achieved by FaST-LMM-Select was lower than we expected given its cross-validation report (≈30% reduction in MSE).

The calibration results (mean χ^2 at null SNPs) also show that in these simulations, the choice of FaST-LMM-Select to use ≈100–300 SNPs in the mixed model led to mild inflation due to inadequate stratification correction. Accordingly, the LD Score intercept for FaST-LMM-Select is inflated relative to the other methods. (Note that the LD Score intercepts for the other methods, which are properly calibrated, are slightly greater than 1 due to attenuation bias; see Supplementary Note Section 1.5 for a detailed discussion.) Normalizing FaST-LMM-Select to have the same LD Score intercept as BOLT-LMM-inf (which is how BOLT-LMM is calibrated) produces a properly calibrated statistic: e.g., in (a) we get 1.021 / (1.026/1.008) = 1.003, which is similar to the other methods. Note, however, that calibration alone does not solve the problem of potential false positives at unusually differentiated markers.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 7. Correlations between mixed model statistics computed by various mixed model methods under the infinitesimal model

R^2 (s.e.m.)	GCTA-LOCO	GEMMA
BOLT-LMM-inf	0.999471 (0.000017)	0.962928 (0.000916)
EMMAX	0.963326 (0.000969)	0.999997 (0.000000)

Squared correlation coefficients between χ^2 association statistics over all SNPs computed by various mixed model methods using the simulation setup of Fig. 2 (real genotypes from the WTCCC2 data set, simulated phenotypes with 5060 causal SNPs explaining 0.52 of phenotypic variance). Means and standard errors over 5 simulations are reported. BOLT-LMM-inf and GCTA-LOCO statistics both avoid proximal contamination via LOCO analysis but are expected to differ slightly because GCTA-LOCO is the standard prospective statistic whereas BOLT-LMM-inf is a retrospective statistic. EMMAX and GEMMA are both susceptible to proximal contamination but are expected to differ slightly because GEMMA is an exact statistic whereas EMMAX is an approximate statistic.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 8. Comparison of χ^2 statistics computed by various methods at known SNPs for WGHS phenotypes

Phenotype	# known SNPs	# tagged SNPs	# $p < 0.05$ PCA & LMM	Mean log fold χ^2 LMM / PCA	# $p < 0.05$ PCA & BOLT	Mean log fold χ^2 BOLT / PCA
ApoB	59 (ref. [55])	57	40	2.3 (1.4)	40	9.9 (2.1)
LDL	59 (ref. [55])	57	45	1.4 (1.1)	45	8.7 (1.7)
HDL	72 (ref. [55])	72	55	3.4 (1.2)	55	9.4 (1.8)
Cholesterol	73 (ref. [55])	71	53	1.8 (1.1)	54	5.5 (1.2)
Triglycerides	41 (ref. [55])	39	29	-0.6 (1.8)	29	1.7 (2.0)
Height	180 (ref. [56])	170	131	6.0 (1.5)	134	6.2 (1.6)
BMI	32 (ref. [57])	31	18	4.2 (1.8)	18	3.8 (1.8)
SystolicBP	20 (ref. [58])	20	11	2.6 (1.9)	11	2.6 (1.9)
DiastolicBP	19 (ref. [58])	19	9	3.2 (1.8)	9	3.0 (2.0)

The first two data columns provide additional information about known SNPs used in the power comparison of Supplementary Table 9: # known SNPs, number of genome-wide significant associated SNPs reported in largest GWAS to date; # tagged SNPs, number of such SNPs with an $R^2 \geq 0.2$ tagging SNP typed in WGHS. (Note that for the ApoB phenotype, we used known SNPs for the closely related LDL phenotype.) Sums of χ^2 statistics compared in Supplementary Table 9 were computed across the WGHS-typed tagging SNPs.
We also report here one additional metric for increase of power: Mean log fold-change in χ^2 statistics at tagging SNPs. This metric weights all tagging SNPs evenly (Supplementary Table 9), whereas comparing sums of χ^2 statistics weights stronger associations more heavily. Because log fold-change is sensitive to noise from non-replicating SNPs, we restrict to tagging SNPs with at least nominal significance ($p < 0.05$) according to both methods being compared.
Methods: PCA, linear regression using 10 principal components as covariates; LMM, BOLT-LMM-inf; BOLT, BOLT-LMM. Errors, s.e.m.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 9. BOLT-LMM increases power to detect associations for WGHS phenotypes

Phenotype	h_g^2	Prediction R^2				χ^2 incr. at known loci	
		PCA (%)	LMM (%) Actual, *Expected*	BOLT (%)	p-value BOLT>PCA	LMM vs. PCA (%)	BOLT vs. PCA (%)
ApoB	0.24	0.0	1.5 (0.1), *1.7*	10.4 (0.3)	4.3E–6	1.5 (0.6)	10.0 (1.2)
LDL	0.18	0.1	1.0 (0.1), *1.0*	7.5 (0.2)	2.8E–6	1.4 (0.6)	8.1 (0.8)
HDL	0.22	0.1	1.7 (0.1), *1.5*	6.7 (0.2)	2.3E–5	1.7 (1.5)	7.1 (2.2)
Cholesterol	0.20	0.1	1.1 (0.1), *1.1*	5.9 (0.1)	1.3E–6	1.7 (0.7)	6.9 (0.8)
Triglycerides	0.18	0.0	1.2 (0.1), *1.0*	4.4 (0.3)	0.00014	0.4 (1.0)	3.1 (1.2)
Height	0.47	1.9	7.7 (0.4), *6.4*	9.5 (0.3)	1.9E–6	6.4 (1.3)	7.6 (1.4)
BMI	0.23	0.0	1.6 (0.1), *1.5*	2.1 (0.1)	8.5E–5	1.3 (1.4)	1.3 (1.5)
SystolicBP	0.17	0.2	1.0 (0.1), *0.8*	1.0 (0.1)	0.00014	3.6 (2.2)	3.6 (2.2)
DiastolicBP	0.14	0.0	0.7 (0.0), *0.5*	0.7 (0.0)	0.00017	2.9 (1.7)	2.7 (1.6)

We report relative power of different association tests using two roughly equivalent metrics: comparison of χ^2 statistics at known loci, a direct but noisy approach, and out-of-sample prediction R^2 based on the underlying model (both plotted in Figure 3. For reference, we also report the heritability parameter h_g^2 estimated by BOLT-LMM and expected prediction R^2 for LMM [34]. Methods: PCA, linear regression using 10 principal components; LMM, standard (infinitesimal) mixed model; BOLT, BOLT-LMM Gaussian mixture model. Actual prediction R^2 values are from 5-fold cross-validation: predictions for each left-out fold were computed by fitting all SNP effects simultaneously (for mixed model methods) or estimating covariate effects (for PCA) using the training folds. Note in particular that for PCA, only covariate effects, not SNP effects, were used for prediction. Standard errors are across folds. Expected prediction R^2 for LMM was computed using $N = 23,294 \times 4/5$ (taking into account the 5-fold cross-validation), M_{eff} = 60,000, and h_g^2 estimated by BOLT-LMM, given that the WGHS data set contains little relatedness. (Note that for height, actual prediction R^2 for LMM is slightly higher than expected because PCs explain a non-negligible amount of variance.) Significance for BOLT-LMM>PCA prediction R^2 was assessed using a one-sided paired t-test across folds. Percent increases in χ^2 statistics computed by various methods across known loci are comparisons between sums of χ^2 statistics over typed SNPs in highest LD with published associated SNPs from the largest GWAS to date (Supplementary Table 8). Standard errors are jackknife estimates.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 10. Effect size mixture parameters chosen by BOLT-LMM for WGHS phenotypes

Phenotype	h_g^2	f_2	p
ApoB	0.24	0.1	0.01
LDL	0.18	0.1	0.01
HDL	0.22	0.1	0.01
Cholesterol	0.20	0.1	0.01
Triglycerides	0.18	0.1	0.01
Height	0.47	0.3	0.02
BMI	0.23	0.5	0.01
SystolicBP	0.17	0.5	0.5
DiastolicBP	0.14	0.5	0.01

We report the best-fit mixture-of-Gaussians prior on SNP effect sizes determined by BOLT-LMM using cross-validation to optimize out-of-sample prediction R^2. The spike and slab mixture of Gaussians is parameterized by the total variance attributed to the combined Gaussian mixture along with two mixture parameters: f_2, the proportion of variance allotted to the spike component (small-effect SNPs), and p, the probability that a SNP effect is drawn from the slab component (large-effect SNPs). 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.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 11. Calibration of χ^2 statistics computed by various methods for WGHS phenotypes

	Mean χ^2				λ_{GC}			
Phenotype	LR	PCA	LMM	BOLT-LMM	LR	PCA	LMM	BOLT-LMM
ApoB	1.110	1.098	1.103	1.109	1.061	1.049	1.055	1.054
LDL	1.100	1.079	1.084	1.091	1.055	1.039	1.042	1.045
HDL	1.108	1.096	1.100	1.100	1.073	1.066	1.068	1.065
Cholesterol	1.108	1.087	1.090	1.094	1.063	1.044	1.045	1.046
Triglycerides	1.098	1.092	1.089	1.090	1.069	1.059	1.058	1.057
Height	1.565	1.212	1.227	1.232	1.503	1.159	1.170	1.169
BMI	1.104	1.101	1.099	1.099	1.105	1.101	1.100	1.098
SystolicBP	1.102	1.063	1.076	1.076	1.102	1.061	1.074	1.074
DiastolicBP	1.073	1.056	1.062	1.062	1.074	1.058	1.061	1.062

Mean χ^2 test statistics across all SNPs (left columns) and genomic inflation factors λ_{GC} (right columns) using four methods: LR, linear regression; PCA, linear regression using 10 principal components as covariates; LMM, BOLT-LMM-inf; and BOLT-LMM. In all cases, both mean χ^2 and λ_{GC} exceed 1 because of polygenicity. PCA, LMM, and BOLT-LMM are consistently calibrated, whereas uncorrected linear regression suffers inflation due to population stratification, especially for height. Standard errors for differences in mean χ^2 statistics between pairs of methods are all at most 0.001, with the exception of uncorrected linear regression vs. other methods in the height analysis, for which the standard error is 0.003. Standard errors of the mean χ^2 statistics themselves are 0.003–0.004.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 12. Control for stratification: Height *p*-values at lactase computed by various methods

	LR	PCA	LMM	BOLT-LMM
χ^2 statistic	33.09	3.66	6.56	6.07
p-value	9×10^{-9}	0.06	0.01	0.01

We report χ^2 statistics and p-values computed at rs2011946, the SNP typed in WGHS with highest LD ($R^2 = 0.64$ in 1000 Genomes reference samples) to lactase-associated SNP rs4988235. Methods: LR, linear regression; PCA, linear regression using 10 principal components as covariates; LMM, BOLT-LMM-inf; and default BOLT-LMM.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 13. Power of mixed model association vs. linear regression for simulated case-control ascertained traits

(a) $h_g^2 = 0.25$ (for liabilities underlying case-control phenotype)

	Case prevalence				
Method	50%	10%	5%	1%	0.1%
Linear regression	2.828 (0.010)	3.723 (0.015)	4.401 (0.015)	6.206 (0.020)	9.179 (0.026)
BOLT-LMM-inf	2.862 (0.010)	3.739 (0.015)	4.384 (0.015)	5.965 (0.018)	8.212 (0.023)
BOLT-LMM	2.887 (0.011)	3.767 (0.015)	4.397 (0.014)	5.859 (0.022)	7.869 (0.025)

(b) $h_g^2 = 0.5$ (for liabilities underlying case-control phenotype)

	Case prevalence				
Method	50%	10%	5%	1%	0.1%
Linear regression	4.661 (0.014)	6.452 (0.021)	7.756 (0.023)	11.360 (0.030)	17.290 (0.043)
BOLT-LMM-inf	4.904 (0.016)	6.582 (0.022)	7.662 (0.021)	10.202 (0.025)	13.252 (0.033)
BOLT-LMM	5.165 (0.018)	6.798 (0.026)	7.826 (0.024)	9.816 (0.034)	12.326 (0.037)

We compare mean χ^2 statistics over causal SNPs for linear regression, BOLT-LMM-inf, and BOLT-LMM analysis of simulated case-control traits with prevalences ranging from 50%–0.1%. We simulated genotypes by generating individuals as mosaics of up to 100 random "ancestors" from the WTCCC2 data set, resampling ancestors every 500 SNPs. We restricted the SNP set to the first 2,500 SNPs on each autosome, for a total of M = 55,000 SNPs (so that $M_{\text{effective}} \approx$ 10,000 independent SNPs). We simulated case-control phenotypes using a liability threshold model in which we first generated continuous phenotypes with **(a)** $h_g^2 = 0.25$ or **(b)** $h_g^2 = 0.5$ explained by M_{causal} = 1,000 markers and then defined cases as individuals with phenotypes exceeding a threshold corresponding to the desired prevalence. In each simulation, we ascertained 5,000 cases and 5,000 controls for a total of N = 10,000 simulated individuals. We then ran association analysis on the ascertained samples. Errors, s.e.m. over 100 simulations.
The results show that for non-ascertained traits, power of linear regression $<$ BOLT-LMM-inf $<$ BOLT-LMM, consistent with our findings for quantitative traits. This trend reverses for ascertained traits at a prevalence near 5%. For the infinitesimal mixed model vs. linear regression, this result is consistent with the findings of ref. [12], which also observes that the relative reduction in test statistics increases with h_g^2 and the ratio N / $M_{\text{effective}}$ in addition to ascertainment severity. In our simulations, $N \approx M_{\text{effective}}$, so these simulations correspond to $N \approx$ 60,000 in real data sets, which have $M_{\text{effective}} \approx$ 60,000 [12]. For data sets with fewer than 60,000 samples, case-control ascertainment will present less of a problem for mixed model methods than indicated above.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 14. Precision of infinitesimal mixed model statistic calibration

Phenotype	Mean prospective stat at 30 random SNPs	Mean retrospective stat (pre-calib.) at same SNPs	Calibration factor c_{inf}, equation (9)
ApoB	1.285	1.281	0.996 (0.001)
LDL	1.013	1.011	0.998 (0.000)
HDL	1.026	1.024	0.997 (0.001)
Cholesterol	0.989	0.986	0.997 (0.000)
Triglycerides	0.881	0.879	0.998 (0.000)
Height	0.783	0.777	0.993 (0.001)
BMI	0.900	0.897	0.997 (0.001)
Systolic BP	0.853	0.849	0.996 (0.001)
Diastolic BP	0.770	0.769	0.998 (0.001)

We report the BOLT-LMM-inf calibration factor for the nine analyzed WGHS phenotypes. This calibration is similar to GRAMMAR-Gamma [10] but is estimated by computing statistics at only 30 SNPs. The standard errors (jackknife estimates) show that 30 SNPs are enough to achieve high calibration precision.

Nature Genetics: doi:10.1038/ng.3190

Supplementary Table 15. LD Score calibration factors in BOLT-LMM analyses of WGHS phenotypes

Phenotype	Calibration factor c, equation (11)
ApoB	1.004 (0.003)
LDL	1.001 (0.002)
HDL	1.001 (0.002)
Cholesterol	0.999 (0.002)
Triglycerides	1.000 (0.001)
Height	0.997 (0.001)
BMI	0.998 (0.001)
Systolic BP	0.996 (0.000)
Diastolic BP	0.998 (0.000)

We report the BOLT-LMM calibration factor for the nine analyzed WGHS phenotypes, which equals the ratio of the LD Score intercepts of the uncalibrated BOLT-LMM statistic to that of the BOLT-LMM-inf statistic. Standard errors are jackknife estimates leaving out each chromosome in turn.

Nature Genetics: doi:10.1038/ng.3190