nature genetics

Efficient Bayesian mixed-model analysis increases association power in large cohorts

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

Linear mixed models are a powerful statistical tool for identifying genetic associations and avoiding confounding. However, existing methods are computationally intractable in large cohorts and may not optimize power. All existing methods require time cost $O(MN^2)$ (where N is the number of samples and M is the number of SNPs) and implicitly assume an infinitesimal genetic architecture in which effect sizes are normally distributed, which can limit power. Here we present a far more efficient mixed-model association method, BOLT-LMM, which requires only a small number of $O(MN)$ time iterations and increases power by modeling more realistic, non-infinitesimal genetic architectures via a Bayesian mixture prior on marker effect sizes. We applied BOLT-LMM to 9 quantitative traits in 23,294 samples from the Women's Genome Health Study (WGHS) and observed significant increases in power, consistent with simulations. Theory and simulations show that the boost in power increases with cohort size, making BOLT-LMM appealing for genome-wide association studies in large cohorts.

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

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

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA. [5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [6]Department of Endocrinology, Children's Hospital Boston, Boston, Massachusetts, USA. [7]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [8]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. Correspondence should be addressed to P.-R.L. (loh@hsph.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu).

Received 22 July 2014; accepted 16 December 2014; published online 2 February 2015; doi:10.1038/ng.3190

We provide an efficient software implementation of our algorithm, BOLT-LMM, and demonstrate its computational efficiency on simulated data sets of up to 480,000 individuals. Our simulations also show that BOLT-LMM achieves increased association power over standard infinitesimal mixed-model analysis of traits driven by a few thousand causal SNPs. We also applied BOLT-LMM in the mixed-model analysis of 9 quantitative traits in 23,294 samples from the WGHS[25] and observed increased association power equivalent to up to a 10% increase in effective sample size. We demonstrate through theory and simulations that the power boost increases with cohort size, making BOLT-LMM a promising approach for large-scale GWAS.

RESULTS

Overview of methods

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

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

Table 1 Comparison of fast mixed-model association methods that model all SNPs

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

[a]For methods that have been updated over multiple publications, we cite and list the characteristics of the latest published version. [b]If $M < N$, FaST-LMM and FaST-LMM-Select can complete in $O(M^2N)$ time. [c]FaST-LMM-Select models non-infinitesimal genetic architectures by restricting the mixed model to a subset of SNPs; a caveat of this approach is that it may incur susceptibility to confounding from stratification[12]. [d]GRAMMAR-Gamma requires $O(MN^2)$ time for only the initial computation of the GRM but not for computing association test statistics. For a detailed breakdown of the computational complexity per algorithmic step, see Table 1 of ref. 12.

Computational cost of BOLT-LMM versus existing methods

To analyze the computational performance of BOLT-LMM, we simulated data sets of sizes ranging from $N = 3{,}750$ to 480,000 individuals and $M = 300{,}000$ SNPs. We used the genotypes from the Wellcome Trust Case Control Consortium 2 (WTCCC2) data set[30] analyzed in ref. 12, which contains 15,633 individuals of European ancestry, to form mosaic chromosomes, and we used a phenotype model in which 5,000 SNPs explained 20% of phenotypic variance (**Supplementary Note**).

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

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

Power and false positive control of BOLT-LMM in simulations

To assess the power of BOLT-LMM to detect associated loci, we performed additional simulations using real genotypes from the

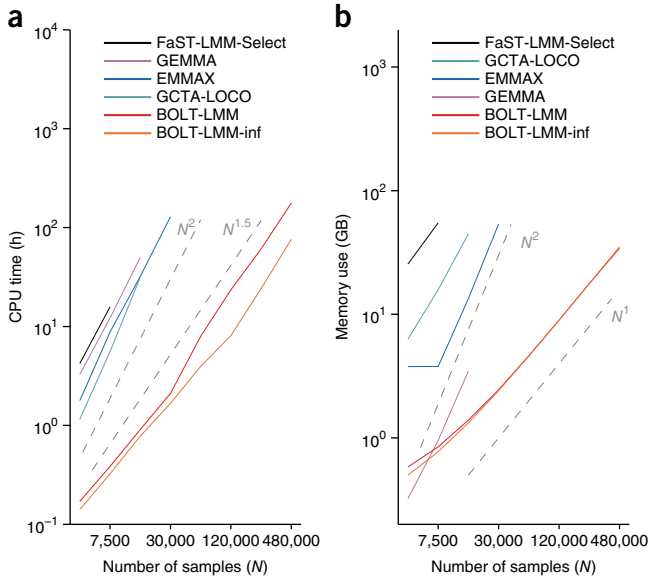

Figure 1 Computational performance of mixed-model association methods. (**a**,**b**) Log-log plots of run time (**a**) and memory use (**b**) as a function of sample size (N). The slopes of the curves correspond to exponents of power law scaling with N. Benchmarking was performed on simulated data sets in which each sample was generated as a mosaic of genotype data from 2 random 'parents' from the WTCCC2 data set ($N = 15{,}633$, $M = 360{,}000$) and phenotypes were simulated with $M_{causal} = 5{,}000$ SNPs explaining $h^2_{causal} = 0.2$ of phenotypic variance. The reported run times are the medians of five identical runs using one core of a 2.27-GHz Intel Xeon L5640 processor. We caution that run time comparisons may vary by a small constant factor as a function of the computing environment. FaST-LMM-Select (GCTA-LOCO and EMMAX) memory usage exceeded the 96 GB available at $N = 15{,}000$ (30,000 and 60,000, respectively). GEMMA encountered a runtime error (segmentation fault) at $N = 30{,}000$. Software versions: FaST-LMM-Select, v2.07; GCTA-LOCO, v1.24; EMMAX, v20120210; GEMMA, v0.94. Numerical data are provided in **Supplementary Table 1**.

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

We computed χ^2 association statistics using linear regression with ten principal components[32], GCTA-LOCO[12], BOLT-LMM-inf and BOLT-LMM. We were unable to test FaST-LMM-Select[15] on this data set because of its memory requirements (**Fig. 1**). For each method, we computed the means of the χ^2 statistics over the SNPs of standardized effect and compared these means across simulations involving different numbers of causal SNPs (**Fig. 2a** and **Supplementary Table 2**).

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

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

To verify that BOLT-LMM is correctly calibrated and robust to confounding, we also computed mean χ^2 statistics across SNPs on the second halves of chromosomes, simulated to all have zero effect ('null' SNPs). Because our simulated phenotypes included an ancestry effect, the statistics from linear regression without correcting for population stratification suffered an inflation of 35%. In contrast, the BOLT-LMM and BOLT-LMM-inf statistics were both well calibrated (**Supplementary Fig. 4b** and **Supplementary Tables 3** and **4**). We further verified that type I error was properly controlled (Online Methods and **Supplementary Table 5**) and that the distribution of statistics at null SNPs did not deviate noticeably from

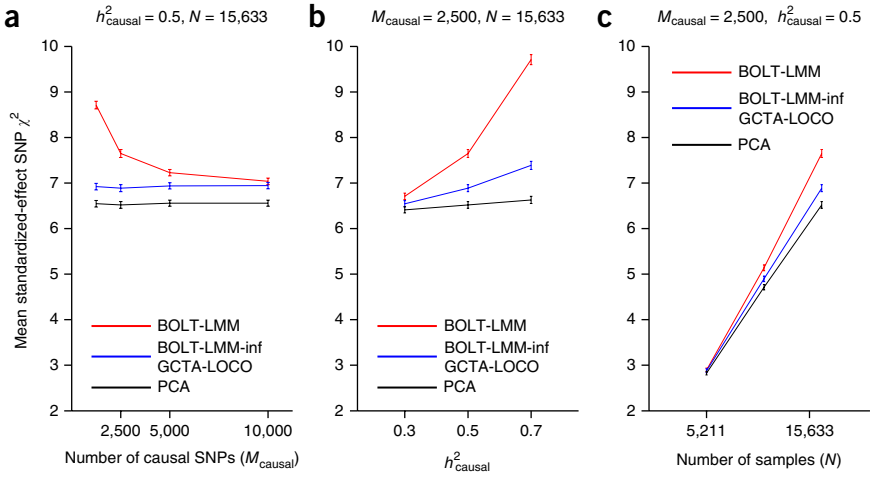

Figure 2 BOLT-LMM increases the power to detect associations in simulations. (**a**–**c**) Mean χ^2 statistics at SNPs of standardized effect as a function of the number of causal SNPs (**a**), the proportion of variance explained by causal SNPs (**b**) and the number of samples (**c**). Simulations used real genotypes from the WTCCC2 data set (N = 15,633, M = 360,000) and simulated phenotypes with the specified number of causal SNPs explaining the specified proportion of the phenotypic variance and 60 additional SNPs of standardized effect explaining an additional 2% of the variance. Error bars, s.e.m.; 100 simulations. We verified on the first five simulations that the BOLT-LMM-inf and GCTA-LOCO statistics were nearly identical (**Supplementary Table 7**). Numerical data are provided in **Supplementary Table 2**.

a 1-degree-of-freedom χ^2 distribution (**Supplementary Fig. 6a**,**b**). The genomic inflation factors[33] for BOLT-LMM and BOLT-LMM-inf exceeded 1 in these simulations (**Supplementary Fig. 6c**,**d**), consistent with the polygenicity of the simulated phenotype and the use of a mixed-model statistic that successfully avoids proximal contamination[12,13]. In contrast, EMMAX and GEMMA had deflated test statistics (**Supplementary Fig. 4b**).

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

Lastly, we investigated the similarity between the BOLT-LMM-inf mixed-model statistic and the statistics for existing methods at the level of individual SNPs. Despite BOLT-LMM-inf using an infinitesimal model, its statistic was not identical to any statistic computed by an existing mixed-model method because it is an approximate test statistic and avoids proximal contamination (**Table 1** and Online Methods). Nonetheless, we observed that BOLT-LMM-inf statistics very nearly matched GCTA-LOCO statistics (from a standard prospective model), with $R^2 > 0.999$ (**Supplementary Fig. 7** and **Supplementary Table 7**).

Application of BOLT-LMM to WGHS phenotypes

To assess the efficacy of Gaussian mixture-model analysis for increasing power with real phenotypes, we analyzed 9 phenotypes in the WGHS (N = 23,294 samples and M = 324,488 SNPs after quality control) (Online Methods). These phenotypes consisted of five lipid phenotypes, height, body mass index and two blood pressure phenotypes; we chose to analyze these phenotypes because of the availability of large-scale GWAS results.

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

BOLT-LMM achieved higher power than PCA for all traits studied (**Fig. 3** and **Supplementary Table 9**). Most of the increase in power was due to gains over infinitesimal mixed-model analysis, with the magnitude of this gain in power increasing with the inferred concentration of genetic effects at a few loci (**Supplementary Table 10**). Standard errors of the direct method of assessing improvement (estimating mean χ^2 statistics at known loci) were somewhat high (0.6–2.2%; **Fig. 3a** and **Supplementary Table 9**), so the improvement was statistically significant ($P < 0.05$) for only six of nine traits. According to the prediction R^2 metric, the improvements were statistically significant for all traits ($P < 0.0002$) (**Fig. 3b** and **Supplementary Table 9**). The largest gains in power were achieved for lipid traits; for the levels of ApoB, a lipoprotein closely related to low-density lipoprotein (LDL) cholesterol, BOLT-LMM analysis achieved a 10% increase in mean χ^2

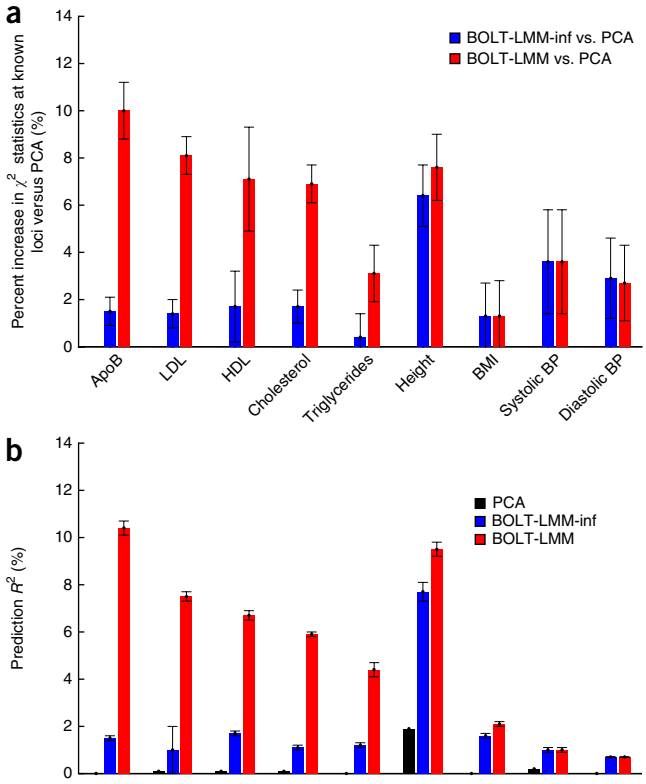

Figure 3 BOLT-LMM increases the power to detect associations for WGHS phenotypes. We compare the power (measured using two roughly equivalent metrics) of linear regression using ten principal components, standard (infinitesimal) mixed-model analysis and BOLT-LMM Gaussian mixture-model analysis. (**a**) Percent increases in χ^2 statistics across known associated loci using mixed-model methods versus PCA: values shown are the ratios of the sums of χ^2 statistics over the typed SNPs in highest linkage disequilibrium with the published associated SNPs. (**b**) Prediction R^2 values from fivefold cross-validation: each fold was left out in turn, and predictions were computed by fitting all SNP effects simultaneously (for mixed-model methods) or estimating covariate effects (for PCA) using the training folds. (Note that BOLT-LMM-inf is equivalent to best linear unbiased prediction (BLUP) here.) We show PCA in **b** because the small amount of variance that the principal components explain (due to population stratification) provides a baseline that allows the translation of prediction R^2 values into the gain in power for mixed-model association versus regression with principal-component covariates. That is, the correspondence between association power and prediction accuracy is such that the red bars in **a** roughly correspond to the differences between the red and black bars in **b**, and the situation is analogous for the blue bars in **a** (Online Methods). Error bars show jackknife standard error over known loci in **a** (**Supplementary Table 8**) and five cross-validation folds in **b**. Numerical data are provided in **Supplementary Table 9**. HDL, high-density lipoprotein; BMI, body mass index; BP, blood pressure. X axis labels are the same in panels **a**,**b**.

statistics versus PCA and a 9% increase versus infinitesimal mixed-model analysis at known loci. To verify that these increases were not merely driven by a few loci with the largest effects, we also computed flat averages across loci of improvements in χ^2 statistics (restricting to loci replicating in the WGHS with at least nominal significance of $P < 0.05$ to reduce statistical noise) and obtained consistent results (**Supplementary Table 8**). Simulations showed that these improvements will increase with sample size (**Fig. 2c** and **Supplementary Fig. 5**).

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

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

DISCUSSION

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

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

Bayesian methods have previously been developed that apply non-infinitesimal models to improve the accuracy of genetic risk prediction. These methods extend in principle to association testing, although the Bayes factors and posterior inclusion probabilities that are naturally produced by Bayesian analysis do not directly translate into customary GWAS frequentist test statistics[37]. The variational Bayes spike regression (vBsr) method[38] is a recent step toward addressing this issue, proposing a z statistic heuristically calibrated by assuming that the

vast majority of variants are unassociated (as in genomic control[33]), but such a technique is prone to deflation when large sample sizes cause inflation due to polygenicity[13,24]. BOLT-LMM sidesteps this difficulty via its hybrid approach of leaving each chromosome out in turn, fitting a Bayesian model on the remaining SNPs and then applying a retrospective hypothesis test for association of the left-out SNPs with the residual phenotype. In contrast to modeling all SNPs simultaneously and assessing evidence for association using Bayesian posterior inference[37], our approach generalizes existing mixed-model methods that are widely used, and we believe its ability to harness the power of Bayesian analysis while still computing frequentist statistics will be useful to GWAS practitioners. Additionally, such a hybrid approach lends itself readily to efficiently testing millions of imputed SNP dosages for association while including only typed SNPs in the mixed model, which we recommend to limit computational costs.

Although BOLT-LMM improves upon existing mixed-model association methods in both speed and power, it still has limitations. First, the gain in power that BOLT-LMM offers over existing methods via its more flexible prior on SNP effect sizes is contingent on the true genetic architecture being sufficiently non-infinitesimal and the sample size being sufficiently large (**Supplementary Fig. 5**). Second, BOLT-LMM, like existing mixed-model methods, is susceptible to loss of power when used to analyze large ascertained case-control data sets in diseases of low prevalence[12]. We recommend BOLT-LMM for randomly ascertained quantitative traits, ascertained case-control studies of diseases with a prevalence of ≥5% (**Supplementary Table 13**)—for example, type 2 diabetes, heart disease, common cancers, hypertension and asthma—and studies of rarer diseases in large, non-ascertained population cohorts[39,40]. For large ascertained case-control studies of rarer diseases, we are developing a method of modeling ascertainment using posterior mean liabilities (LTMLM)[41]; applying the techniques of BOLT-LMM to these posterior mean liabilities is an avenue for future research. Third, although mixed-model analysis is effective in correcting for many forms of confounding, performing careful data quality control remains critical to avoid false positives. Fourth, our work does not attempt to estimate the extent to which the heritability parameter estimated by BOLT-LMM (denoted as h_g^2) might be influenced by population structure or relatedness, nor does it conduct or evaluate genetic prediction in external validation samples from an independent cohort[34]. Fifth, we have not studied the performance of mixed-model methods in data sets dominated by family structure[23]. Sixth, the running time of BOLT-LMM scales with the number of phenotypes analyzed; for data sets with a very large number of phenotypes (P), the GRAMMAR-Gamma method[10], which has a running time of $O(MN^2 + MNP)$ (reviewed in ref. 12), may be faster. Seventh, we have only tested BOLT-LMM in human data sets, which have very different patterns of linkage disequilibrium and genetic architectures from plant and animal data. In particular, given that some approximations we make may be violated in non-human data sets (for example, treating the denominator of the prospective test statistic as nearly constant[10]), we are unsure whether the BOLT-LMM statistic is valid in these scenarios. Similarly, these assumptions should be viewed with caution when testing very rare variants. Finally, we have developed fast mixed-model analysis for a mixed model with one random genetic effect; extending the algorithm to model multiple variance components[42] is a direction for future work.

URLs. BOLT-LMM software and source code, http://www.hsph.harvard.edu/alkes-price/software/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

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

AUTHOR CONTIRBUTIONS

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

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **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. *Nat. Genet.* **42**, 348–354 (2010).
4. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
5. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
6. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
7. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**, 825–830 (2012).
8. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071 (2012).
9. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. 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. *Nat. Genet.* **44**, 1166–1170 (2012).
11. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat. Genet.* **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. *Nat. Genet.* **46**, 100–106 (2014).
13. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
14. Stahl, E.A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
15. Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep.* **3**, 1815 (2013).
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.H., Hayes, B.J. & Goddard, M.E. 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 Genet.* **9**, e1003264 (2013).
20. Meuwissen, T.H., 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**, 2 (2009).
21. Carbonetto, P. & Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **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. *Am. J. Hum. Genet.* **92**, 652–666 (2013).
24. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* doi:10.1038/ng.3211 (2 February 2015).
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. *Clin. Chem.* **54**, 249–255 (2008).
26. García-Cortés, L.A., Moreno, C., Varona, L. & Altarriba, J. Variance component estimation by resampling. *J. Anim. Breed. Genet.* **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. *J. Dairy Sci.* **91**, 360–366 (2008).
29. VanRaden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **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**, 214–219 (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. *Nat. Genet.* **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. *Nat. Rev. Genet.* **14**, 507–515 (2013).
35. Campbell, C.D. *et al.* Demonstrating stratification in a European American population. *Nat. Genet.* **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* **197**, 1045–1049 (2014).
37. Stephens, M. & Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **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* **497**, 517–520 (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 Genet.* **7**, e1002141 (2011).
41. Hayeck, T. *et al.* Mixed model with correction for case-control ascertainment increases association power. *bioRxiv* doi:10.1101/008755 (2014).
42. Speed, D. & Balding, D.J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).

ONLINE METHODS

Standard mixed-model association methods. Standard methods employ a model

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

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

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

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

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

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

where σ_g^2 is a variance parameter. Environmental effects are assumed to be independently and identically distributed normally such that e is also multivariate normal with

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

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

In practice, the variance parameters σ_g^2 and σ_e^2 are unknown. Several existing methods[3,10,12] therefore take a two-step approach to computing association statistics, first estimating the variance parameters (with the SNP x_{test} removed from the model) using restricted maximum-likelihood (REML) and then computing the prospective χ^2 (1-degree-of-freedom) test statistic (as previously proposed in family-based tests[43]) as

$$\chi^2_{\text{LMM}} = \frac{(x_{\text{test}}'V^{-1}y)^2}{x_{\text{test}}'V^{-1}x_{\text{test}}} \quad (5)$$

where

$$V = \text{cov}(y) = \sigma_g^2 K + \sigma_e^2 I \quad (6)$$

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

$$\chi^2_{\text{LMM–LOCO}} = \frac{(x_{\text{test}}'V_{\text{LOCO}}^{-1}y)^2}{x_{\text{test}}'V_{\text{LOCO}}^{-1}x_{\text{test}}} \quad (7)$$

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

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

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

$$\chi^2_{\text{BOLT-LMM-inf}} = \frac{(x_{\text{test}}'V_{\text{LOCO}}^{-1}y)^2}{c_{\text{inf}}} \quad (8)$$

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

$$c_{\text{inf}} = \frac{\text{mean}\,(x_{\text{test}}'V_{\text{LOCO}}^{-1}y)^2}{\text{mean}\,\chi^2_{\text{LMM–LOCO}}} \quad (9)$$

such that

$$\text{mean}(\chi^2_{\text{BOLT-LMM-inf}}) = \text{mean}(\chi^2_{\text{LMM-LOCO}}) \quad (10)$$

In practice, for computational efficiency, we take means over 30 pseudorandom SNPs not significantly associated with the phenotype ($\chi^2 < 5$ estimated with the GRAMMAR statistic[44]). We have observed empirically that 30 random SNPs are sufficient to estimate the calibration factor to within 1% (**Supplementary Table 14**).

We can view the BOLT-LMM-inf statistic either as an approximation of the standard prospective statistic (which treats phenotypes as random) or as a retrospective statistic (which treats genotypes as random and builds a null model on SNPs). The first perspective is motivated by the observation that, in human genetics applications, the denominator of the prospective statistic in equation (5), $x_{\text{test}}'V^{-1}x_{\text{test}}$, is nearly independent of the SNP x_{test} being tested[10]. From this perspective, BOLT-LMM-inf is similar to GRAMMAR-Gamma[10], with two key differences: (i) BOLT-LMM-inf is computed via much faster algorithms for performing initial variance parameter estimation and estimating the calibration constant and (ii) BOLT-LMM-inf avoids proximal contamination via LOCO analysis. Alternatively, we can also view BOLT-LMM-inf as a retrospective quasi-likelihood score test similar to $T^{\text{SCORE–R}}$ (ref. 45) and MASTOR[23] (**Supplementary Note**).

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

We may generalize this approach by defining

$$\chi^2_{\text{BOLT-LMM}} = \frac{(x_{\text{test}}'y_{\text{residual–LOCO}})^2}{c} \quad (11)$$

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

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

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

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

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

and environmental effects e_n (for n indexing samples) are independently drawn from

$$e_n \sim N(0,\sigma_e^2)$$

Performing best linear unbiased prediction amounts to computing the posterior mean of the genetic effect $X_{\text{LOCO}}\beta_{\text{LOCO}}$.

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

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

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

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

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

$$y_{\text{residual-LOCO}} = y - X_{\text{LOCO}}\beta_{\text{LOCO}} \tag{15}$$

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

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

Step 1a: estimate variance parameters. A key feature of BOLT-LMM is the estimation of the variance parameters σ_g^2 and σ_e^2 using only linear-time iterations without building or decomposing any covariance matrices. We use a Monte Carlo REML approach[26,27] that eliminates all $O(MN^2)$ and $O(N^3)$ time matrix computations, requiring only the solution of linear systems of mixed-model equations. We solve the mixed-model equations using conjugate gradient iteration, which requires only $O(MN)$ time matrix vector products[28,29] (**Supplementary Note**).

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

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

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

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

We reparameterize the remaining 2 degrees of freedom using the parameters p and f_2, where f_2 denotes the proportion of the total mixture variance within the second Gaussian (the spike component that models small genome-wide effects)

$$f_2 = \frac{(1-p)\sigma_{\beta,2}^2}{p\sigma_{\beta,1}^2 + (1-p)\sigma_{\beta,2}^2} \tag{17}$$

Because the model fit is insensitive to the precise values of the mixture parameters, we test a discrete set of model parameter combinations: $f_2 \in \{0.5, 0.3, 0.1\}$, $p \in \{0.5, 0.2, 0.1, 0.05, 0.02, 0.01\}$. Note that $f_2 = 0.5$, $p = 0.5$ corresponds to the infinitesimal model: when $f_2 = 1 - p$, the two Gaussians are identical and the mixture is degenerate. We bound f_2 from below to ensure that at least a small amount (10%) of the mixture variance is assigned to the spike component, protecting against confounding from genome-wide effects. We bound p from below to prevent the model from trying to fit too strongly to a few SNPs, which makes model-fitting computationally difficult and also increases the susceptibility to confounding. BOLT-LMM performs model selection among the 18 possible parameter pairs (f_2, p) by performing cross-validation to optimize mean-squared prediction R^2.

BOLT-LMM uses a variational approximation to fit Bayesian linear regressions with Gaussian mixture priors. Approximation methods are necessary for Bayesian inference in this setting because exact posterior means involve intractable integrals. We apply a fully factored variational approximation[21,22,38] that repeatedly loops through the SNPs, updating the estimated effect size of each SNP with its posterior mean conditional on current estimates of all other SNP effects. This iteration has also previously been termed 'iterative conditional expectation' (ICE)[20]. The variational Bayes framework puts this iteration on a sound theoretical footing as an optimization of an approximate log-likelihood

doi:10.1038/ng.3190

function; the iteration monotonically increases this function and is guaranteed to converge[46]. In fact, we show that the optimization can be reformulated as cyclic coordinate descent applied to a penalized regression problem arising from Bayesian linear regression using a transformed prior (**Supplementary Note**). The approximate log likelihood also serves as a convenient convergence criterion: BOLT-LMM stops the iteration when the increase in approximate log likelihood over one full update cycle drops below 0.01.

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

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

WGHS data set. The WGHS is a prospective cohort of initially healthy, female North American healthcare professionals. We analyzed 23,294 individuals with self-reported European ancestry with genotyping at 324,488 SNPs after quality control (**Supplementary Note**)[25].

Interpretation of the heritability parameter. The heritability parameter (denoted h_g^2) estimated by BOLT-LMM may in general include some contribution from cryptic relatedness or population structure[47] and thus may not strictly correspond to the heritability explained by genotyped SNPs[48]. Kang *et al.*[3] refer to this parameter as 'pseudo-heritability' for this reason. Because the WGHS samples that we primarily analyze here do not contain substantial relatedness or population structure, we have simply used the notation h_g^2 to avoid complicating the discussion.

43. Chen, W.-M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
44. 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).
45. Chen, W.-M., Manichaikul, A. & Rich, S.S. A generalized family-based association test for dichotomous traits. *Am. J. Hum. Genet.* **85**, 364–376 (2009).
46. Boyd, S.P. & Vandenberghe, L. *Convex Optimization* (Cambridge University Press, 2004).
47. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
48. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).