nature genetics

Advantages and pitfalls in the application of mixed-model association methods

Jian Yang[1,2,8], Noah A Zaitlen[3,8], Michael E Goddard[4,9], Peter M Visscher[1,2,9] & Alkes L Price[5–7,9]

Mixed linear models are emerging as a method of choice for conducting genetic association studies in humans and other organisms. The advantages of the mixed-linear-model association (MLMA) method include the prevention of false positive associations due to population or relatedness structure and an increase in power obtained through the application of a correction that is specific to this structure. An underappreciated point is that MLMA can also increase power in studies without sample structure by implicitly conditioning on associated loci other than the candidate locus. Numerous variations on the standard MLMA approach have recently been published, with a focus on reducing computational cost. These advances provide researchers applying MLMA methods with many options to choose from, but we caution that MLMA methods are still subject to potential pitfalls. Here we describe and quantify the advantages and pitfalls of MLMA methods as a function of study design and provide recommendations for the application of these methods in practical settings.

Mixed-model association methods prevent false positive associations and increase power

Mixed linear models are an emerging method of choice when conducting association mapping in the presence of sample structure, including geographic population structure, family relatedness and/or cryptic relatedness[1–12]. The basic approach involves building a genetic relationship matrix (GRM) that models genome-wide sample structure, estimating the contribution of the GRM to phenotypic variance using a random-effects model (with or without additional fixed effects) and computing association statistics that account for this component of phenotypic variance (methods are provided in the **Supplementary Note**). We note that mixed linear models can also be used to estimate components of heritability explained by genotyped markers[13,14] and to predict complex traits using genetic data[15,16].

MLMA methods are effective in preventing false positive associations due to sample structure in studies of humans and model organisms[1–6]. In particular, simulations show that the correction for confounding is nearly perfect for common variants, even when geographic population structure, a fixed effect, is modeled as a random effect on the basis of overall covariance[6,17–19] (however, rare variants pose a greater challenge for all methods owing to differential confounding of rare and common variants[20]). MLMA methods also increase power to detect causal variants by applying a correction that is specific to sample structure[1–6]. In the case of geographic population structure, markers with large differences in allele frequency between populations receive a larger correction. In the case of relatedness structure, the contribution of related individuals to test statistics is reduced, preventing overweighting of redundant information due to correlation structure.

An underappreciated point is that MLMA can also increase power in studies without sample structure by implicitly conditioning on associated loci other than the candidate locus that do not show genome-wide significant association in the data being analyzed[8]. For example, a GRM computed from all markers can be used to approximate the set of causal markers (implicitly assuming that all markers are causal), but this approximation can be generalized. The increase in power scales with the ratio N/M of the number of samples (N) to the effective number of independent markers (M), as the information about unknown associated loci depends on the number of samples. In simulations of a quantitative trait with no sample structure and no linkage disequilibrium (LD) between markers (methods are provided in the **Supplementary Note**), application of MLMA instead of linear regression increases average $-\log_{10}$ (P values) at causal markers from 2.89 to 2.94 (1.8% increase) when $N = 10,000$ and $M = 100,000$, and from 2.92 to 3.46 (18% increase) when $N = 10,000$ and $M = 10,000$. We note that this improvement is contingent on exclusion of the candidate marker from the GRM.

Reducing the computational cost of mixed-model association analysis

In initial implementations of MLMA, the component of phenotypic variance explained by the GRM was estimated separately when testing for the association of each candidate marker. This approach accounts for the fact that the total variance explained by all markers except the candidate marker may vary across candidate markers in the case of markers of large effect[1–3]. Even for efficient implementations[3], this variance is computationally demanding, requiring a computation time of $O(MN^3)$, where

[1]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [2]University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia. [3]Department of Medicine, Lung Biology Center, University of California, San Francisco, San Francisco, California, USA. [4]Faculty of Land and Food Resources, University of Melbourne, Parkville, Victoria, Australia. [5]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [6]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [7]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [8]These authors contributed equally to this work. [9]These authors jointly directed this work. Correspondence should be addressed to P.M.V. (peter.visscher@uq.edu.au) or A.L.P. (aprice@hsph.harvard.edu).

Received 23 August 2013; accepted 30 December 2013; published online 29 January 2014; doi:10.1038/ng.2876

Table 1 Computational cost of EMMAX, FaST-LMM, GEMMA, GRAMMAR-Gamma and GCTA

Method	Building GRM	Variance components	Association statistics
EMMAX	$O(MN^2)$	$O(N^3)$	$O(MN^2)$
FaST-LMM[a]	$O(MN^2)$	$O(N^3)$	$O(MN^2)$
GEMMA	$O(MN^2)$	$O(N^3)$	$O(MN^2)$
GRAMMAR-Gamma	$O(MN^2)$	$O(N^3)$	$O(MN)$
GCTA	$O(MN^2)$	$O(N^3)$	$O(MN^2)$

For each method, we list the computational cost of each step.
[a]If $M < N$, the computational cost of FaST-LMM can be reduced to $O(M^2N)$.

M is the number of markers and N is the number of samples, because variance component estimation is repeated for each candidate marker.

Several computational speedups have subsequently been developed. First, two independent studies observed that, if markers have small effects, variance components can be approximated by estimating them only once using all markers (as previously proposed in family-based tests[21]), making MLMA feasible on large data sets[4,5]. Two subsequent studies developed computationally efficient exact methods[7,11], which do not require variance components to be the same for all candidate markers. Each of these methods enables exact MLMA analysis in a computation time of $O(MN^2 + N^3)$. The difference between approximate and exact methods was reported to be large in a mouse data set with pervasive relatedness and large effect sizes but was negligible in a human data set[11]. Another fast approximate method has recently been described[12]. Several of these methods use a single eigendecomposition of the GRM to rotate the data, thereby removing its structure[7,11,12].

The computation time of each method can be broken down into three steps: (i) building the GRM, (ii) estimating variance components and (iii) computing association statistics for each SNP. In **Table 1**, we list the computational cost of each of these steps for the EMMAX[4], FaST-LMM[7], GEMMA[11] and GRAMMAR-Gamma[12] implementations, as well as for our GCTA[22] implementation (methods are provided in the **Supplementary Note**; see URLs). GRAMMAR-Gamma has the advantage that the cost of step (iii) is reduced from $O(MN^2)$ to $O(MN)$, greatly reducing the cost of analyzing a large number of phenotypes. To quantify the computational cost in data sets of realistic size, we benchmarked the running time and memory usage of GCTA using simulations of a quantitative trait without sample structure (**Supplementary Table 1** and **Supplementary Note**).

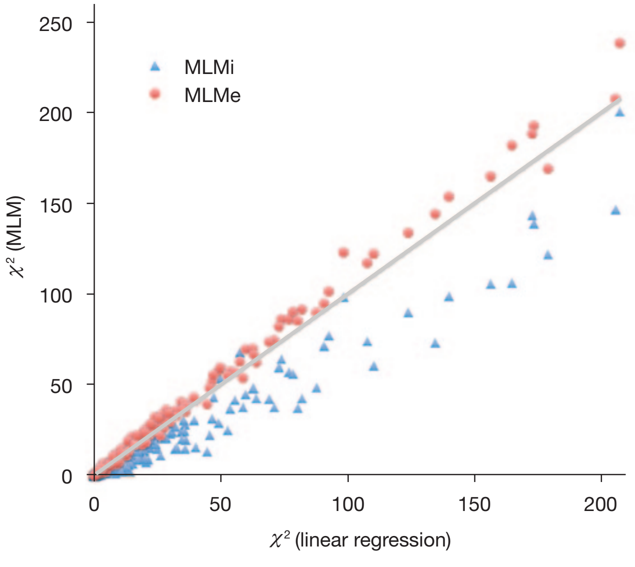

Figure 1 MLMe increases power and MLMi decreases power compared to linear regression. We report χ^2 association statistics at 500 causal markers for MLMi versus linear regression and for MLMe versus linear regression for a simulation with genotype data from ref. 25 and simulated phenotypes for $N = 10{,}000$ samples. For MLMe, the 500 causal markers were always excluded from computing the GRM.

Pitfall: loss in power when the candidate marker is included in the GRM

Recent work has shown that inclusion of the candidate marker in the GRM can lead to loss in power[7,8,23]. This decreased power is due to double-fitting of the candidate marker in the model, both as a fixed effect tested for association and as a random effect as part of the GRM. Listgarten *et al.*[8], who referred to this phenomenon as "proximal contamination," demonstrated that a mixed linear model with the candidate marker excluded (MLMe) is the mathematically correct approach and provided an elegant and efficient algorithm for MLMe analysis (implemented in FaST-LMM software). However, owing to computation time or memory constraints (and complexities of LD), the mixed linear model with the candidate marker included (MLMi) is more commonly applied in practice[23]. It is of interest to quantify the power loss of MLMi relative to MLMe to help guide this choice of method. In this section, we provide new analytical derivations, validated by simulations, to quantify the reduction in test statistics when MLMi is applied.

Analytical derivations of mean association statistics. We assumed a set of unrelated samples without population structure or other artifacts. We let N denote the number of samples, M denote the number of markers and h_g^2 denote the heritability explained by genotyped and/or imputed markers[13]. We assumed that markers were unlinked, but the same derivations apply to linked markers if M denotes the effective number of independent markers, which for humans is approximately 60,000 (ref. 24; **Supplementary Note**). We emphasize that it is the effective number of independent markers (not the total number of markers) that matters. Details of each derivation below are provided in the **Supplementary Note**.

For linear regression (LR), the expected mean of χ^2 association statistics (λ_{mean}) is

$$\lambda_{mean}\,(LR) = 1 + Nh_g^2\,/\,M \qquad (1)$$

regardless of the genetic architecture of the trait[24].

For MLMi, the λ_{mean} value at markers used to construct the GRM is

$$\lambda_{mean}\,(MLMi) = 1 \qquad (2)$$

Equation (2) highlights the dangers of using λ_{mean} (or λ_{median}) to assess the presence of population stratification or other artifacts. A researcher who observes lower λ_{mean} (or λ_{median}) values for MLMi than for linear regression might conclude that this difference is due to correction for confounding, but this result is in fact expected, even in the absence of any confounding.

Finally, for MLMe,

$$\lambda_{mean}\,(MLMe) = 1 + \frac{Nh_g^2 M}{1 - r^2 h_g^2} \qquad (3)$$

where $r^2 \approx Nh_g^2/M$ when $M > N$. The ratio of λ_{mean} between MLMe and MLMi is also

$$1 + \frac{Nh_g^2 M}{1 - r^2 h_g^2}$$

which is consistent for causal, null and all markers (**Supplementary Table 2**). If $M >> N$ (i.e., r^2 is small), this ratio is only slightly greater than $1 + Nh_g^2\,/\,M$. The difference between MLMe and MLMi is that MLMe tests the null hypothesis that the candidate marker has no effect, whereas MLMi tests the null hypothesis that the candidate marker has an effect size drawn from a normal distribution $N(0, h_g^2/M)$.

Simulations. We compared the results of linear regression, MLMi and MLMe in simulations of a quantitative trait without sample structure for various values of N and M (methods are provided in the **Supplementary Note**). Our results showed that MLMe increased power relative to linear regression, but MLMi reduced power (**Table 2** and **Supplementary Fig. 1**). The magnitude of these effects was proportional to the N/M ratio, consistent with our derivations (**Table 2** and **Supplementary Table 3**); the difference in power was small at $N = 10{,}000$ and $M = 100{,}000$, but it is increasingly common for genome-wide association studies (GWAS) to be performed at sample sizes considerably larger than $N = 10{,}000$. In all simulations, linear regression and MLMe had λ_{mean} values at null markers equal to 1.00, but MLMi had λ_{mean} values at all markers equal to 1.00, such that MLMi had λ_{mean} values for null markers that were less than 1.00.

We also conducted simulations based on real genotypes for 133,036 SNPs on chromosomes 1–3 in 10,000 unrelated individuals from data analyzed in ref. 25 (methods are provided in the **Supplementary Note**). For simplicity, when running MLMe, we excluded from the GRM all the SNPs on a chromosome where the candidate SNP was located. Results again showed that MLMe increased power relative to linear regression but that MLMi reduced power (**Fig. 1**, **Table 2** and **Supplementary Table 3**), consistent with our derivations. Effects were magnified because only a portion of the genome was analyzed, but analogous effects in proportion to N/M are expected at other values of N and M.

In summary, we recommend the use of MLMe in preference to MLMi. An efficient implementation of MLMe via a leave-one-chromosome-out analysis is provided in GCTA software (GCTA-LOCO; methods are provided in the **Supplementary Note**). An efficient implementation is also provided in FaST-LMM software[7,8].

Pitfall: using a small subset of markers in the GRM can compromise correction for stratification

Three recent papers have advocated choosing a subset of markers to include in the GRM when employing MLMA methods[7,8,26]. FaST-LMM[7] uses an equally spaced subset of 4,000 (or 8,000) random markers (M_R) in the GRM, motivated by a computational speedup that reduces computational cost to $O(M_R^2N)$ when $M_R < N$. FaST-LMM-Select[8,26] uses the top markers (M_T) with the most significant linear regression P values in the GRM, with M_T markers chosen on the basis of either the first local minimum of the genomic control factor $\lambda_{\mathrm{median}}$[8] or the global maximum of out-of-sample prediction accuracy using the resulting GRM[26]. The latter approach allows for the possibility of including all markers in the GRM ($M_T = M$) but is computationally intensive (with running time >10 times that for MLMA using all markers) owing to the high cost of computing out-of-sample prediction accuracy using all markers; an alternative (described on page 10 of the FaST-LMM version 2.05 user manual) is to choose M_T markers using the first local maximum of out-of-sample prediction accuracy. These approaches have allowed the valuable observation that a substantial increase in power can be attained by implicitly conditioning only on loci that are relatively likely to be truly associated, motivating a thorough investigation of the impact on correcting for stratification. Below, we evaluate the impact of these choices on both false positive associations and power. In all of these simulations, we excluded the candidate marker from the GRM (MLMe), consistent with refs. 7,8,26.

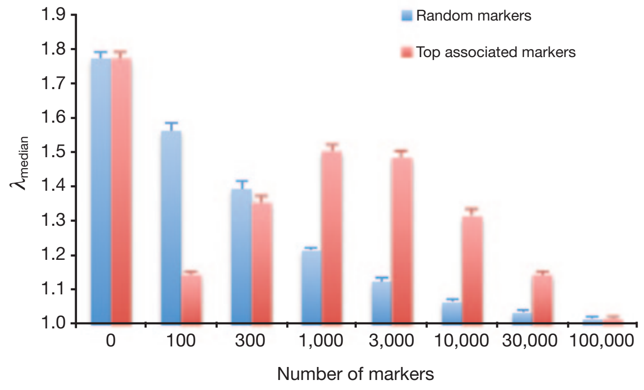

Figure 2 Effectiveness of mixed linear models using random or top associated markers in correcting for stratification. We report average $\lambda_{\mathrm{median}}$ (± s.e.m.) in 100 simulations with population stratification based on $N = 10{,}000$ samples, $M = 100{,}000$ markers, 2 discrete subpopulations with fixation index (F_{ST}) = 0.005 and a mean trait difference of 0.25 s.d. between subpopulations. Calibration of small P values is reported in **Supplementary Table 4**.

To investigate the number of random markers (M_R) needed to correct for stratification, we conducted simulations of a quantitative trait with population stratification (methods are provided in the **Supplementary Note**). Our results indicate that, when there is subtle population stratification, a few thousand random markers are not sufficient to provide a thorough correction for stratification (**Fig. 2**, **Supplementary Table 4** and **Supplementary Note**), consistent with previous studies[11,27].

We also investigated use of the top M_T associated markers to correct for stratification. Our results indicate that using the top M_T associated markers selected on the basis of the first local minimum of the genomic control factor $\lambda_{\mathrm{median}}$[8] may not be effective in correcting for stratification and can lead to a local minimum in $\lambda_{\mathrm{median}}$ that is different from the global minimum[26] (**Fig. 2** and **Supplementary Note**). In contrast, the approach from ref. 26 of using the top M_T associated markers selected

Table 2 MLMe increases power and MLMi decreases power compared to linear regression

Number of samples (N)	Number of markers (M)	Linear regression (expected value)	MLMi (expected value)	MLMe[a] (expected value)
Scenario I: simulated unlinked markers				
10,000	10,000	10.93 ± 0.03 (11.00)	9.81 ± 0.01 (10.05[b])	13.27 ± 0.03 (13.36)
10,000	100,000	11.20 ± 0.03 (11.00)	10.97 ± 0.02 (10.09[b])	11.40 ± 0.03 (11.25)
Scenario II: simulations based on real genotype data[25]				
10,000	133,036	26.99 ± 0.10 (26.00)	21.44 ± 0.05 (19.85)	28.42 ± 0.10 (29.08)

In scenario I, we report average χ^2 association statistics (± s.e.m.) at 500 candidate causal markers for linear regression, MLMi and MLMe averaged across 100 simulations based on simulated genotypes. In scenario II, we report average χ^2 association statistics (± s.e.m.) at 200 causal markers for simulations based on genotype data from ref. 25 with simulated phenotypes. In both scenarios, expected values based on theoretical derivations are given in parentheses. More details, including simulations at other values of N and M, results for all markers, power to detect significant associations at different P-value thresholds and the equations to calculate the expected values, are provided in **Supplementary Table 3**.

[a] The 500 candidate causal markers were always excluded from calculating the GRM in scenario I, and the MLMe analysis was performed using GCTA-LOCO in scenario II. [b] For MLMi, the h_g^2 for markers included in the GRM is 100%, and the derivation is much less accurate. However, the derivation is much more accurate at lower values of h_g^2 (**Supplementary Table 3a**).

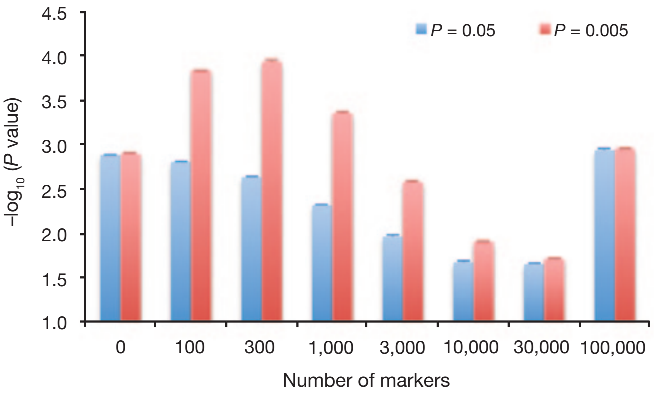

Figure 3 Effectiveness of mixed linear models using top associated markers in increasing study power. We report average $-\log_{10}$ (P values) (± s.e.m.) at causal markers in 100 simulations based on N = 10,000 samples, M = 100,000 markers and fraction $P = 0.05$ or $P = 0.005$ for causal markers. Power to detect significant associations at different P-value thresholds is reported in **Supplementary Table 5**.

on the basis of the global maximum of out-of-sample prediction accuracy resulted in $M_T = M$ and thus provided an effective correction for stratification in these simulations.

We then turned to the question of power. We generalized our simulations without sample structure, with fraction $P = 0.05$ or $P = 0.005$ of causal markers, to consider the impact of using the top M_T associated markers in the GRM for various values of M_T (**Fig. 3**). When $P = 0.05$, there are a large number of causal markers with small effect sizes, such that the top M_T associated markers did not correspond with the true set of causal markers, and including all markers in the GRM ($M_T = M$) performed best. When $P = 0.005$, there are a smaller number of causal markers with larger effect sizes, such that the top M_T associated markers more closely reflected the true set of causal markers, and including only a small subset of top markers in the GRM performed best. Results at other parameter settings showed that the optimal strategy depended on both the sample size and the genetic architecture of the trait (**Supplementary Table 5** and **Supplementary Note**). The approach from ref. 26 of using the top M_T associated markers selected on the basis of the global maximum of out-of-sample prediction accuracy yielded the value of M_T that maximized power in each of these simulations, achieving the optimal strategy.

Finally, we explored using the top M_T associated markers in simulations with both stratification and causal markers. Results for stratification correction were similar to in our simulations without causal markers (**Supplementary Table 6**), and results for power were similar to in our simulations with no sample structure (**Supplementary Table 7**). The approach from ref. 26 of using the top M_T associated markers selected on the basis of the global maximum of out-of-sample prediction accuracy again yielded the value of M_T that maximized power in these simulations, often at the cost of effective stratification correction. For example, for N = 10,000, M = 100,000 and $P = 0.005$, this approach selected $M_T = 300$, leading to a λ_{median} of 1.26 (compared to λ_{median} of 1.00 if including all markers in the GRM). This result highlights the challenge that efforts to maximize power can compromise effective stratification correction.

In summary, on the basis of methods published so far, we recommend that studies of randomly ascertained quantitative traits in which population stratification is a key concern should generally include all markers (except for the candidate marker and markers in LD with the candidate marker) in the GRM. In contrast, the approach from ref. 26 is expected to perform well when the primary goals are to maximize power and correct for cryptic relatedness, and this method may also prove useful when there is differential confounding of rare variants due to spatially localized stratification[28].

Pitfall: loss of power in ascertained case-control studies

All methods for mixed-model association analysis published so far assume that study samples are randomly ascertained with respect to the phenotype of interest. Although this is usually true for quantitative phenotypes, it is not true for case-control studies, which generally oversample disease cases to increase study power. Recent work has highlighted the loss of power that occurs in ascertained case-control studies when genetic or clinical covariates are modeled as fixed effects without accounting for ascertainment, and a subset of these studies have developed new methods to address this problem[29–33]. However, the issue of power loss due to ascertainment has not previously been investigated for MLMA, which models both known and unknown associated markers as random effects.

We conducted simulations to investigate the use of existing MLMA methods in ascertained case-control studies. We extended our simulations without sample structure to simulate different values of disease prevalence f via the liability threshold model[34] (methods are provided in the **Supplementary Note**). Results for MLMe compared to linear regression showed that, for large N/M ratios and small f values, MLMe could suffer a substantial loss in power (**Table 3**). Similar results were obtained for different values of p (the fraction of non-candidate markers that are causal) and the proportion of variance explained by each candidate marker (**Supplementary Table 8**). We further note that, for large N/M ratios and small f values, the heritability explained by genotyped markers (h^2_g) is misestimated by MLMe, even after accounting for the observed versus liability scale with correction for case-control ascertainment[35]. However, using the correct value of h^2_g does not ameliorate the loss of power (**Supplementary Table 9** and **Supplementary Note**).

In summary, MLMA can suffer a severe loss of power due to case-control ascertainment, motivating further research on MLMA methods in case-control samples. The choice of whether to apply MLMA or other methods should be a function of sample size and the severity of case-control ascertainment.

Table 3 MLMe decreases power compared to linear regression under case-control ascertainment

Number of samples (N)	Number of markers (M)	Disease prevalence (f)	Linear regression	MLMe
10,000	10,000	0.001	3.06 ± 0.15	2.22 ± 0.12
10,000	10,000	0.01	3.04 ± 0.16	2.64 ± 0.14
10,000	10,000	0.1	3.04 ± 0.17	3.06 ± 0.17
10,000	100,000	0.001	2.96 ± 0.16	2.78 ± 0.16
10,000	100,000	0.01	2.66 ± 0.14	2.54 ± 0.13
10,000	100,000	0.1	3.24 ± 0.16	3.26 ± 0.16

We report average $-\log_{10}$ (P values) (± s.e.m.) at causal markers for linear regression and MLMe averaged across 100 simulations with $P = 0.05$ and with each candidate marker explaining $10/N$ of observed-scale variance. Results for different values of N, M, f and the proportion of variance explained by each candidate marker are reported in **Supplementary Table 8**, which also reports the power to detect significant associations at different P-value thresholds.

Table 4 Empirical results in multiple sclerosis and ulcerative colitis data sets

	LR	PCA	MLMi	MLMe[a]	FaST-4K	FaST-Top	FaST-TopX
Multiple sclerosis, 360,557 SNPs (λ_{median})	3.95 (3.86)	1.25 (1.23)	0.99 (0.97)	1.23 (1.20)	1.86 (1.80)	1.42 (1.39)	1.41 (1.39)
Multiple sclerosis, 75 published SNPs	18.50	10.20	8.90	11.30	13.98	10.99	10.56
Ulcerative colitis, 458,560 SNPs (λ_{median})	1.16 (1.16)	1.11 (1.10)	1.00 (0.99)	1.10 (1.09)	1.14 (1.13)	1.08 (1.09)	1.16 (1.15)
Ulcerative colitis, 24 published SNPs	14.06	13.63	12.11	13.43	13.99	10.75	14.09

We report average χ^2 association statistics for all markers (λ_{median} in parentheses) and for published associated markers for each method. The FaST-Top method selected M_T = 2,000 top markers for multiple sclerosis and M_T = 400 top markers for ulcerative colitis, and the FaST-TopX method selected M_T = 2,800 top markers for multiple sclerosis and M_T = 3 top markers for ulcerative colitis. LR, linear regression.

[a] The MLMe analysis was performed using GCTA-LOCO.

Advantages and pitfalls of MLMA in two empirical case-control studies

We investigated the advantages and pitfalls of MLMA in 2 recent GWAS of multiple sclerosis and ulcerative colitis involving over 20,000 samples[23,36]. We chose these studies for several reasons. First, the multiple sclerosis study was the first large GWAS conducted using MLMA methods. Second, the authors of that study recognized that inclusion of candidate markers in the GRM (MLMi) was a potential pitfall, although analyses were conducted using MLMi with available methods and software. Third, owing to the large sample sizes of these studies, these data sets were ideal for exploring the issues highlighted by our simulations.

We analyzed data from 10,204 multiple sclerosis cases and 5,429 controls genotyped on Illumina arrays[23] (methods are provided in the **Supplementary Note**). These subsets of cases and controls were not matched for ancestry (in contrast to in ref. 23) and exhibited substantial population stratification. We retained unmatched samples to maximize the sample size, which we believe is appropriate for these analyses. We also analyzed data from 2,697 ulcerative colitis cases and 5,652 controls genotyped on Affymetrix arrays[36] (methods are provided in the **Supplementary Note**).

We compared seven methods of computing association statistics: linear regression, linear regression with 5 principal-component covariates[27] (PCA), MLMi, MLMe, FaST-LMM using M_R = 4,000 random markers[7] (FaST-4K), FaST-LMM-Select using top M_T markers selected on the basis of the first local minimum for λ_{median}[8] (FaST-Top) and FaST-LMM-Select using top M_T markers selected on the basis of the first local maximum for out-of-sample prediction accuracy using the resulting GRM (FaST-TopX). For each method, we computed average χ^2 association statistics at all markers and at 75 and 24 known associated markers for multiple sclerosis and ulcerative colitis, respectively (methods are provided in the **Supplementary Note**).

We first considered genome-wide average χ^2 values (**Table 4** and **Supplementary Table 10**). For multiple sclerosis, the genome-wide value of 0.994 for MLMi was consistent with our derivations and simulations (**Table 2**), as was the value of 1.232 for MLMe if the effective number of markers was M = 60,000 and the heritability explained by genotyped markers was h^2_g = 0.266 on the liability scale (0.757 on the observed scale[35], assuming disease prevalence of 0.1%), which is a plausible value given that liability-scale h^2_g was estimated at 0.30 ± 0.03 in ref. 37 using independent data. For ulcerative colitis, we observed a genome-wide value of 0.998 for MLMi and 1.100 for MLMe, consistent with h^2_g = 0.244 on the liability scale (0.695 on the observed scale) given the lower sample size. The average χ^2 value from PCA was similar to that from MLMe for both traits. Thus, for both MLMe and PCA, the observed inflation in test statistics is consistent with polygenic effects according to our derivations, simulations and independently obtained estimates of h^2_g. A higher value for PCA than for MLMi does not necessarily imply that PCA failed to correct for population structure (as suggested in ref. 23), as our derivations and simulations showed that correctly calibrated test statistics are expected to have a higher average χ^2 value than in MLMi under a polygenic model. In contrast, FaST-4K, FaST-Top and FaST-TopX generally yielded average χ^2 values that were higher than those derived with PCA and MLMe, consistent with incomplete correction for stratification (**Fig. 2** and **Supplementary Tables 4** and **6**). Although it is theoretically possible that the higher average χ^2 values for these methods could be entirely due to higher average χ^2 values at causal markers, this is unlikely given that the methods attained relatively similar average χ^2 values at known associated markers.

We also tried running FaST-LMM-Select using top M_T markers selected on the basis of the global maximum of out-of-sample prediction accuracy[26]. Our runs failed to complete because the 96-GB memory limit was exceeded. The authors of ref. 26 have reported that running this approach to completion on the same data resulted in every marker being selected for both multiple sclerosis and ulcerative colitis and obtained results identical to MLMe (D. Heckerman and O. Weissbrod, personal communication).

We next considered χ^2 values at known associated markers (**Table 4** and **Supplementary Table 11**). Of the methods attaining complete correction for stratification, MLMe consistently produced higher χ^2 values than MLMi for both multiple sclerosis (70 of 75 markers; $P = 1 \times 10^{-15}$) and ulcerative colitis (24 of 24 markers; $P = 1 \times 10^{-7}$), consistent with the simulations (**Fig. 1**). MLMe also produced higher χ^2 values than FaST-Top for ulcerative colitis (18 of 24 markers; P = 0.02); this was the only instance in which FaST-4K, FaST-Top or FaST-TopX attained complete correction for stratification. However, comparison of MLMe and PCA was inconclusive, with MLMe producing higher values for multiple sclerosis (43 of 75 markers; P = 0.25) and PCA producing higher values for ulcerative colitis (13 of 24 markers; P = 0.84). Owing to the weaker correlation of these methods, these comparisons were noisy, and analyses of additional data sets will be needed to conclusively distinguish the performance of MLMe and PCA on empirical data. We note that the much lower χ^2 values for MLMi compared to PCA at known associated markers was attributed by ref. 23 to structure that was not captured by PCA. However, the pitfalls of MLMi (**Fig. 1**) provide an alternative explanation.

Recommendations and future directions

MLMA methods can prevent false positive associations and increase power at reasonable computational cost. However, our theoretical derivations, simulations and application to empirical data identify potential pitfalls such as inclusion of the candidate marker in the GRM, use of a small subset of markers in the GRM and effects from case-control ascertainment.

We recommend excluding candidate markers from the GRM (MLMe) in preference to including them (MLMi). This approach can be efficiently implemented via a leave-one-chromosome-out analysis[7] implemented in GCTA software (GCTA-LOCO; methods are provided in the

Supplementary Note). An efficient implementation is also provided in FaST-LMM software[7,8]. Our analytical derivations demonstrate the advantages of MLMe relative to MLMi and also quantify the expected inflation in MLMe test statistics in the absence of confounding, potentially eliminating the need to apply an additional round of genomic control[38] correction as in many recent studies[39]. However, distinguishing between polygenic effects and incomplete correction for stratification is an important direction of future research (B. Bulik-Sullivan, N. Patterson, A.L.P., M. Daly and B. Neale, unpublished data).

We recommend that studies of randomly ascertained quantitative traits should generally include all markers (except for the candidate marker and markers in LD with the candidate marker) in the GRM, except as follows. First, the set of markers included in the GRM can be pruned by LD to reduce running time (with association statistics still computed for all markers). Second, genome-wide significant markers of large effect should be conditioned out as fixed effects[4,9]. Third, when population stratification is less of a concern, we recommend the approach from ref. 26 of using the top M_T associated markers selected on the basis of the global maximum from out-of-sample prediction accuracy. (This approach may choose either a subset of markers ($M_\mathrm{T} < M$) or all markers ($M_\mathrm{T} = M$), but computational constraints may preclude the latter choice.) Finally, a potentially appealing way to capture the power advantages of selecting a subset of SNPs to include in the GRM while addressing concerns about stratification is to employ FaST-LMM-Select with principal components (G. Tucker, A.L.P. and B. Berger, unpublished data and D. Heckerman, C. Lippert, J. Listgarten and O. Weissbrod, personal communication).

Ascertained case-control studies present a special challenge due to the potential loss in power with standard MLMA methods. When sample size is small or disease prevalence is high, standard MLMA methods can be used (**Table 3**). Otherwise, in data sets with no relatedness structure, PCA can be used[27]. (In this case, conditioning on genome-wide significant markers or other covariates of large effect can be either omitted[32] or retained using methods that explicitly model case-control ascertainment to increase power[29–31,33].) In ascertained case-control data sets with relatedness structure, we know of no good alternative to MLMA.

We conclude by highlighting three areas in mixed-model association analysis in which there is a pressing need for the development of new methods. First, there is a need for MLMA methods for ascertained control traits that do not suffer a loss of power. Second, there is a need for MLMA methods that use mixture distributions of prior effect sizes to increase their power, mirroring advances in phenotypic prediction for livestock and human traits using Bayesian methods[40–42]. Third, further work is needed to develop and assess methods for rare variants, which pose a greater challenge for all methods[20,28].

URLs. GCTA software (GCTA-MLMi and GCTA-LOCO), http://ctgg.qbi.uq.edu.au/software/gcta/mlmassoc.html; EMMAX software, http://genetics.cs.ucla.edu/emmax/; Fast-LMM software, http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/Fastlmm/; GEMMA software, http://stephenslab.uchicago.edu/software.html; GRAMMAR-Gamma, in GenABEL package of the GenABEL project (also see MixABEL package of GenABEL project), http://www.genabel.org/; National Human Genome Research Institute GWAS catalog, http://www.genome.gov/gwastudies/. Wellcome Trust Case Control Consortium 2 data were downloaded from http://www.wtccc.org.uk/ccc2/.

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

ACKNOWLEDGMENTS
We are grateful to N. Patterson, D. Heckerman, J. Listgarten, C. Lippert, E. Eskin, B. Vilhjalmsson, P. Loh, T. Hayeck, T. Frayling, A. McRae, L. Ronnegard, O. Weissbrod, G. Tucker and the GIANT Consortium for helpful discussions and to A. Gusev and S. Pollack for assistance with the multiple sclerosis and ulcerative colitis data sets. We are grateful to two anonymous referees for their helpful comments. This study makes use of data generated by the Wellcome Trust Case Control Consortium and data from the database of Genotypes and Phenotypes (dbGaP) under accessions phs000090.v2.p1 and phs000091.v2.p1 (see the **Supplementary Note** for the full set of acknowledgments for these data). This research was supported by US National Institutes of Health (NIH) grants R01 HG006399, P01 GM099568 and R01 GM075091, by the Australian Research Council (DP130102666) and by the Australian National Health and Medical Research Council (APP1011506 and APP1052684).

AUTHOR CONTRIBUTIONS
All authors conceived the project and designed the analyses. J.Y., N.A.Z. and A.L.P. performed the analyses. J.Y., M.E.G. and P.M.V. provided the theoretical derivations. J.Y. wrote the GCTA software. J.Y., N.A.Z. and A.L.P. wrote the manuscript with edits from all authors.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

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. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
3. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
4. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
5. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
6. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
7. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
8. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).
9. 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).
10. 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).
11. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
12. 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).
13. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
14. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Hum. Genet.* **131**, 1655–1664 (2012).
15. Henderson, C.R. Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447 (1975).
16. de los Campos, G., Gianola, D. & Allison, D.B. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* **11**, 880–886 (2010).
17. Sul, J.H. & Eskin, E. Mixed models can correct for population structure for genomic regions under selection. *Nat. Rev. Genet.* **14**, 300 (2013).
18. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. Response to Sul and Eskin. *Nat. Rev. Genet.* **14**, 300 (2013).
19. Wang, K., Hu, X. & Peng, Y. An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. *Hum. Hered.* **76**, 1–9 (2013).
20. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
21. Chen, W.M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
22. Yang, J. *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
23. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
24. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
25. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).

26. Lippert, C. *et al*. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep.* **3**, 1815 (2013).
27. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
28. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat. Genet.* **45**, 470–471 (2013).
29. Mefford, J. & Witte, J.S. The Covariate's Dilemma. *PLoS Genet.* **8**, e1003096 (2012).
30. Zaitlen, N. *et al.* Analysis of case-control association studies with known risk variants. *Bioinformatics* **28**, 1729–1737 (2012).
31. Clayton, D. Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation. *Genet. Epidemiol.* **36**, 409–418 (2012).
32. Pirinen, M., Donnelly, P. & Spencer, C.C. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat. Genet.* **44**, 848–851 (2012).
33. Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet.* **8**, e1003032 (2012).
34. Falconer, D.S. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann. Hum. Genet.* **31**, 1–20 (1967).
35. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
36. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
37. Lee, S.H. *et al.* Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum. Mol. Genet.* **22**, 832–841 (2013).
38. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
39. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
40. 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).
41. Erbe, M. *et al.* Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* **95**, 4114–4129 (2012).
42. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).