nature genetics

# An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations

Vincent Segura[1,2,4], Bjarni J Vilhjálmsson[1,3,4], Alexander Platt[1,3], Arthur Korte[1], Ümit Seren[1], Quan Long[1] & Magnus Nordborg[1,3]



**Population structure causes genome-wide linkage disequilibrium between unlinked loci, leading to statistical confounding in genome-wide association studies. Mixed models have been shown to handle the confounding effects of a diffuse background of large numbers of loci of small effect well, but they do not always account for loci of larger effect. Here we propose a multi-locus mixed model as a general method for mapping complex traits in structured populations. Simulations suggest that our method outperforms existing methods in terms of power as well as false discovery rate. We apply our method to human and *Arabidopsis thaliana* data, identifying new associations and evidence for allelic heterogeneity. We also show how a priori knowledge from an *A. thaliana* linkage mapping study can be integrated into our method using a Bayesian approach. Our implementation is computationally efficient, making the analysis of large data sets ($n > 10{,}000$) practicable.**

With the increasing availability of genomic polymorphism data, genome-wide association studies (GWAS) are becoming the default method for investigating the genetics of quantitative traits. Typically, GWAS are carried out using single-locus tests to identify associations between polymorphisms and traits in either case-control populations or cohorts. However, both study designs are subject to confounding by population structure, leading to an inflation of test statistics and a high false positive rate[1,2]. Several methods have been proposed to address this issue, including genomic control[3], structured association[4], principal-components analysis[5] and mixed linear models[6]. Genomic control scales the test statistics uniformly, so that the observed median test statistic equals the expected one. Even though this approach reduces the inflation of test statistics globally, it does not change the rank of the polymorphisms, as they are subject to the same correction. In the structured association and principal-component analysis approaches, population structure is taken into account by including covariates in the association model that represent the cluster memberships and principal-component loadings of the individuals, respectively. Whereas these approaches are expected to perform well when the population structure is simple, they may perform poorly when the structure is more complex: for example, when individuals show a continuum of relatedness[7]. An additional improvement has been made with the use of mixed linear models, which are based on the insight that confounding can be caused by the genetic background of causal variants in the presence of population structure. The mixed model controls for the genetic background through a random polygenic term with a covariance structure described by a relationship matrix, so that correlations in phenotype mirror relatedness[8], as predicted by Fisher's classical model[9]. This approach has been shown to perform well in plants, animals and humans[6,10–12], and methods have been developed to allow the analysis of large GWAS data sets in a reasonable amount of time[11,13,14].

All these approaches are based on single-locus tests combined with some kind of diffuse genomic background. However, for complex traits controlled by several large-effect loci, these approaches may not be appropriate, especially in the presence of population structure[12] (indeed, a substantial inflation of single-locus test statistics is expected for complex traits, even in the absence of population structure)[15]. Explicit use of multiple cofactors in the statistical model is an obvious alternative and is indeed standard in traditional linkage mapping, where both multiple–quantitative trait locus (QTL) mapping and composite interval mapping have been shown to outperform simple interval mapping[16,17]. In GWAS, the case for including multiple loci is arguably even stronger, as the confounding effects of background loci may be present across the genome (due to linkage disequilibrium) rather than only locally (due to linkage)[18]. Thus, whereas conditioning on known causative factors in GWAS has typically been conducted on a local scale to help identify multiple alleles and clarify complex associations[12,19,20], we believe that it should be done on a genome-wide basis. As shown, conditional analysis on a genome-wide scale may well lead to higher power and a lower false discovery rate (FDR) than single-locus approaches (**Fig. 1**). Similarly, in the context of human genetics,

[1]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna, Austria. [2]Institut National de la Recherche Agronomique (INRA), UR0588, Orléans, France. [3]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to M.N. (magnus.nordborg@gmi.oeaw.ac.at).

Received 16 November 2011; accepted 4 May 2012; published online 17 June 2012; doi:10.1038/ng.2314

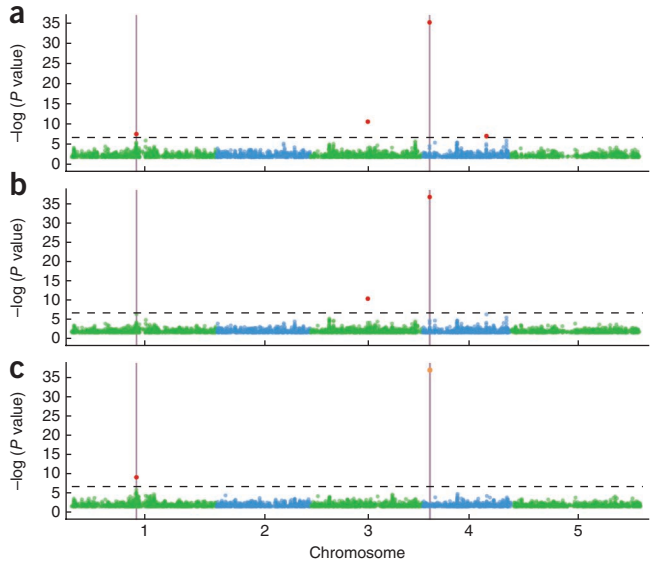


**Figure 1** A GWAS for a simulated trait with two causal SNPs randomly chosen from a real *A. thaliana* SNP data set. Random error was added to the trait to fix the heritability at 25%. Causal SNPs are marked by vertical lines. (**a**) A single-SNP linear regression scan detects four significantly associated SNPs (red circles) at a Bonferroni-corrected threshold of 0.05 (dashed horizontal line). Half of these SNPs are false positives, and the other half are true positives, leading to FDR of 50% and power of 100%. (**b**) A single-SNP mixed-model[11,14] scan eliminates one false positive but also one true positive, leading to similar FDR (50%) and decreased power of 50% compared to the model in **a**. (**c**) Adding the most significant SNP as a cofactor to the mixed model (orange circle) recovers the second causal SNP, while eliminating the last false positive, leading to the perfect scenario with FDR of 0% and power of 100%.

it has been suggested that conditioning on major-effects loci, like the major histocompatibility (MHC) region, may improve power[11].

However, automatically including cofactors is challenging when the number of predictors is large compared to the number of observations. This is particularly problematic in GWAS, where the number of polymorphisms ($p$) can reach millions but where the number of phenotyped and genotyped individuals ($n$) is rarely more than tens of thousands. Such 'large $p$, small $n$' problems are very challenging: the model space is usually too large to explore exhaustively, and the maximum number of polymorphisms that can be fitted at a time must be less than the number of individuals. In addition, identifying causative polymorphisms by fitting more than one polymorphism at a time is complicated by the presence of linkage disequilibrium. Several approaches have been proposed to address these issues, including stepwise regression[21] and penalized regression with different penalty functions, such as ridge regression, normal exponential gamma, elastic net and LASSO[22–26]. These approaches have been shown to perform better than single-locus approaches, but most are either computationally unfeasible in GWAS[27] or do not explicitly address the problems posed by population structure. As an alternative, we propose using a simple, stepwise mixed-model regression with forward inclusion and backward elimination, which, despite being limited in terms of exploring the model space, has the advantage of being computationally efficient and therefore applicable to GWAS. To effectively address the population structure issue, we make use of an approximate version of the mixed model[11,14] in which we re-estimate genetic and error variances at each step of the regression (Online Methods). As the variance attributed to the random polygenic term decreases when cofactors are added to the model, we propose to use the heritable variance estimate as a criterion to stop forward inclusion. Then, backward elimination is performed from the last forward model for a more thorough exploration of the model space. We evaluate various model selection criteria through simulations, which suggest that the proposed multi-locus mixed-model (MLMM) method performs well in terms of FDR and power. Finally, we show the usefulness of our approach by applying it to human and *A. thaliana* data.

## RESULTS

### Simulations

GWAS data were simulated by adding phenotypic effects to real genotypic data from *A. thaliana*[28] under two different scenarios: a 2-locus model and a 100-locus model. For the latter, additivity was assumed, whereas, for the former, different types of interactions were explored (Online Methods).

We compared our proposed MLMM method with three other mapping methods: a single-locus approximate mixed model that corrects for population structure but does not take into account other major loci (MM)[11,14]; a stepwise linear model that takes other major loci into account but does not correct for population structure (SWLM); and a single-locus linear model that does neither (LM). The four methods were compared in terms of their statistical power and FDR. For single-locus methods, SNPs were considered to be detected if their $P$ values were below a defined threshold, whereas, for the multi-locus methods, detected SNPs were those belonging to the most complex model in which the marginal $P$ values of cofactors were all below a defined threshold.

The results for the 100-locus model are shown (**Fig. 2** and **Supplementary Figs. 1–4**), and can be summarized as follows. First, methods that use a kinship term to correct for population structure

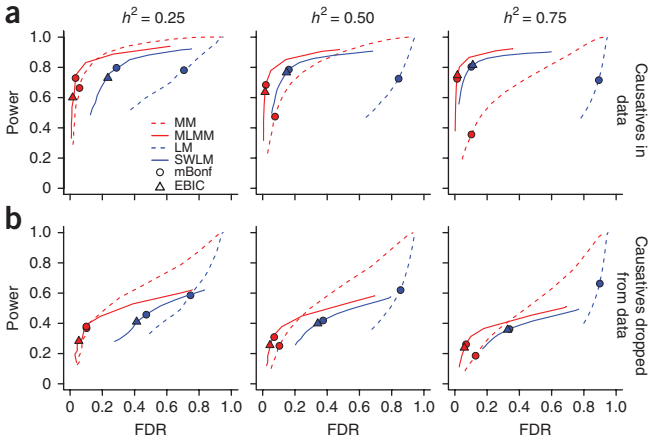


**Figure 2** Power and FDR in 100-locus model simulations for four different mapping methods: LM, SWLM, MM and MLMM. (**a**,**b**) For the purpose of computing power and FDR, a causal SNP was considered to be detected if a SNP within 25 kb on either side was determined to have a significant association (results for other window sizes are given in **Supplementary Fig. 3**), and only causal SNPs that were detectable in principle (that were marginally significant at a Bonferroni-corrected threshold of 0.05 in a simple linear model) were considered. For clarity, only the backward path of the multi-locus methods (SWLM and MLMM) is shown (comparison between forward and backward paths is given in **Supplementary Fig. 4**). Circles and triangles represent the best-fitting model according to the mBonf and EBIC model selection criteria, respectively. Power and FDR were estimated with (**a**) and without (**b**) the causal loci included. Three phenotypic heritabilities were used in the simulations: 0.25 (left), 0.5 (middle) and 0.75 (right).

always outperform comparable methods that do not (MM and MLMM versus LM and SWLM, respectively). There is simply too much structure in these data for it to be ignored without paying a very heavy price in terms of increased FDR (**Supplementary Fig. 1**). Second, multi-locus methods generally outperform comparable single-locus methods (SWLM and MLMM versus LM and MM, respectively), as long as the causative sites are included in the data (**Fig. 2a**). The advantage increases with increasing heritability, because, under our simulation scheme, increased heritability implies more loci of large effect and, hence, greater confounding (**Supplementary Figs. 1** and **2**). If the causative sites themselves are excluded from the data, the single-locus mixed model (MM) may have greater power than the multi-locus version (MLMM) but only at the cost of greatly increased FDR (**Fig. 2b**).

The two-locus simulations allowed us to examine the advantages of including cofactors in the mixed model under several scenarios of population structure and/or epistasis (Online Methods). Regardless of the scenario considered, MLMM consistently performed at least as well as the other methods when restricted to a small FDR (**Fig. 1** and **Supplementary Fig. 5**). When two causal sites were chosen at random, the improvement in power observed for MLMM over that in the single-marker MM was almost entirely attributed to increased power to detect the second causal site (**Supplementary Fig. 6).**

A serious problem when employing multi-locus models is knowing how many loci to include. We propose two model selection criteria: the extended Bayesian information criterion (EBIC)[29] and the multiple-Bonferroni criterion (mBonf), defined as the largest model in which all cofactors have a *P* value below a Bonferroni-corrected threshold (we used a threshold of 0.05). Our simulations showed that both criteria are consistent in bounding the FDR for the MLMM method, regardless of the simulation scenario, with EBIC being slightly more stringent than mBonf (**Fig. 2** and **Supplementary Fig. 5**). In addition, the genome-wide *P* values in the models selected by both criteria were uniformly distributed, showing the ability of mixed models to control confounding by population structure in a multi-locus setting (**Supplementary Fig. 1**). Furthermore, both criteria performed appropriately in extreme scenarios where there was no detectable signal in the data, as might occur when an external confounding variable interacts nonlinearly with a single causal locus[18]. In this case, MLMM with one of the proposed criteria correctly selected a model without any SNPs, whereas the other methods tested would identify only false positives (**Supplementary Fig. 5**). In summary, MLMM with the conservative FDR provided by the proposed model selection criteria consistently outperformed the other methods in all scenarios that we examined.

For completeness, we also compared MLMM to other single-locus mixed-model implementations, including the exact mixed model[30] and the approximate mixed model with compression[14], as they have been shown to perform better than the approximate method. These methods did indeed perform slightly better than the approximate method in our simulations but were still far from the performance achieved by MLMM (**Supplementary Fig. 7**).

**Table 1 SNPs identified in multi-locus mixed-model analysis of NFBC1966 traits**

| | | | | *P* value | | Previous identification | |
|---|---|---|---|---|---|---|---|
| SNP | Chr. | Position | Gene | EBIC | mBonf | Sabatti *et al.* | Kang *et al.* |
| Associated with triglyceride levels (mM) | | | | | | | |
| rs673548 | 2 | 21091049 | *APOB* | | $5.1 \times 10^{-8}$ | + | + |
| rs1260326 | 2 | 27584444 | *GCKR* | $1.5 \times 10^{-10}$ | $7.9 \times 10^{-11}$ | + | + |
| rs10096633 | 8 | 19875201 | *LPL* | $1.6 \times 10^{-8}$ | $2.4 \times 10^{-8}$ | + | + |
| Associated with HDL levels (mM) | | | | | | | |
| rs1532085 | 15 | 56470658 | *LIPC* | $9.2 \times 10^{-12}$ | $8.0 \times 10^{-12}$ | + | + |
| rs3764261 | 16 | 55550825 | *CETP* | $2.7 \times 10^{-32}$ | $3.7 \times 10^{-23}$ | + | + |
| rs7499892 | 16 | 55564091 | *CETP* | | $9.5 \times 10^{-8}$ | – | – |
| rs255049 | 16 | 66570972 | *LCAT* | $1.3 \times 10^{-8}$ | $4.8 \times 10^{-8}$ | + | + |
| rs1800961 | 20 | 42475778 | *HNF4A* | | $1.5 \times 10^{-7}$ | – | – |
| Associated with LDL levels (mM) | | | | | | | |
| rs646776 | 1 | 109620053 | *CELSR2* | $4.2 \times 10^{-16}$ | $4.2 \times 10^{-16}$ | + | + |
| rs693 | 2 | 21085700 | *APOB* | $7.1 \times 10^{-12}$ | $7.1 \times 10^{-12}$ | + | + |
| rs11668477 | 19 | 11056030 | *LDLR* | $1.0 \times 10^{-9}$ | $1.0 \times 10^{-9}$ | + | + |
| rs157580 | 19 | 50087106 | *TOMM40-APOE* | $2.2 \times 10^{-17}$ | $2.2 \times 10^{-17}$ | + | – |
| rs405509 | 19 | 50100676 | *TOMM40-APOE* | $1.3 \times 10^{-12}$ | $1.3 \times 10^{-12}$ | – | – |
| Associated with CRP levels (mM) | | | | | | | |
| rs2369146 | 1 | 157934819 | *CRP* | $4.5 \times 10^{-9}$ | $2.8 \times 10^{-9}$ | – | – |
| rs2794520 | 1 | 157945440 | *CRP* | $1.1 \times 10^{-29}$ | $6.6 \times 10^{-30}$ | + | + |
| rs2650000 | 12 | 119873345 | *HNF1A* | $1.3 \times 10^{-12}$ | $1.0 \times 10^{-12}$ | + | + |
| rs8106922 | 19 | 50093506 | *TOMM40-APOE* | | $1.6 \times 10^{-12}$ | – | – |
| rs439401 | 19 | 50106291 | *TOMM40-APOE* | | $2.2 \times 10^{-9}$ | – | – |
| Associated with glucose levels (mM) | | | | | | | |
| rs560887 | 2 | 169471394 | *G6PC2* | $2.2 \times 10^{-13}$ | $2.2 \times 10^{-13}$ | + | + |
| rs2971671 | 7 | 44177862 | *GCK* | $3.2 \times 10^{-9}$ | $3.2 \times 10^{-9}$ | – | + |
| rs3847554 | 11 | 92308474 | *MTNR1B* | $4.7 \times 10^{-11}$ | $4.7 \times 10^{-11}$ | –[a] | – |
| Associated with SBP | | | | | | | |
| rs782602 | 2 | 55702813 | *SMEK2* | | $1.4 \times 10^{-7}$ | – | –[b] |

Models were selected using either EBIC or mBonf. Chr., chromosome.
[a]This SNP was not reported by Sabatti *et al.*, but they reported two other SNPs located in the same gene. [b]Kang *et al.* did not report this association because they used a *P*-value threshold slightly more stringent than the Bonferroni-corrected threshold of 0.05 used for mBonf.

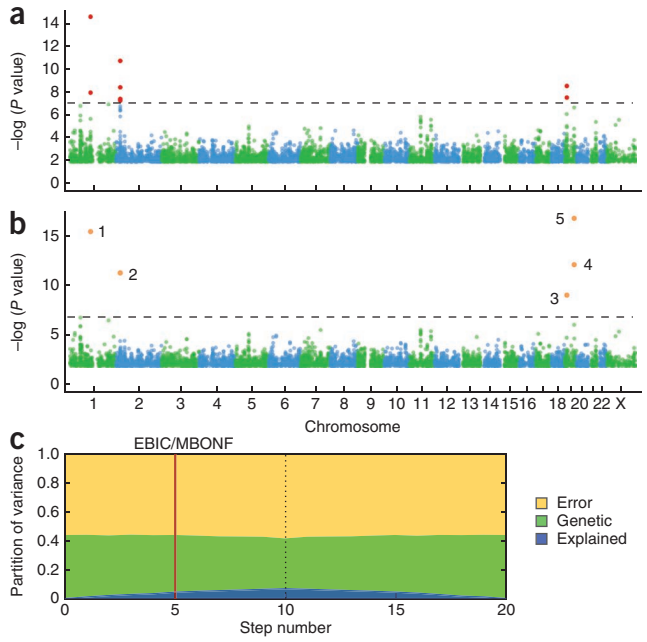


**Figure 3** GWAS for LDL levels in the NFBC1966 data set. (**a**) A single-locus mixed model identifies seven SNPs in three genes (red circles) at a Bonferroni-corrected threshold of 0.05 (dashed horizontal line). (**b**) MLMM identifies five SNPs in four genes (orange circles numbered in the order in which they were included in the model). (**c**) Partition of variance at each step of MLMM (ten forward and ten backward) into variance explained by the SNPs included in the model, kinship and noise.

## Application to a human data set

To show the feasibility as well as the usefulness of MLMM, we applied it to a previously published data set of metabolic traits in the Northern Finland Birth Cohort (NFBC1966)[31]. The data were previously reanalyzed to show the usefulness of the mixed model[11], and we used the same settings for mixed-model estimation here. The SNPs identified using MLMM are listed in **Table 1**. As predicted by our simulations, EBIC was more stringent than mBonf, resulting in the selection of models that were either similar to or nested within the models selected by mBonf. Using the less-conservative mBonf criterion, we identified all the associations previously detected with the single-locus mixed model[11] and nine additional associations. Of the newly identified association signals, three were located near genes previously reported using the same data[31] (two in the *TOMM40-APOE* cluster for low-density lipoprotein (LDL) levels and one in *MTNR1B* for glucose levels), and four were located in gene regions not previously reported with this data set (one in *HNF4A* for high-density lipoprotein (HDL) levels, one in *SMEK2* for systolic blood pressure (SBP) and two in the *TOMM40-APOE* cluster for C-reactive protein (CRP) levels). The remaining two associations were additional SNPs in genes that had already been reported (*CETP* for HDL and *CRP* for CRP levels). The detected association in *HNF4A* for HDL levels (rs1800961) in a gene region not previously reported with this data set has been replicated in two meta-analyses of 30,714 and 99,900 individuals each[32,33].

Multiple SNPs with significant association signals within or near a single gene suggest either allelic heterogeneity or the presence of an untyped causal variant that is partially represented by multiple SNPs (or both). In the case of the associations located in the *TOMM40-APOE* cluster (for both LDL and CRP levels), we observed a marked decrease in the *P* values for the two selected SNPs when they were both included in the model (**Fig. 3** and **Supplementary Fig. 8**), which presumably explains why they were not identified using the single-locus mixed model. This type of situation is expected when loci mask each other, for example, when alleles of compensatory effect are correlated, as seems to be the case here ($R^2 = 0.33$ and 0.25 for LDL and CRP levels, respectively).

We show the percentage of variance explained by the SNPs included in the model and the percentages of unexplained genetic and residual variance at the different steps of the MLMM for LDL levels (**Fig. 3** and **Supplementary Fig. 9**). It is notable that most of the heritable phenotypic variation remains unexplained.

## Application to an *A. thaliana* data set

Sodium accumulation in the leaves of *A. thaliana* has been shown to be strongly associated with genotype and expression levels of the $Na^+$ transporter *AtHKT1;1* (ref. 34). In particular, a SNP located in the first exon of the gene (chromosome 4: 6,392,280) shows a highly significant association (*P* value = $6.33 \times 10^{-14}$ using an approximate mixed model). We reanalyzed these data using MLMM and found that the sole SNP previously reported[34] only explains part of the signal in the associated region (**Fig. 4**).

Instead, the optimal model obtained with MLMM (according to both EBIC and mBonf) included three SNPs, which together explained 42.3% of the phenotypic variation. This model included the previously reported SNP, which explained 27.7% of the variation, and a second SNP only 22 kb away from the gene, suggesting that there might be multiple causal variants in the gene. To further investigate the associations in this particular region, we applied our method locally, using only the 508 SNPs located within 100 kb of the gene. Using EBIC,

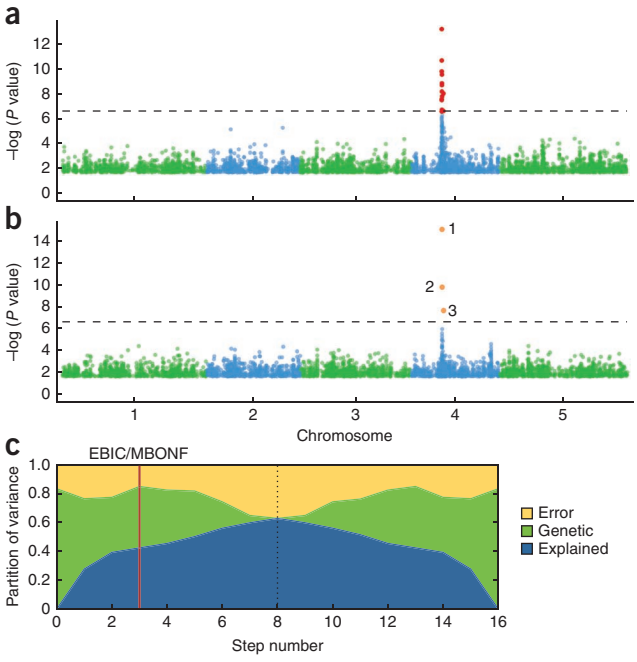


**Figure 4** GWAS for sodium accumulation in *A. thaliana*. (**a**) A single-locus mixed model identifies a strong peak of significantly associated SNPs on chromosome 4 (red circles) at a Bonferroni-corrected threshold of 0.05 (dashed horizontal line). (**b**) MLMM identifies three SNPs (orange circles numbered in the order in which they were included in the model). (**c**) Partition of variance at each step of MLMM (eight forward and eight backward) into variance explained by the SNPs included in the model, kinship and noise.

six SNPs were included in the model, all within 25 kb of *AtHKT1;1*, which explained 52.6% of the phenotypic variation (**Supplementary Fig. 10**), leaving 20.5% of the heritable fraction of the total variance unexplained. As noted, this suggests either allelic heterogeneity or the presence of one or more untyped causal variants. However, as the largest possible fraction of variance explained by a single binary SNP (which would have a minor allele frequency of 0.32) is 47.6%, we conclude that there is evidence for allelic heterogeneity in this case.

## DISCUSSION

The problem of population structure in GWAS is best viewed as one of model mis-specification. Single-locus tests of association are the wrong model to use in cases where the trait is not attributable to a single locus. Ignoring the genetic background may be defensible in some circumstances but is clearly not when causative alleles are correlated across loci due to population structure and/or selection[12], resulting in biased estimates of effect sizes. The problem has long been recognized by animal breeders, who developed a mixed linear model to reduce bias[8]. This approach works well but assumes that the phenotypic covariance between individuals can be predicted by their relatedness, as estimated by genotypes at SNPs across the genome. As demonstrated by Fisher close to 100 years ago[9], this approximation is reasonable if the genetic background is sufficiently smooth, but it is clear that loci of relatively large effect may make this approach invalid[18]. We therefore propose to extend the mixed model for GWAS to include multiple loci, in parallel to what is routinely done in QTL linkage mapping[16,17].

Our proposed method includes significant effects in the model via a forward-backward stepwise approach, while re-estimating the variance components of the model at each step. If the fixed effects included are real, they can reduce the unexplained heritable variance and effectively lower the restraints posed by the mixed model on other markers that correlate with population structure. As demonstrated by simulations, our MLMM implementation shows promising performance in terms of power and FDR in comparison with a single-marker scan and a stepwise linear regression, especially when applying a conservative threshold, which can be achieved with one of the proposed model quality criteria. In particular, MLMM performed much better than the other methods tested for structured samples and traits involving several loci with moderate to large effect.

Applying MLMM to real data from humans and *A. thaliana*, we identified interesting new associations as well as evidence for allelic heterogeneity. Indeed, as it includes multiple loci in the model, MLMM helps identify evidence for allelic heterogeneity in addition to interactions, although it is difficult to exclude the possibility that multiple associated SNPs within a region are detected because of partial linkage disequilibrium with an untyped causal variant[12,18,20]. However, with the rapid development of DNA sequencing[35], it is increasingly likely that causal variants will be typed. As seen in our simulations, all tested methods, especially MLMM, will benefit greatly from this. Applied here to the analysis of quantitative traits, MLMM can also be applied to the study of disease heritability. Indeed, it is possible to analyze a disease phenotype with an approximate mixed model by considering a binary quantitative response corresponding to case-control status[11]. MLMM partitions the phenotypic variance into genetic, random and explained variance at each step, suggesting a natural stopping criterion (genetic variance of 0) for including cofactors. This allows the user to obtain estimates of the explained and unexplained heritable variance, as well as gives insights into trait architecture.

MLMM is far from a panacea, however. The greedy forward-backward inclusion of SNPs is clearly limited in exploring the huge model space.

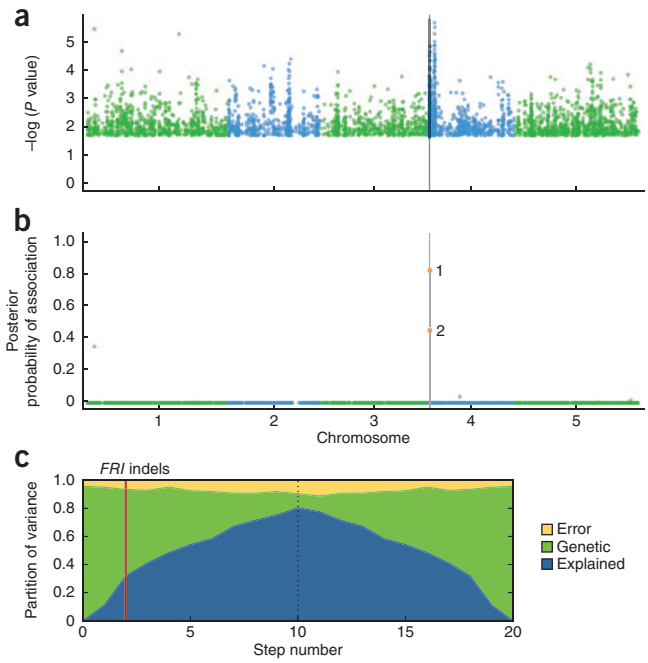


**Figure 5** An example of Bayesian MLMM for the analysis of *FLC* expression in *A. thaliana*. (**a**) An approximate mixed-model scan for *FLC* expression, with the *FRIGIDA* gene marked by a vertical line. (**b**) The posterior probability of association scan after the Bayesian MLMM has included two loci in the model (orange circles), which incidentally are the two causative indels previously identified. (**c**) Partition of phenotypic variance for each forward inclusion (ten steps) and backward elimination (ten steps after the dashed line). The vertical red line marks the model with the two causative indels included in the model.

More sophisticated algorithms, like LASSO[36], are worth exploring. However, similar to other penalized methods, LASSO typically assumes independence between markers, which would not be appropriate for structured data. In the context of structured data, LASSO might give a large effect size to a marker that is in linkage disequilibrium with many other markers, whereas a mixed model would down-weight such markers. A potential improvement on this would be to use LASSO in conjunction with a mixed model[26]. Although this approach is potentially very promising, it is currently too computationally demanding for GWAS data sets. Another promising approach is resample model averaging[37], which has been applied successfully to joint linkage association analysis[38]. However, it is important to realize that the problem is fundamentally very difficult. For example, we have previously shown that linkage disequilibrium between two known causal alleles of the *A. thaliana* flowering locus *FRIGIDA* (*FRI*) and the genomic background give rise to a very complicated pattern of association in a GWAS of *FLOWERING LOCUS C* (*FLC*) expression[12]. None of the methods tested here identified the causal sites. This is not unexpected, as there are many spurious one- and two-locus models that fit the data better than those involving the true causal loci. In cases like this, we think it is unlikely that progress will be made without independent data to help prioritize variants. As MLMM is based on a linear model, it can easily be extended for Bayesian analysis[39,40] and allows for the integration of previous knowledge into the model. Indeed, returning to the *FLC* example, by placing a 100-fold prior on all markers within 10 kb of *FRI*, we allow MLMM to include the two known causal variations as the first two cofactors in the model, showing how priors can help identify causal loci and improving the model (**Fig. 5**).

**URLs.** INRA MIGALE platform, http://migale.jouy.inra.fr/; R version of MLMM, https://cynin.gmi.oeaw.ac.at/home/resources/mlmm; Python version of MLMM, https://github.com/bvilhjal/mixmogam; Scientific Tools for Python (SciPy) package, http://www.scipy.org/.

## METHODS

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

*Note: Supplementary information is available in the online version of the paper.*

### ACKNOWLEDGMENTS

We acknowledge the NFBC1966 Study investigators for allowing us to use their phenotype and genotype data in our study. The NFBC1966 Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the Broad Institute, the University of California, Los Angeles (UCLA), the University of Oulu and the National Institute for Health and Welfare in Finland. This manuscript was not prepared in collaboration with the investigators from the NFBC1966 Study and does not necessarily reflect the opinions or views of these investigators or those at the collaborating institutes. We thank N.B. Freimer and S.K. Service for their help in pre-processing the NFBC1966 data. We would also like to thank P. Forai for excellent information technology and cluster support at GMI, the INRA MIGALE bioinformatics platform for additional computational resources and D.V. Conti, D.J. Balding and S. Srivastava for useful discussions on the topic. Finally, we would like to thank the anonymous reviewers for their helpful comments on the manuscript. This work was supported by grants from the Ecologie des Forêts, Prairies et milieux Aquatiques (EFPA) department of INRA to V.S. and Deutsche Forschungsgemeinschaft (DFG) to A.K. and by grants from the US National Institutes of Health (P50 HG002790) and the European Union Framework Programme 7 (TransPLANT, grant agreement 283496) to M.N., as well as by the Austrian Academy of Sciences through GMI.

### AUTHOR CONTRIBUTIONS

All authors contributed to designing the study. V.S. and B.J.V. ran the simulations and analyzed the data. V.S., B.J.V. and M.N. wrote the manuscript with input from A.P., A.K., Ü.S. and Q.L.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2314.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.



1. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
2. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
3. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
4. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
5. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
6. 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).
7. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
8. Henderson, C.R. *Application of Linear Models in Animal Breeding* (University of Guelph, Guelph, Canada, 1984).
9. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
10. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
11. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
12. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
13. 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).
14. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
15. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
16. Jansen, R.C. Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211 (1993).
17. Zeng, Z.B. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468 (1994).
18. Platt, A., Vilhjalmsson, B.J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045–1052 (2010).
19. Allen, A.S., Satten, G.A., Bray, S.L., Dudbridge, F. & Epstein, M.P. Fast and robust association tests for untyped SNPs in case-control studies. *Hum. Hered.* **70**, 167–176 (2010).
20. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
21. Cordell, H.J. & Clayton, D.G. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am. J. Hum. Genet.* **70**, 124–141 (2002).
22. Hoggart, C.J., Whittaker, J.C., De Iorio, M. & Balding, D.J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
23. Malo, N., Libiger, O. & Schork, N.J. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am. J. Hum. Genet.* **82**, 375–385 (2008).
24. Croiseau, P. & Cordell, H.J. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. *BMC Proc.* **3**, S61 (2009).
25. Cho, S. *et al.* Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann. Hum. Genet.* **74**, 416–428 (2010).
26. Wang, D., Eskridge, K.M. & Crossa, J. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *J. Agric. Biol. Environ. Stat.* **16**, 170–184 (2011).
27. Ayers, K.L. & Cordell, H.J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* **34**, 879–891 (2010).
28. Horton, M.W. *et al.* Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
29. Chen, J.H. & Chen, Z.H. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771 (2008).
30. Astle, W. & Balding, D.J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
31. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (2009).
32. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
33. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
34. Baxter, I. *et al.* A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter *AtHKT1;1*. *PLoS Genet.* **6**, e1001193 (2010).
35. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
36. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc., B* **58**, 267–288 (1996).
37. Valdar, W., Holmes, C.C., Mott, R. & Flint, J. Mapping in structured populations by resample model averaging. *Genetics* **182**, 1263–1277 (2009).
38. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
39. Stephens, M. & Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
40. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).

## ONLINE METHODS

**Data.** Both *A. thaliana* and human data were used for the examples. The genotype data for *A. thaliana* included 1,307 individual plants genotyped at 214,051 SNPs using a 250K Affymetrix SNP chip[28]. The two *A. thaliana* phenotype data sets used were (i) sodium levels averaged over 6 replicates of 342 accessions[34] and (ii) *FLOWERING LOCUS C* (*FLC*) expression measured in 166 accessions[12]. For *FLC* expression, the genotype data used were the same as those previously described[12], which come from a subset of the 1,307 individual plants and contain 216,130 markers, including 3 indels within or near the *FRIGIDA* (*FRI*) gene. For priors, we gave every marker that was within 10 kb of the *FRI* gene a 100-fold greater prior than the base prior. We then scaled them so that the sum of the priors over all the SNPs was 1.

The human data set used was the 1966 North Finland Birth Cohort NFBC1966 composed of 5,402 individuals having both phenotypic and genotypic data[31]. Phenotypic data consisted of measures for 10 quantitative traits, and genotypic data were available for 368,177 SNP markers. We were able to obtain the exact same data set, including 5,326 individuals and 331,475 SNPs after filtering, that was used previously[11]. The proportion of missing genotypes was <1%; we imputed the missing genotypes with the corresponding average per SNP to speed up the mixed-model computations.

**Simulations.** Using the *A. thaliana* genotypic data[28], we simulated two types of traits: simple ones controlled by one or two causal loci and complex ones controlled by 100 loci. For the simple traits, two randomly chosen SNPs or one randomly chosen SNP and one binary latent variable were used to generate phenotypes with three phenotypic models (additive, and/or, xor; **Supplementary Table 1**). The latent binary variable was designed by dividing the accessions in half on the basis of their latitude of origin, which we refer to as the latent north-south variable, to generate substantial covariance between the phenotypes and population structure. An additional random deviation was added, drawn from a multivariate normal distribution having a mean of zero and a scaled identity matrix as covariance to fix the trait heritability to 0.1. We simulated 1,000 phenotypes for each simulation type (two causative SNPs or one causative SNP and the latent binary variable), phenotypic model and phenotypic heritability. For complex traits, we used an additive model with 100 randomly sampled SNPs having effect sizes drawn from an exponential distribution with a rate of 1. An additional random deviation was added, drawn from a normal distribution with a mean of zero and scaled identity matrix as covariance matrix to fix the trait heritability to 0.25, 0.5 and 0.75. For each phenotypic heritability, 500 phenotypes were simulated. All simulated phenotypes have been analyzed with the four methods presented in the main text. For completeness, another single-locus approximate mixed model was used to analyze the phenotypes simulated under the 100-locus model. To control some potential confounding from population structure that was not accounted for by the random term, this approach uses as covariates the ten first principal components from a principal-component analysis of the standardized genotypic data. As no obvious difference was observed between this additional approach and the approximate mixed model, only the latter was presented (**Supplementary Fig. 11**).

**Linear mixed model.** Following Fisher's[9] polygenic model and adopting similar notation as was used previously[41], the phenotypic value of the $i$th individual can be denoted as

$$y_i = \mu + \sum_{j=1}^{m} x_{ij} a_j + e_i$$

where $m$ is the total number of causal loci, $x_{ij}$ is the genotype (coded in numerical terms) of the $j$th causal locus to the $i$th individual, $a_j$ is the effect size of the $j$th locus and $e_i$ is the error. If we assume that there are a large number of independent causal loci and that their effects are drawn from a Gaussian distribution (Fisher's infinitesimal model), we can sum them and approximate them with a Gaussian random variable. We therefore modeled the trait using a mixed model[8], where the phenotype can be denoted in vector notation as $\mathbf{y} = X\beta + \mathbf{g} + \mathbf{e}$, where $X$ is a matrix of fixed effects (for example, SNPs), $\beta$ is a vector of effect sizes, $\mathbf{g}$ is a vector of random polygenic effects with covariance matrix $\sigma_g^2 K^*$ and $\mathbf{e}$ is a vector of random independent effects with variance $\sigma_e^2$ modelling the residual error. Both the random terms, $\mathbf{g}$ and $\mathbf{e}$, are assumed to have a Gaussian distribution with mean of 0. Here, $K^*$ denotes the adjusted kinship matrix, where the loci included as fixed effects are excluded from kinship matrix estimation. If $M \gg n$, where $M$ is the number of causal loci and $n$ is the number of individuals, then $K^* \approx K$. Different assumptions lead to different kinship matrices that can be used for the mixed model as described in the **Supplementary Note**.

**Multiple loci mixed model.** We used forward-backward stepwise linear mixed-model regression, where the variance components $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ are estimated before each step. The variance estimates are used to obtain generalized least-square (GLS) effect size estimates and F-test $P$ values for each SNP. The SNP with the most significant association is then added to the model as a cofactor for the next step, and the $P$ values for all cofactors are re-estimated together with the variance components. For stopping criteria for the forward regression, we suggest stopping when the $\hat{\sigma}_g^2 / \mathrm{var}(\mathbf{y})$ estimate is close to zero or when a maximum number of forward steps is reached. After stopping the forward stepwise regression, a backward stepwise regression is performed by dropping the least significant cofactor in the model at each step. The variance components and $P$ values of all cofactors are again re-estimated at each step. For variance component estimation at each forward and backward step, the markers included as cofactors in the model can be excluded from the kinship matrix calculation, although we did not do this, as their effect on kinship is arguably negligible.

We made use of the Gram-Schmidt process[41], which makes each step as fast as the first one when $M \gg n$ (when the number of SNPs is much greater than the number of individuals). At each step, we obtained the QR decomposition of the cofactor matrix to obtain the $Q$ matrix, and we used this to calculate the marginal inverse-variance matrix as

$$M^{-\frac{1}{2}} = (I - Q'Q)'V^{-\frac{1}{2}}$$

where $V = \sigma_g^2 K + \sigma_e^2 I$ is the covariance matrix estimated at each step.

We explored several model selection criteria to select the most appropriate model. The classic Bayesian information criterion (BIC) is too tolerant in the context of GWAS, allowing for too many loci in the model, and is therefore not recommended. As an alternative, we used extended BIC, initially defined as the BIC penalized by the model space dimension[29]. We also propose and define a new criterion, the multiple Bonferroni criterion (mBonf), which selects the model with the most loci that all have $P$ values below the Bonferroni threshold. This criterion enables the user to specify the $P$-value threshold if one wants to allow for a higher FDR or restrict to a lower one. The computational complexity of our implementation is described in the **Supplementary Note**.

**Employing priors on loci.** As described[39], it is possible to employ priors on loci in a Bayesian model, where the Bayes factor is calculated for each locus. Calculating the Bayes factor, however, is not always easy, as it requires integrating out the model parameters that have some specified prior distributions. In our case, the model parameters of interest were the effect sizes of the loci in the model. A rough approximation can be achieved using the Schwarz criterion, which allowed us to avoid defining priors on the effect sizes and evaluating the integral[42]. We define the approximate Bayes factor (ABF) as

$$\log ABF = \log P(D \mid \beta, M_1) - \log P(D \mid \beta, M_0) - \frac{1}{2}(d_1 - d_0)\log n$$

where $n$ in the number of individuals, $D$ is the observed data, $M_i$ is the $i$th model and $d_i$ is the degree of freedom in the $i$th model. Using this approximation together with a prior probability $\pi$ for the causal locus, we define the approximate posterior probability of association (APPA).

$$\mathrm{APPA} = \frac{\mathrm{ABF} \times \pi}{1 - \pi(1 + \mathrm{ABF})}$$

We note that this quantity should be treated more as a score than a probability, as it is a rough estimate of the actual probability.

**Software availability.** MLMM has been implemented in two programming languages, Python and R, which have been made available (see URLs). The R

version relies on the original EMMA implementation[10]. The Python version relies heavily on the SciPy package, which can be compiled with different basic linear algebra subprograms (BLAS) versions, including GotoBLAS and the Intel Math Kernel Library (MKL).


41. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, New York, 2009).
42. Kass, R.E. & Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).