HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2013 January 01.

Published in final edited form as:

Nat Genet. ; 44(7): 825–830. doi:10.1038/ng.2314.

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

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

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria

[2]Institut National de la Recherche Agronomique, UR0588, F-45075 Orléans, France

[3]Molecular and Computational Biology, University of Southern California, Los Angeles, California, United States of America

Abstract

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 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 novel associations in known candidates as well as evidence for allelic heterogeneity. We also demonstrate 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 datasets (n > 10000) practicable.

INTRODUCTION

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 designs are subject to confounding by population structure, leading to an

*These authors contributed equally to this work.

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 paper with input from A.P., A.K., Ü.S., and Q.L.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

URLs MLMM has been implemented in two programming languages, Python and R. The R version relies on the original EMMA implementation[10] and can be obtained at https://cynin.gmi.oeaw.ac.at/home/resources/mlmm. The Python version can be obtained at https://github.com/bvilhjal/mixmogam. The Python version relies heavily on the scipy package (http://www.scipy.org/) which can be compiled with different basic linear algebra subprograms (BLAS) versions, including GotoBLAS and intel math kernel library (MKL).

inflation of test statistics and a high false positive rate[1,2]. Several methods have been proposed to deal with 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 representing the cluster memberships and principal component loadings of the individuals, respectively. While these approaches are expected to perform well when the population structure is simple, they may perform poorly when the structure is more complex, *e.g.*, when individuals display a continuum of relatedness[7]. A further improvement has been made with the use of mixed linear models, which are based on the insight that confounding will be caused by the genetic background of causal variants in the presence of population structure. The mixed model controls for this through a random polygenic term having a covariance structure described by a relationship matrix so that correlation in phenotype mirrors relatedness[8], as predicted by Fisher's classical model[9]. This approach has been shown to perform well in plants, animals and humans[6,10,11,12] and methods have been developed to allow the analysis of large GWAS datasets 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. For complex traits controlled by several large-effect loci, this may not be appropriate, especially in presence of population structure[12] (indeed, a substantial inflation of single-locus test statistics can be 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 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 genome-wide (due to linkage disequilibrium) rather than just local (due to linkage)[18]. Thus, while conditioning on known causative factors in GWAS has typically been done 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 illustrated in Fig. 1, a conditional analysis at a genome-wide scale may well have higher power and lower false discovery rate (FDR) than single-locus approaches. Similarly, in the context of human genetics, it has been suggested that conditioning on major effects loci, like the *MHC*, 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 for GWAS, where the number of polymorphisms can reach millions, but the number of phenotyped and genotyped individuals 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 polymorphism that can be fitted at a time must be lower than the number of individuals. In addition, identifying the 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-, elastic net and LASSO[22,23,24,25,26]. These approaches have been shown to perform better than single locus approaches, but most are either computationally infeasible 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 handle the population structure issue effectively we make use of an approximate version of the mixed model[11,14], where re-estimate the genetic and error variances at each step of the regression (see **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 the forward inclusion; and then to perform a backward elimination from the last forward model for a more thorough exploration of the model space. We evaluate various model-selection criteria through simulations, which also suggest that the proposed multi-locus mixed-model (MLMM) method performs well in terms of false discovery rate and power. Finally, we demonstrate the utility 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 two-locus model, and a 100-locus model. For the latter, additivity was assumed, whereas for the former, different types of interactions were explored. For details, see **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 other major loci into account (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 their false discovery rate (FDR). For single-locus methods, SNPs were considered detected if their p-values were below the defined threshold; while for the multi-locus methods, the detected SNPs were those belonging to the most complex model whose cofactors' marginal p-values were all below the defined threshold.

The results for the 100-locus model are shown in 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 always outperform comparable methods that do not (MM and MLMM vs 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 vs LM and MM, respectively) as long as the causative sites themselves are included in the data (Fig. 2a–c). 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. 1–2). 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. 2d–f).

The 2-locus simulations allowed us to examine the advantages of including cofactors in the mixed model under several scenarios of population structure and/or epistasis (for details, see **Methods**). Regardless of the scenario considered, MLMM consistently performed at least as well as the other methods when restricted to small FDR (Supplementary Fig. 5, see also Fig. 1). When there are two random randomly chosen causal sites, the improvement in power observed for MLMM compared to the single marker MM is 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 criteria (EBIC)[29], and the multiple-Bonferroni criterion (mBonf) defined as the largest model whose cofactors all have a p-value below a Bonferroni-corrected threshold (we used 0.05). Our simulations show that both criteria are consistent in bounding the FDR for the MLMM method regardless of the simulation scenario, 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, demonstrating the ability of mixed models to control confounding by population structure in a multi-locus setting (Supplementary Fig. 1). Furthermore, both criteria perform appropriately in extreme scenarios where there is no detectable signal in the data, as when an external confounding variable interacts non-linearly with a single causal locus[18]. In this case, MLMM with one of the proposed criteria correctly selects a model without any SNPs whereas the other methods tested would identify false positives only (Supplementary Fig. 5). In summary, MLMM, with the conservative FDR provided by the proposed model-selection criteria, consistently outperforms the other methods in all scenarios we have simulated.

For completeness, we also compared MLMM to other single-locus mixed model implementations, 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 used above. These did indeed perform slightly better than the approximate method in our simulations, but were still far from the performances achieved by MLMM (Supplementary Fig. 7).

Application to humans

To illustrate the feasibility as well as the utility of MLMM, we applied it to a previously published dataset of metabolic traits in the Northern Finland Birth Cohort (NFBC1966)[31]. The data were previously reanalyzed to demonstrate the utility of the mixed-model[11], and we used the same settings for the 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. Focusing on the less conservative mBonf criterion, we identify all the associations previously detected using the single-locus mixed-model[11] and

nine additional associations. Of the latter, three were located near genes previously reported using the same data[31] (two in the *TOMM40- APOE* cluster for low density lipoprotein (LDL) and one in *MTNR1B* for glucose (GLU), while four were located in gene regions not previously reported with this dataset (*HNF4A* for high density lipoprotein (HDL), *SMEK2* for systolic blood pressure (SBP), and two in the *TOMM40-APOE* cluster for C-reactive protein (CRP)). The remaining two were additional SNPs in genes that had already been identified (*CETP* for HDL, and *CRP* for CRP). The detected association in *HNF4A* for HDL (rs1800961) — a gene region not previously reported with this dataset — has been replicated in two meta-analyses based on 30,714 and 99,900 individuals respectively[32,33].

Multiple significant SNPs within or near a single gene suggest either allelic heterogeneity or the presence of an untyped causal variant that is partially tagged by multiple SNPs (or both). In the case of the associations located in the *TOMM40-APOE* cluster (for both LDL and CRP), we observed a dramatic decrease of 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 expected when loci mask each other, for example by when alleles of compensatory effect are correlated, as appears to be the case here ($R^2 = 0.33$ and $R^2 = 0.25$ for LDL and CRP respectively).

Fig. 3 also shows the percentage of variance explained by the SNPs included into the model as well as the percentages of unexplained genetic and residual variance at the different steps of MLMM for LDL (for the other phenotypes, see Supplementary Fig. 9). It is notable that most of the heritable phenotypic variation remains unexplained.

Application to *A. thaliana*

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*[34]. In particular, a SNP (chr4:6392280) located in the first exon of the gene has a highly significant association (p-value = 6.33×10^{-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 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 22kb 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, *i.e.* using only the 508 SNPs located within 100kb of the gene. Using the EBIC, six SNPs were included in the model, all within 25kb of *AtHKT1;1*, explaining 52.6% of the phenotypic variation (Supplementary Fig. 10), leaving 20.5% as unexplained heritable fraction of the total variance. As noted above, this suggests either allelic heterogeneity or the presence of one or more untyped causal variants. However, since 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.

DISCUSSION

The problem of population structure in GWAS is best viewed as one of model mis-specification. When carrying out single-locus tests of association, we are using the wrong model unless the trait is actually due 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 the bias[8]. This approach works well, but assumes that the phenotypic covariance between individuals can be predicted by their relatedness, as estimated by genome-wide SNPs. As demonstrated by Fisher close to 100 years ago[9], this approximation is reasonable if the genetic background is sufficiently smooth, but it is easy to see that loci of relatively larger effect may make it 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 which correlate with population structure. As demonstrated by simulations, our implementation (MLMM) displays 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 novel associations as well as evidence for allelic heterogeneity. Indeed, as it includes multiple loci in the model, MLMM helps identify evidence for allelic heterogeneity as well as interactions, although it is difficult to exclude that multiple associated SNPs within a region are due to 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, and especially MLMM, will benefit greatly from this. While applied here to quantitative traits, MLMM can also be applied to diseases. Indeed, it is possible to analyze a disease phenotype with an approximate mixed-model by considering a binary quantitative response corresponding to the case-control status[11]. Finally, MLMM partitions the phenotypic variance into genetic, random and explained variance at each step, suggesting a natural stopping criteria (genetic variance of 0) for including cofactors. This allows the user to obtain estimates of the explained and unexplained heritable variance, as well as give insights into the 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. More sophisticated algorithms, like LASSO[36], are worth exploring. However, as other penalized methods, LASSO assumes

independence between markers, which is obviously not appropriate for structured data. This might cause LASSO to give a large effect size to a marker that is in LD with many other markers, whereas a mixed model would down-weight such markers. A potential improvement on this would be use LASSO in conjunction with a mixed-model[26]. While this approach is potentially very promising, it is currently too computationally demanding for GWAS datasets. 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 hard. 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 in this paper identify the causal sites—not surprisingly because 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 us prioritize variants. Since MLMM is based on a linear model it can easily be extended for Bayesian analysis[39,40] and allow for the integration of prior information into the model. Indeed, returning to the *FLC* example, by placing a 100-fold prior on all markers within 10kb of *FRI* we allow MLMM to include the two known causal variation as the first two cofactors in the model, demonstrating how prior knowledge can help identifying causal loci, and improving the model (Fig. 5).

METHODS

Data

Both *A. thaliana* and human data were used for the examples. The genotype data for *A. thaliana* included 1,307 individuals genotyped at 214,051 SNPs using a 250K Affymetrix SNP chip[28]. The two *A. thaliana* phenotype datasets 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 was the same as used by Atwell *et al.*, which is a subset of the 1,307 individuals and contains 216,130 markers, including three indels within or near the *FRIGIDA* (*FRI*) gene. As priors we gave every marker which were within 10kb from the *FRI* gene a 100 fold greater prior over the base prior. We then scaled them so that the sum of the priors over all the SNPs was 1.

The human dataset 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 10 quantitative traits, and genotypic data in 368,177 SNP markers. We were able to obtain the exact same dataset, *i.e.* 5,326 individuals and 331,475 SNPs after filtering, as used in[11]. The proportion of missing genotypes was < 1% which we imputed with its 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 using the three phenotypic models (additive, “and/or”, “xor”) described in Supplementary Table 1. The latent binary variable was designed by splitting the accessions in half on the basis of their latitude of origin, which we refer to as the latent north-south variable, to generate some 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. 1,000 phenotypes were simulated for each simulation type (*i.e.*, 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 has been used to analyze the phenotypes simulated under the 100-locus model. To control some potential population structure confounding 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 extra approach and the approximate mixed-model, only the latter was presented in the results (Supplementary Fig. 11).

Linear mixed model

Following Fisher's[9] polygenic model and adopting similar notation as in Yang *et al.*[41], the phenotypic value of the *i*'th individual can be written 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 (re- 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 a large number of the independent causal loci are and that their effects are drawn from a Gaussian distribution (Fisher's infinitesimal model) we can sum them up and approximate them with a Gaussian random variable. We therefore model the trait using a mixed model[8], where the phenotype can be written in vector notation as

$$\mathbf{y} = X\beta + \mathbf{g} + \mathbf{e},$$

where β are the effect sizes of fixed effects (e.g. SNPs), **g** is a vector of random polygenic effects and has distribution $\mathbf{g} \tilde{} N\left(0, \sigma_g^2 K^*\right)$ and **e** is the residual (error) and has distribution $\mathbf{e} \tilde{} N\left(0, \sigma_e^2 I\right)$. Here K^* denotes the adjusted kinship matrix, where the loci included as fixed effects are excluded from the 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 (MLMM)

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 most significant SNP 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. As a 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 step 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 component and p-values of all cofactors again re-estimated at each step. For the variance components 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 since their effect on the kinship is arguably negligible.

We make use of the Gram-Schmidt process[41] which makes each step as fast as the first one when $M \gg n$ (i.e. the number of SNPs is much greater the number individuals). At each step we obtain the QR-decomposition of the cofactor matrix to obtain the $\mathcal{Q}$ matrix and use it to calculate the marginal inverse variance matrix

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

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 the extended BIC, initially defined by Chen and Chen[29] as the BIC penalized by the model space dimension. We also propose and define a new criterion, the multiple Bonferroni criterion (mBonf) which selects the model with most loci for which all have p-values below the Bonferroni threshold. This criteria enables the user to specify the p-value threshold if one wants to allow for a higher false discovery rate (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 in[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 is however not always easy, as it requires integrating out the model parameters which have some specified prior distributions. In our case the model parameters of interest are the effect sizes of the loci in the model. A rough approximation can be achieved using the Schwarz criterion which

allows us to avoid having to define priors on the effect sizes and evaluate the integral[42]. We define the approximate Bayes factor (ABF) as

$$logABF = logP(D|\beta, M_1) - logP(D|\beta, M_0) - \frac{1}{2}(d_1 - d_0)\, logn,$$

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

$$APPA = \frac{ABF \cdot \pi}{1 - \pi\,(1 + ABF)}.$$

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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, UCLA, University of Oulu, and the National Institute for Health and Welfare in Finland. This manuscript was not prepared in collaboration with investigators of the NFBC1966 Study and does not necessarily reflect the opinions or views of the NFBC1966 Study Investigators, the Broad Institute, UCLA, University of Oulu, National Institute for Health and Welfare in Finland or the NHLBI. We furthermore thank Nelson B. Freimer and Susan K. Service for their help in pre-processing the NFBC1966 data. We would also like to thank Petar Forai for excellent IT and cluster support at the GMI, the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for further computational resources, and David V. Conti, David J. Balding, and Sudeep Srivastava for useful discussions on the topic. Finally, we would like to thank the anonymous reviewers for helpful comments on the manuscript. This work was supported by grants from the EFPA department of INRA to V.S. and the DFG to A.K., and by grants from the United States National Institutes of Health (P50 HG002790) and the EU Framework Programme 7 (“TransPLANT”, grant agreement number 283496) to M.N., as well as by the Austrian Academy of Sciences through the GMI.

REFERENCES

1. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003; 361:598–604. [PubMed: 12598158]
2. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nature Genetics. 2004; 36:512–517. [PubMed: 15052271]
3. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]
4. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. American Journal of Human Genetics. 2000; 67:170–181. [PubMed: 10827107]
5. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38:904–909. [PubMed: 16862161]
6. Yu JM, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics. 2006; 38:203–208. [PubMed: 16380716]

7. Zhao KY, et al. An Arabidopsis example of association mapping in structured samples. Plos Genetics. 2007; 3:12.
8. Henderson, CR. Application of Linear Models in Animal Breeding. Unversity of Guelph, Guelph, Ont; 1984.
9. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh. 1918; 52:399–433.
10. Kang HM, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178:1709–1723. [PubMed: 18385116]
11. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010; 42:348–354. [PubMed: 20208533]
12. Atwell S, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010; 465:627–631. [PubMed: 20336072]
13. Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007; 177:577–585. [PubMed: 17660554]
14. Zhang ZW, et al. Mixed linear model approach adapted for genome-wide association studies. Nature Genetics. 2010; 42:355–360. [PubMed: 20208535]
15. Yang J, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011; 19:807–812. [PubMed: 21407268]
16. Jansen RC. Interval mapping of multiple quantitative trait loci. Genetics. 1993; 135:205–211. [PubMed: 8224820]
17. Zeng ZB. Precision mapping of quantitative trait loci. Genetics. 1994; 136:1457–1468. [PubMed: 8013918]
18. Platt A, Vilhjalmsson BJ, Nordborg M. Conditions under which genome-wide association studies will be positively misleading. Genetics. 2010; 186:1045–1052. [PubMed: 20813880]
19. Allen AS, Satten GA, Bray SL, Dudbridge F, Epstein MP. Fast and robust association tests for untyped SNPs in case-control studies. Human Heredity. 2010; 70:167–176. [PubMed: 20689309]
20. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. Plos Biology. 2010; 8:e100029.
21. Cordell HJ, Clayton DG. 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. American Journal of Human Genetics. 2002; 70:124–141. [PubMed: 11719900]
22. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. Plos Genetics. 2008; 4:e1000130. [PubMed: 18654633]
23. Malo N, Libiger O, Schork NJ. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. American Journal of Human Genetics. 2008; 82:375–385. [PubMed: 18252218]
24. Croiseau P, Cordell HJ. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. BMC Proceedings. 2009; 3:S61. [PubMed: 20018055]
25. Cho S, et al. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. Annals of Human Genetics. 2010; 74:416–428. [PubMed: 20642809]
26. Wang D, Eskridge KM, Crossa J. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. Journal of Agricultural Biological and Environmental Statistics. 2011; 16:170–184.
27. Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic Epidemiology. 2010; 34:879–891. [PubMed: 21104890]
28. Horton MW, et al. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet. 2012; 44:212–216. [PubMed: 22231484]
29. Chen JH, Chen ZH. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008; 95:759–771.

30. Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Statistical Science. 2009; 24:451–471.
31. Sabatti C, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature Genetics. 2009; 41:35–46. [PubMed: 19060910]
32. Kathiresan S, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nature Genetics. 2009; 41:56–65. [PubMed: 19060906]
33. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]
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 Genetics. 2010; 6:e1001193.
35. Altshuler DL, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
36. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological. 1996; 58:267–288.
37. Valdar W, Holmes CC, Mott R, Flint J. Mapping in structured populations by resample model averaging. Genetics. 2009; 182:1263–1277. [PubMed: 19474203]
38. Tian F, et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics. 2011; 43:159–U113162. [PubMed: 21217756]
39. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics. 2009; 10:681–690.
40. Servin B, Stephens M. Imputation-based analysis of association studies: Candidate regions and quantitative traits. Plos Genetics. 2007; 3:e114. [PubMed: 17676998]

ADDITIONAL REFERENCES

41. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. Springer; New York: 2009.
42. Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995; 90:773–795.

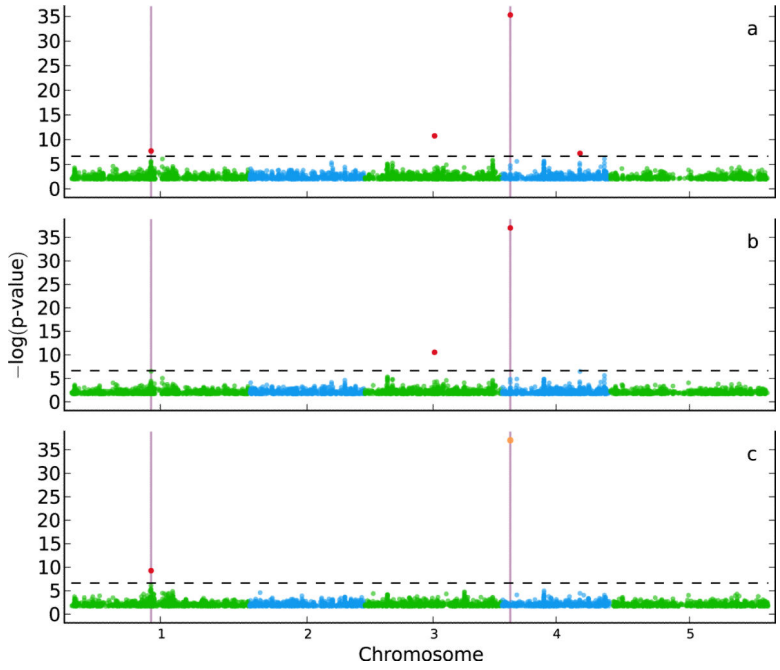

Figure 1.

A GWAS for a simulated trait with two causal SNPs (marked by vertical lines), randomly chosen from a real *A. thaliana* SNP dataset[28]. Random error was added to the trait to fix the heritability at 25%. (a) A single-SNP linear regression scan detects four significantly associated SNPs (at a Bonferroni-corrected threshold of 0.05; dashed horizontal line) marked in red. Half of these SNPs are false positives and the other half true positives, leading to a false discovery rate (FDR) of 50% and a power of 100%. (b) A single-SNP mixed-model[11,14] scan eliminates one false positive but also one true positive, leading to a similar (50%) FDR while decreasing the power to 50%. (c) Adding the most significant SNP as a cofactor to the mixed model (marked in orange) recovers the second causal SNP while eliminating the last false positive, leading to the perfect case of a FDR of 0% and a power of 100%.

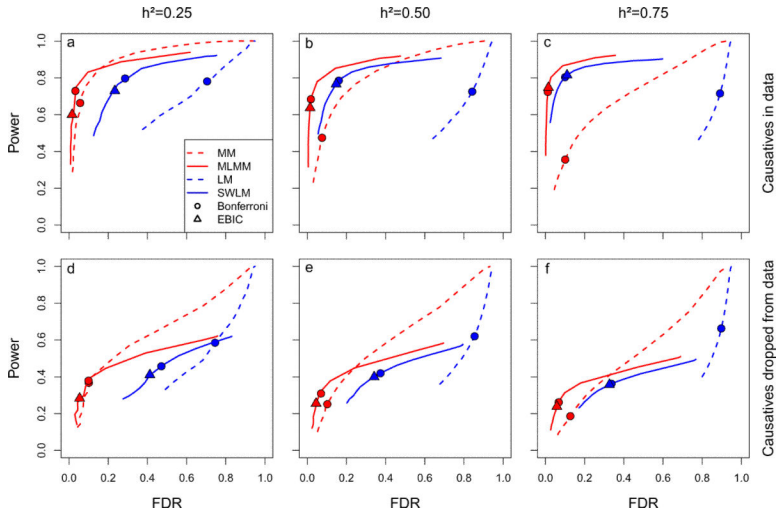

Figure 2.

Power and false discovery rate (FDR) in the 100-locus model simulations for four different mapping methods: linear model (LM), stepwise linear model (SWLM), mixed-model (MM), and multi-locus mixed-model (MLMM). For the purpose of computing power and FDR, a causal SNP was considered detected if a SNP within 25kb on either side was declared significant (results for other window sizes are given in Supplementary Fig. 3), and only causal SNPs that were in principle detectable (i.e., 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: a comparison between forward and backward paths is given in Supplementary Fig. 4. Circles and triangles denote the best-fitting model according to the Bonferroni and EBIC model-selection criteria, respectively. Three phenotypic heritabilities were used in the simulations: 0.25 (a, d), 0.50 (b, e), and 0.75 (c, f). Power and FDR was estimated with (a–c) and without (d–f) the causal loci included.

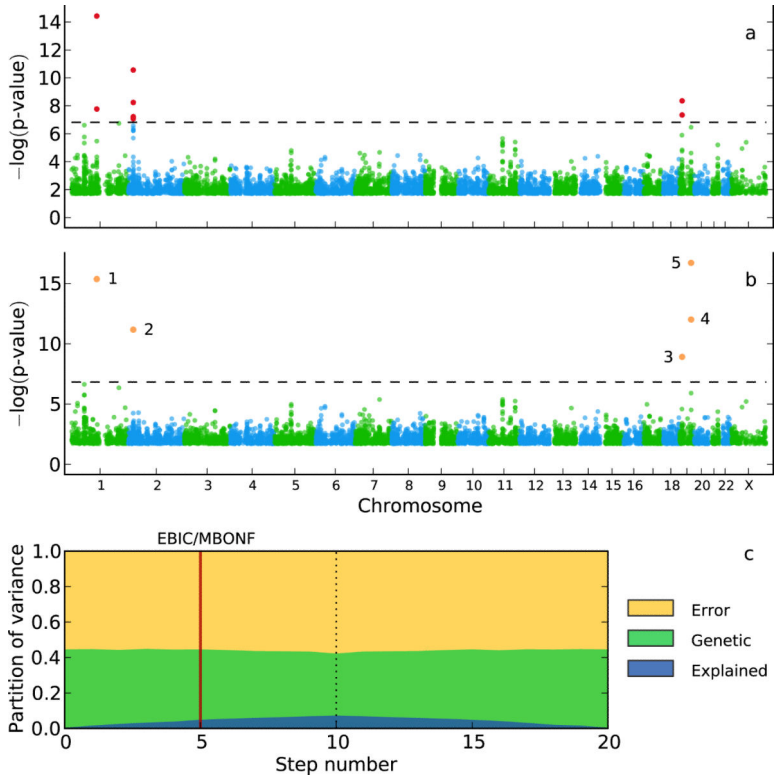

Figure 3.

GWAS for low-density lipoprotein (LDL) in the NFBC1966 dataset. (a) A single-locus mixedmodel identifies seven SNPs in three genes (marked in red; Bonferroni-corrected threshold of 0.05; dashed horizontal line). (b) A multi-locus mixed-model (MLMM) identifies five SNPs in four genes (marked in orange, and numbered in the order they were included in the model). (c) Partition of variance at each step of MLMM (10 forward and 10 backward) into variance explained by: the SNPs included in the model (blue); kinship (green); and noise (yellow).

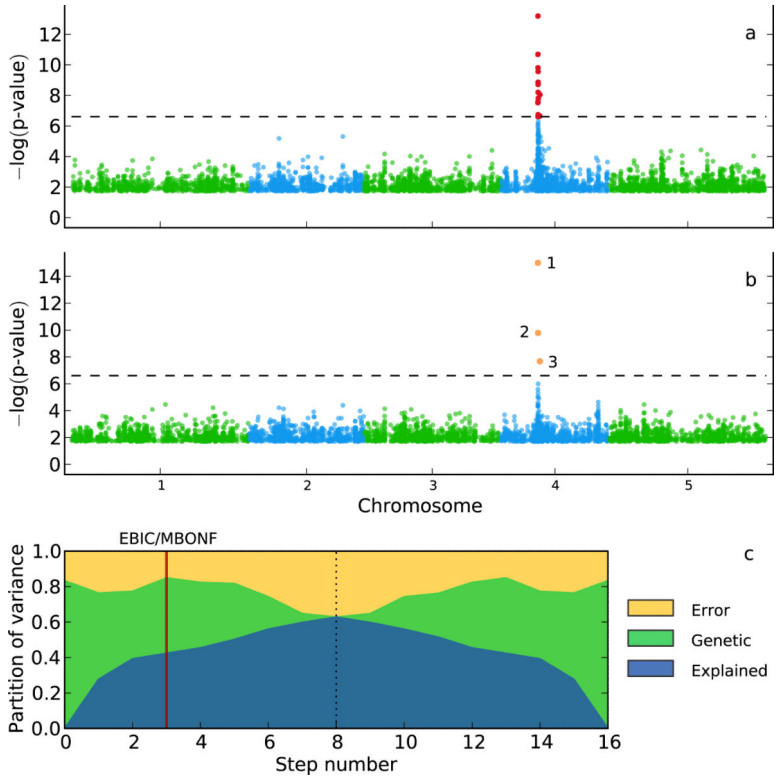

Figure 4.

GWAS for Na^+ accumulation in *A. thaliana.* (a) A single-locus mixed-model identifies a strong peak of significantly associated SNPs on chromosome 4 (marked in red; Bonferroni-corrected threshold of 0.05; dashed horizontal line). (b) Multi-locus mixed-model (MLMM) identifies three SNPs (marked in orange, and numbered in the order they were included in the model). (c) Partition of variance at each step of MLMM (8 forward and 8 backward) into variance explained by: the SNPs included in the model (blue); kinship (green); and noise (yellow).

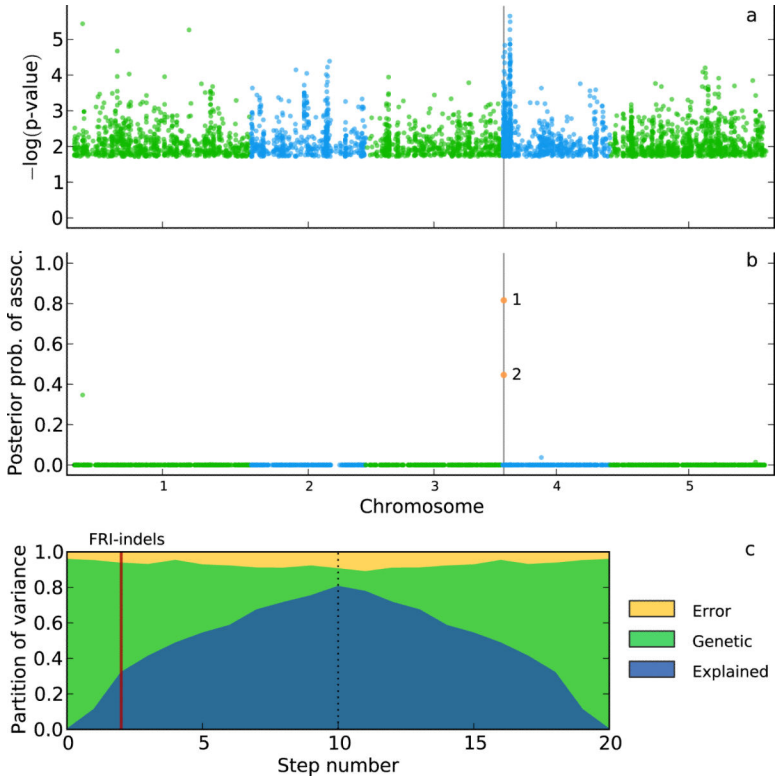

Figure 5.

An example of Bayesian multi-locus mixed-model (MLMM) for the analysis of *FLOWERING LOCUS C (FLC)* expression in *A. thaliana*. (a) An approximate mixed-model scan for *FLC* expression, marking the *FRIGIDA* gene with a vertical grey line. (b) The posterior probability of association scan after the Bayesian MLMM has included two loci into the model, which incidentally are the two previously identified causative indels. (c) Partition of phenotypic variance for each forward inclusion (10 steps) and backwards elimination (10 steps after the dotted line). The vertical red line marks the model with the two causative indels in the model.

Table 1

SNPs identified in multi-locus mixed-model analysis of the NFBC1966 traits

				p-value		Previously identified by	
SNP	**Chr.**	**Position**	**Gene**	**EBIC**	**mBonf**	**Sabatti *et al.*[31]**	**Kang *et al.*[11]**
Associated with mmol/l TG							
rs673548	2	21091049	*APOB*		5.1×10^{-8}	Y	Y
rs1260326	2	27584444	*GCKR*	$1.5 \times 1O^{-10}$	7.9×10^{-11}	Y	Y
rs10096633	8	19875201	*LPL*	1.6×10^{-8}	2.4×10^{-8}	Y	Y
Associated with mmol/l HDL							
rs1532085	15	56470658	*LIPC*	9.2×10^{-12}	8.0×10^{-12}	Y	Y
rs3764261	16	55550825	*CETP*	2.7×10^{-32}	3.7×10^{-23}	Y	Y
rs7499892	16	55564091	*CETP*		9.5×10^{-8}	N	N
rs255049	16	66570972	*LCAT*	1.3×10^{-8}	4.8×10^{-8}	Y	Y
rs1800961	20	42475778	*HNF4A*		1.5×10^{-7}	N	N
Associated with mmol/l LDL							
rs646776	1	109620053	*CELSR2*	4.2×10^{-16}	4.2×10^{-16}	Y	Y
rs693	2	21085700	*APOB*	7.1×10^{-12}	7.1×10^{-12}	Y	Y
rs11668477	19	11056030	*LDLR*	1.0×10^{-9}	1.0×10^{-9}	Y	Y
rs157580	19	50087106	*TOMM40-APOE*	2.2×10^{-17}	2.2×10^{-17}	Y	N
rs405509	19	50100676	*TOMM40-APOE*	1.3×10^{-12}	1.3×10^{-12}	N	N
Associated with mmol/l CRP							
rs2369146	1	157934819	*CRP*	4.5×10^{-9}	2.8×10^{-9}	N	N
rs2794520	1	157945440	*CRP*	1.1×10^{-29}	6.6×10^{-30}	Y	Y
rs2650000	12	119873345	*HNF1A*	1.3×10^{-12}	1.0×10^{-12}	Y	Y
rs8106922	19	50093506	*TOMM40-APOE*		1.6×10^{-12}	N	N
rs439401	19	50106291	*TOMM40-APOE*		2.2×10^{-9}	N	N
Associated with mmol/l GLU							
rs560887	2	169471394	*G6PC2*	2.2×10^{-13}	2.2×10^{-13}	Y	Y
rs2971671	7	44177862	*GCK*	3.2×10^{-9}	3.2×10^{-9}	N	Y
rs3847554	11	92308474	*MTNR1B*	4.7×10^{-11}	4.7×10^{-11}	N[a]	N

				p-value		Previously identified by	
SNP	**Chr.**	**Position**	**Gene**	**EBIC**	**mBonf**	**Sabatti *et al.*[31]**	**Kang *et al.*[11]**
Associated with SBP							
rs782602	2	55702813	*SMEK2*		1.4×10^{-7}	N	N[b]

Models were selected using either extended BIC (EBIC) or the multiple Bonferroni criterion (mBonf). Phenotype abbreviations are TG, triglyceride; HDL, high-density lipoprotein; LDL, low density lipoprotein; CRP, C-reactive protein; GLU, glucose; SBP, systolic blood pressure.

[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 then the Bonferroni-corrected threshold of 0.05 used for mBonf.