BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 2 2013, pages 206–214
doi:10.1093/bioinformatics/bts669

Genetics and population analysis

Advance Access publication November 22, 2012

A Lasso multi-marker mixed model for association mapping with population structure correction

Barbara Rakitsch[1,*], Christoph Lippert[1,*], Oliver Stegle[1,*] and Karsten Borgwardt[1,2]

[1]Machine Learning and Computational Biology Research Group, Max Planck Institute for Intelligent Systems and Max Planck Institute for Developmental Biology and [2]Zentrum für Bioinformatik, Eberhard Karls Universität, 72076 Tübingen, Germany

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Exploring the genetic basis of heritable traits remains one of the central challenges in biomedical research. In traits with simple Mendelian architectures, single polymorphic loci explain a significant fraction of the phenotypic variability. However, many traits of interest seem to be subject to multifactorial control by groups of genetic loci. Accurate detection of such multivariate associations is non-trivial and often compromised by limited statistical power. At the same time, confounding influences, such as population structure, cause spurious association signals that result in false-positive findings.

Results: We propose linear mixed models LMM-Lasso, a mixed model that allows for both multi-locus mapping and correction for confounding effects. Our approach is simple and free of tuning parameters; it effectively controls for population structure and scales to genome-wide datasets. LMM-Lasso simultaneously discovers likely causal variants and allows for multi-marker–based phenotype prediction from genotype. We demonstrate the practical use of LMM-Lasso in genome-wide association studies in *Arabidopsis thaliana* and linkage mapping in mouse, where our method achieves significantly more accurate phenotype prediction for 91% of the considered phenotypes. At the same time, our model dissects the phenotypic variability into components that result from individual single nucleotide polymorphism effects and population structure. Enrichment of known candidate genes suggests that the individual associations retrieved by LMM-Lasso are likely to be genuine.

Availability: Code available under http://webdav.tuebingen.mpg.de/u/karsten/Forschung/research.html.

Contact: rakitsch@tuebingen.mpg.de, ippert@microsoft.com or stegle@ebi.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 22, 2012; revised on November 12, 2012; accepted on November 13, 2012

1 INTRODUCTION

Although many quantitative traits in humans, plants and animals have been observed to be heritable, a comprehensive understanding of the underlying genetic architecture is still missing. In some cases, genome-wide association studies and linkage mapping have already revealed individual causal variants that control trait variability, for example, genetic mapping yielded insights into the genetic architecture of global-level traits in plants (Atwell *et al.*, 2010) and mouse (Valdar *et al.*, 2006), as well as the risks for important human diseases, such as type 2 diabetes (Craddock *et al.*, 2010). Nevertheless, the statistical analysis of these genetic data has proven to be challenging, not least because single-genetic variants rarely explain larger fractions of phenotype variability; hence, individual effect sizes are small (Mackay *et al.*, 2009; McCarthy *et al.*, 2008). An inherent limitation of power to map weak effects is because of confounding relatedness between samples. Population structure can induce false association patterns with large numbers of loci being correlated with the phenotype. To understand the true genetic architecture of complex traits, it is necessary to address both of these challenges, taking population structure into account and joint modelling of true multifactorial associations.

If multiple variants contribute to phenotype variation in an additive fashion, simple methods that assess the significance of individual loci independently are likely to fall short: masking effects between causal single nucleotide polymorphisms (SNPs) can limit mapping power, with relevant loci not reaching genome-wide significance levels (McCarthy *et al.*, 2008). These shortcomings have been widely addressed in multivariate regression, explicitly modelling the additive effects of multiple SNPs. The corresponding methods either fit sparse predictors of all genome-wide SNPs, using a shrinkage prior or use stepwise forward selection (Yang *et al.*, 2012). Applying a Laplace prior leads to the Lasso (Li *et al.*, 2011), and related priors have also been considered (Hoggart *et al.*, 2008). With the same ultimate goal to capture the genetic effects of groups of SNPs, variance component models have recently been proposed to quantify the heritable component of phenotype variation explainable by an excess of weak effects (Yang *et al.*, 2010).

Population structure induces spurious correlations between genotype and phenotype, complicating the genetic analysis. A major source of these effects can be understood as deviation from the idealized assumption that the samples in the study population are unrelated. Instead, population structure in the sample is difficult to avoid, and even in a seemingly stratified sample, the extent of hidden structure cannot be ignored (Newman *et al.*, 2001). Models that account for the presence of such structure are routinely applied and have been shown to greatly reduce the impact of this confounding source of variability. For instance, EIGENSTRAT builds on the idea of extracting the major axes of population differentiation using a principal component analysis

*To whom correspondence should be addressed.

decomposition of the genotype data (Price *et al*., 2006), and subsequently including them into the model as additional covariates. Linear mixed models (Kang *et al*., 2008; Kang *et al*., 2010; Lippert *et al*., 2011; Yu *et al*., 2006; Zhang *et al*., 2010) provide for more fine-grained control by modelling the contribution of population structure as a random effect, providing for an effective correction of family structure and cryptic relatedness.

Although both, correction for population structure and joint mapping of multiple weak effects, have been addressed in isolation, few existing approaches are capable of addressing both aspects jointly. In line with EIGENSTRAT, Hoggart *et al*. (2008) and Li *et al*. (2011) add principal components to the model to correct for population structure. In parallel to our work, Segura *et al*. (2012) have proposed a related multi-locus mixed model approach, however, using stepwise forward selection instead of using the Lasso.

Here, we propose a novel analysis approach that combines multivariate association analysis with accurate correction for population structure. Our method allows for joint identification of sets of loci that individually have small effects and at the same time accounts for possible structure between samples. This joint modelling explains larger fractions of the total phenotype variability while dissecting it in variance components specific to individual SNP effects and population effects.

Our approach bridges the advantages of linear mixed models with Lasso regression; hence, modelling complex genetic effects while controlling for relatedness in a comprehensive fashion. The proposed linear mixed models (LMM)-Lasso is conceptually simple, computationally efficient and scales to genome-wide settings. Experiments on semi-empirical data show that the rigorous combination of Lasso and mixed modelling approaches yields greater power to detect true causal effects in a large range of settings. In retrospective analyses of studies from *Arabidopsis* and mouse, we show that through joint modelling of population structure and individual SNP effects, LMM-Lasso results in superior models of the genotype to phenotype map. These yield better quantitative predictions of phenotypes while selecting only a moderate number of SNP with individual effects. Additional evidence of the effects uncovered by LMM-Lasso likely being real is given by an enrichment analysis, suggesting that the hits obtained are often in the vicinity of genes with known implication for the phenotype.

2 MULTIVARIATE LINEAR MIXED MODELS

Our approach builds on multivariate statistics, explaining the phenotype variability by a sum of individual genetic effects and random confounding variables. In brief, the phenotype of m samples $\mathbf{y} = (y_1, \ldots, y_m)$ is expressed as the sum of n SNPs $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$

$$\mathbf{y} = \underbrace{\sum_{j=1}^{n} \beta_j \mathbf{s}_j}_{\text{genetic factors}} + \underbrace{\mathbf{u}}_{\text{confounding}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}} \tag{1}$$

Here, $\boldsymbol{\psi}$ denotes observation noise and $\mathbf{u}$ are confounding influences. Confounding influences in genetic mapping are typically not directly observed; however, their Gaussian covariance $\mathbf{K}$ can in many cases be estimated from the observed data. To account for confounding by population structure, $\mathbf{K}$ can be reliably estimated from genetic markers, for example, using the realized relationship matrix which captures the overall genetic similarity between all pairs of samples (Hayes *et al*., 2009). Similarly, in genetic analyses of gene expression, $\mathbf{K}$ can be fit to capture and correct for the confounding effect of gene expression heterogeneity (Fusi *et al*., 2012; Listgarten *et al*., 2010). Marginalizing over the random effect $\mathbf{u}$ results in a Gaussian marginal likelihood model (Kang *et al*., 2008) whose covariance matrix accounts for confounding variation and observation noise.

The resulting mixed model is typically considered in the context of single candidate SNPs, that is, restricting the sum in Equation (1) to a particular SNP while ignoring all others (Kang *et al*., 2008; Kang *et al*., 2010; Lippert *et al*., 2011; Yu *et al*., 2006; Zhang *et al*., 2010). Although computationally efficient and easy to interpret, this independent analysis can be compromised by complex genetic architectures with some genetic factors masking others (Platt *et al*., 2010a).

Some improvement can be achieved by stepwise regression or forward selection, which has recently been extended to the mixed model framework (Segura *et al*., 2012; Yang *et al*., 2012). However, these approaches are often caught in suboptimal modes, as they are order dependent (Segura *et al*., 2012).

As an alternative, we propose an efficient approach to carry out joint inference in the model implied by Equation (1). Our approach assesses all SNPs at the same time while accounting for their interdependencies and without making any assumptions on their ordering. To allow for applications to genome-wide SNP data, we place a Laplacian shrinkage prior over the fixed effects β_j, assigning zero-effect size to the majority of SNPs as done in the classical Lasso (Tibshirani, 1996).

We call this approach LMM-Lasso, as it combines the advantages of established LMM with sparse Lasso regression. The resulting model allows for dissecting the explained phenotype variance into individual SNP effects and effects caused by population structure.

2.1 Linear mixed model Lasso

Let $\mathbf{S}$ denote the $m \times n$ matrix of n SNPs for m individuals, $\mathbf{s}_j$ is then the $m \times 1$ vector representing SNP j. We model the phenotype for m individuals, $\mathbf{y} = (y_1, \ldots, y_m)$ as the sum of genetic effects β_j of SNPs $\mathbf{s}_j$ and confounding influences $\mathbf{u}$ [see Equation (1)]. The genetic effects are treated as fixed effects, whereas the confounding influences are modelled as random effects. The genetic effect terms are summed over genome-wide polymorphisms, where the great majority of SNPs have zero-effect size, that is, $\beta_j = 0$, which is achieved by a Laplace shrinkage prior on all weights. The random variable $\mathbf{u}$ is not observed directly. Instead, we assume that the distribution of $\mathbf{u}$ is Gaussian with covariance $\mathbf{K}$, $\mathbf{u} \sim \mathcal{N}(0, \sigma_g^2 \mathbf{K})$.

Assuming Gaussian noise, $\boldsymbol{\psi} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$, and marginalizing over the random variable $\mathbf{u}$, we can write down the conditional posterior distribution over the weight vector $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{S}, \mathbf{K}, \sigma_g^2, \sigma_e^2, \lambda) \propto \underbrace{\mathcal{N}(\mathbf{y}|\sum_{j=1}^{n} \beta_j \mathbf{s}_j, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})}_{\text{marginal likelihood}} \underbrace{\prod_{j=1}^{n} e^{-\frac{\lambda}{2}|\beta_j|}}_{\text{prior}} \tag{2}$$

Here, λ denotes the sparsity hyperparameter of the Laplace prior, σ_e^2 is the residual noise variance and σ_g^2 denotes the variance of the random effect components.

2.2 Parameter inference

Learning the hyperparameters $\mathbf{\Theta} = \{\lambda, \sigma_g^2, \sigma_e^2\}$ and the weights $\boldsymbol{\beta}$ jointly is a hard non-convex optimization problem. Here, we propose a combination of fitting some of these parameters on the null model with the individual SNP effects excluded and reduction to a standard Lasso regression problem.

Null-model fitting To obtain a practical and scalable algorithm, we first optimize σ_g^2, σ_e^2 by maximum likelihood under the null model, ignoring the effect of individual SNPs. The analogous procedure is widely used in single-SNP mixed models and has been shown to yield near-identical results to an exact approach (Kang *et al.*, 2010). To speed up the computations needed, we optimize the ratio of the random effect and the noise variance, $\delta = \sigma_e^2/\sigma_g^2$, which can be optimized efficiently by using computational tricks proposed elsewhere (Lippert *et al.*, 2011):

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{S}, \mathbf{K}, \sigma_g^2, \delta, \lambda) \propto \mathcal{N}(\mathbf{y}|\sum_{j=1}^{n} \beta_j \mathbf{s}_j, \sigma_g^2(\mathbf{K} + \delta\mathbf{I})) \prod_{j=1}^{n} e^{-\frac{\lambda}{2}|\beta_j|} \quad (3)$$

Briefly, we compute the eigendecomposition of the covariance $\mathbf{K}$=Udiag($\mathbf{d}$)U^{T}, which can be used to rotate the data such that the covariance matrix of the normal distribution is isotropic. We carry out 1D numerical optimization of the marginal likelihood [Equation (3)] with respect to δ, whereas σ_g^2 can be optimized in closed form in every evaluation.

Reduction to standard Lasso problem Having fixed δ, we use the eigendecomposition of $\mathbf{K}$ again to rotate our data such that the covariance matrix becomes isotropic:

$$p(\boldsymbol{\beta}|\widetilde{\mathbf{y}}, \widetilde{\mathbf{S}}, \mathbf{K}, \sigma_g^2, \lambda) \propto \mathcal{N}(\widetilde{\mathbf{y}}|\sum_{j=1}^{n} \beta_j \widetilde{\mathbf{s}}_j, \sigma_g^2\mathbf{I}) \prod_{j=1}^{n} e^{-\frac{\lambda}{2}|\beta_j|} \quad (4)$$

Here, $\widetilde{\mathbf{S}}$ denotes the rotated and rescaled genotypes, and $\widetilde{\mathbf{y}}$ denotes the respective phenotypes:

$$\tilde{\mathbf{S}} = (\mathrm{diag}(\mathbf{d}) + \delta\mathbf{I})^{-\frac{1}{2}}\mathbf{U}^{\mathbf{T}}\mathbf{S},$$
$$\tilde{\mathbf{y}} = (\mathrm{diag}(\mathbf{d}) + \delta\mathbf{I})^{-\frac{1}{2}}\mathbf{U}^{\mathbf{T}}\mathbf{y}$$

Using this transformation, the task of determining the most probable weights in Equation (4) is now equivalent to the Lasso regression model, as maximizing the conditional posterior with respect to $\boldsymbol{\beta}$ is equivalent to minimizing the negative log of Equation (4):

$$\min_{\boldsymbol{\beta}} \frac{1}{\sigma_g^2}||\widetilde{\mathbf{y}} - \widetilde{\mathbf{S}}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$$

A related algorithm for combining random effects with the Lasso has been proposed in Schelldorfer *et al.* (2011), which includes generalized linear mixed models with ℓ_1-penalty at the cost of higher computational complexity. An appropriate setting of λ can be found by cross-validation to maximize the overall predictive performance or stability selection (Meinshausen and Bühlmann, 2010).

The computational efficiency of the two-stage procedure proposed here depends on the approximation to fit δ on the null model, allowing for the reduction of the problem to standard Lasso regression. For univariate single-SNP mixed models, efficient optimization of δ for each SNP can be done by recently proposed computational tricks (Lippert *et al.*, 2011; Zhou and Stephens, 2012). Unfortunately, these techniques cannot be directly applied in the multivariate setting. In principle, it is possible to extend the cross-validation to optimize over pairs (δ, λ). However, this remains impracticable for most datasets because of the additional computational cost implied; hence, we consider optimizing δ on the null model in the experiments (Kang *et al.*, 2010).

2.3 Phenotype prediction

Given a trained LMM-Lasso model on a set of genotypes and phenotypes, we can predict the unobserved phenotype of test individuals. The predictive distribution can be derived by conditioning the joint distribution over all individuals on the training individuals (Rasmussen and Williams, 2006), resulting in a Gaussian predictive distribution

$$p(\mathbf{y}^\star|\mathbf{y}, \mathbf{S}^\star, \mathbf{S}) = \mathcal{N}(\mathbf{y}^\star|\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star) \quad (5)$$

with

$$\boldsymbol{\mu}^\star = \underbrace{\mathbf{S}^\star\boldsymbol{\beta}}_{\text{Lasso prediction}} + \underbrace{\mathbf{K}_{\mathbf{S}^\star\mathbf{S}}(\mathbf{K} + \delta\mathbf{I})^{-1}(\mathbf{y} - \mathbf{S}\boldsymbol{\beta})}_{\text{Random effect prediction}}$$
$$\boldsymbol{\Sigma}^\star = \sigma_g^2(\mathbf{K}_{\mathbf{S}^\star\mathbf{S}^\star} + \delta\mathbf{I}) - \sigma_g^2\mathbf{K}_{\mathbf{S}^\star\mathbf{S}}(\mathbf{K} + \delta\mathbf{I})^{-1}\mathbf{K}_{\mathbf{S}\mathbf{S}^\star}s \quad (6)$$

The mean prediction is a sum of contributions from the Lasso component and the random effect part, which is similar to best linear unbiased prediction (Robinson, 1991). The matrix $\mathbf{K}_{\mathbf{S}^\star\mathbf{S}}$ denotes the covariance matrix between the test individuals $\mathbf{S}^\star$ and the train individuals $\mathbf{S}$, $\mathbf{K}_{\mathbf{S}^\star\mathbf{S}^\star}$ is the covariance matrix between all test individuals and $\mathbf{K} := \mathbf{K}_{\mathbf{SS}}$ is the covariance matrix between all training individuals, which with slight abuse of notation are denoted by their genetics $\mathbf{S}$.

2.4 Choice of the random effect covariance to account for population structure

Depending on the application, the random effect covariance $\mathbf{K}$ can be chosen in a variety of ways. Here, we discuss specific options to account for population structure.

Choice of genetic similarity matrix For the identity by descent matrix, an entry is defined as the predicted proportion of the genome that is identical by descent given the pedigree information. In contrast, the identity by state matrix simply counts the number of loci on which the samples agree, whereas the realized relationship matrix (RRM) is calculated as the linear kernel between the SNPs (Hayes *et al.*, 2009). In subsequent experiments, we have used the RRM. An example for the RRM-matrix derived from the *Arabidopsis thaliana* dataset is given in Supplementary Figure S1.

Realized relationship matrix and relationship to Bayesian linear regression From a Bayesian perspective, using the RRM as the covariance matrix is equivalent to integrating over all SNPs in a

linear-additive model with an independent Gaussian prior over the weights $\mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \sigma_g^2\mathbf{I})$ (Goddard *et al.*, 2009). The choice of a Gaussian prior leads to a dense posterior distribution, reflecting the a priori belief that a large fraction of SNPs jointly contribute to phenotype variability. This prior choice is in sharp contrast to the generally accepted opinion that most SNPs are not causal.

Thus, choosing this particular covariance matrix **K** can be regarded as modelling genetic effects that are confounded because of population structure or are small additive infinitesimal effects, whereas single SNPs that have a sufficiently large effect size are directly included in the Lasso model.

2.5 Scalability and runtime

The appeal of the LMM-Lasso is a runtime performance comparable with the standard LASSO. The difference is a one-time off cubic cost for the decomposition of the random effect matrix **K** to rotate the genotype and phenotype data (see Section 2.2).

To demonstrate the applicability to genome-wide datasets, we have empirically measured the runtime for computing the complete path of sparsity regularizers on the synthetic dataset, consisting of 1196 plants and 213 624 SNPs. On a single core of a Mac Pro (3 GHz, 12 MB L2-Cache, 16 GB memory), the Lasso required 145 min central processing unit (CPU) time and the LMM-Lasso 146 min of CPU time.

If needed, the runtime of LMM-Lasso could be improved in several ways. First, if the number of samples is large ($m > 10^5$), the runtime is dominated by the decomposition of **K** and rotating the data for the optimization of δ. As shown in Lippert *et al.* (2011), reducing the covariance **K** to a low-rank representation calculated from a small subset of n_s SNPs, yields similar results while reducing the runtime from $O(m^2n)$ to $O(mn_s^2)$. Second, the runtime of the ℓ_1-solver is heavily dependent on the optimization method used. Fortunately, the development of new and efficient ℓ_1-solvers is still an active area of research. New approaches include parallelized coordinate descent algorithms (Bradley *et al.*, 2011) and screening tests that are able to prune away SNPs that are guaranteed to have zero weights (Xiang *et al.*, 2011), avoiding to load the complete genotype matrix into the working memory.

3 METHODS AND MATERIAL

3.1 Arabidopsis thaliana

We obtained genotype and phenotype data for up to 199 accessions of *A.thaliana* from Atwell *et al.* (2010). Each genotype comprises 216 130 SNPs per accession. We study the group of phenotypes related to the flowering time of the plants. We excluded phenotypes that were measured for <150 accessions to avoid possible small sample size effects, resulting in 20 flowering phenotypes that were considered. The relatedness between individuals ranges in a wide spectrum leading to a complex population structure (Platt *et al.*, 2010b).

3.2 Mouse inbred population

We also obtained genotype and phenotype data for 1940 mice from a multi-parent inbred population (Valdar *et al.*, 2006). Each individual genotype comprises of 12 226 SNPs. All mice were derived from eight inbred strains and were crossed to produce a heterogenous stock. The phenotypes span a large variety of different measurements ranging from biochemical to behavioural traits. Here, we focused on 273 phenotypes that have numeric or binary values.

3.3 Semi-empirical data

To assess the accuracy of alternative methods for variable selection, we considered a semi-empirical example based on the extended *A.thaliana* dataset (Horton *et al.*, 2012) consisting of 1196 plants. We considered real phenotype data to obtain realistic background signal that is subject to population structure. In addition to this empirical background, we added simulated associations with different effect sizes and a range of complexities of the genetic models. For full details of the simulation procedure and the evaluation of associations recovered by different methods, see Supplementary Text.

3.4 Preprocessing

We standardized the SNP data, which have the effect that the prior on the effect size is dependent on the minor allele frequency: for instance, SNPs with a low minor allele frequency require a smaller weight to have the same effect on the phenotype, and, hence, will be more likely driven to zero at the maximum a posteriori (MAP)-solution. On the phenotypes, we performed a Box–Cox transformation (Sakia, 1992) and subsequently standardized the data.

3.5 Model selection

Variation of the model complexity of Lasso methods can either be done by choosing the number of active SNPs or equivalently by varying the hyperparameter λ explicitly. For the benefit of direct interpretability, we chose to vary the number of active SNPs. For a fixed number of selected SNPs, we found the corresponding hyperparameter λ by a combination of bracketing and bisection as done in Wu *et al.* (2009).

To select which of these Lasso models is most suitable, we considered alternative strategies, depending on the objective.

(i) **Phenotype prediction** To predict phenotypes, we used 10-fold cross-validation. We split the data randomly into 10-folds. Each fold was once picked as test dataset, with all other folds being used for training the model. The model was selected to maximize the explained variance on the test set. In this comparison, we considered models with different numbers of SNPs, varying from {0,1,2,...,10,20,30,...,100,150,200,250}, with the additional constraint that the number of active SNPs should not exceed the number of samples.

(ii) **Variable selection** To assess the significance of individual features, we considered stability selection (Meinshausen and Bühlmann, 2010). Here, we fixed the number of active SNPs to 20 and drew randomly 90% of the data 100 times. To accommodate the limited sample size, we did not use 50% of the samples for each draw as proposed in the original article. We selected all SNPs that were found in >50% of all restarts. We used the smallest threshold possible to also detect SNPs that have a small effect size.

In consequence, we allowed selection of a modest number of false-positive results. Significance estimates can be deduced from the selection frequency of individual SNPs (Meinshausen and Bühlmann, 2010).

To obtain a complete ranking of features, as used to evaluate models in the simulation study, we used the LASSO regularization path and ranked features by the order of inclusion into the model.

4 RESULTS

4.1 Semi-empirical setting with known ground truth

We assessed the ability of LMM-Lasso to recover true genotype to phenotype associations in a semi-empirical simulated dataset. To ensure realistic characteristics of population structure, we simulated confounding such that it borrows key characteristics from *A.thaliana*, a strongly structured population.

To compare our method with existing techniques, we considered the standard Lasso, which models all SNPs jointly but without correcting for population structure, as well as univariate LMM, which effectively control for confounding, but consider each SNP in isolation. As a baseline, we also considered a standard univariate linear model (LM), which neither accounts for confounding nor considers joint effects because of complex genetic architectures. Both, the standard Lasso and LMM-Lasso were fit in identical ways (see Section 3.5). For LMM and the LMM-Lasso, we used the RRM as covariance matrix and fit δ on the null model. For univariate models, the ranking of individual SNPs was done according to their *P*-values; for multivariate models, we considered the order of inclusion into the model. A fair comparison between the univariate and multivariate methods is difficult, as the univariate methods select blocks of linked markers, whereas the multivariate methods select only representative markers per block. The evaluation criterion used here accounts for these subtleties (see Supplementary Text S1, Section 1).

LMM-Lasso ranks causal SNPs higher than alternative methods First, we compared the alternative methods in terms of their accuracy in recovering SNPs with a true simulated association (Fig. 1a). Methods that account for population structure (LMM-Lasso, LMM) are more accurate than their counterparts, with LMM-Lasso performing best. Although the linear mixed model performs well at recovering strong associations, the independent statistical testing falls short in detecting weaker associations that are likely masked by stronger effects (Supplementary Fig. S2a). Comparing methods that account for population structure and naïve methods, we observe that accounting for this confounding effect avoids the selection of SNPs that merely reflect relatedness without a causal effect (Supplementary Fig. S2b). An alternative evaluation, which considers the receiver operating characteristic curve, given in Figure 1b, yields identical conclusions.

Next, we explored the impact of variable simulation settings. As common in the literature, we used the area under the precision-recall curve as a summary performance measure to compare different algorithms. Precision and recall both depend on the decision threshold, above which a marker is predicted to be activated. By varying this threshold, one obtains a precision-recall curve.

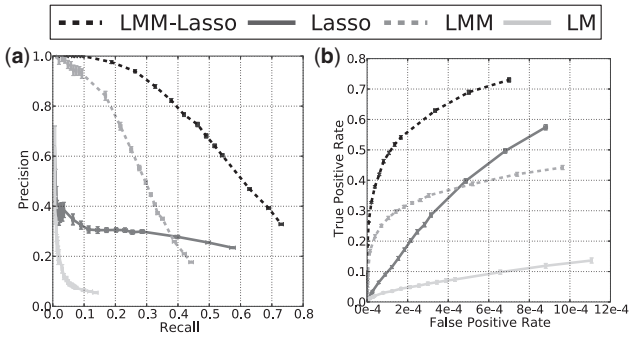

Fig. 1. Evaluation on semi-empirical genome-wide association study (GWAS) datasets, mimicking population structure as found in *A.thaliana*. **(a)** Precision-recall curve for recovering simulated causal SNPs using alternative methods. Shown is precision [TP/(TP + FP)] as a function of the recall [TP/(TP + FN)]. **(b)** Alternative evaluation of each method on the identical dataset using receiver operating characteristics (ROC). Shown is the true-positive rate (TPR) as a function of the false-positive rate (FPR)

Figure 2a shows the area under the precision recall curve as a function of an increasing ratio of population structure and independent environmental noise. When the confounding population structure is weak, both the Lasso and the LMM-Lasso perform similarly. As expected, the benefits of population structure correction in LMM-Lasso are most pronounced in the regime of strong confounding. We also examined the ability of each method to recover genetic effects for increasing complexities of the genetic model, varying the number of true causal SNPs while keeping the overall genetic heritability fixed (Fig. 2b). LMM-Lasso performs better than alternative methods for the whole range of considered settings, with the difference in accuracy being the largest for genetic architectures of medium complexity. These results show that, in the regime of a larger number of true weak associations, it is advantageous to include a genetic covariance **K** that accounts for some of the weak effects (Yang *et al.*, 2010). The identical effect is observed when varying the ratio between true genetic signal versus confounding and noise (Fig. 2c). Again, the performance of the LMM-Lasso is superior to all other methods, and the strengths are particularly visible for medium signal to noise ratios.

4.2 LMM-Lasso explains the genetic architecture of complex traits in model systems

Having shown the accuracy of LMM-Lasso in recovering causal SNPs in simulations, we now demonstrate that the LMM-Lasso better models the genotype-to-phenotype map in *A.thaliana* and mouse (Valdar *et al.*, 2006). Here, we focus on 20 flowering time phenotypes for *A.thaliana*, which are well characterized, and 273 mouse phenotypes, which are relevant to human health.

LMM-Lasso more accurately predicts phenotype from genotype and uncovers sparser genetic models First, we considered

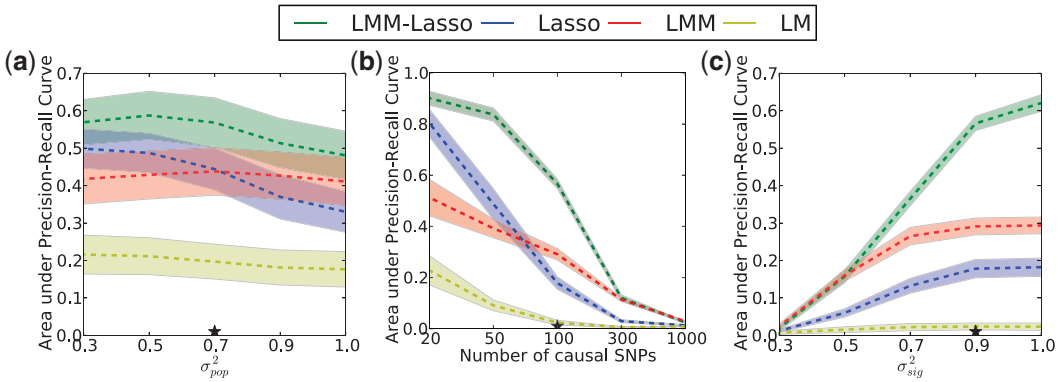

Fig. 2. Evaluation of alternative methods on the semi-empirical GWAS dataset for different simulation settings. Area under precision-recall curve for finding the true simulated associations. Alternative simulation parameters have been varied in a chosen range. **(a)** Evaluation for different relative strength of population structure σ^2_{pop}. **(b)** Evaluation for true simulated genetic models with increasing complexity (more causal SNPs). **(c)** Evaluation for variable signal to noise ratio σ^2_{sig}

phenotype prediction to investigate the capability of alternative methods to explain the joint effect of groups of SNPs on phenotypes. To measure the predictive power, we assessed which fraction of the total phenotypic variation can be explained by the genotype using different methods (Ober *et al.*, 2012). Explained variance is defined as the fraction of the total variance of the phenotype that can be explained by the model and in our experiments equals one minus the mean squared error, as we preprocessed the data to have zero mean and unit variance. We avoided prediction on the training data, as for all methods, this leads to anti-conservative estimates of variance explained because of overfitting (see Supplementary Fig. S4 for a comparison).

Figure 3a and b show the explained variance of the two methods on the independent test dataset for each phenotype in the two datasets. For both model organisms, LMM-Lasso explained at least as much variation as the Lasso. We omitted the univariate methods, as their performance is generally lower because of the simplistic assumption of a single causal SNP (see Supplementary Fig. S4 for comparative predictions in *A.thaliana*). In a fraction of 85.00% of the *A.thaliana* and 91.58% of the mouse phenotypes, LMM-Lasso was more accurate in predicting the phenotype, and thus explained a greater fraction of the phenotype variability from genetic factors than the Lasso. In contrast, Lasso achieved better performance in only 15% of the *A.thaliana* and 8.42% of the mouse phenotypes. Beyond an assessment of the genetic component of phenotypes, LMM-Lasso dissects the phenotypic variability into the contributions of individual SNPs and of population structure. Figure 3c and d show the number of SNPs selected in the respective genetic models for prediction. With the exception of two phenotypes, LMM-Lasso selected substantially fewer SNPs than the Lasso, suggesting that the Lasso includes additional SNPs into the model to capture the effect of population structure through an additional set of individual SNPs. This observation is in line with the insights derived from the simulation setting where the majority of excess SNPs selected by Lasso are indeed driven by population effects (Supplementary Fig. S2b). Although the genetic models fit by LMM-Lasso are substantially sparser, they nevertheless suggest complex genetic

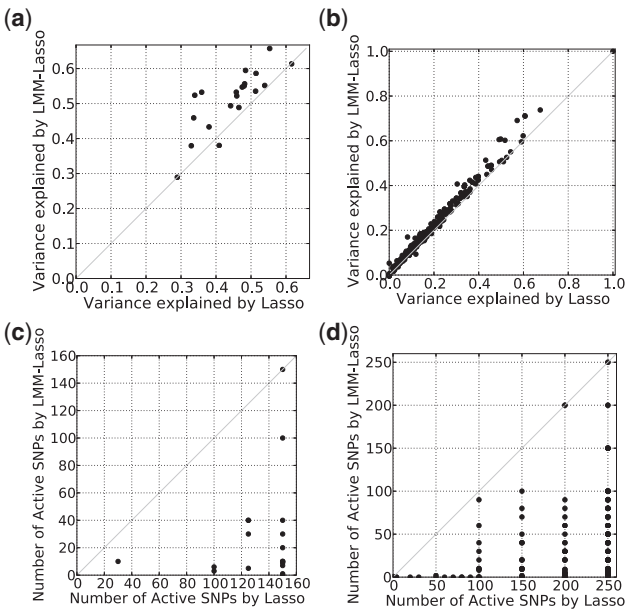

Fig. 3. Predictive power and sparsity of the fitted genetic models for Lasso and LMM-Lasso applied to quantitative traits in model systems. Considered were flowering phenotypes in *A.thaliana* and bio-chemical and physiological phenotypes with relevance for human health profiled in mouse. Comparative evaluations include the fraction of the phenotypic variance predicted and the complexity of the fitted genetic model (number of active SNPs). **(a)** Explained variance in *A.thaliana*. **(b)** Explained variance in mouse. **(c)** Complexity of fitted models in *A.thaliana*. **(d)** Complexity of fitted models in mouse

control by multiple loci. In 90.00% of *A.thaliana* and in 66.06% of the mouse phenotypes, LMM-Lasso selected more than one SNP, in 40.00/45.49% of the cases, the number of SNPs in the model was >10.

LMM-Lasso allows for dissecting individual SNP effects from global genetic effects driven by population structure Next, we investigated the ability of LMM-Lasso to differentiate between

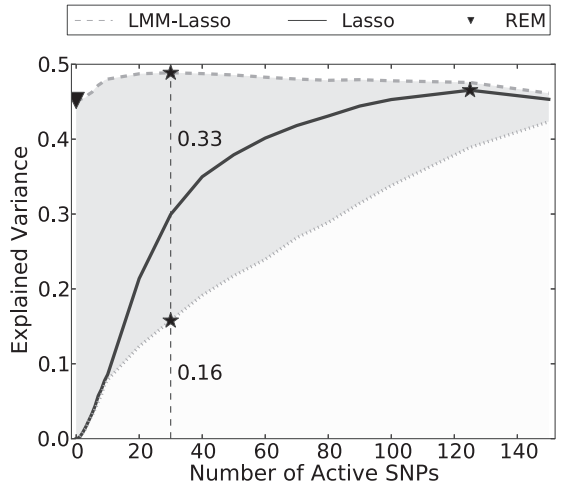

Fig. 4. Variance dissection into individual SNP effects and global genetic background driven by population structure. Shown is the explained variance on an independent test set as a function of the number of active SNPs for the flowering phenotype (10°C) in *A.thaliana*. The predictive test set variance of the Lasso as a function of the number of SNPs is shown in black (solid), the total predictive variance of LMM-Lasso is shown in grey (dashed). The shaded area indicates the fraction of variance LMM-Lasso explains by means of population structure (dark grey) and individual SNP effects (light grey). LMM-Lasso without additional SNPs in the model corresponds to a genetic random effect model (black triangle)

individual genetic effects and effects caused by population structure. Figure 4 shows the explained variances for the phenotype flowering time (measured at 10°C) for *A.thaliana*. Again, these estimates were obtained using a cross validation approach. It is known (Zhao *et al*., 2007) that flowering is strikingly associated with population structure, which explains why the LMM-Lasso already captured a substantial fraction (45.17%) of the phenotypic variance, when using realized relationships alone (number of active SNPs = 0). Because of the small sample size, cross-validation can underestimate the true explained variance (Hastie *et al*., 2003). Nevertheless, cross-validation is fair for comparison and conservative, as it avoids possible overfitting.

For increasing number of SNPs included in the model, the explained variance of LMM-Lasso gradually shifted from the kernel to the effects of individual SNPs. In this example, the best performance (48.87%) was reached with 30 SNPs in the model, where the relative contribution of the random effect model was 33.10% and of the individual SNPs are 15.77%. In comparison, Lasso explained at most 46.53% of the total variance, when 125 SNPs were included in the model.

Associations found by LMM-Lasso are enriched for SNPs in proximity to known candidate genes Finally, we considered the associations retrieved by alternative methods in terms of their enrichment near candidate genes with known implications for flowering in *A.thaliana*. It can be advantageous to remove the SNP of interest from the population structure covariance [see also discussion in Lippert *et al*. (2011)]. Thus, we applied LMM-Lasso on a per-chromosome basis estimating the effect of population structure from all remaining chromosomes.

Table 1. Associations close to known candidate genes[a]

Phenotype	LMM-Lasso	Lasso
LD	**5/54**	4/69
LDV	**5/63**	3/69
SD	**3/55**	2/61
SDV	**5/54**	2/60
FT10	1/48	**4/67**
FT16	3/51	**4/68**
FT22	**2/54**	1/64
2W	**3/53**	2/65
8W	2/51	**4/59**
FLC	**5/52**	3/53
FRI	3/43	3/46
8WGHFT	**4/59**	2/66
8WGHLN	1/48	**4/58**
0WGHFT	**4/58**	3/63
FTField	**4/61**	3/69
FTDiameterField	1/49	1/51
FTGH	1/49	**2/61**
LN10	**3/50**	2/67
LN16	2/58	**3/64**
LN22	**4/54**	2/65

[a]We report true positive/positive results (TP/P) for LMM-Lasso and Lasso for all phenotypes related to flowering time in *A.thaliana*. P are all activated SNPs, and TP are all activated SNPs that are close to candidate genes.
A bold entry indicates that the method finds more TP than its competitor. Explanations for the phenotype abbreviations can be found in the supplement of Atwell (2010).

To obtain a comparable cut-off of significance, we used stability selection for both the LMM-Lasso and Lasso (see Section 3.5).

Table 1 shows that the LMM-Lasso found a greater number of SNPs linked to candidate genes for 12 phenotypes, whereas Lasso retrieved a greater number for only 6 phenotypes. In the remaining two phenotypes, both methods performed identically (for a complete list of candidate genes found by LMM-Lasso, see Supplementary Table S1). We also investigated to what extent the solution is affected by different selection thresholds (see Supplementary Fig. S6). Reassuringly, the LMM-Lasso outperformed the standard Lasso over a large range of different values. It is difficult to compare the multivariate approaches with univariate techniques in a quantitative manner, as the univariate models tend to retrieve complete LD-Blocks. Thus, we revert to reporting the *P*-values of the univariate methods for the SNPs detected by the LMM-Lasso.

We also considered to what extent the findings provide evidence for allelic heterogeneity or the existence of an imperfectly tagged causal locus. Overall, 14.75% of the SNPs linked to candidate genes and selected by the LMM-Lasso appear as adjacent pairs (Supplementary Table S2), that is, having a distance <10 kb from each other, whereas 5.56% of the SNPs selected by the Lasso do. From all activated SNPs, 8.18% selected by LMM-Lasso and 18.96% selected by the Lasso have at least a second active SNP in close proximity. A simulated example, illustrating how the Lasso methods can detect genetic heterogeneity is shown in Supplementary Text S1, Section 3.

5 DISCUSSION

We have presented a Lasso multi-marker mixed model (LMM-Lasso) for detecting genetic associations in the presence of confounding influences, such as population structure. The approach combines the attractive properties of mixed models that allow for elegant correction for confounding effects and those of multi-marker models that consider the joint effects of sets of genetic markers rather than one single locus. As a result, LMM-Lasso is able to better recover true genetic effects, even in challenging settings with complex genetic architectures, weak effects of individual markers or presence of strong confounding effects.

LMM-Lasso is relevant for genome-wide association studies of complex phenotypes, particularly the large number of phenotypes whose genetic basis is conjectured to be multifactorial (Flint and Mackay, 2009). Here, we have demonstrated such practical use through retrospective analysis of *A.thaliana* and data from inbred mouse lines. First, we found that the combination of random effect modelling and multivariate linear models as done in LMM-Lasso improves the prediction of phenotype from genotype, suggesting that the underlying model that accounts for both population structure effects and multi-locus effects is a better fit to real genetic architectures. It is widely accepted that parts of the missing heritability in single-locus genome-wide association mapping can often be explained by a large number of loci that have a joint effect on the phenotype (Yang *et al*., 2010) while leading only to weak signals of association if considered independently. In addition to recovering greater fractions of the heritable component of quantitative traits, LMM-Lasso allows for differentiating between variation that is broad-scale genetic and, hence, likely caused by population structure and individual genetic effects. In *A.thaliana* and mouse, this approach revealed substantially sparser genetic models than naïve Lasso approaches. In addition, LMM-Lasso retrieves genetic associations that are enriched for known candidate genes. In line with the findings in Yang *et al*. (2012), we retrieved an increased rate of physically adjacent SNPs selected in proximity to candidate genes.

Neither the concept of accounting for population structure nor multivariate modelling of the genetic data is novel per se. An approach for distinct populations based on multi-task learning is presented in Puniyani *et al*. (2010). Here, the samples are first divided into populations, and are subsequently assumed to be independent inside the populations. The different populations are then coupled via a shared regularization term.

There is a vast amount of literature using a ℓ_1-regularized approach for genome-wide associations studies (Kim and Xing, 2009; Lee and Xing, 2012; Wu *et al*., 2009). In Foster *et al*. (2007), a sparse random effect model is proposed, in which the markers are modelled as random Lasso effects. In Hoggart *et al*. (2008) and Li *et al*. (2011), the authors suggest adding principal components to the model to correct for population structure. Although these approaches can be effective in some settings, principal components cannot account for family structure or cryptic relatedness (Price *et al*., 2010). Importantly, none of these approaches consider including random effects to control for confounding. A notable exception is the general ℓ_1-mixed model framework by Schelldorfer *et al*. (2011), who consider a random effect component but do not provide a scalable algorithm that is applicable to genome-wide settings.

The proposed model is also closely related to existing mixed models; however, these are predominantly considering individual SNPs in isolation. An exception is the work in parallel of Segura *et al*. (2012), who propose a joint model of multiple large-effect loci in a mixed model using a stepwise regression approach. An important difference to our work is the sequential selection of SNPs, which implies an effect because of ordering, whereas LMM-Lasso selects all SNPs jointly.

As sample sizes increase, the power of detecting multifactorial effects will quickly rise. Moreover, larger datasets improve the feasibility to estimate accurate *P*-values of individual markers by using an approach related to stability selection (Meinshausen *et al*., 2009), which also involves multiple randomized splitting of the dataset.

However, controlling the type 1 error is still not possible when the size of the coefficients is small [more details can be found in Bühlmann (2012)]. Our results suggest that ℓ_1-regularized methods can indeed be an attractive tool for fitting multifactorial effects in genetic settings; however, assessing the statistical significance remains a challenge for future research on Lasso methods in general.

LMM-Lasso addresses the problem that multi-marker mapping is inherently linked to the challenge of some markers being picked up by the model because of their correlation with a confounding variable, such as population structure. In a pure Lasso regression model, it is unclear which markers merely reflect these hidden confounders. LMM-Lasso on the other hand explains confounding explicitly as random effect, and thus, helps to resolve the ambiguity between individual genetic effects and phenotype variability because of population structure. In summary, we, therefore, deem the LMM-Lasso a useful addition to the current toolbox of computational models for unravelling genotype–phenotype relationships.

ACKNOWLEDGEMENT

The authors thank Bjarni J. Vilhjalmsson and Yu Huang for providing the list of genes that are involved in flowering of *A.thaliana* and Nicolo Fusi for preprocessing of the mouse data.

Funding: B.R., C.L. and K.B. were funded by the Max Planck Society. O.S. was supported by a Marie Curie FP7 fellowship.

Conflict of Interest: none declared.

REFERENCES

Atwell,S. *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Bradley,J.K. *et al.* (2011) Parallel coordinate descent for l1-regularized loss minimization. *ICML*, 321–328.

Bühlmann,P. (2012) Statistical significance in high-dimensional linear models. arXiv:1202.1377v2.

Craddock,N. *et al.* (2010) Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.

Flint,J. and Mackay,T.F. (2009) Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res.*, **19**, 723–733.

Foster,S. *et al.* (2007) Incorporating lasso effects into a mixed model for quantitative trait loci detection. *J. Agric. Biol. Environ. Stat.*, **12**, 300–314.

Fusi,N. *et al.* (2012) Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.*, **8**, e1002330.
Goddard,M.E. *et al.* (2009) Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.*, **24**, 517–529.
Hastie,T. *et al.* (2003) *The Elements of Statistical Learning*. Corrected edition. Springer New York Inc., New York, NY, USA.
Hayes,B.J. *et al.* (2009) Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb.)*, **91**, 47–60.
Hoggart,C.J. *et al.* (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.
Horton,M.W. *et al.* (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.*, **44**, 212–216.
Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
Kang,H. *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.
Kim,S. and Xing,E.P. (2009) Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet.*, **5**, e1000587.
Lee,S. and Xing,E.P. (2012) Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs. *Bioinformatics*, **28**, i137–i146.
Lippert,C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.
Listgarten,J. *et al.* (2010) Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 16465–16470.
Li,J. *et al.* (2011) The bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**, 516–523.
Mackay,T.F.C. *et al.* (2009) The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.*, **10**, 565–577.
McCarthy,M. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.*, **72**, 417–473.
Meinshausen,N. *et al.* (2009) P-values for high-dimensional regression. *J. Am. Stat. Assoc.*, **104**, 1671–1681.
Newman,D. *et al.* (2001) The importance of genealogy in determining genetic associations with complex traits. *Am. J. Hum. Genet.*, **69**, 1146–1148.
Ober,U. *et al.* (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet.*, **8**, e1002685.
Platt,A. *et al.* (2010a) Conditions under which genome-wide association studies will be positively misleading. *Genetics*, **186**, 1054–1052.
Platt,A. *et al.* (2010b) The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.*, **6**, e1000843.
Price,A.L. *et al.* (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 459–463.
Price,A. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
Puniyani,K. *et al.* (2010) Multi-population GWA mapping via multi-task regularized regression. *Bioinformatics*, **26**, i208–i216.
Rasmussen,C.E. and Williams,C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA.
Robinson,G. (1991) That blup is a good thing: the estimation of random effects. *Stat. Sci.*, **6**, 15–32.
Sakia,R.M. (1992) The box-cox transformation technique: a review. *Statistician*, **41**, 169.
Schelldorfer,J. *et al.* (2011) Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scand. Stat. Theory Appl.*, **38**, 197–214.
Segura,V. *et al.* (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.*, **44**, 825–830.
Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, **58**, 267–288.
Valdar,W. *et al.* (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.*, **38**, 879–887.
Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.
Xiang,Z.J. *et al.* (2011) Learning sparse representations of high dimensional data on large scale dictionaries. In Shawe-Taylor,J. *et al.* (ed.) *Advances in Neural Information Processing System*. Vol. 24, Granada, Spain, pp. 900–908.
Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.
Yang,J. *et al.* (2012) Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, **44**, 369–375.
Yu,J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Methods*, **38**, 203–208.
Zhang,Z. *et al.* (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, **42**, 355–360.
Zhao,K. *et al.* (2007) An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.*, **3**, e4.
Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.