An Efficient Genome-Wide Multilocus Epistasis Search

Hanni P. Kärkkäinen,* Zitong Li,*,† and Mikko J. Sillanpää*,†,1
*Department of Biology and Biocenter Oulu, and †Department of Mathematical Sciences, University of Oulu, Oulu FIN-90014, Finland

ABSTRACT There has been a continuing interest in approaches that analyze pairwise locus-by-locus (epistasis) interactions using multilocus association models in genome-wide data sets. In this paper, we suggest an approach that uses sure independence screening to first lower the dimension of the problem by considering the marginal importance of each interaction term within the huge loop. Subsequent multilocus association steps are executed using an extended Bayesian least absolute shrinkage and selection operator (LASSO) model and fast generalized expectation-maximization estimation algorithms. The potential of this approach is illustrated and compared with PLINK software using data examples where phenotypes have been simulated conditionally on marker data from the Quantitative Trait Loci Mapping and Marker Assisted Selection (QTLMAS) Workshop 2008 and real pig data sets.

KEYWORDS epistasis; sure independence screening; multilocus association model; extended Bayesian LASSO; genome-wide data

GIVEN the fast development of high-throughput laboratory and statistical marker imputation techniques, the currently available number of markers in genome-wide association studies or genomic prediction studies can be tens of millions of markers (Georges 2014; 1000 Genomes Project Consortium 2012). Screening through these enormous sets using single-locus association techniques may be time-consuming because of slowed computation times when correcting the data with respect to cryptic relatedness (Kang *et al.* 2010). The number of markers to be analyzed becomes ultrahigh dimensional if the set of markers is extended to include all pairwise locus-by-locus interaction terms as pseudo-markers (Sillanpää 2009; Li and Sillanpää 2012). In fact, this creates a tradeoff between execution time and accuracy in breeding-value estimation (Hu *et al.* 2011).

Use of multilocus association models to analyze genome-wide associations or to estimate genomic breeding values necessitates variable selection. This is often done using Bayesian variable selection methods (O'Hara and Sillanpää 2009). Computation times required to estimate parameters in these models are generally much higher than when single-locus models are used. Therefore, one may prefer to apply faster maximum a posteriori probability (MAP) estimation tools rather than Markov chain Monte Carlo (MCMC) techniques (Kärkkäinen and Sillanpää 2012a; Knürr *et al.* 2013).

To reduce the dimension of this problem and to make it more regularized (possibly with a higher rank and reduced multimodality) before variable selection, a sure independence screening method (Fan and Lv 2008) has been proposed. In this method, a subset (say, 5000–10,000) of best-ranked marginal associations is selected, and this marker subset is subjected to Bayesian variable selection using modern variable selection methods (Kärkkäinen and Sillanpää 2012a; Knürr *et al.* 2013). Even though this strategy seems to work efficiently in practice, there are still two camps: groups accepting this approach (*e.g.*, Kärkkäinen and Sillanpää 2012a; Knürr *et al.* 2013) and groups proposing even fancier MCMC or MAP estimation algorithms of QTL effects from even larger original marker sets (*e.g.*, Peltola *et al.* 2012; Gao *et al.* 2013). The common problem of the latter approaches is the prolonged execution time and thus difficulty in monitoring the convergence of the algorithms in the ultrahigh-dimensional space.

In this paper, we illustrate with real-world data sets and simulated genetic architectures that we are able to map large parts of QTL with epistatic genetic architectures using our own specification of sure independence screening and extended

doi: 10.1534/genetics.115.182444
Manuscript received December 8, 2014; accepted for publication September 15, 2015; published Early Online September 23, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.182444/-/DC1
[1]Corresponding author: Departments of Mathematical Sciences and Biology, P.O. Box 3000, University of Oulu, Oulu FIN-90014, Finland. E-mail: mjs@rolf.helsinki.fi

Bayesian LASSO (Kärkkäinen and Sillanpää 2012b). The sizes of the illustrated problems are originally of order of 280 million markers or even more.

Materials and Methods

Conventional single-locus epistasis model

Single-locus approaches have been used widely for estimating the epistasis effects of both quantitative and binary traits (Wei *et al.* 2014). A standard single-locus model to search pairwise G × G interactions is defined by

$$y_i = \beta_0 + \beta_j x_{ij} + \beta_k x_{ik} + \beta_{jk} x_{ij} x_{ik} + e_i \qquad (1)$$

where y_i denotes the phenotypic value of the n individuals, x_{ij} or x_{ik} is the genotype value of individual i of marker j or k ($j, k = 1, \ldots, p, j \neq k$, where p represents the total number of loci), β_0 is the population intercept, β_j and β_k are the main effects of the loci j and k, β_{jk} is the pairwise interaction effect of loci j and k, and e_i corresponds to the residuals, assumed to be independently Gaussian distributed as $e_i \sim \mathcal{N}(0, \sigma_e^2)$. The marker genotypes are coded so that 1 and −1 refer to the common and rare homozygotes, respectively, while 0 refers to the heterozygote. Based on the model, one can calculate the *t*-statistic and the corresponding *P*-value of the interaction effect β_{jk} as evidence for declaring the significance. The approach has been implemented in several software tools for genome-wide association mapping such as PLINK (Purcell *et al.* 2007).

The multilocus method

Complex traits are often controlled by multiple genes and interactions among them, and such a process cannot be adequately described by a single-locus model. A multilocus method that simultaneously estimates the additive effects of multiple SNPs in one computational procedure may better mimic the true genetic mechanism under a complex trait.

A multilocus Gaussian association model is defined by

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \qquad (2)$$

where y_i denotes the phenotypic value of the n individuals, x_{ij} is the genotype value of individual i of marker j, β_0 is the population intercept, and e_i corresponds to the residuals, assumed to be independently Gaussian distributed as $e_i \sim \mathcal{N}(0, \sigma_e^2)$.

When the number of markers p is larger than the number of individuals n, Equation (2) becomes an oversaturated model, so variable selection and shrinkage estimation are required to provide a valid and unique solution to the equation. We have selected a generalized expectation-maximization (GEM) version of the extended Bayesian LASSO (EBL) to perform the shrinkage estimation [see Mutshinda and Sillanpää (2010) for the EBL and Kärkkäinen and Sillanpää (2012b) for the GEM algorithm). Following common Bayesian LASSO (*e.g.*, Park and Casella 2008), the EBL sets a Laplace prior density for the p marker effects β_j in a linear Gaussian association model. Similar to the hierarchical Bayesian LASSO, the Laplace prior is expressed as a scale mixture of Gaussian densities with an exponential mixing distribution. A Gaussian prior with independent locus-specific variances is assigned to the marker effects $\beta_j|\sigma_j^2 \sim \mathcal{N}(0, \sigma_j^2)$, and a further exponential prior is assigned to the variances σ_j^2. However, under the EBL, the variances σ_j^2 have independent exponential prior densities $\sigma_j^2|\lambda_j^2 \sim \mathrm{Exp}(\lambda_j^2/2)$, where the regularization or LASSO parameters λ_j^2 are locus specific. The LASSO parameter is divided into two parts; *i.e.*, $\lambda_j^2 = \delta^2\eta_j^2$, where δ^2 is the model sparsity common to all loci, and η_j^2 is a locus-specific deviation representing the shrinkage at locus j. Gamma(κ, ξ) and Gamma(φ, υ) hyperpriors are given for the components of the LASSO parameters δ^2 and η_j^2, respectively. The rate parameters of the Gamma densities ξ and υ affect the model sparsity and need to be tuned to a data-specific value. The shape parameters κ and φ are set to unity. The population intercept and the residual variance are given uninformative prior densities $p(\beta_0) \propto 1$ and $p(\sigma_e^2) \propto 1/\sigma_e^2$, respectively. The parameter estimation is performed by a GEM algorithm that finds the maximum point of the joint posterior density by updating the parameters, one at the time, to the expected values of the fully conditional posterior densities (Kärkkäinen and Sillanpää 2012a, b).

Model (2) can be extended to include the pairwise interaction terms, which are defined by

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k<1}^{p} \beta_{kl} x_{ik} x_{il} + e_i \qquad (3)$$

where β_{kl} represents the pairwise interaction effect of marker pair (k, l). Recoding the indexes of Equation (3) can lead to an expression that is similar to that in Equation (2). Alternatively, the interaction terms can be interpreted as pseudo-markers, and the interaction effects can be estimated simultaneously with the main effects (Li and Sillanpää 2012).

The critical point in the estimation is managing the ultra-high number of explanatory variables (*i.e.*, the dimension of the model). The number of possible interactions between marker loci will get wildly out of hand very rapidly with increasing marker sets: already with merely 1400 marker loci, there are 1 million pairwise interactions; with 50,000 markers, the number of interactions exceeds 1 billion. It is unlikely that any multilocus model could handle such a number of variables without any dimensional reduction.

Our proposed strategy for efficiently searching G × G interactions

Our proposed approach is based on a combination of preselection of the variables and a Bayesian multilocus association model. The dimensions of the model are first reduced by selecting a predetermined number of interactions based on their marginal correlation with the trait value [sure independence screening by Fan and Lv (2008)]. Similar strategies have been applied for main effect QTL detection in association

studies (Li *et al.* 2011; Fang *et al.* 2014). Originally, the sure independence screening is performed by multiplying a $p \times n$ matrix of genotypes by an n-vector of trait values and selecting d largest values of a p-vector of correlation coefficients, leading to computational complexity $O(np)$. However, because of the overwhelming number of interaction terms (p^2), we compile the pseudomarkers and compute the correlations one at the time, saving only the d currently most highly correlated ones. This procedure increases the computation time because the d-vector has to be sorted during each round, but it decreases memory usage to a feasible level because the $p^2 \times n$ pseudomarker matrix is not held in memory. The preselection by sure independence screening is expected to select a set of variables that includes all the important ones to the actual multilocus model, while the multilocus model performs the final pruning and estimates the marker effects. The number of variables selected for the multilocus analysis depends on the number of individuals in the data set. Generally, the maximum number is assumed to be 10 times the number of individuals (Hoti and Sillanpää 2006), but in our experience, an even smaller proportion may be optimal (Kärkkäinen and Sillanpää 2012a).

We argue that when markers and pseudomarkers are simultaneously subjected to variable selection, the procedure may erroneously favor single-interaction effects instead of selecting two main effects owing to the strong prior toward a small number of trait-associated terms in the model [see example analysis in Sillanpää (2009)]. Thus, to prevent the interaction effects from masking the main effects, we begin by estimating only the main effects, after which the interaction effects are estimated from the residuals. Because the two-step approach consists of two separate estimation procedures, it also enables a higher total number of variables to be included into the multilocus analysis.

The procedure can be summarized as follows:

1. Use sure independence screening to select the markers most correlated with the phenotype (this step is obsolete if marker number is low compared to the number of individuals).
2. Estimate the main effects with a multilocus model.
3. Residual = phenotype − sum of the estimated main effects.
4. Use sure independence screening to select the pseudomarkers most correlated with the residual.
5. Estimate the interaction effects with a multilocus model using the residual as the response variable.

The Matlab codes for simulating the phenotype data and implementing the method are provided in the Supporting Information, File S1.

Data sets

The first set of genotypes originates from simulated data introduced in the XII Quantitative Trait Loci Mapping and Marker Assisted Selection (QTLMAS) Workshop 2008 (Lund *et al.* 2009; Crooks *et al.* 2009). The data set consists of 5865 individuals from seven generations of half-sib families, with information on 6000 biallelic SNP loci evenly distributed over six chromosomes of length 100 cM each.

The second data set consists of real genotypes of 933 F_2 pigs from a White Duroc × Erhualian intercross (Ma *et al.* 2013), genotyped on the Illumina Porcine SNP60 Beadchip. After removing markers with missing genotypes or a minor allele frequency (MAF) < 0.05, the number of markers in the analyses was 23,491. The genotype data are available at the Dryad repository (http://dx.doi.org/10.5061/dryad.7kn7r).

In both cases, the genetic architecture of the trait and the corresponding phenotype data were simulated based on the genotypes. We created 50 QTLMAS and 10 pig data sets. Each simulation began by randomly selecting 10 marker loci with a MAF > 0.4 to act as causal variants. The high MAF limit was introduced to produce detectable interaction effects. Two of the randomly selected loci were set to have main effects only, while two had both main and mutual interaction effects, and the remaining six loci had only locus-by-locus interaction effects, resulting a total of eight genetic effects (four main and four interaction effects). The genetic effects then were drawn from a Gamma density with shape 4 and scale 0.5. The phenotypic value was constructed as the sum of the genetic effects and a random residual drawn from a normal distribution with mean zero and variance set to produce heritability 0.5 within the QTLMAS data sets and 0.8 within the pig data sets. Note that the causal loci were excluded from the QTLMAS marker set but were included in the pig data given to the EBL.

Data availability

File S2 contains SNP data originally simulated in the QTLMAS 2008 workshop.

Example Analyses/Results

In the QTLMAS genotype set, the number of markers was roughly the same as the number of individuals, and there was no need for preselection (the first of the preceding steps was obsolete). Hence, all the markers with a MAF > 0.05 (total of 5702) were included in the main effects EBL (step 2). The number of possible locus-by-locus interaction terms in this case was $5702^2/2 \approx 16$ million, of which 10,000 pseudomarkers were selected for the interaction EBL (steps 4 and 5). For both EBL estimations, the shrinkage-inducing hyperprior parameters ξ and υ were set to 0.1. The hyperprior parameters were selected to yield a reasonable, although not necessarily the best, result by arbitrarily testing several different values.

In the pig genotype set, the number of markers was 25 times the number of individuals, and the marker set was pruned based on the marginal correlation between the markers and the phenotype (step 1). We selected 3000 (representing ~1/8 of the markers) most correlated markers for the main effects EBL (step 2). The number of locus-by-locus interactions possible with the pig data was 280 million, of which 5000 pseudomarkers were selected for the interaction EBL (steps 4 and 5). The hyperprior parameters for both the EBLs were

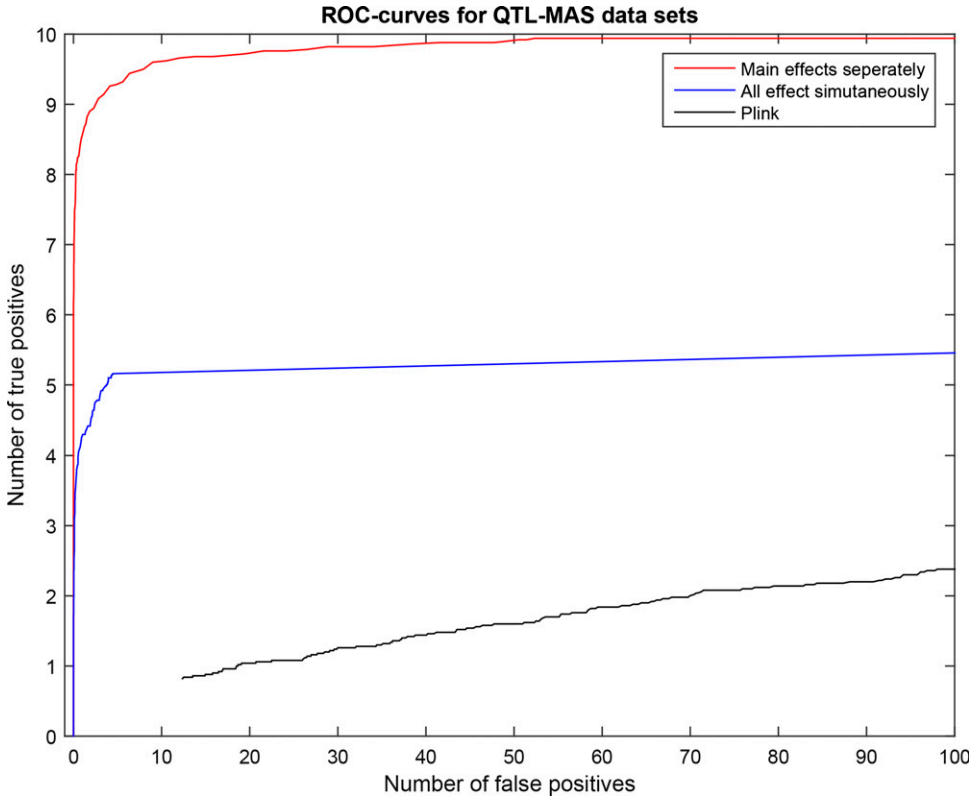

Figure 1 QTLMAS data sets. ROC curves acquired with the proposed method including sure independence screening and the EBL when the main and interaction effects are estimated separately (red) and when all the effects are estimated simultaneously (blue). The black curve corresponds to the standard single-marker-based PLINK analysis for interactions.

set to $\xi = 1$ and $\upsilon = 1$. Additionally, for comparison purposes, the classic single-locus interaction search method was implemented on both data sets using PLINK software. For the PLINK analyses, we included all causative SNPs to the marker sets to be analyzed.

A causal locus was considered to have been correctly identified if one or more signals were reported within a 10-cM window (5 cM on both sides; 1 Mb $\cong$ 1 cM) around a simulated locus. The number of true positives n_{tp} was the number of windows consisting of one or more signals, while the number of false negatives n_{fn} was the number of windows without a reported signal. Signals outside the windows were considered to be false positives n_{fp}. The number of true negatives n_{tn} was the number of markers and pseudomarkers that fell outside the windows around the simulated loci minus the number of false positives. These quantities were used to examine the performance of the methods.

Averaged over the 50 simulated QTLMAS-based data sets, our multilocus method was able to identify 8.72 of the 10 causal loci with a confidence level corresponding to an average number of false positives of <1. Closer examination revealed that, on average, 1.68 of the 2 main effects loci, 5.28 of the 6 interaction effects loci, and 1.76 of the 2 loci with both effects were found. The performance of the method is illustrated in Figure 1 as the red ROC curve (Fawcett 2006). Our proposed approach clearly outperforms the PLINK software in terms of the power to detect true QTL and the ability to exclude false positives.

Regarding to the pig data sets, which are almost 18 times the number in the QTLMAS data sets, our multilocus method detected, on average, 5.50 of the 10 causal loci, consisting 1.30 of the 2 main effects, 2.50 of the 6 interaction effects, and 1.70 of the 2 loci with both effects. At the same time, our method also reported five false-positive loci. The ROC curve presented in Figure 2 illustrates the performance of the method with the 10 pig data sets. As in the QTLMAS data, the performance of our suggested approach was clearly superior to that of the PLINK software.

The computation time of the algorithm, implemented with Matlab v. 7.10.0, was $\sim$40 min for the QTLMAS data and a few hours for the pig data with a 64-bit Windows 7 desktop computer with a 4.20-GHz Intel(i7) CPU and 16.0 GB of RAM. The time-consuming part of the estimation was the sure independence screening for the interaction terms; the rest took only few seconds.

The PLINK analysis, conducted by a 16-core high-performance server, took $\sim$15 hr for the QTLMAS data sets and 52 hr for the pig data sets. This indicates that our suggested approach for epistasis analysis also has certain computational advantages over the conventional single-locus method.

Although the primary motive for this work was to determine whether it is possible to perform a genome-wide interaction search with a multilocus model, there is an additional point of interest concerning estimation of the interaction effects separately from the residual. Based on our numerical experiments, it seems that it is useful to perform the estimation separately for the main and interaction effects. We

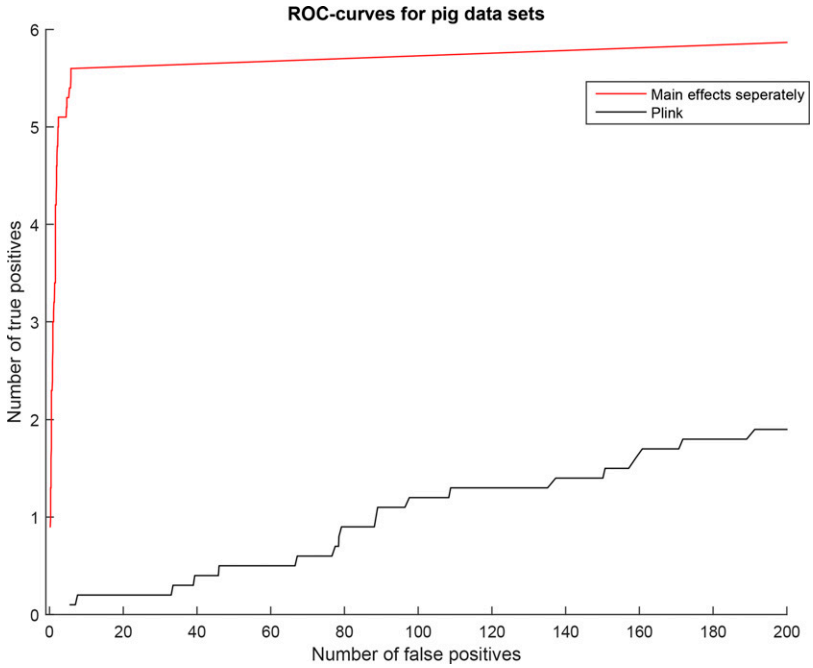

Figure 2 Pig data sets. The ROC curve (red) acquired with the proposed method including sure independence screening and the EBL and estimating the main and interaction effects separately. The black curve corresponds to the standard single-marker-based PLINK analysis for interactions.

tested simultaneous estimation of the main and interaction effects with the 50 QTLMAS data sets. A total of 10,000 variables were selected by sure independence screening for the multilocus analysis based on their correlation with the phenotype. Only 20 of the selected variables were original markers (*i.e.*, selected owing to main effect), while the remaining 9980 were pseudomarkers. As earlier, the hyperprior parameters ξ and υ for the EBL estimation were set to 0.1. With these settings, only 4.34 of the 10 causal loci (*i.e.*, specifically, 0.94 of the 2 main effects loci, 2.38 of the 6 interaction effects loci, and 1.02 of the 2 loci with both effects) were identified. The average performance of the simultaneous estimation is illustrated by the blue ROC curve in Figure 1.

By closely examining the results of the PLINK analysis of both data sets, we found the following interesting phenomenon: when marker *A* and marker *B* are highly correlated with each other, they might be detected as an interaction pair, although, in reality, they should represent a single locus with a main effect instead of two separate loci with an interaction effect (*cf*. Wood *et al.* 2014). This phenomenon partially explain why the number of false positives reported by the PLINK software in our analysis is high. Therefore, our strategy of estimating the interaction effects based on the residual also might help the PLINK software to avoid misclassifying the main and interaction effects. This point needs to be validated with real data sets in future research.

Discussion

Direct application of multilocus association models is questionable with genomic data sets of tens of millions of markers. The situation gets even worse when all pairwise locus-by-locus interaction terms are also included in the model. The prevailing practice in epistasis studies is to consider the interaction terms hierarchically—*i.e.*, only between the loci with significant main effects. Even if such a practice can reduce the number of terms to be considered in the models drastically, it is possible that a trait may exhibit strong pairwise locus-by-locus interaction effects in the presence of negligible main effects (*e.g.*, Frankel and Schork 1996). In this paper, we have considered another strategy, in which all possible pairwise interaction terms are considered, but the efficient dimension-reducing step makes the problem more suitable to multilocus association models so as to find a small subset of important predictors. Even if the dimension-reducing step is huge, reducing an original 280 million discrete predictors to 5000 important candidates, the sure independence screening seems to work surprisingly well in including a significant number of relevant predictors in the chosen subset. This is not a trivial task because it is complicated by the discrete nature of marker data as well as the apparent collinearity among them. The EBL is thus finally picking up the most important few from the 5000 candidates but cannot do anything if those 5000 do not happen to include the important candidates in the first place. Therefore, sure independence screening can be thought to be the most important ingredient here for success. Thus, this initial work hopefully will lead in a direction in which dimension reduction is seen as a necessary first step in genome-wide analysis.

In the case studies, our multilocus approach was evaluated mainly based on empirical evidence. Further effort is needed to make the method truly suitable for practical use. For example, this work simply suggested that a marker was significant if its effect was larger than a certain threshold. In real data analysis, it is often necessary to adopt a methodologically more sound decision rule for QTL judgment. To formally declare significance of a (pseudo) marker in a multilocus

model, phenotype permutation can be used (Xu 2003; Li and Sillanpää 2012). Because necessary reanalyses in phenotype permutation are done only after sure independent screening, the computation time needed may be reasonable. However, because phenotype permutation is conservative, our ability to find a significant association using the EBL and phenotype permutation highly depends on the level of collinearity among markers (which came from the sure independence screening step). A high level of collinearity will increase the tendency of multilocus models to distribute the effect over several neighboring loci, leaving individual signals undetected. High collinearity in the original marker set also may lead to the situation that prescreening will select markers very unevenly—some genomic regions may have too many representatives, and others may have no representation at all. Therefore, it is a future challenge to find the appropriate changes needed for the sure independence screening step to provide a good representation of markers from different trait-associated genome regions and where between-marker dependency is moderate. Some putative solutions could be (1) constraining the prescreening step so that all the included loci must have a correlation that is lower than a given threshold or the number of included loci from a single genomic region is limited or (2) developing multilocus inference methods to be more robust to collinearity [*e.g.*, see Heuven and Janss (2010) and Pasanen *et al.* (2015) for alternative ways to combine/reconstruct the distributed signal over two or more neighboring loci in the MCMC context].

Acknowledgments

We thank the editor and two anonymous reviewers for their valuable comments that improved the manuscript. We are also grateful to Osmo Hakosalo for inspiring discussions. This work was supported by research funding from Biocenter Oulu.

Literature Cited

Crooks, L., G. Sahana, D.-J. de Koning, M. S. Lund, and O. Carlborg, 2009 Comparison of analyses of the QTLMAS XII common dataset. II. Genome-wide association and fine mapping. BMC Proc. 3(Suppl. 1): S2.

Fan, J., and J. Lv, 2008 Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. B 70: 849–911.

Fang, M., W. Fu, D. Jiang, Q. Zhang, D. Sun *et al.*, 2014 A multiple-SNP approach for genome-wide association study of milk production traits in Chinese Holstein cattle. PLoS One 9: e99544.

Fawcett, T., 2006 An introduction to ROC analysis. Pattern Recognit. Lett. 27: 882–891.

Frankel, W. N., and N. J. Schork, 1996 Who's afraid of epistasis? Nat. Genet. 14: 371–373.

Gao, H., G. Su, L. Janss, Y. Zhang, and M. S. Lund, 2013 Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. J. Dairy Sci. 96: 4678–4687.

1000 Genomes Project Consortium, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Georges, M., 2014 Towards sequence-based genomic selection of cattle. Nat. Genet. 46: 808–809.

Heuven, H. C. M., and L. L. G. Janss, 2010 Bayesian multi-QTL mapping for growth curve parameters. BMC Proc. 4(Suppl. 1): S12.

Hoti, F., and M. J. Sillanpää, 2006 Bayesian mapping of genotype × expression interactions on quantitative and qualitative traits. Heredity 97: 4–18.

Hu, Z., Y. Li, Z. Song, Y. Han, X. Cai *et al.*, 2011 Genomic value prediction for quantitative traits under the epistatic model. BMC Genet. 12: 15.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. E. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42: 348–354.

Kärkkäinen, H. P., and M. J. Sillanpää, 2012a Back to basics for Bayesian model building in genomic selection. Genetics 191: 969–987.

Kärkkäinen, H. P., and M. J. Sillanpää, 2012b Robustness of Bayesian multilocus association models to cryptic relatedness. Ann. Hum. Genet. 76: 510–523. [Corrigenda: Ann. Hum. Genet. 77: 275 (2013)].

Knürr, T., E. Läärä, and M. J. Sillanpää, 2013 Impact of prior specifications in a shrinkage-inducing Bayesian model for quantitative trait mapping and genomic prediction. Genet. Sel. Evol. 45: 24.

Li, J., K. Das, G. Fu, R. Li, and R. Wu, 2011 The Bayesian LASSO for genome-wide association studies. Bioinformatics 27: 516–523.

Li, Z., and M. J. Sillanpää, 2012 Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190: 231–249.

Lund, M. S., G. Sahana, D.-J. de Koning, G. Su, and O. Carlborg, 2009 Comparison of analyses of the QTLMAS XII common dataset. I. Genomic selection. BMC Proc. 3(Suppl. 1): S1.

Ma, J., J. Yang, L. Zhou, Z. Zhang, H. Ma *et al.*, 2013 Genome-wide association study of meat quality traits in a White Duroc × Erhualian F2 intercross and chinese Sutai pigs. PLoS One 8: e64047.

Mutshinda, C. M., and M. J. Sillanpää, 2010 Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186: 1067–1075.

O'Hara, R. B., and M. J. Sillanpää, 2009 A review of Bayesian variable selection methods: what, how and which. Bayesian Anal. 4: 85–118.

Park, T., and G. Casella, 2008 The Bayesian LASSO. J. Am. Stat. Assoc. 103: 681–686.

Pasanen, L., L. Holmström, and M. J. Sillanpää, 2015 Bayesian LASSO, scale space and decision making in association genetics. PLoS One 10: e0120017.

Peltola, T., P. Marttinen, and A. Vehtari, 2012 Finite adaptation and multistep moves in the Metropolis-Hastings algorithm for variable selection in genome-wide association analysis. PLoS One 7: e49445.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Sillanpää, M. J., 2009 Detecting interactions in association studies by using simple allele recoding. Hum. Hered. 67: 69–75.

Wei, W.-H., G. Hemani, and C. S. Haley, 2014 Detecting epistasis in human complex traits. Nat. Rev. Genet. 15: 722–733.

Wood, A. R., M. A. Tuke, M. A. Nalls, D. Hernandez, S. Bandinelli *et al.*, 2014 Another explanation for apparent epistasis. Nature 514: E3–E5.

Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. Genetics 163: 789–801.

Communicating editor: G. A. Churchill

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.182444/-/DC1

An Efficient Genome-Wide Multilocus Epistasis Search

Hanni P. Kärkkäinen, Zitong Li, and Mikko J. Sillanpää

DOI: 10.1534/genetics.115.182444

File S1

Matlab codes for implementing the method

Available for download at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.182444/-/DC1