Heredity (2014) 112, 351–360
© 2014 Macmillan Publishers Limited All rights reserved 0018-067X/14
www.nature.com/hdy

ORIGINAL ARTICLE

Combined linkage disequilibrium and linkage mapping: Bayesian multilocus approach

P Pikkuhookana[1,2,3,4] and MJ Sillanpää[2,3,4]

Quantitative trait loci (QTL) affecting the phenotype of interest can be detected using linkage analysis (LA), linkage disequilibrium (LD) mapping or a combination of both (LDLA). The LA approach uses information from recombination events within the observed pedigree and LD mapping from the historical recombinations within the unobserved pedigree. We propose the Bayesian variable selection approach for combined LDLA analysis for single-nucleotide polymorphism (SNP) data. The novel approach uses both sources of information simultaneously as is commonly done in plant and animal genetics, but it makes fewer assumptions about population demography than previous LDLA methods. This differs from approaches in human genetics, where LDLA methods use LA information conditional on LD information or the other way round. We argue that the multilocus LDLA model is more powerful for the detection of phenotype–genotype associations than single-locus LDLA analysis. To illustrate the performance of the Bayesian multilocus LDLA method, we analyzed simulation replicates based on real SNP genotype data from small three-generational CEPH families and compared the results with commonly used quantitative transmission disequilibrium test (QTDT). This paper is intended to be conceptual in the sense that it is not meant to be a practical method for analyzing high-density SNP data, which is more common. Our aim was to test whether this approach can function in principle.

Heredity (2014) **112,** 351–360; doi:10.1038/hdy.2013.111; published online 20 November 2013

Keywords: Bayes; multilocus LDLA analysis; SNP; linkage analysis; linkage disequilibrium

INTRODUCTION

There are two main approaches of finding susceptibility genes (quantitative trait loci, QTL) that influence quantitative traits with the aid of molecular markers. One is variance-component-based linkage analysis (LA), where information comes from recombination events occurring between markers within the pedigree (Blangero and Almasy, 1997; Almasy and Blangero, 1998). The other is linkage disequilibrium (LD) mapping, also known as population-based association mapping, which is based on historical recombination events (George and Elston, 1987; Lander and Schork, 1994). In both analyses, one is interested in finding a signal due to close linkage between a marker and a QTL, but they use different sources of information from the data. LA uses the information that exists within families/pedigrees and follows co-segregation of loci, which is broken down by recombination during few generations of the collected pedigree. Association/LD mapping requires a marker allele to be in considerable LD (*viz.*, non-random allelic association) with a QTL allele across the whole population. In this case, the marker and QTL need to be closely linked in the same chromosome in order for LD to persist after several generations.

LA works with known pedigrees and assumes that loci measured at pedigree founders are in linkage equilibrium (i.e., requiring markers to be unlinked). This assumption is often incorrect. LA also needs extended pedigrees and large family size in fine mapping (Darvasi and Soller, 1995). Even if pedigree or family size can be sometimes very large (e.g., in plants), LA-based fine mapping has a low resolution owing to the limited number of recombinants.

In association/LD mapping, the interest is in finding the statistical association between marker loci and trait values. It relies on the assumption that all alleles that affect the trait are inherited from a single or very few common ancestor(s) (Terwilliger and Weiss, 1998). This analysis needs dense markers in order to find the marker that is very closely linked to the QTL. Another drawback of the association analysis is that significant association may be due to population stratification or other confounding factors rather than LD between a considered marker and a near trait locus (Conti and Witte, 2003; Marchini *et al.*, 2004). Combining these two approaches (i.e., LD and LA) into a single analysis, known as LDLA, yields statistically more powerful and robust analysis because linkage information confirms only real association signals (Hernández-Sánchez *et al.*, 2009). In addition, LDLA could improve the mapping resolution (Meuwissen *et al.*, 2002).

Variance-component-based LA uses locus-specific identity-by-descent (IBD) matrix containing sole linkage information (George *et al.*, 2000). This approach has been generalized to LDLA analysis in animal and plant genetics literature. Such variance-component-based LDLA uses a similar model, where linkage and association information are combined in a single IBD matrix (e.g., Meuwissen *et al.*, 2002; Hernández-Sánchez *et al.*, 2009) by assuming that loci measured at pedigree founders can be in LD (i.e., allowing markers to be linked). This is very different from LDLA analyses of human genetics literature, which use mainly quantitative transmission disequilibrium test (QTDT)-based methods (Abecasis *et al.*, 2000a; Ott *et al.*, 2011)

[1]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland; [2]Department of Biology, University of Oulu, Oulu, Finland; [3]Department of Mathematical Sciences, University of Oulu, Oulu, Finland and [4]Biocenter Oulu, University of Oulu, Oulu, Finland
Correspondence: Professor MJ Sillanpää, Department of Biology and Department of Mathematical Sciences, University of Oulu, P.O.Box 3000, FIN-90014 Oulu, Finland.
E-mail: mjs@rolf.helsinki.fi

Received 22 April 2013; revised 2 September 2013; accepted 27 September 2013; published online 20 November 2013

testing linkage conditional on LD or the other way round. The overview of previous LDLA and LA methods for quantitative traits has been presented in Table 1. See Ott *et al.* (2011) for a review of LDLA methods.

Multilocus association analysis is presumably statistically more powerful than single-locus association testing (Zhang *et al.*, 2011) and can reduce the upward bias of effect estimates (Allison *et al.*, 2002). Multiple testing problems can also be avoided when using multilocus models (Kilpikari and Sillanpää, 2003). Note also that Bayesian multilocus association models without a polygenic term have been found to be more robust to the presence of family and population structure than single-locus association models (e.g., Pikkuhookana and Sillanpää, 2009; Kärkkäinen and Sillanpää, 2012).

In addition to the benefits associated with multilocus model, there are several benefits in our LDLA model framework itself. First, our LDLA model yields separate LA and LD signals in addition to combined LDLA signal. In previous variance-component-based LDLA analyses, in order to separate signal sources, one needs to run extra LD and LA analyses. Second, our LDLA model is free of assumptions about parameters of population model such as effective population size, number of generations since base population and known haplotypes for each individual unlike many previous LDLA approaches (e.g., Meuwissen and Goddard, 2001; Meuwissen *et al.*, 2002; Gasbarra *et al.*, 2009; Hernández-Sánchez *et al.*, 2009). For example, supplying strongly incorrect population parameters in LDLA approaches based on the Wright–Fisher model may be detrimental to LDLA (Hernández-Sánchez *et al.*, 2009). Finally, because our Bayesian LDLA model includes additive genotype effects from multilocus association model and additive linkage effects from variance-component-based LA model, most properties of the original models still hold for this new LDLA model. For example, LA-based IBD matrices needed in our LDLA method can be estimated with existing packages, for example, Merlin (Abecasis *et al.*, 2002), SimWalk2 (Sobel and Lange, 1996), Genehunter2 (Kruglyak *et al.*, 1996), LOKI (Heath, 1997; Thompson and Heath, 1999) or SOLAR (Almasy and Blangero, 1998). Moreover, the proposed Bayesian multilocus LDLA model can provide a firm protection against confounding due to family and population structure, because multilocus association model part (see e.g., Kärkkäinen and Sillanpää, 2012) and LA model part (by definition) are both robust to this problem.

MODELS AND METHODS

We assume that earlier LA or LD studies have already found genetic activity in small chromosomal regions that suggest the presence of QTL (*cf.* Meuwissen and Goddard, 2004; Cantor *et al.*, 2005). Our LDLA approach is designed to be used as a secondary approach to filter best SNPs (single-nucleotide polymorphisms) as putative QTLs from these regions. In our model, the term QTL is used for the locus, which is associated or linked with actual trait locus, and we consider that QTL exists only in positions where we have genotyped SNP markers. Let us have n individuals from independent families with known pedigrees and from the same population. Let N_M be the number of preselected sets of SNP loci, $m = (m_{1,1}, m_{1,2}, \ldots, m_{n,N_M})$ SNP genotypes and $y = (y_1, \ldots, y_n)^T$ the observed continuous phenotypes, where y_i denotes continuous phenotype of the ith individual. We summarize the genotypes as $x_{i,j} = -1$, when genotype is AA homozygote, $x_{i,j} = 0$ for the heterozygote AB and $x_{i,j} = 1$ for the other homozygote. We follow the notation of Pikkuhookana and Sillanpää (2009) and assume that N_M is relatively small (a few hundreds at most). A continuous phenotype y_i is explained with genetic factors using the

Table 1 Overview of previous LA and LDLA methods

Paper	*Model*	*Data needed*	*Experimental design*	*Estimation*	*1 locus/ multilocus*	*IBD matrix*
Meuwissen and Goddard (2004)	LDLA	Haplotypes No. of generations Effective population size	General pedigree	Bayes MCMC	Multilocus	Yes
Lund *et al.* (2003)	LDLA	Haplotypes	Granddaughter design	Classic ML	1 or 2 locus	Yes
Fulker *et al.* (1999)	LDLA/ TDT	Dense map of diallelic markers Sib-pairs with parental genotypes	Sib-pair	Classic	1 locus	No
George *et al.* (1999)	TDT	Pedigree	General pedigree	Classic	1 locus	No
Hernández-Sánchez *et al.* (2009)	LDLA	Genotypes/haplotypes Pedigree No. of generations Effective population size	Any population structure	Classic	1 locus	Yes
Pérez-Enciso (2003)	LDLA	Haplotypes Pedigree	General pedigree	Bayes MCMC	1 locus	No
Abecasis *et al.* (2000b)	TDT	Genotypes	Pedigree	Classic	1 locus	No
Farnir *et al.* (2002)	LDLA	Haplotypes	Half sib	Classic EM	1 locus	No
Göring and Terwilliger (2000)	LDLA	Genotypes	Any population structure	Classic	1 locus	No
Lee and Van der Werf (2006)	LDLA	Haplotypes	General pedigree	Empirical Bayes/stochastic EM	Multilocus	Yes
Yi and Xu (2000)	LA	Genotypes	General pedigree	Bayes	Multilocus	Yes
George *et al.* (2000)	LA	Genotypes	General pedigree	Classic	1 locus	Yes
Lange *et al.* (2013)	LDLA	Pedigree	General pedigree	Classic	1 locus	No

Abbreviations: LDLA, linkage disequilibrium linkage analysis; MCMC, Markov Chain Monte Carlo methods; TDT, transmission disequilibrium test.

following linear LDLA model

$$y_i = \mu + \sum_{j=1}^{N_M} \left(I_j^L \beta_{i,j}^L + I_j^A \beta_j^A x_{i,j} \right) + \varepsilon_i, \tag{1}$$

where μ is the population intercept and $\varepsilon_i \sim N(0, \sigma_0^2)$ is a normally distributed residual term with mean zero and variance σ_0^2. Here, $\beta_{i,j}^L$ is the additive coefficient for the linkage effect of marker *j*. It is an element in multivariate normally distributed effect vector (over individuals *i*) with mean $\tilde{0}$ (vector of zeroes) and covariance matrix $\Pi_j \sigma_{Lj}^2$, that is, $\beta_j^L \mid \Pi_j, \sigma_{Lj}^2 \sim \mathrm{MVN}(\tilde{0}, \Pi_j \sigma_{Lj}^2)$. The additive coefficient for the association effect of marker *j* is β_j^A. The linkage and association indicators for marker *j* are I_j^L and I_j^A, respectively, whose value 1 corresponds to the inclusion, and value 0 to the exclusion of the effect in the model. With restriction $I_j^L = 0$ for all *j*, model (1) corresponds to common multilocus association model. In addition, when restricting $I_j^A = 0$ for all *j*, model (1) reduces to a typical multilocus variance component-based LA model. Association indicators can be collected to a vector $I^A = (I_1^A, I_2^A, \ldots, I_{N_M}^A)$ and unknown association effects to matrix β^A whose diagonal elements are $\beta_1^A, \beta_2^A, \ldots, \beta_{N_M}^A$ and zero elsewhere. Similarly, $\beta_j^L = (\beta_{j,1}^L, \ldots, \beta_{j,n}^L)$, $j = 1, \ldots, N_M$ is the vector of additive (random) genetic effects due to *j*th putative QTL with LA covariance $\Pi_j \sigma_{Lj}^2$. Here, Π_j is the additive IBD matrix of the putative QTL based on LA information. We use Merlin (Abecasis *et al.*, 2002), which uses information from multiple marker loci simultaneously, to estimate multipoint IBD probabilities (p_0, p_1, p_2) at each SNP position for each pair of individuals. However, one can use any package for estimating IBD probabilities, such as LOKI or SOLAR (for a review of common methods, see Mao and Xu, 2005). We can calculate the expected values of elements of the additive IBD matrix using these probabilities as $E(IBD) = (p_1 \times 1 + p_2 \times 2)/2$, where p_1 is the probability that individuals share one allele IBD and p_2 is the probability that individuals share both alleles IBD. The genotypic data of the *j*th SNP marker can be written as vector $x_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,n})$, where $j = 1, \ldots, N_M$, and further as matrix $X = (x_1, x_2, \ldots, x_{N_M})^T$. The incidence matrix *Z* that associates additive linkage effects with phenotypic observations is identified here as an identity matrix $Z = I$. Matrix *Z* may also be other than identity matrix for example in cases where there are individuals in pedigree which have no phenotypic values. Now, we can rewrite the linear LDLA-model (1) as

$$y = \mu + \sum_{j=1}^{N_M} I_j^L Z \beta_j^L + I^A \beta^A X + \varepsilon. \tag{2}$$

To simplify calculations, Mrode and Thompson (1989) and Waldmann *et al.* (2008) applied transformation for covariance (additive genetic relationship) matrix to obtain prior independence structure for transformed variables. The former standardized variance into 1, whereas the latter restored variance in its original scale. Following the similar principle, we introduced transformation for locus-specific IBD matrices. With transformations $F_j = Z_j \Pi_j^{\frac{1}{2}}$ and $\gamma_j = \Pi_j^{-\frac{1}{2}} \beta_j^L$, model (2) can be rewritten as

$$y = \mu + \sum_{j=1}^{N_M} I_j^L F_j \gamma_j + \sum_{j=1}^{N_M} I_j^A \beta_j^A x_j + \varepsilon. \tag{3}$$

With this transformation, we obtain $\gamma_j \mid \sigma_{Lj}^2 \sim \mathrm{MVN}(\tilde{0}, I\sigma_{Lj}^2)$ when $j = 1, \ldots, N_M$. The construction of square-root matrices is shown in Appendix 1.

HIERARCHICAL MODEL

Prior distributions

In Bayesian analysis, one specifies prior distributions for the unknown parameters. For association effects, a prior was assigned as $p(\beta^A \mid \sigma_A^2) = \prod_{j=1}^{N_M} p(\beta_j^A \mid \sigma_{Aj}^2)$. The functional form of $p(\beta_j^A \mid \sigma_{Aj}^2)$ is normal density with mean zero and effect-specific variance σ_{Aj}^2. For the effect-specific variances, Jeffreys' prior was assigned as $p(\sigma_{Aj}^2) \propto 1/\sigma_{Aj}^2$. The same shrinkage-based variable selection as described in Pikkuhookana and Sillanpää (2009) is used here to find sparse trait-associated sets of SNPs. This is a double-shrinkage model with two sources of sparseness (indicator-variables and effect shrinkage). Thus, $P(I^A) = \prod_{j=1}^{N_M} P(I_j^A)$ and $P(I^L) = \prod_{j=1}^{N_M} P(I_j^L)$. Here, $P(I^A = 1)$ represents prior probability of the indicator being one and has been selected to be small. For the effect-specific variances of linkage effects, a Jeffreys' prior was assigned as $p(\sigma_{Lj}^2) \propto 1/\sigma_{Lj}^2$. For the transformed linkage effects, a prior was assigned as $p(\gamma \mid \sigma_L^2) = \prod_{j=1}^{N_M} p(\gamma_j \mid \sigma_{Lj}^2)$, where $p(\gamma_j \mid \sigma_{Lj}^2) = 1 \Big/ \left((2\pi)^{n/2} \mid I\sigma_{Lj}^2 \mid^{1/2}\right) \exp\left[-\frac{1}{2}\gamma_j'(I\sigma_{Lj}^2)^{-1}\gamma_j\right]$. Prior for μ is $p(\mu) \propto 1$ and prior density for $\sigma_0^2 = \mathrm{var}(\varepsilon_i)$ is $p(\sigma_0^2) \propto 1/\sigma_0^2$. Implementational details are provided in Appendix 1.

Missing data model

In the current simulation study, it is assumed that no phenotypic data are missing. Note that SNP data come to our analysis in two forms: (1) as IBD matrices to the LA model and (2) as coded SNP genotypes to the LD model. Thus, it is natural that in the linkage part of the model the method that is used to estimate multipoint IBD probabilities (e.g., Merlin) also handles in its own way the missing values in genotype data (contributing to the linkage signal). In the association part of the model, missing values are handled as random variables via Bayesian inference following Pikkuhookana and Sillanpää (2009). In this process, the genotypic values of the individuals' parents are accounted for. The joint probability distribution of the marker *j* over individuals is given by $p(m_j) = \prod_{i \in \mathrm{Founders}} p(m_{i,j}) \prod_{\substack{i \in \mathrm{non} \\ \mathrm{Founders}}} p(m_{i,j} \mid m_{m,j}, m_{f,j})$, where $m_j = (m_{1,j}, \ldots, m_{n,j})^t$ is the genotype pattern at marker *j*. Transmission probabilities follow the Mendelian rules of inheritance. In our genotype data contributing to the association part of the model, we omit the recombination aspect in handling of missing values. However, linkage between markers has been taken into account in estimating IBD probabilities. Thus, the the prior density function of the genetic data is

$$p(m) \propto \prod_{j=1}^{N_M} \left(\prod_{i \in \mathrm{Founders}} p(m_{i,j}) \prod_{\substack{i \in \mathrm{non} \\ \mathrm{Founders}}} p(m_{i,j} \mid m_{m,j} m_{f,j}) \right)$$

More details have been provided in the Appendix 1.

Posterior distributions

Posterior distributions $p(\theta \mid \mathrm{data})$ for the parameters θ are derived from the likelihood of the data $p(\mathrm{data} \mid \theta)$ and the prior distributions $p(\theta)$ (Gelman *et al.*, 2004). From Bayes formula, we obtain the joint posterior density of parameters from likelihood of the data and the prior distributions as $p(\theta \mid \mathrm{data}) \propto p(\mathrm{data} \mid \theta) p(\theta)$. Here, θ represents all unknown parameters and data represents the given data including phenotypes, genotypes and given IBD matrices. The posterior distributions are estimated using Markov Chain Monte Carlo (MCMC) methods.

EXAMPLES

Data

We use the same genotype data sample of 15 CEPH families as in our previous study (Pikkuhookana and Sillanpää, 2009). Data were edited to remove 5 individuals with missing genotypes for all markers. This was done to minimize the influence of missing data on the simulation of the phenotype and analysis of the association part of the model. Thus, the number of individuals in our data was 205. Our quality control criteria for choosing markers in the study was Hardy–Weinberg equilibrium, minor allele frequency larger than 5% and only a single missing genotype at a marker. Hardy–Weinberg equilibrium was tested with χ^2-test (Balding, 2006) with a P-value threshold of 0.001. We collected 22 SNP markers that fulfilled all the criteria from chromosome 1 (Table 2). Our selection criteria differ from typical cases of few SNP markers such that we concentrated on markers with few missing genotypes rather than certain small areas. Genotypes that were missing in CEPH database were also missing in our analysis.

Simulated data replicates

To assess average performance of the methods and exclude the influence of sampling variation in the analysis results, we simulated 25 replicated phenotypic data sets conditionally on the genotypic data of selected CEPH families. We used WinBUGS 1.4.3 software to simulate data (Spiegelhalter *et al.*,1999; Lunn *et al.*, 2000). Simulated QTL positions, with influence on the phenotypic value, were placed on a few such markers that had no missing genotypes. This way, we avoided making any changes on existing linkage patterns in the genotype data that would have an effect on phenotype. Genotypic effects were simulated on the locus between markers 4 and 5 and were (-4, 1, 6) for genotypes (AA, Aa, aa), respectively, and on marker 16 they were (0, 7, 14) for genotypes (BB, Bb, bb), respectively. Simulated QTL between markers 4 and 5 had a smaller effect on phenotype, and that particular marker was removed from the data before the LDLA analysis. Thus, this QTL is expected to be hard to identify with any mapping methods. Simulated heritability of the data replicates varied from 0.418 to 0.584 with a mean value of 0.503. Moreover, the mean QTL heritability of the left QTL was $\sim$0.064 and that of the right QTL was $\sim$0.305, while the remainder was due to the polygenes. All replicated data sets were analyzed with both our Bayesian LDLA model and quantitative transmission disequilibrium test (Abecasis *et al.*, 2000a, b).

In addition, we simulated other sets of 25 data replicates. Here, the phenotypic value was simulated using the same QTL positions as in a previous experiment. The stronger genotypic effect (-10, 1, 12) was simulated in the locus between markers 4 and 5 for genotypes (AA, Aa, aa), respectively. Similarly as before, this locus was not included in the data that were used in the analysis. The weaker genotypic effect (0, 3, 6) was simulated on marker 16 for genotypes (BB, Bb, bb), respectively. Thus, the mean QTL heritability of the left QTL was $\sim$0.308 and that of the right QTL was $\sim$0.056, whereas overall heritability remained on the same range as in earlier replicated data sets. In the content that follows, we will refer to the above experiments as scenario A (major QTL at position 16) and scenario B (minor QTL at position 16).

Table 2 An overview of selected SNP markers

SNP locus	*NCBI Entrez link*	*location cM*
1	rs594535	4502
2	rs874515	23 752
3	rs726344	33 106
4	rs760594	33 231
QTL	rs1566599	37 481
5	rs905111	37 488
6	rs926830	40 495
7	rs1003323	55 233
8	rs927612	58 007
9	rs556161	61 097
10	rs976574	62 399
11	rs1551331	77 591
12	rs15864	110 094
13	rs930548	112 309
14	rs876694	157 413
15	rs744168	166 361
16-QTL	rs764180	170 067
17	rs725164	170 114
18	rs726427	182 932
19	rs489472	192 344
20	rs1125953	196 235
21	rs1001737	232 484

The name of the SNP on the CEPH database and location on centiMorgans.

Analyses

We analyzed our replicated data sets with the proposed method (Bayesian LDLA) using WinBUGS 1.4.3 for parameter estimation. We used MCMC chain of length 10 000, with burn-in 1000 and thinning of 10; that is, we discarded the first 1000 MCMC samples from the chain and only every 10th MCMC sample was stored and used for estimation. Convergence was assessed by visual inspection of trace plots of different parameters. We summarized our results as the mean posterior occupancy probabilities for SNP j, $P(I_j = 1 \mid \text{data})$, obtained by averaging occupancy probability estimates over 25 replicates. Posterior occupancy probability (or simply QTL-probability) can be calculated as a fraction of MCMC samples where different combinations of linkage and association indicators equal to one. For association $P(I_j^A = 1 \mid \text{data})$, for linkage $P(I_j^L = 1 \mid \text{data})$, for linkage in presence of association $P(I_j^A = 1 \text{ and } I_j^L = 1 \mid \text{data})$ and for linkage or association $P(I_j^A = 1 \text{ or } I_j^L = 1 \mid \text{data})$. Posterior occupancy probability indicates posterior probability that the corresponding SNP effect is included in the model. Mean posterior occupancy probability provides information about average performance. We also calculated Bayes Factor (BF) of association as

$$\mathrm{BF}_j^A = \frac{P\left(I_j^A = 1 \mid \text{ data}\right)/P\left(I_j^A = 0 \mid \text{ data}\right)}{P(I_j^A = 1)/P(I_j^A = 0)}$$

for each marker j (see Kass and Raftery, 1995). Bayes factor of linkage, that of linkage and association, as well as BF of linkage or association (BF_j^L, $\mathrm{BF}_j^{A\cap L}$ and $\mathrm{BF}_j^{A\cup L}$) are calculated in a similar principle. Bayes factor is a useful statistic because the Bayes factor scale is independent of prior odds. The following categories have been suggested for Bayes factors according to the strength of evidence provided by data in favor of 'QTL presence' as opposed to 'no QTL' (see Jeffreys, 1961). The first class is evidence 'not worth more than bare mention' when BF is between 1 and 3. The second class is 'substantial' evidence with BF between 3 and 10. The third class is 'strong' evidence with BF between 10 and 100. The final class is 'decisive' evidence with BF above 100. Before MCMC analysis, we first used the Merlin software to estimate IBD matrices for each loci. Three families had one missing founder. As Merlin demands both founders at each family to be present, the missing founder can be created or the existing one removed. We decided to remove the existing founders in these families. Thus, in the linkage part of the model, the number of individuals in the data is three less than that in the association part of the model.

For comparison of the Bayesian LDLA analyses of the current simulated data replicates, we also analyzed the same data with the QTDT program (Abecasis *et al.*, 2000a, b). The QTDT performs LDLA analysis as a joint analysis of means and covariance matrices by using a maximum likelihood and a common association model for pedigree data. The same estimated IBD matrices were used in QTDT as in the Bayesian LDLA analyses with WinBUGS. Three different models analyzed with QTDT included (1) testing the presence of association, (2) testing linkage and (3) testing linkage while simultaneously considering association.

To understand better how our Bayesian LDLA model works, we analyzed the same replicated data sets (scenario A) also with multilocus association model (with restriction $I_j^L = 0$ for all j in model 1) and with multilocus variance component-based LA model (with restriction $I_j^A = 0$ for all j in model 1). Details of MCMC estimation and the posterior summaries used were identical as above for Bayesian LDLA analysis.

RESULTS

Bayesian LDLA analysis with replicated data sets

To assess the average performance of the Bayesian LDLA analysis, the results were averaged over 25 replicated data sets (Figure 1).

Scenario A. The LD model of Bayesian LDLA on average detected the simulated stronger QTL at position 16 accurately (top left, Figure 1). In the same analysis, the simulated weaker QTL on the left (QTL between markers 4 and 5) did not show on average almost any association signal at its genomic region. When inspecting the average linkage signal provided by the LA model of Bayesian LDLA (second left, Figure 1), posterior occupancy probabilities were slightly higher in a correct region around the simulated weaker QTL on the left. The average combined linkage or association signal from Bayesian LDLA analysis was very strong at the correct QTL position 16, whereas the simulated weaker QTL on the left, at region between position 4 and 5, showed only a minor peak in this combined analysis (bottom left, Figure 1). Again, the average combined linkage and association signal from Bayesian LDLA analysis was very strong at correct QTL position 16, whereas that of the simulated weaker QTL on the left was only modest (third left, Figure 1).

Scenario B. Here, the LD model of the Bayesian LDLA captured the weaker QTL at position 16 very well (top right, Figure 1). As in scenario A, the left QTL showed no association signal. LA model of Bayesian LDLA found the left QTL region (between markers 4 and 5) quite well (second right, Figure 1). It also showed a small peak at position 16. When inspecting combined signals of linkage or association (bottom right, Figure 1), both QTL regions stand out, even if there are elevated levels of QTL probabilities on every position. The combined signal of linkage and association (third right, Figure 1) shows similar results as the combined signal of linkage or association.

Scenarios A and B. To assess sensitivity of Bayes factor thresholds to judge QTLs, the number of replicated Bayesian LDLA analyses exceeding the given BF threshold by using two hypothetical BF levels 1 and 1.5 was illustrated (Figure 2). Almost none (three or less) of the replicated Bayesian LDLA analyses exceeded the BF threshold of 3 in the vicinity of the weaker QTL on the left (scenario A). Thus, we argue that the BF threshold of substantial evidence 3 is too strict for the weakest QTL in these data. This may be partially due to the fact that indicators act as an additional source of shrinkage and may induce a downward bias on the resulting BFs (Sillanpää *et al.*, 2012). On the other hand, when inspecting Figure 2, the BF threshold of 1.5 seems to protect against false positives while still maintaining real combined LDLA signals around the simulated QTL on the left (especially on panels on bottom, Figure 2). In the figure, BF threshold of 1.0 seems to suffer from false positives, and the QTL evidence is not so clear either.

QTDT analysis

The QTDT program can be used to test the presence of association, linkage or both association and linkage simultaneously. Interpretation of the combined QTDT analysis differs from the Bayesian LDLA analysis presented here. Although the Bayesian combined linkage and association method can separate or combine two sources of signals freely (as was shown in Figure 1), signal detection in QTDT is based on comparing separate analyses of association, linkage and linkage given association (Figure 3). If the QTDT signal of linkage given association seems to disappear (as in bottom left, Figure 3) compared with the signal of the linkage model (middle left, Figure 3) or of the association model (top left, Figure 3), the QTDT signal can be interpreted as the signal of true QTL (Fulker *et al.*, 1999).

To assess the sensitivity of *P*-values to judge QTLs, the number of replicated QTDT analyses exceeding the given threshold of *P*-values 0.05 and 0.1 was illustrated (Figure 3).

Scenario A. In the analyses of replicated data, the use of the QTDT model on average detected the simulated stronger QTL at position 16 accurately, with averaged *P*-values varying from $0.2{*}10^{-8}$ to $0.2{*}10^{-19}$. The weaker simulated QTL on the left (in region between markers 4 and 5) was not detected with QTDT too well. Figure 3 clearly illustrates that there is practically no hope to detect the weaker simulated QTL on the left using QTDT. Linkage test found significant *P*-values only twice in marker 3 (*P*-values 0.084 and 0.094) and marker 4 (*P*-values 0.083 and 0.095) and once in marker 5 (*P*-value 0.067) of the 25 replicated data sets.

Scenario B. QTL at position 16 seems to be found on average correctly by association test in QTDT, even if the QTL heritability at that position is smaller in this scenario. Averaged *P*-values vary from 0.06 to $0.4{*}10^{-6}$. The minor QTL at position 16 was not detected by linkage test or subsequent QTDT test of linkage given association (right, Figure 3). Linkage test in QTDT found on average elevated signals around the correct region of major QTL on the left, although this region is quite wide (from SNP 1 to SNP 10). Because this major QTL region on the left showed no signal in the association test, this QTL cannot be detected by QTDT test of linkage given association either.

Restricted analyses

Results of the restricted analyses of scenario A data sets are shown in Figure 4. Multilocus association analysis found correct QTL at position 16 (left, Figure 4). No association signal of the weaker QTL between positions 4 and 5 can be found. Unlike LDLA analyses, multilocus variance component-based LA analysis showed linkage signal around position 16 (right, Figure 4). As a drawback, LA analysis suffers from a false positive at position 20, and its signal at the region of the weaker QTL did not seem to be significantly higher than the general signal level.

DISCUSSION

Owing to the limited number of recombination events, LA seems to find QTL regions fine, but it does not separate true QTL from correlated markers. On the other hand, population-based association analysis has a high resolution, but without including LA it tends to give false positives owing to structured data. Thus, we have proposed a conceptual approach of combining LA and LD methods into the Bayesian multilocus LDLA framework. Previous variance component-based LDLA methods developed in plant and animal genetics literature (Table 1) are relying to a certain extent on known

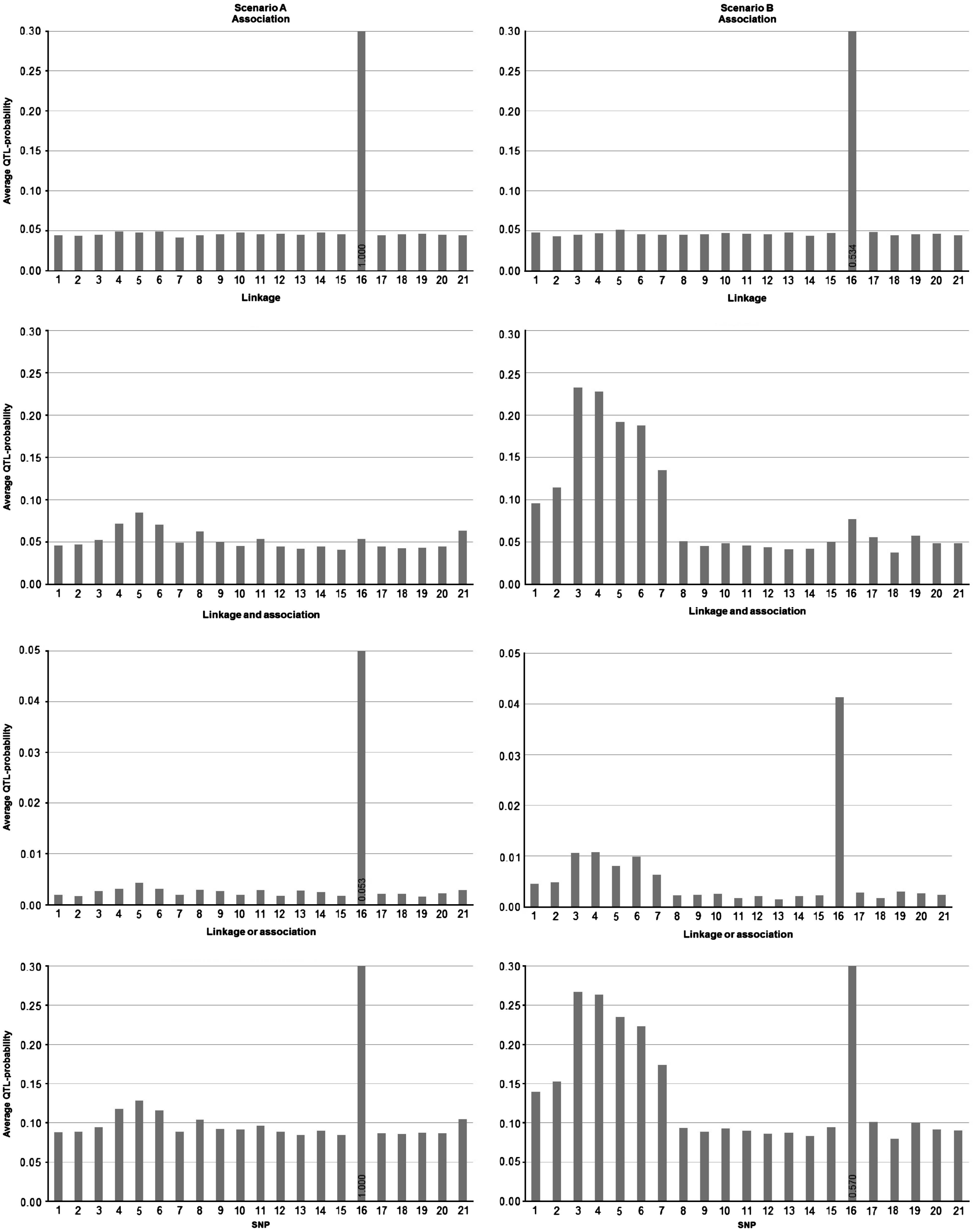

Figure 1 Mean posterior occupancy probabilities for SNPs obtained from Bayesian LDLA analysis by averaging QTL-probability estimates over 25 replicated data sets in scenario A (left) and in scenario B (right). Index of SNPs on x-axis and average QTL probabilities on *y*-axis calculated from indicator variables in the MCMC sample for association (top panel), linkage (second panel), linkage and association (third panel), linkage or association (bottom panel). Note that *y*-axis scales in panels depend on prior occupancy probabilities. In cases where QTL probability exceeds the upper limit of the *y*-axis, its value is presented in the bottom of each column of the bar chart.

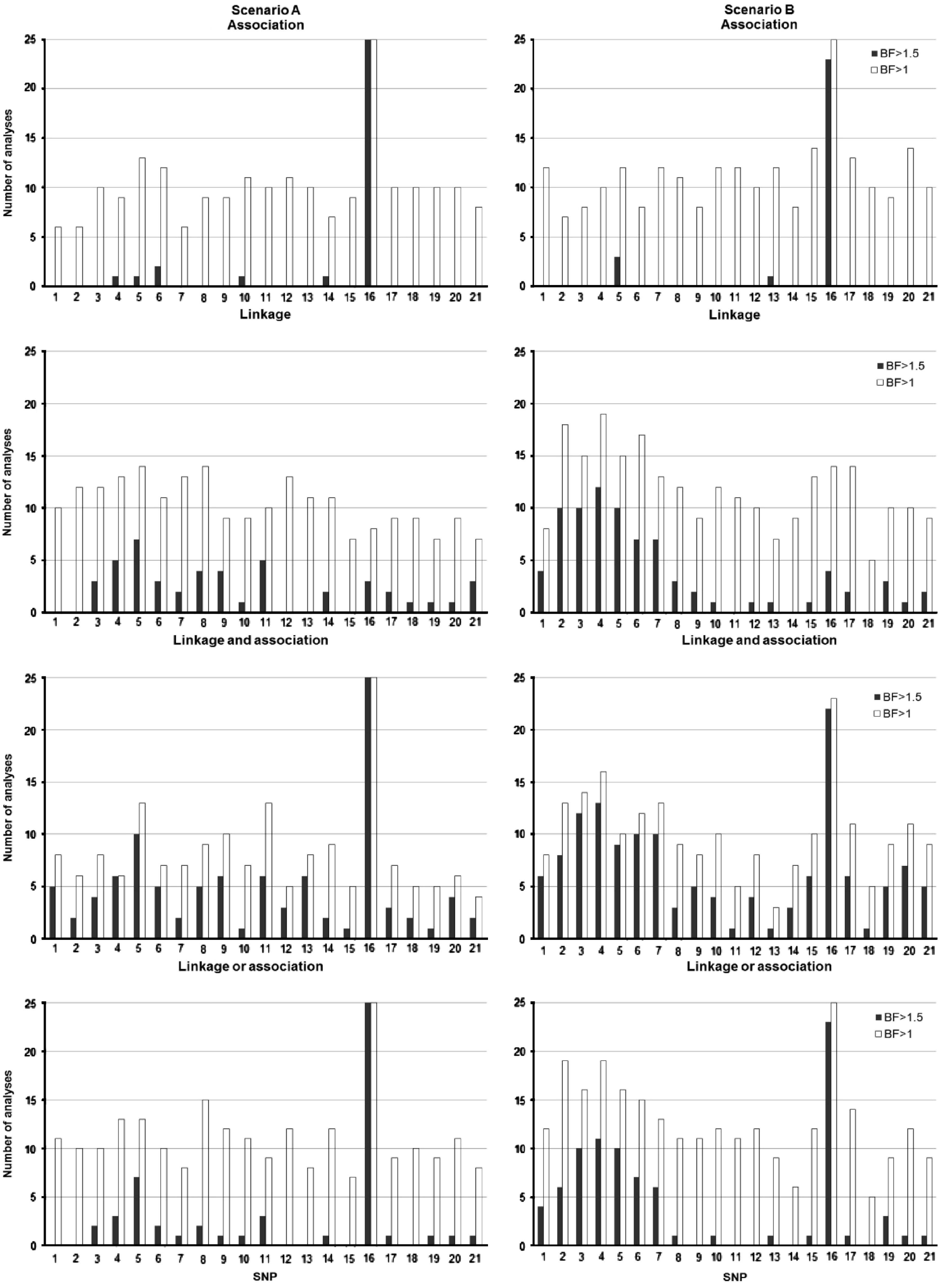

Figure 2 Number of Bayesian LDLA analyses where BF exceeds a given threshold. Index of SNPs on the *x*-axis and the number of analyses where BF exceeds values 1 and 1.5 on the *y*-axis. Scenario A is shown on the left and scenario B on the right. The first panel shows results from association analysis and the second panel from linkage analysis. The third panel shows results when either linkage or association signal is found in the model, and the case when both association and linkage signals are found simultaneously in the model is in the fourth panel.

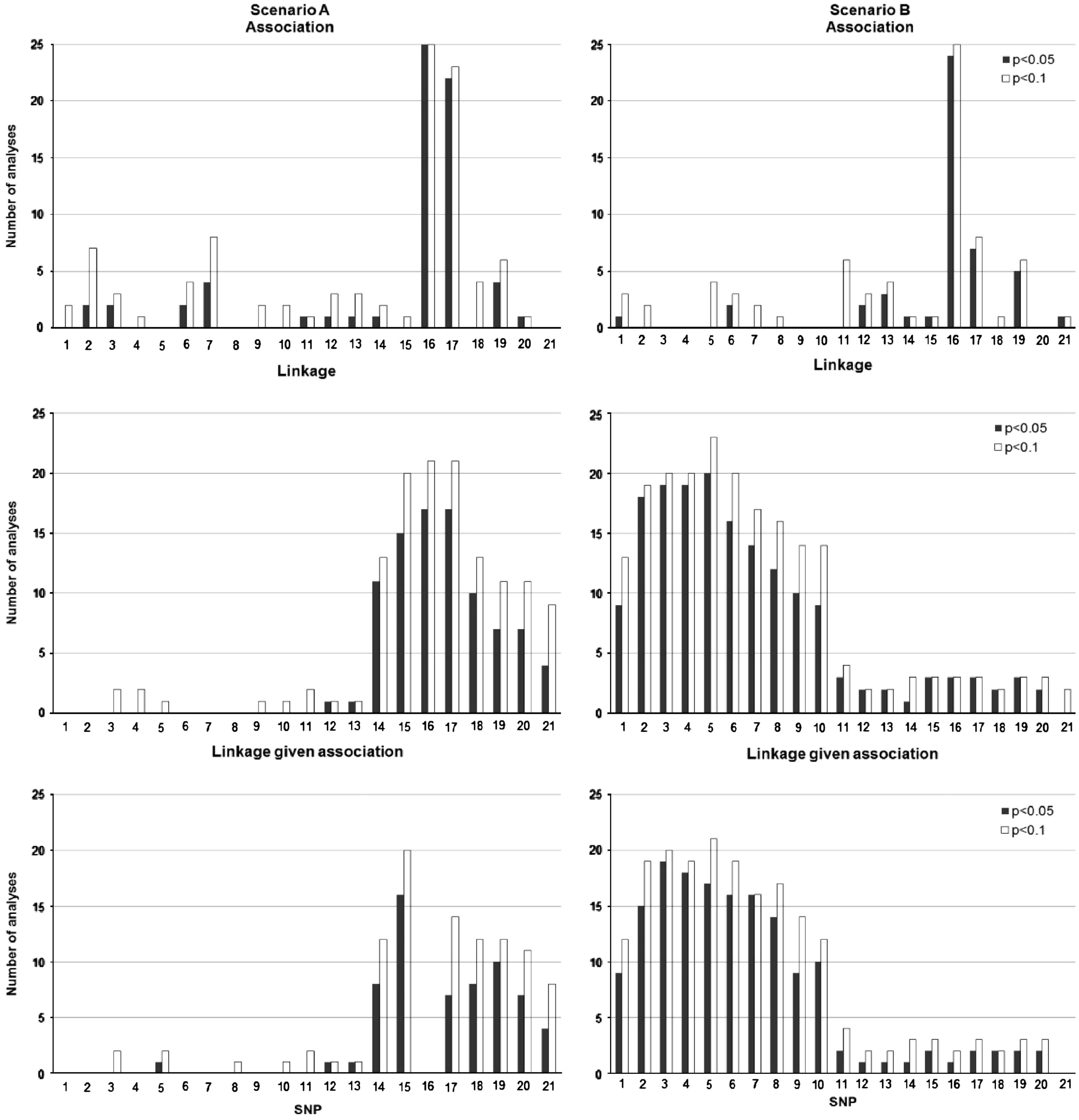

Figure 3 The number of QTDT analyses with significant *P*-values where *P*-value exceeds the given values of *P*-value threshold for significant result of 0.05 and 0.1. Scenario A is shown on the left and scenario B on the right. The panels contain results from sole association analysis (top), sole linkage analysis (middle) and combined association and linkage analysis (bottom).

haplotypes, which are costly/expensive to collect, and/or they need assumptions about population demography (Hernández-Sánchez *et al.*, 2009). Similar to LDLA approaches of human genetics, our method makes fewer such assumptions than previous methods developed in plant and animal genetics literature (Table 1), and it takes advantage of existing packages on IBD estimation. Combining LA and LD methods together provides more accurate analysis. Combined analysis using QTDT can find true simulated position of QTL, whereas sole association and LA gives less accurate results. The weaker-simulated QTL (scenario A) on the left (in region between markers 4 and 5) was not found at all with QTDT. In addition, Bayesian LDLA had difficulties in finding that QTL, but it showed at least a weak signal. The other simulated QTL was found with Bayesian LDLA with probability one. Poor identification of the smaller effect QTL may be due to the fact that the simulated effect on that QTL was much weaker than the effect of the other QTL. Our Bayesian LDLA outperformed QTDT in replicated simulation analysis of scenario A. It showed better power to identify also the smaller effect QTL. In scenario B, QTDT showed either clear linkage signal or clear association signal for each QTL but not both signals for the same QTL. Thus, QTDT test of linkage given association could not detect the QTLs similarly as Bayesian LDLA analysis.

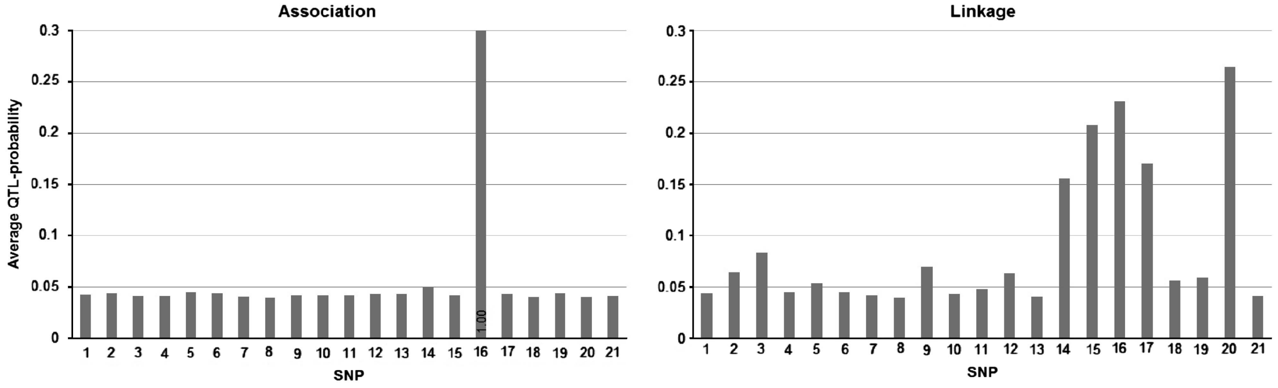

Figure 4 Mean posterior occupancy probabilities for SNPs obtained from the restricted analyses of scenario A data. Sole association analysis is on the left and sole linkage analysis on the right. In cases where QTL probability exceeds the upper limit of the y-axis, its value is presented in the bottom of each column of the bar chart.

The interesting difference between these two models was that the region showing linkage signal around major QTL (between positions 4 and 5) was few SNP markers wider in QTDT than in Bayesian LDLA analysis.

Restricted analyses of scenario A give hint of the reason why Bayesian LDLA model found no linkage signal at position 16 (second left, Figure 1). It seems that this is a consequence of including various kinds of genetic (LA and LD) effects simultaneously to the model and performing Bayesian variable selection to them jointly. Similar behavior has been seen earlier by Bhattacharjee *et al.* (2008) and Pikkuhookana and Sillanpää (2009), where SNP marker information and gene expression information were considered jointly in the model, and the stronger information type alone explained most of the variation in the genomic region. In Bayesian LDLA model with scenario A data, the association signal is so strong that it leaves nothing to be explained for linkage signal (second left, Figure 1). The above kind of ‘winner takes it all’ phenomenon is suggested by analyzing the same data with Bayesian LDLA model, where association indicators are restricted to zero (with restriction $I_j^A = 0$ for all *j*). Here, one can see clear elevated levels of linkage signal around QTL position 16 (right panel, Figure 4) as in QTDT.

Our results are in agreement with previous observations that simultaneous modeling of the association and linkage gives more accurate results than modeling them separately. Generally, analyzing data with the LDLA model does not suffer from spurious associations due to population demography and relatedness, which are the main difficulties of the LD model. This is because LA is robust to structured data and gives good control for the LD part of the model (Sillanpää, 2011). Earlier studies (Pikkuhookana and Sillanpää, 2009; Kärkkäinen and Sillanpää, 2012) have shown that the multilocus association model seems to be robust to cryptic relatedness, and thus our Bayesian multilocus LDLA method introduces robustness on each part of the model.

ELECTRONIC-DATABASE INFORMATION

CEPH genotype database, http://www.cephb.fr/cephdb/

NCBI GenBank, http://www.ncbi.nlm.nih.gov/Genbank/

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are grateful to Mahlako Makgahlela, Hanni P. Kärkkäinen and an anonymous reviewer for their constructive comments on this paper. This work was supported by the Finnish Graduate School of Population Genetics and the research grants from the Academy of Finland, the University of Helsinki’s Research Funds, Biocenter Oulu and Oulun läänin talousseuran maataloussäätiö.

Abecasis GR, Cardon LR, Cookson WO (2000a). A General test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**: 279–292.

Abecasis GR, Cookson WO, Cardon LR (2000b). Pedigree tests of transmission disequilibrium. *Eur J Hum Genet* **8**: 545–551.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002). Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**: 79–101.

Allen AS, Rathouz PJ, Satten GA (2003). Informative missingness in genetic association studies: case-parent designs. *Am J Hum Genet* **72**: 671–680.

Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM *et al.* (2002). Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *Am J Hum Genet* **70**: 575–585.

Almasy L, Blangero J (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**: 1198–1211.

Balding DJ (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**: 781–791.

Bhattacharjee M, Botting CH, Sillanpää MJ (2008). Bayesian biomarker identification based on marker-expression-proteomics data. *Genomics* **92**: 384–392.

Blangero J, Almasy L (1997). Multipoint oligogenic linkage analysis of quantitative traits. *Genet Epidemiol* **14**: 959–964.

Cantor RM, Chen GK, Pajukanta P, Lange K (2005). Association testing in a linked region using large pedigrees. *Am J Hum Genet* **76**: 538–542.

Conti DV, Witte JS (2003). Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet* **72**: 351–363.

Darvasi A, Soller M (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* **141**: 1199–1207.

Farnir F, Grisart B, Coppieters W, Riquet J, Berzi P, Cambisano N *et al.* (2002). Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* **161**: 275–287.

Fulker DW, Cherny SS, Sham PM, Hewitt JK (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* **64**: 259–267.

Gasbarra D, Pirinen M, Sillanpää MJ, Arjas E (2009). Bayesian QTL mapping based on reconstruction of recent genetic histories. *Genetics* **183**: 709–721.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis*, 2nd ed. Chapman and Hall: London.

George V, Tiwari HK, Zhu X, Elston RC (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet* **65**: 236–245.

George AV, Visscher PM, Haley CS (2000). Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156**: 2081–2092.

George V, Elston RC (1987). Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet Epidemiol* **4**: 193–201.

Heath SC (1997). Markov Chain segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.

Göring HH, Terwilliger JD (2000). Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot accurately specified. *Am J Hum Genet* **66**: 1310–1327.

Hernández-Sánchez J, Grunchec J-A, Knott S (2009). A web application to perform linkage disequilibrium and linkage analyses on a computational grid. *Bioinformatics* **25**: 1377–1383.

Hosmer DW, Lemeshow S (1989). *Applied Logistic Regression*. John Wiley & Sons: New York.

Jeffreys H (1961). *Theory of Probability*, 3rd ed. Clarendon Press: Oxford, UK.

Kass RE, Raftery AE (1995). Bayes Factors. *J Am Stat Ass* **90**: 773–795.

Kilpikari R, Sillanpää MJ (2003). Bayesian analysis of multilocus association in quantitative and qualitative traits. *Genet Epidemiol* **25**: 122–135.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**: 1347–1363.

Kärkkäinen HP, Sillanpää MJ (2012). Robustness of Bayesian multilocus association models to cryptic relatedness. *Ann Hum Genet* **76**: 510–523.

Lander ES, Schork NJ (1994). Genetic dissection of complex traits. *Science* **265**: 2037–2048.

Lange K, Cantor R, Horvath S, Papp JC, Sabatti C, Sinsheimer JS *et al*. (2013). *Mendel 13.2 Documentation*. UCLA School of Medicine: Los Angeles, Available at, http://www.genetics.ucla.edu/software/download?file=172

Lee SH, Van der Werf JHJ (2006). Using dominance relationship coefficients based on linkage disequilibrium and linkage with general complex pedigree to increase mapping resolution. *Genetics* **174**: 1009–1016.

Lund MS, Sørensen P, Guldbrandtsen B, Sorensen DA (2003). Multitrait fine mapping in quantitative trait loci using combined linkage disequilibria and linkage analysis. *Genetics* **163**: 405–410.

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000). WinBUGS – A Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* **10**: 325–337.

Mao Y, Xu S (2005). A Monte Carlo algorithm for computing IBD matrices using incomplete marker information. *Heredity* **94**: 305–315.

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004). The effects of human population structure on large genetic association studies. *Nat Genet* **36**: 512–517.

Meuwissen THE, Goddard ME (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genet Sel Evol* **33**: 605–634.

Meuwissen THE, Karlsen A, Lien S, Olsaker I, Goddard ME (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* **161**: 373–379.

Meuwissen THE, Goddard ME (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet Sel Evol* **36**: 261–279.

Mrode R, Thompson R (1989). An alternative algorithm for incorporating the relationships between animals in estimating variance components. *J Anim Breed Genet* **106**: 89–95.

Nicodemus KK, Luna A, Shugart YY (2007). An evaluation of power and type I error of single-nucleotide polymorphism transmission/disequilibrium–based statistical methods under different family structures, missing parental data, and population stratification. *Am J Hum Genet* **80**: 178–185.

Ott J, Kamatani Y, Lathrop M (2011). Family-based designs for genome-wide association studies. *Nat Rev Genet* **12**: 465–474.

Pérez-Enciso M (2003). Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics* **163**: 1497–1510.

Pikkuhookana P, Sillanpää MJ (2009). Correcting for relatedness in Bayesian models for genomic data association analysis. *Heredity* **103**: 223–237.

Rubin DB (1976). Inference and missing data. *Biometrika* **63**: 581–592.

Sillanpää MJ (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* **106**: 511–519.

Sillanpää MJ, Pikkuhookana P, Abrahamsson S, Knürr A, Fries E, Lerceteau P, Waldmann P, Garcia-Gil MR (2012). Simultaneous estimation of multiple quantitative trait loci and growth curve parameters through hierarchical Bayesian modeling. *Heredity* **108**: 134–146.

Sobel E, Lange K (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am J Hum Genet* **58**: 1323–1337.

Spiegelhalter DJ, Thomas A, Best NG (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit: Cambridge, UK.

Terwilliger JD, Weiss KM (1998). Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol* **9**: 578–594.

Thompson EA, Heath SC (1999). Estimation of conditional multilocus gene identity among relatives. Statistics in Molecular Biology and Genetics: selected proceedings of a 1997 joint AMS-IMS-SIAM summer conference on statistics in molecular biology. *IMS Lecture Note-Monograph Series* **33**: 95–113.

Tipping ME (2001). Sparse Bayesian learning and the relevance vector machine. *J Mach Learn Res* **1**: 211–244.

Waldmann P, Hallander J, Hoti F, Sillanpää MJ (2008). Efficient Markov Chain Monte Carlo implementation of Bayesian analysis of additive and dominance genetic variances in noninbred pedigrees. *Genetics* **179**: 1101–1112.

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

Yi N, Xu S (2000). Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics* **156**: 411–422.

Yi N, Xu S (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045–1055.

Zhang F, Guo X, Deng HW (2011). Multilocus association testing of quantitative traits based on partial least-squares analysis. *PLoS One* **6**: e16739.

APPENDIX 1

Handling of the missing values of genotypic data in the association part of the model is given here. We assumed Mendelian transmission, uniform allele frequencies and that genotypes are missing at random (Rubin, 1976). Note that assuming missing data mechanism to be random in TDT is known to be a poor choice, as it tends to inflate linkage signals (e.g., Allen *et al.*, 2003; Nicodemus *et al.*, 2007). However, because the new Bayesian LDLA model has been constructed from two separate (LD and LA) models and assuming random missing data mechanism is not a problem in either of them, most properties of the original models still hold for this new LDLA model. In the WinBUGS software, we created two artificial individuals who are parents of all founders. These individuals are heterozygotes in all their markers allowing missing founder genotypes to have all possible allele combinations. Owing to total probability, the genotypes of the founders are dependent on the genotypes of the rest of the pedigree, creating both downward and upward dependencies.

Construction of diagonalizing transformations for the IBD matrices:

Singular value decomposition on symmetric matrix A is $A = USU'$, where U is the orthonormal matrix ($U'U = I$ and $\det(U) = 1$) and S is the diagonal matrix ($S \ = \mathrm{diag}(s_1, s_2, \ldots, s_n)$). Required square-root matrix is achieved by equation $A^{\frac{1}{2}} = US^{\frac{1}{2}}U'$ and inverse square-root matrix by $A^{-\frac{1}{2}} = US^{-\frac{1}{2}}U'$.

Jeffreys' improper prior $p(\sigma_j^2) \propto 1/\sigma_j^2$ is used to induce sparseness into our model (Xu, 2003). Jeffreys' prior equals Inv-Gamma(0,0) prior, which is also improper (Yi and Xu, 2008). Assuming Inv-Gamma(0,0) prior for effect-specific variance components is a special case of scale mixture parameterization of Student's *t*-distribution, which is assumed marginally for β_j^A (Tipping, 2001). WinBUGS program does not allow the use of improper priors, and thus we approximated the improper prior with a proper one. One solution is to use Inv-Gamma(a,b) prior with a and b close to zero. We have earlier found that this leads to numerical instability producing 'trap messages' in WinBUGS (Pikkuhookana and Sillanpää, 2009). Let $\tau = 1/\sigma^2$ be the precision parameter. We used transformation $\phi = \log(\tau)$ for the precision parameters. Prior $p(\tau) \propto 1/\tau$ leads to flat prior for $p(\phi) = p(\tau)|\frac{\partial\tau}{\partial\phi}| \propto \frac{1}{\tau} \ \tau = 1$ (see Gelman *et al.* 2004, p. 65). Flat prior is also improper, but when we restrict it to some finite range, it will become a proper prior. Flat prior for μ is also improper, and we approximate that with flat normal density with zero mean and large variance. With small modifications, our quantitative model can be used for binary phenotype. Logistic regression (see e.g. Hosmer and Lemeshow, 1989) is a good tool for modeling binary traits.