Published March, 2009

Simulation Appraisal of the Adequacy of Number of Background Markers for Relationship Estimation in Association Mapping

Jianming Yu,* Zhiwu Zhang, Chengsong Zhu, Dindo A. Tabanao, Gael Pressoir, Mitchell R. Tuinstra, Stephen Kresovich, Rory J. Todhunter, and Edward S. Buckler

Abstract

Complex trait dissection through association mapping provides a powerful complement to traditional linkage analysis. The genetic structure of an association mapping panel can be estimated by genomewide background markers and subsequently accounted for in association analysis. Deciding the number of background markers is a common issue that needs to be addressed in many association mapping studies. We first showed that the adequacy of markers in relationship estimation influences the maximum likelihood of the model explaining phenotypic variation and demonstrated this influence with a series of computer simulations with different trait architectures. Analyses and computer simulations were then conducted using two different data sets: one from a diverse set of maize (*Zea mays* L.) inbred lines with a complex population structure and familial relatedness, and the other from a group of crossbred dogs. Our results showed that the likelihood-based model-fitting approach can be used to quantify the robustness of genetic relationships derived from molecular marker data. We also found that kinship estimation was more sensitive to the number of markers used than population structure estimation in terms of model fitting, and a robust estimate of kinship for association mapping with diverse germplasm requires a certain amount of background markers (e.g., 300–600 biallelic markers for the simulated pedigree materials, >1000 single nucleotide polymorphisms or 100 simple sequence repeats [SSRs] for the diverse maize panel, and about 100 SSRs for the canine panel). Kinship construction with subsets of the whole marker panel and subsequent model testing with multiple phenotypic traits could provide ad hoc information on whether the number of markers is sufficient to quantify genetic relationships among individuals.

Published in The Plant Genome 2:63–77. Published 18 Mar. 2009.
doi: 10.3835/plantgenome2008.09.0009

677 S. Segoe Rd., Madison, WI 53711 USA
An open-access publication

ASSOCIATION MAPPING provides a powerful complement to traditional linkage analysis for understanding the genetic basis of complex traits (Lander and Schork, 1994; Risch and Merikangas, 1996; Mackay, 2001; Doerge, 2002; Darvasi and Shifman, 2005). Association mapping is now being performed in many species (Thornsberry et al., 2001; Neale and Savolainen, 2004; Aranzana et al., 2005; Lindblad-Toh et al., 2005; Breseghello and Sorrells, 2006), far beyond the human disease studies (Hirschhorn and Daly, 2005; Wang et al., 2005) from which the method originated. In contrast to pedigree-based samples or designed mapping populations, many populations used in association mapping studies often have obscure complex geographies and histories (Hey and Machado, 2003; Yu et al., 2006). This inherent genetic structure of an association mapping population, if unaccounted for, may lead to an excess of spurious results (Voight and Pritchard, 2005; Price et al., 2006; Yu et al., 2006; Zhao et al., 2007). Many statistical methods have been proposed to account for population structure and familial relatedness: structured association (Pritchard and Rosenberg, 1999; Pritchard et al., 2000; Falush et al., 2003); genomic control (Devlin and Roeder, 1999; Devlin et al., 2001, 2004); mixed-model approach (Yu et al., 2006); and principal

J. Yu and C. Zhu, Dep. of Agronomy, Kansas State Univ., Manhattan, KS 66506; Z. Zhang and S. Kresovich, Institute for Genomic Diversity, Cornell Univ., Ithaca, NY 14853; D.A. Tabanao, Philippine Rice Research Institute, Maligaya, Muñoz 3119, Nueva Ecija, Philippines; G. Pressoir, Hispaniola Center for Biofuels and Sustainable Agriculture, Port-au-Prince, Haiti; M.R. Tuinstra, Dep. of Agronomy, Purdue Univ., West Lafayette, IN 47907; R.J. Todhunter, College of Veterinary Medicine, Cornell Univ., Ithaca, NY 14853; E.S. Buckler, USDA-ARS, Institute for Genomic Diversity, Cornell Univ., Ithaca, NY 14853. Received 12 Sept. 2008. *Corresponding author (jyu@ksu.edu).

Abbreviations: BIC, Bayesian Information Criterion; DI, distraction index; DLS, dorsolateral subluxation; PIC, polymorphism information content; QTL, quantitative trait locus; SNP, single nucleotide polymorphism; SSR, simple sequence repeat.

component approach (Price et al., 2006). The essence of these approaches is to exploit information from random molecular markers across the genome to account for genetic relatedness in deriving the test statistics explicitly or through ad hoc adjustment.

One important concern in carrying out an association mapping study with diverse germplasm is whether adequate background marker information was used to account for genetic relatedness. In many model species, recent advances in genomic technologies have enabled researchers to score thousands of single nucleotide polymorphisms (SNPs) in high-throughput fashion (Shendure et al., 2005; Syvanen, 2005). However, for many genetic studies performed in nonmodel species, the number of molecular markers scored across the association mapping panel is still limited by the cost of genotyping, the availability of the markers, and the complexity of the genome. In contrast, simple sequence repeats (SSRs), though much less abundant than SNPs, are another type of genetic markers that has been widely used. Among other factors, the balance between the greater discriminatory power of SSRs and the lower cost of SNPs may vary among species (Weir et al., 2006). Even in species for which the genome sequence and millions of SNPs are available, the combination of genotyping throughput and sample throughput is still a challenge (Syvanen, 2005). Nevertheless, this constraint should not hamper initiating association mapping studies with a smaller set of SSRs or SNPs. As was proposed in human, a multistage strategy (i.e., small-scale probing studies followed by a large-scale validation) would be a method of choice both scientifically and economically (Hirschhorn and Daly, 2005).

Genetic relatedness analysis based on molecular marker information has been well studied because of its importance in many areas such as ecology, human genetics, agriculture, and forensics (Weir et al., 2006). Many different methods have been proposed for a variety of genetic relatedness estimates (Rousset, 2002; Blouin, 2003). Relative kinship estimates (Loiselle et al., 1995; Ritland, 1996a) have been successfully used to account for the relatedness in diverse association mapping panels (Yu et al., 2006; Zhao et al., 2007) because they provide both inter- and intra-individual estimates in a symmetric matrix analogous to the traditional pedigree-based coancestry matrix used in mixed models (Henderson, 1984). While the robustness of population structure estimates from random background markers has been previously studied (Pritchard et al., 2000) and validated in many empirical studies (Evanno et al., 2005; Camus-Kulandaivelu et al., 2007), the robustness of kinship estimates with varied numbers of background markers would provide further insight into the application of the unified mixed-model approach in the context of association mapping.

In the current study, we first laid out the theoretical reasoning of using model testing to assess the adequacy of background markers for relationship estimation under maximum likelihood framework and validated with computer simulations. To our limited knowledge, this is the first study where the assessment of the accuracy of cofactors and variance–covariance structure estimated from molecular markers was conducted through the likelihood-based model fitting of quantitative traits. We then demonstrated with two data sets the usefulness of evaluating robustness of genetic relatedness via model testing and variance components analysis. For theoretical computer simulations and analyses with two empirical data sets, different complex traits were examined to cover a range of scenarios in which genetic relatedness has different levels of effect on phenotypic variation.

MATERIALS AND METHODS

Mixed Model

The mixed-model equation (Henderson, 1975, 1984) can be expressed as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e} \qquad [1]$$

where $\mathbf{y}$ is a vector of phenotypic observations; $\boldsymbol{\beta}$ is a vector of fixed effects; $\mathbf{u}$ is a vector of random polygenic background effects; $\mathbf{e}$ is a vector of residuals; and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices of 1s and 0s relating $\mathbf{y}$ to $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively. The variances of the random effects are assumed to be $\mathrm{Var}(\mathbf{u}) = 2\mathbf{K}V_g$, and $\mathrm{Var}(\boldsymbol{e}) = \mathbf{R}V_R$ (Yu et al., 2006), where $\mathbf{K}$ is an $n \times n$ matrix of relative kinship coefficients (obtained from SPAGeDi [Hardy and Vekemans, 2002]) that define the degree of genetic covariance between a pair of individuals; $\mathbf{R}$ is an $n \times n$ matrix with the off-diagonal elements being zero and the diagonal elements being the reciprocal of the number of observations for which each phenotypic data point was obtained; V_g is the genetic variance; and V_R is the residual variance.

For model testing, maximum likelihood rather than restricted maximum likelihood can be used to solve the mixed model and obtain the variance component estimates of V_g and V_R because comparisons among models with different random or fixed variables were conducted (Littell et al., 2006). However, if variance component estimation is the focus, restricted maximum likelihood should be used to remove the bias associated with the fixed model in maximum likelihood approach (Lynch and Walsh, 1998). We have verified that the choice of maximum likelihood or restricted maximum likelihood did not alter the pattern of changes in our analysis. The –2 log-likelihood function for estimating the parameter V_g and V_R is

$$l(\mathrm{y} \mid V_g, V_R) = \log|\mathbf{V}| + \mathbf{r}`\mathbf{V}^{-1}\mathbf{r} + n\log(2\pi) \qquad [2]$$

where $\mathbf{r} = \mathrm{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathrm{y}$, and $\mathbf{V} = \mathbf{Z}(2\mathbf{K}V_g)\mathbf{Z}' + \mathbf{R}(V_R)$. It can be seen that the accuracy in estimating $\mathbf{K}$ affects two nonconstant terms (i.e., $\log|\mathbf{V}|$ and $\mathbf{r}'\mathbf{V}^{-1}\mathbf{r}$) in Eq. [2] and subsequently influences the maximum likelihood value under given $\mathbf{K}$ and the associated values of V_g and V_R at convergence point.

Alternatively, we can view this in terms of deviation of marker-based $\mathbf{K}_M$ from the true $\mathbf{K}_T$, $\mathbf{K}_M = \mathbf{K}_T + \Delta\mathbf{K}$,

where $\mathbf{\Delta K}$ is the deviation associated with the estimation. Then, based on marker-estimated relationship ($\mathbf{K_M}$), we have the variance of trait y ($\mathbf{V_M}$) and

$$\mathbf{V_M} = \mathbf{Z}(2\mathbf{K_M}V_g)\mathbf{Z}' + \mathbf{R}(V_R) = \mathbf{Z}[2(\mathbf{K_T} + \mathbf{\Delta K})V_g]\mathbf{Z}' + \mathbf{R}(V_R)$$

If all elements of $\mathbf{\Delta K}$ approach zero, then $\mathbf{V_M}$ will approach $\mathbf{V_T}$. And subsequently, with Eq. [2] the likelihood value $l(\mathbf{K_M})$ approaches $l(\mathbf{K_T})$. To our knowledge, this would be very difficult, if not impossible, to prove mathematically, particularly given the mixed model itself is solved through a parameter (V_g and V_R) searching process. However, extensive simulations with different trait architectures can be conducted to demonstrate that this is the case.

With known $\mathbf{K}$, the maximum likelihood of model, $l(\mathrm{y} \mid \hat{V}_g, \hat{V}_R)$, finds $\hat{V}_g$ and $\hat{V}_R$ at a convergence point (Henderson, 1984; SAS Institute, 1999). With different $\mathbf{K}$ being estimated from different sets of markers, the maximum likelihood process converges at different point with different $\hat{V}_g$ and $\hat{V}_R$. In addition, if $\mathbf{X\beta}$ contains covariates, such as fixed population structure ($\mathbf{Q}$) effects, the accuracy in estimating these covariates also influences the maximum likelihood (i.e., through the term $\mathbf{r} = \mathrm{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathrm{y}$ in Eq. [2]). Relating to genetics, a more accurate estimate of $\mathbf{K}$ (defining variance–covariance of $\mathbf{u}$) and $\mathbf{Q}$ (relating the covariates) from molecular marker information would result in a better fit of the model in explaining phenotypic variation with genetic relatedness than a less accurate estimate. With –2 residual log-likelihood, Bayesian Information Criterion (BIC) value (Schwarz, 1978) is simply

$$\mathrm{BIC} = l(\mathrm{y} \mid V_g, V_R) + d\log(n) \qquad [3]$$

where d the dimension of the model. The BIC values adjusted the –2 residual log-likelihood values for the number of model parameters and sample size. We presented BIC results in some figures because the models without kinship term (i.e., 0% marker or no $\mathbf{Zu}$ in Eq. [2]) were also examined and no penalty is associated with these models. In the case of analysis without marker information, genetic variance cannot be separated from experimental variance and the difference in model dimension of this model compared with other models equals 1.

Computer Simulations

Due to limited knowledge about population structure in plant and animal species, we focused our simulation on the kinship relationship among a group of 240 inbred lines with different relatedness but without subpopulation structure. The simulation process follows a common breeding history in many plant and animal species, that is, individuals of earlier generations are crossed to form segregating populations from which progeny individuals were derived (Yu et al., 2005). To simulate the genetic relationship, we started with 16 founder inbred lines that were derived from a random-mated ancestral population. A total of 32 inbred lines were then derived from 16 random biparental crosses between founder inbreds, two from each cross. Likewise, another set of 64 inbred lines was obtained at the second round and 128 inbreds were obtained from the third round. These 240 inbred lines were regarded as the simulated association mapping panel to assess the effect of number of background markers on relatedness estimation through the model-fitting procedure. The simulation process, therefore, generated different levels of kinships among individuals but no major population differentiation that would require an estimation of population structure.

For the quantitative trait, we considered a total of $g = 20$ or 50 quantitative trait loci (QTLs) and $m = 400$ or 800 markers. Both marker and QTL had two alleles at each locus and were randomly located across the genome. The genome structure followed a published maize (*Zea mays* L.) linkage map with 1749 cM with 10 chromosomes (Senior et al., 1996). The effects of QTLs were set to be constant (i.e., equal effects, $a^i = 1$) or followed a geometric series (i.e., unequal effects) with the ith QTL having an additive genetic effect of a^i, where $a = 0.90$ for $g = 20$ QTLs or $a = 0.96$ for $g = 50$ QTLs (Lande and Thompson, 1990). Accordingly, each of the two alleles at ith QTL had an effect of either $+a^i$ or $-a^i$. The genotypic value of each inbred line was defined as the sum of genotypic values at all QTLs. The observed phenotypic value of each inbred line was obtained by adding a residual error. Different numbers of QTLs, different effect size distributions, different heritability, and different numbers of starting markers (i.e., assuming 400 or 800 markers are available as the whole set of markers) allowed us to investigate different trait complexities and scenarios that researchers would encounter in real situations. In our simulation, only random drift changes the allele frequency of markers from the expected 0.5 level.

We investigated two simulation settings: general and multiple resampling (Table 1). In the first simulation setting, we focused on the difference among cases and a total of 50 independent runs were conducted for each case (i.e., combination of number of QTLs, QTL effect, total marker number, and heritability). Within each run, new relationships and kinship estimation were generated independently. At

Table 1. Description of theoretical and empirical data–based computer simulations.

Simulation scenario	Study focus	Results presentation
General resampling	Difference among cases (combination of parameters: number of QTL,† QTL effect, total marker number, and heritability) and a total of 50 independent runs were conducted for each case	Fig. 1
Multiple resampling	Multiple resampling of different proportions of markers from each individual case. A single run was randomly chosen for each case and 10 repetitions of resampling were randomly conducted at each marker proportion	Fig. 2–3
Empirical data (maize and canine)	Multiple resampling of different proportions of markers from each dataset and 10 repetitions of resampling were conducted at each marker proportion	Fig. 5–8

†QTL, quantitative trait loci.

each run, five different marker proportions were studied: (i) no marker (0%), (ii) every eight markers (12.5%), (iii) every four markers (25%), (iv) every two markers (50%), and (v) all markers (100%). Heritability of $h^2 = 0.4$ or 0.7 was examined and results are presented in Fig. 1.

In the second simulation setting, we focused on multiple resampling of different proportions of markers from each individual case. A single run was randomly chosen for each case and 10 repetitions of resampling were randomly conducted at each marker proportion (12.5, 25, 50, and 75%). Heritability of $h^2 = 0.4$ or 0.6 was examined and results are presented in Fig. 2 and 3. Switching the higher end heritability from 0.7 to 0.6 is to cover more scenarios. In both simulation settings, additive relationship matrices derived from pedigree information also were examined for comparison. By conducting experiments under both simulation settings, we were able to obtain the general information across independent runs as well as the performance of multiple resampling of a particular run that is comparable to the empirical data analyses.

Maize Association Mapping Panel

A group of 274 diverse maize inbred lines was used for the current study based on the availability of both 912 SNP and 89 SSR marker data. Detailed information about this association mapping panel has been documented in previous publications (Liu et al., 2003; Flint-Garcia et al., 2005; Yu et al., 2006). These maize inbred lines represented the diverse genetic material that is publicly available around the world and has been used for dissection of various complex traits as well as general method development (Thornsberry et al., 2001; Wilson et al., 2004; Yu et al., 2006). Three traits were chosen to cover a range of agronomic and physiologic traits with different heritability estimates and population structure effects: flowering time, ear height, and ear diameter (Flint-Garcia et al., 2005). Flowering time was measured as the number of days to pollen shed; ear height as the distance from the ground to the major ear-bearing node; and ear diameter as the diameter of an ear at the midsection. Field tests were conducted at Clayton, NC (summer nursery), and Homestead, FL (winter nursery), in 2002, and the trait mean of the two field tests was used in the current study. We used trait means rather than observations from individual environments to reduce the environmental error associated with single-environment observation.

For the maize data, three subsets (25, 50, and 75%) of the original SSR and SNP data were randomly selected without replacement to estimate population structure and kinship in the whole set of maize inbred lines (Table 1). This equaled to 228, 456, and 684 SNPs, and 22, 45, and 67 SSRs. For each sampling subset, 10 repetitions were conducted for each population structure and kinship matrix because the population structure estimation through STRUCTURE and mixed-model analyses are both computationally demanding. The average model-fitting statistics across these repetitions and corresponding standard deviations were reported.

Canine Association Mapping Panel

A group of 266 crossbred dogs derived from trait-free Greyhounds and dysplastic Labrador Retrievers was genotyped with 471 SSRs randomly distributed across the genome (Mateescu et al., 2005; Todhunter et al., 2005). Detailed pedigree structure, molecular marker procedures, and QTL mapping results have been previously reported (Mateescu et al., 2005; Todhunter et al., 2005). Briefly, canine hip dysplasia is a complex developmental trait characterized by hip laxity, subluxation, or incongruity of the femoral head and acetabulum in affected hips. Distraction index (DI) was measured as maximum lateral passive hip laxity on both left (DI_left) and right (DI_right) hips, whereas dorsolateral subluxation (DLS) was measured on the dorsoventral radiograph with the hips oriented and loaded in a weight-bearing position on both left (DLS_left) and right (DLS_right) hips (Todhunter et al., 2005).

For the canine data, five subsets (6.25, 12.5, 25, 50, and 75%) of the original SSR data were randomly selected without replacement to estimate kinship among the whole set of dog individuals (Table 1). This equaled to 29, 59, 118, 236, and 353 SSRs used for kinship estimation. The additional smaller subsets (i.e., 6.25 and 12.5%) were examined because of the relatively larger number of the original SSR markers in the canine data (471 SSRs) as compared with maize data (89 SSRs). For each sampling subset, 10 repetitions were conducted for each kinship matrix and average model-fitting statistics across these repetitions and corresponding standard deviations were reported.

Estimation of Genetic Relatedness

Relative kinship (**K**) was calculated using the software SPAGeDi (Hardy and Vekemans, 2002). Since kinship estimates based on Loiselle et al. (1995) and Ritland (1996a) give comparable results, we chose the estimation method of Loiselle et al. (1995) to be consistent with our previous study (Yu et al., 2006). Although these two methods have slightly different calculations, the conceptual formula is $\mathbf{K}_{ij} = (Q_{ij} - Q_m)/(1 - Q_m)$, where Q_{ij} is the probability of identity in state for random loci from individuals i and j, and Q_m is the average probability of identity in state for loci from random individuals from the sample (Hardy and Vekemans, 2002). The kinship estimate at single loci was averaged across loci to give kinship estimates between two individuals (Loiselle et al., 1995; Ritland, 1996a). Negative values were set to zero, as this indicates the relationship of those individuals is less than that of random individuals (Hardy and Vekemans, 2002). Including negative values indicates "negative" covariance among individuals in a genetic variance–covariance matrix. This deviates from the genetic assumption of the traditional mixed model that the variance–covariance among individuals is greater than or equal to zero. We would, therefore, like to only present results with kinships in which the minimum value was set to zero and leave the debate of whether setting negative values to zero or not to future studies.

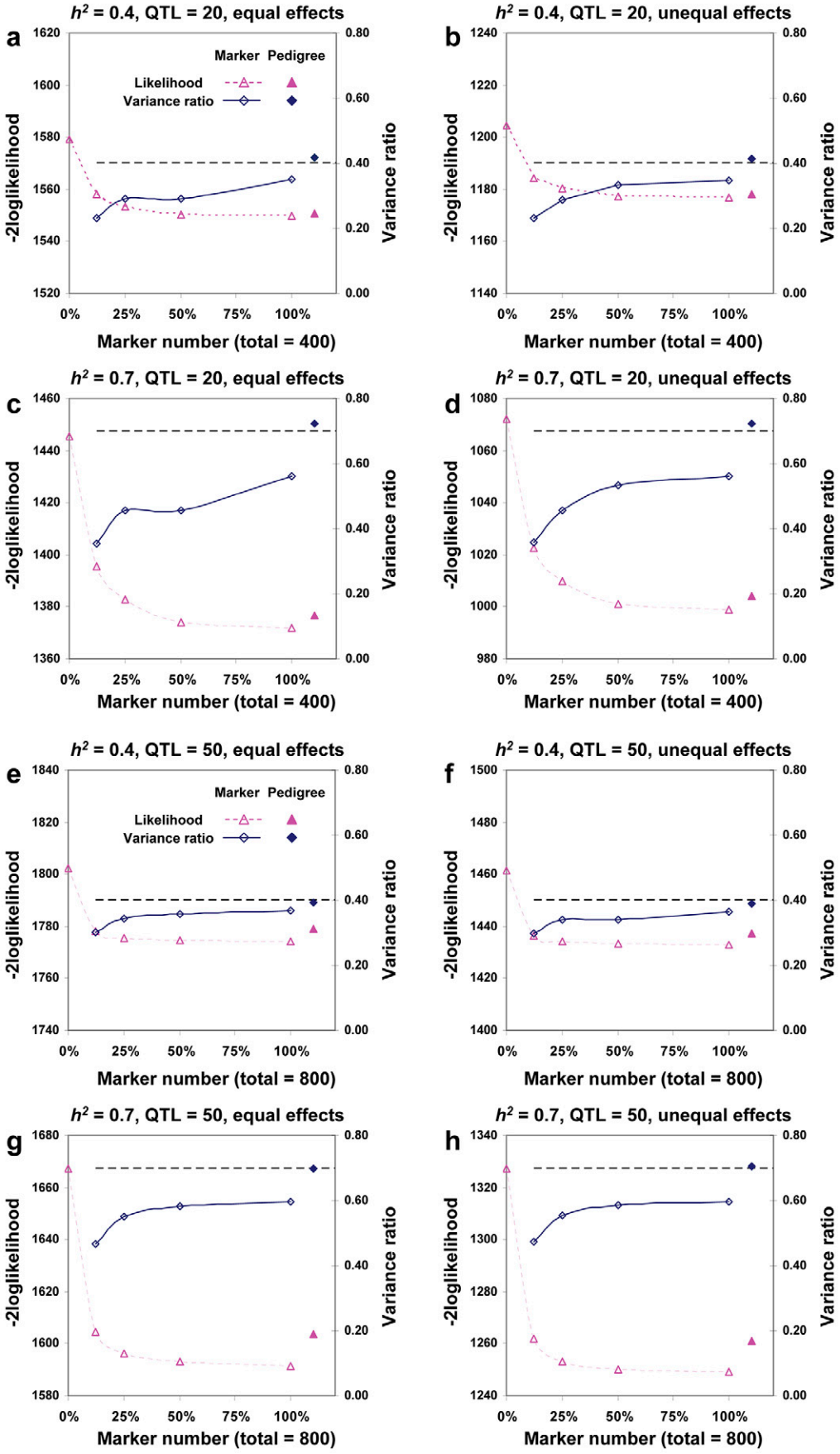

Figure 1. Model fitting of simulated populations with different parameters, independent runs. The –2 residual log-likelihood values of mixed models with different relative kinship (K) matrices and variance ratio [$V_g/(V_g + V_R)$] estimates. Dashed lines represent the simulated heritability (h^2). QTL, quantitative trait loci.

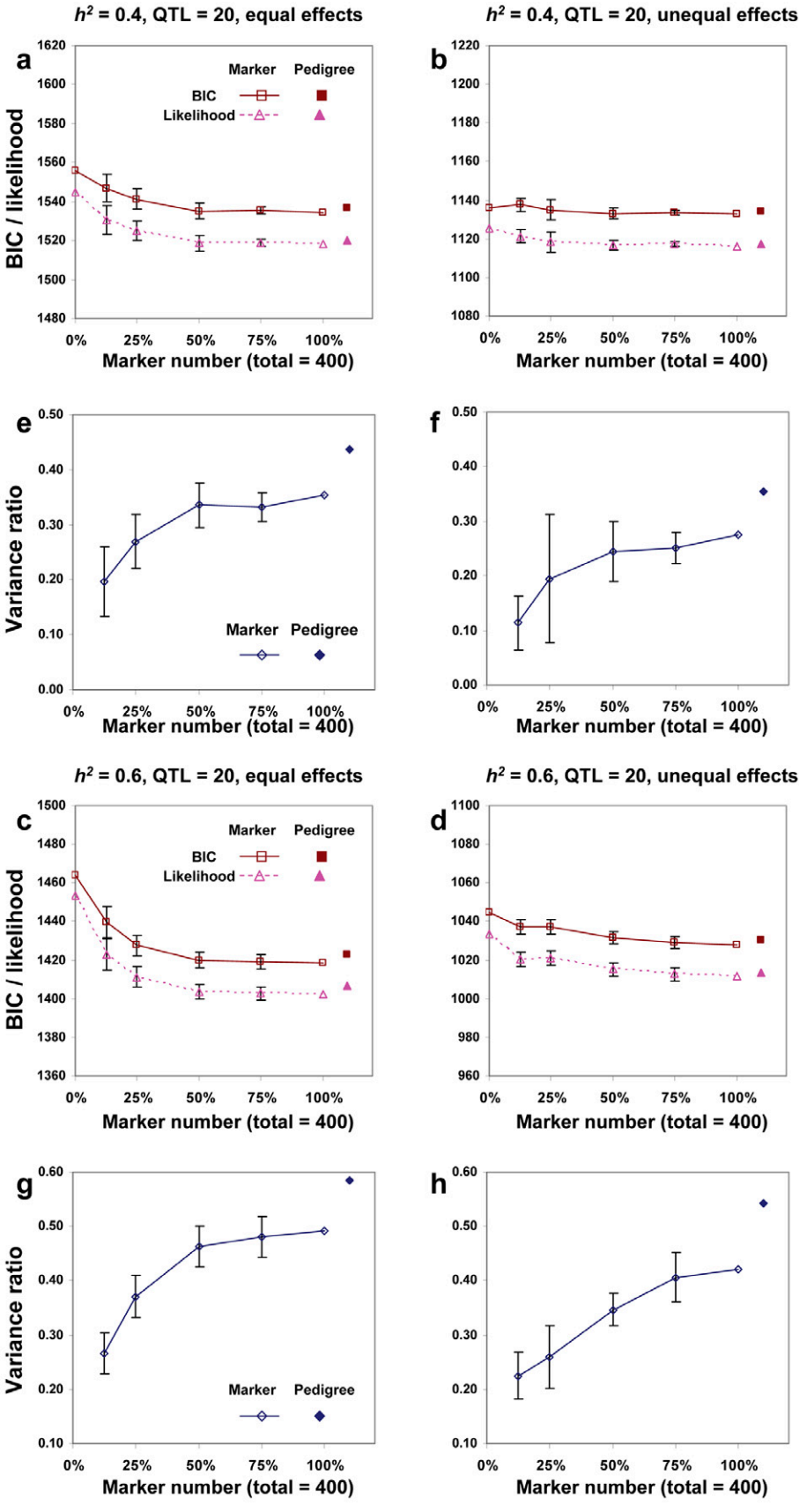

Figure 2. Model fitting of simulated individual populations with different relative kinship (K) with quantitative trait loci (QTL) = 20 and m = 400, multiple resampling at intermediate proportions. (a–d) Bayesian Information Criterion (BIC) and –2 residual log-likelihood values of mixed models with different K matrices. (e–h) Variance ratio [$V_g/(V_g + V_R)$] estimates from mixed models. Standard deviations are shown by vertical bars. h^2, heritability.

Population structure (**Q**) estimates were calculated using the software STRUCTURE (Pritchard et al., 2000; Falush et al., 2003) for maize data as in our previous studies (Flint-Garcia et al., 2005; Yu et al., 2006). We conducted experiments in which the same sets of markers (25, 50, and 75% of original SNPs and SSRs) were used to construct both population structure by STRUCTURE and relative kinship by SPAGeDi. We chose the number of subgroup to be three for running the structure analysis because this has been determined by previous extensive analyses with the whole set of markers and agreed with the breeders' knowledge about these materials (Liu et al., 2003; Flint-Garcia et al., 2005) and also because the main focus of the study was kinship estimation. Three maize subgroups are stiff stalk, non–stiff stalk, and tropical–subtropical. For canine data, no population structure was estimated given the closed breeding structure.

Three general comparison schemes were examined for maize data. In the first scheme, the same subsets of markers were used to estimate both **Q** and **K**. This represented a common scenario when only a single set of markers was available. In the second scheme, we kept the **Q** constant but varied the number of markers for **K** estimation. By having a constant population structure based on the full set of markers, the robustness of kinship estimates with different numbers of markers can be shown directly. In the third scheme, we kept the **K** constant but varied the number of markers for **Q** estimation. Likewise, this provided evidence on the sensitivity of population structure estimation with different numbers of markers. We presented results with the constant **Q** estimated with 89 SSRs for the second scheme, and the constant **K** estimated with 912 SNPs for the third scheme because the results with the constant **Q** with 912 SNPs and the constant **K** with 89 SSRs were identical.

Model Testing Procedure

The mixed-model equations for model testing with the simulated data and canine data are the same as Eq. [1] because there was no population structure simulated or expected. For maize data, the mixed-model (Henderson, 1975, 1984) equation for model testing is expressed as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Qv} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e} \qquad [4]$$

where **v** is a vector of population group effects; and **Q** is a matrix from STRUCTURE relating **y** to **v**. This is an expanded form of Eq. [1] by separating the part of fixed population structure from the rest of the fixed effects for the convenience of explanation. But the statistical properties remain unchanged.

The mixed-model analysis was conducted with Proc Mixed in SAS (SAS Institute, 1999). In our analyses, BIC values allowed a cross comparison of models either without (0% marker) or with **Q** or **K** (different proportion of markers). But for models with the same parameters (i.e., cases other than 0% marker), the penalty terms due to the inclusion of the **Q** or **K** were the same. Some other model selection criteria and their modifications have been recently reviewed (Sillanpaa and Corander, 2002). In our analysis, additional Akaike Information Criterion (AIC) and corrected Akaike Information Criterion (AIC_c) were also given by Proc Mixed in SAS and the only difference with BIC was the scale change in comparing the case with 0% with the rest of the cases.

We chose a consistent scale to present the changes in BIC and likelihood values for three different comparison schemes in maize data (i.e., both **Q** and **K** with different number of markers, **Q** constant but **K** varying, and **K** constant but **Q** varying) to avoid the biases due to the differences in scale in evaluating the significance of these changes. The ranges of changes in BIC and likelihood values were also set the same across different traits for each data set for an unbiased comparison.

For maize data, because the gross population structure was

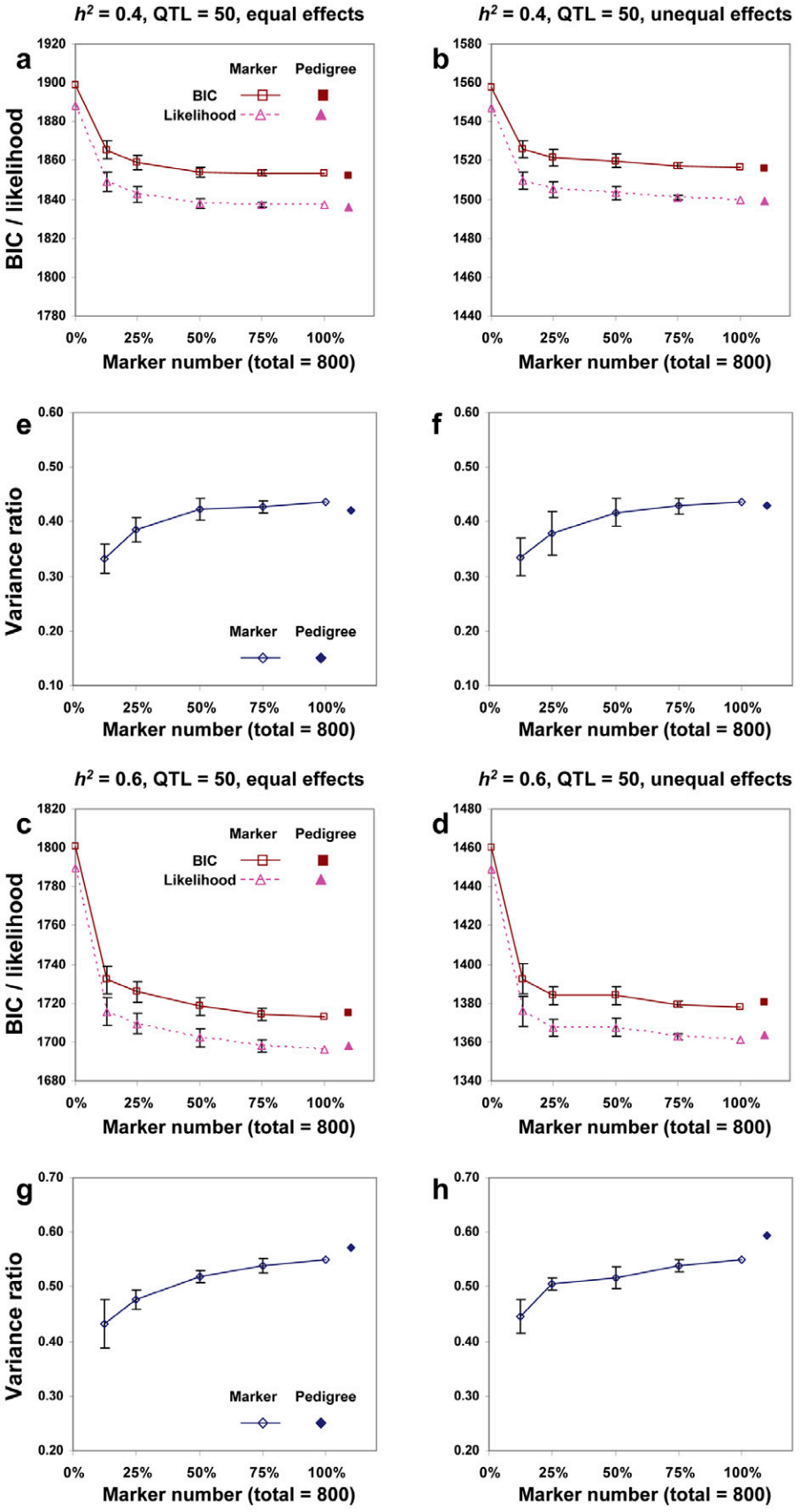

Figure 3. Model fitting of simulated individual populations with different relative kinship (K) with quantitative trait loci (QTL) = 50 and *m* = 800, multiple resampling at intermediate proportions. (a–d) Bayesian Information Criterion (BIC) and –2 residual log-likelihood values of mixed models with different K matrices. (e–h) Variance ratio [$V_g/(V_g + V_R)$] estimates from mixed models. Standard deviations are shown by vertical bars. h^2, heritability.

accounted for by the **Qv** term in the mixed model, the variance ratio $V_g/(V_g + V_R)$ can be interpreted as heritability averaged across three maize subgroups (stiff stalk, non–stiff stalk, and tropical–subtropical from STRUCTURE). We chose variance ratio rather than marker-based heritability because the latter only approaches true heritability (as shown in simulations) and the number of background markers was varied. Variance ratio indicates how much phenotypic variance can be attributed to the genetic variance after accounting for relatedness among individuals for different complex traits.

RESULTS

Computer Simulations

Suppose we have true values of **Q** and **K**, under the maximum likelihood framework, the –2 log-likelihood of the data on convergence value V_g and V_R is $l(y \mid V_g, V_R)$. With $\mathbf{Q}^p$ and $\mathbf{K}^p$ being estimates of **Q** and **K** based on p random markers, the maximum likelihood becomes $l^p(y \mid V_g^p, V_R^p)$. In theory, the inaccuracy in estimating **Q** and **K** decreases the maximum likelihood, which can be expressed as $l^p > l$ in –2 log format. When more background markers are used, the accuracy of $\mathbf{Q}^p$ and $\mathbf{K}^p$ increases and l^p approaches l. However, if the adequate number of markers is reached, we would see a stabilized l^p. The change in BIC follows the same trend, except it imposes a penalty on the model dimension. Meanwhile, the estimates of V_g^p and V_R^p with adequate number of markers approach V_g and V_R of known **Q** and **K**. In terms of variance ratio, $V_g^p/(V_g^p + V_R^p)$ approaches $V_g/(V_g + V_R)$ with an adequate number of markers. Then, if subsets of q or r ($q > r$) markers were sampled from the whole set of p markers and used to obtain a series of ($\mathbf{Q}^q$, $\mathbf{K}^q$) and ($\mathbf{Q}^r$, $\mathbf{K}^r$), the average value of –2 log-likelihood $\bar{l}^q$ should be smaller than $\bar{l}^r$. The use of average over a number of repetitions is to offset the variance in **Q** and **K** estimation process as well as random sampling error, which may lead to $l^q > l^r$ in some cases. Similarly, the average BIC values and the variance ratios for models in which **Q** and **K** are estimated with more markers should be more accurate than those for models in which **Q** and **K** are estimated with fewer markers, and these differences diminish with an adequate number of markers in **Q** and **K** estimation.

Computer simulation results revealed a consistent pattern in changes in likelihood values and variance ratio estimates across different trait complexities for the general simulation setting with independent runs (Fig. 1) and the multiple resampling simulation setting with individual populations (Fig. 2 and 3). For the general simulation setting, resampling was conducted once at each marker proportion and results were averaged across independent runs to obtain the overall pattern. With more markers used in the kinship estimation, both the likelihood value and the variance ratio estimate generally approached a plateau. Pedigree-based additive relationship matrix gave better estimates of variance ratio than any marker-based relationship matrix even though the likelihood value of the model with pedigree information was not always the smallest. As expected, the changes in likelihood value and variance ratio estimate were greater in scale when heritability is higher (i.e., $h^2 = 0.6/0.7$ vs. 0.4), but no obvious differences were observed for different QTL effect distributions (i.e., equal or unequal effects) (Fig. 1–3). In theory, a different weight is given in solving mixed model when incorporating the genetic relationship matrix [i.e., $(2\mathbf{K})^{-1}(V_R/V_g)$] under high heritability than under low heritability.

For given individual populations, multiple resampling at different marker proportion closely resembles the empirical samples and allowed us to examine the fluctuation due to sampling error. The variation due to different subsets of random markers at each marker density generally became smaller as marker number increases (Fig. 2 and 3). Similar to what we observed from the general simulation setting, pedigree-based additive relationship matrix gave better estimates of variance ratio than marker-based relationship matrix. For the simulated association panel with 240 inbred lines, about 300 to 600 biallelic markers, depending on different cases, would give robust kinship estimates to relate to the phenotypic variation of the quantitative trait with 20 or 50 QTLs.

Maize Data

Across 274 maize inbred lines, the mean of major allele frequency was 0.78 for SNP data and 0.33 for SSR data. The average polymorphism information content (PIC; a measurement of informativeness of markers) value (Anderson et al., 1993) for SNPs was 0.24, ranging from 0.004 to 0.375, whereas the average PIC value for SSRs was 0.78, ranging from 0.361 to 0.964 (Fig. 4). Compared with biallelic SNPs, the SSRs had an average allele number of 21.5. Across 266 dogs, the mean major allele frequency was 0.47 for 471 SSRs. The average PIC value was 0.60, ranging from 0.046 to 0.919, and the average allele number was 7.7 (Fig. 4).

The impact of relatedness estimation with molecular markers on model fitting was different across three quantitative traits (Fig. 5). With both BIC and likelihood values being plotted on the same scale across three different quantitative traits, the relative importance of relatedness on model fitting can be shown clearly. The improvement in model fitting was greater for flowering time than for ear height or ear diameter, as more background markers were used in relatedness estimation (**Q** and **K**). Compared with the model without **Q** and **K**, models that account for relatedness had a better fit even when a small number of background markers were used in estimating kinship. For models with **Q** and **K**, the likelihood values showed a parallel pattern as the BIC values because the same penalty terms were applied in calculating BIC from –2 residual log-likelihood. Regardless of the number of markers used, there were two degrees of freedom for **Q** and one degree of freedom for **K**, compared with models without **Q** and **K** (0% marker). Overall, as indicated

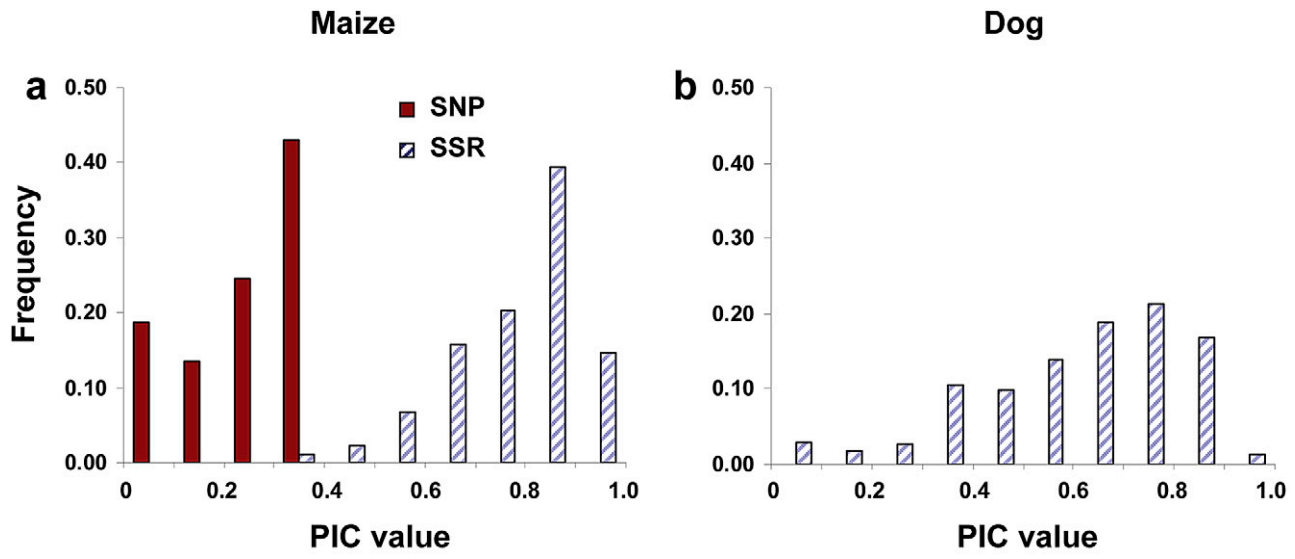

Figure 4. Frequency distribution of polymorphism information content (PIC) for two types of molecular markers. (a) PIC values of 912 single nucleotide polymorphisms (SNPs) and 89 simple sequence repeats (SSRs) scored on 274 diverse maize inbred lines; (b) PIC values of 471 SSRs scored on 266 crossbred dogs.

by the changes in BIC and likelihood values for model fitting and in variance ratio estimates, relatedness estimates did not become stable even with 75% (i.e., 684 SNPs or 67 SSRs) of the whole set of markers in the maize association mapping panel (Fig. 5), indicating that an obvious plateau would require more markers than the current full sets of SSRs and SNPs. This general pattern in model fitting and variance ratio was consistent for all trait–marker combinations. Moreover, as indicated by the large standard deviation bars around the average

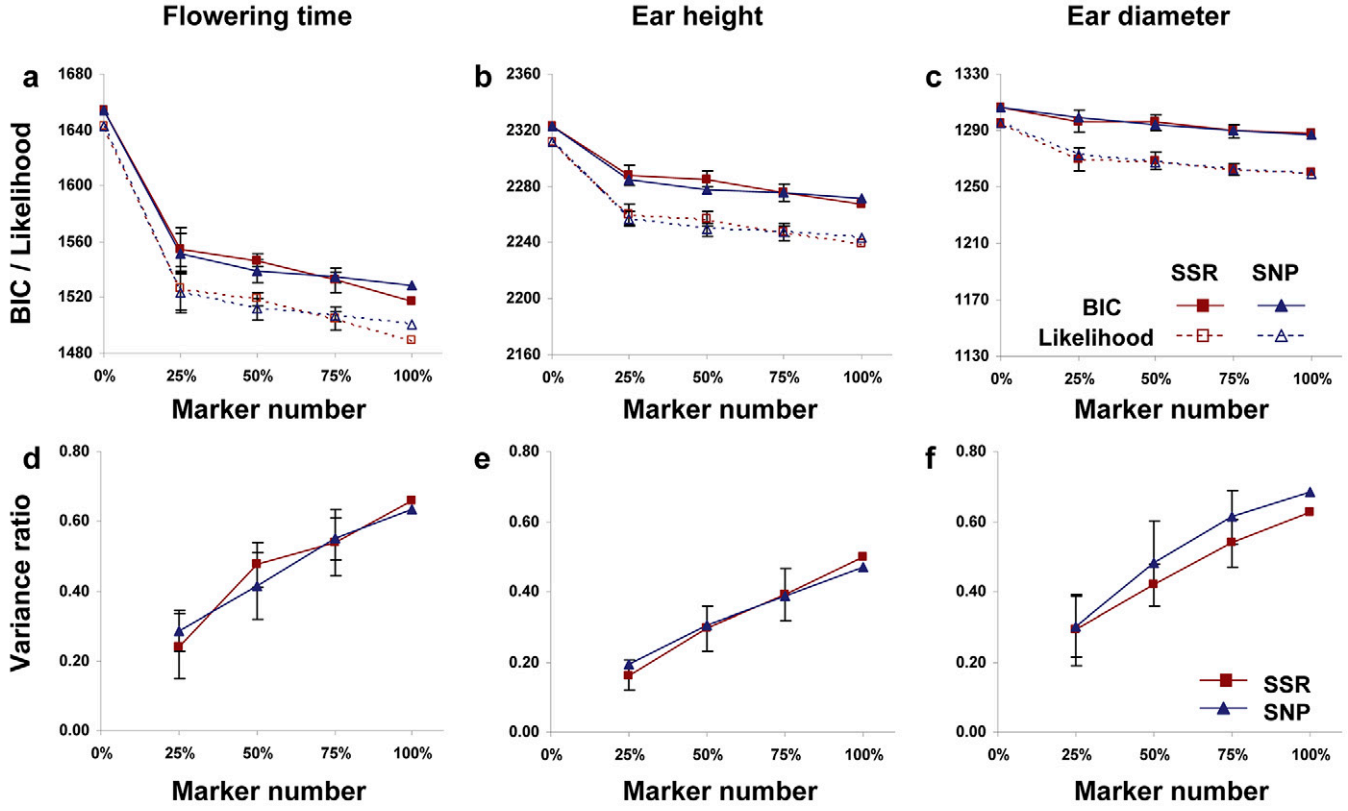

Figure 5. Model fitting for maize quantitative traits with both population structure (Q) and relative kinship (K) being estimated with different numbers of markers. (a–c) Bayesian Information Criterion (BIC) and –2 residual log-likelihood values of mixed models with different Q and K estimates based on different proportion of the whole set of background markers (912 single nucleotide polymorphisms [SNPs] or 89 simple sequence repeats [SSRs]). (d–f) Variance ratio [$V_g/(V_g + V_R)$] estimates from mixed models. Standard deviations are shown by vertical bars.

variance ratios, the variation among different repetitions in each marker subset remained generally high (Fig. 5). The whole set of 89 SSR markers provided roughly the same amount of information as did the whole set of 912 SNP markers for relatedness construction (Fig. 5).

With a constant population structure based on the full set of 89 SSRs across models, the effect of marker number on relative kinship can be shown directly (Fig. 6). In this case, the model with 0% marker number corresponded to the model with only population structure. When more markers were used for kinship estimation, the model fitting and variance ratio estimation improved for all three traits, though at different rates. Except the difference at 0% marker number (due to the constant **Q** in Fig. 6), the general patterns of model fitting and variance ratio estimation with constant **Q** but varying **K** were similar to that with varying **Q** and **K** (as shown in Fig. 5). With a constant kinship based on the full set of 912 SNPs, however, the number of markers on population structure had small to negligible effect on model fitting and variance ratio estimation (Fig. 7). With the penalty term associated with including population structure, the average BIC value for ear diameter was even higher for models with population structure than for models without it (i.e., 0% marker number). The variance ratio estimates were essentially flat with different **Q** based on different numbers of background markers. The results with the constant **Q** with 912 SNPs and the constant **K** with 89 SSRs were very similar to the corresponding results presented. Considering the three scenarios together, our results indicated that the improvement of model fitting and variance component estimation with more background markers was mainly through a more robust estimation of relative kinship (Fig. 5–7).

Canine Data

Similar to the results from maize data, increasing marker number in kinship estimation decreased the BIC values and increased the variance ratios for all four dysplastic traits (Fig. 8). Models with kinship estimates, which account for the genetic relatedness, had better fit than the model without kinship in the canine sample. Average BIC values and variance ratios started reaching a plateau with 118 SSRs (i.e., 25% of the whole set). Further increasing the background markers resulted in minimal improvement in BIC and much smaller increase in variance ratios. The stabilized variance ratios were close between pairs of traits: 0.21 to 0.26 for DI and 0.15 to 0.21 for DLS. Notably, the variation among different repetitions in each marker subset decreased as the number of markers increased, and this reduction in variation was more prominent in BIC than in variance ratios (Fig. 8). There were differences in BIC values between pairs of traits (DI_left vs. DI_right, and DLS_left vs. DLS_right), and Pearson correlation coefficients were significant for phenotypic measurements from left and right hips (0.76 for DI and 0.79 for DLS).

DISCUSSION

Molecular markers have long been used to examine the genetic relationships among individuals with unknown mating records in natural populations and to estimate the genetic distances among different breeding materials (Weir et al., 2006). Ritland (1996b) was among the first scientists who coupled marker-based genetic relatedness with phenotypic similarity to generate estimates of variance components (Lynch and Walsh, 1998). Recent studies have successfully incorporated various marker-based relatedness estimates (e.g., population structure, relative kinship, and principal components) into association analysis with the common goal of alleviating the impact of cryptic relatedness in identifying the causative polymorphisms underlying complex traits (Pritchard et al., 2000; Thornsberry et al., 2001; Falush et al., 2003; Price et al., 2006; Yu et al., 2006). In the current study, we focused on the effect of number of background markers on relationship estimation with computer simulations and analyses of two empirical association mapping populations.

In the traditional application of linear models, the covariate and the variance–covariance structure are usually known from direct measurement or imposed with certain fixed structure. For instance, the variance–covariance of random polygene effect (or breeding value) is defined by the additive relationship matrix based on known pedigree information (Lynch and Walsh, 1998). However, in the context of association mapping, both population structure and relative kinship need to be estimated from molecular marker information due to the diverse origins and complex relationships. Under the maximum likelihood framework, the inaccuracy of the covariates and/or variance–covariance introduced in the process of their estimation with different numbers of molecular markers leads to a decrease in model fit to the data. Fortunately, it has been a common practice for many researchers that phenotypes of general complex traits are often collected and a set of background markers is assayed to estimate the genetic relationship before a more thorough genomewide association analysis or candidate-gene analysis. This makes the assessment approach outlined in our study feasible.

So far, only a few plant association mapping panels with adequate sample size have been thoroughly studied and our knowledge on the general characteristics of these diverse populations is still very limited, preventing us from conducting a thorough examination. For example, a second maize panel has only genomewide SSR data (Camus-Kulandaivelu et al., 2006); the *Arabidopsis* panel contains only 95 diverse lines (Zhao et al., 2007); and the grain sorghum [*Sorghum bicolor* (L.) Moench] diversity panel has only been genotyped with 49 SSR markers (Casa et al., 2008). Accordingly, we have focused our computer simulations on relative kinship rather than on both population structure and relative kinship. At the same time, we chose to conduct a series of marker sampling using two available data sets as the base data:

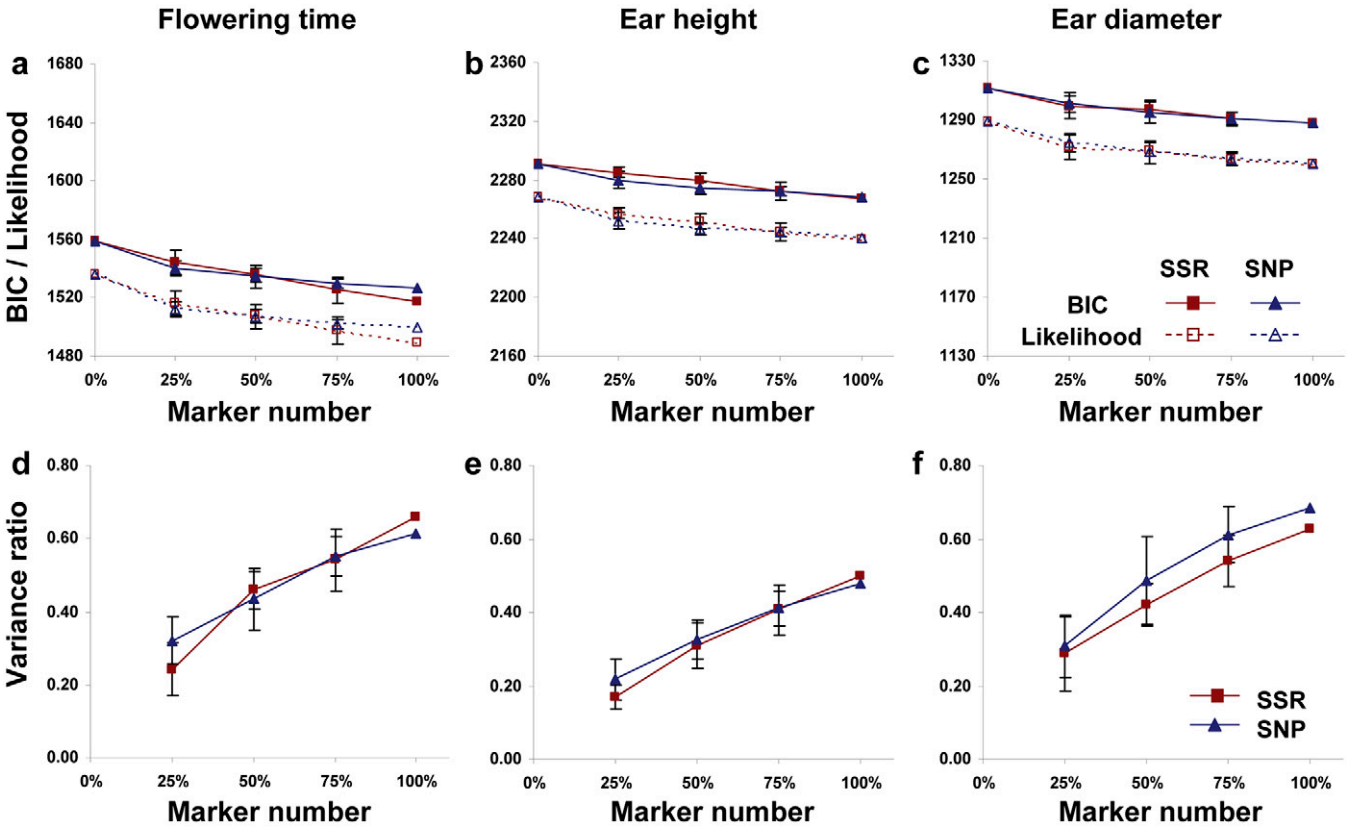

Figure 6. Model fitting for maize quantitative traits with population structure (Q) being constant but relative kinship (K) being estimated with different numbers of markers. (a–c) Bayesian Information Criterion (BIC) and –2 residual log-likelihood values of mixed models with different K based on different proportion of the whole set of background markers (912 single nucleotide polymorphisms [SNPs] or 89 simple sequence repeats [SSRs]). The constant Q was estimated with the whole set of 89 SSR markers. (d–f) Variance ratio [$V_g/(V_g + V_R)$] estimates from mixed models. Standard deviations are shown by vertical bars.

one from a diverse set of maize inbred lines with complex population structure and familial relatedness (Liu et al., 2003; Flint-Garcia et al., 2005; Yu et al., 2006), and the other from a group of crossbred dogs (Mateescu et al., 2005; Todhunter et al., 2005). To our knowledge, the multiple-donor populations derived with both backcross and intercross, as represented by the canine data, are also popular in many plant breeding programs, particularly for disease resistance selection and introgression. We expect a similar genetic data structure in these plant breeding populations as that of the canine data.

With computer simulations of a simulated association mapping panel with only kinship relationships, we demonstrated that the likelihood value and the variance ratio estimate approached their stabilized values as the number of background markers increased. This was consistent under different trait complexities and initial marker number. Interestingly, we also consistently obtained a better variance component estimate with pedigree information than marker information across different simulation scenarios. Since markers allow an a posteriori estimation of identity relative to that expected from pedigree, one could expect a better adjustment of marker-based models. This result points to the probably inevitable limitation of marker-based relationship estimation process. While molecular marker information provides a viable route for relationship quantification when pedigree information is not available or incomplete, there seems to be a limit on how close this marker-based relationship can reach. In our simulations, simulated heritabilities were consistently underestimated by the variance ratio with marker-based relative kinships. Additional research with other relationship estimates (Weir et al., 2006), particularly relating relationship with phenotypic trait variation as demonstrated in our study, is desirable to see whether this underestimation applies to other measures of relatedness. But we speculate that the general pattern in changes in likelihood value and variance ratio estimate with respect to number of markers used in relationship construction remains the same.

Our results with maize data further showed that relative kinship is more sensitive to the number of markers than population structure in terms of model fitting and variance component estimation. This may be explained by the different goals between these two relatedness quantification approaches as well as the way the two different estimates were used in the mixed model. For population structure estimation, group membership and admixture proportion are the main focus, but for relative kinship estimation, pairwise relatedness both inter- and intragroup among all individuals is the focus. In the mixed model, population structure estimates are used to estimate different slopes (as fixed variables) for

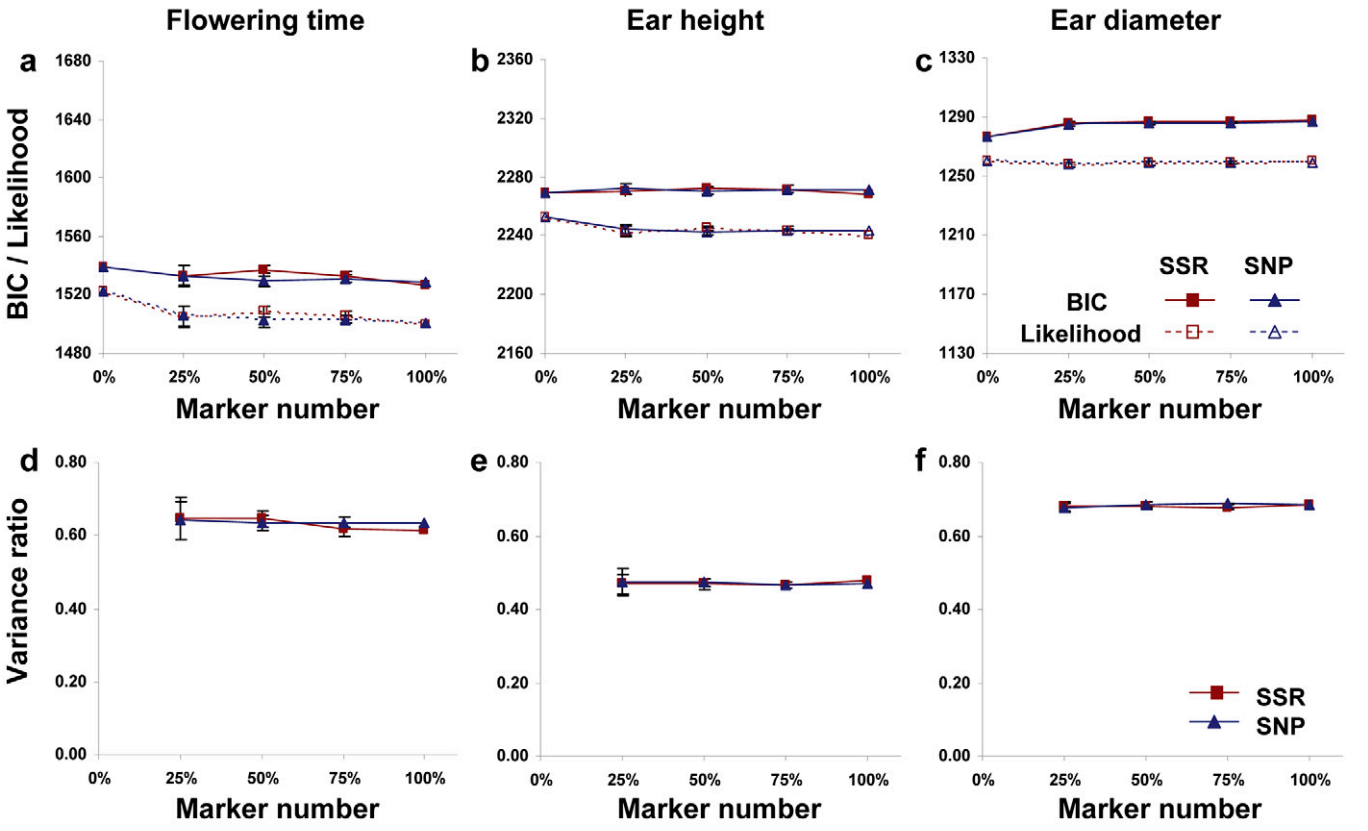

Figure 7. Model fitting for maize quantitative traits with relative kinship (K) being constant but population structure (Q) being estimated with different numbers of markers. (a–c) Bayesian Information Criterion (BIC) and –2 residual log-likelihood values of mixed models with different Q based on different proportion of the whole set of background markers (912 single nucleotide polymorphisms [SNPs] or 89 simple sequence repeats [SSRs]). The constant K was estimated with the whole set of 912 SNP markers. (d–f) Variance ratio [$V_g/(V_g + V_R)$] estimates from mixed models. Standard deviations are shown by vertical bars.

subgroups, while relative kinship estimates are used to estimate the genetic variance (as a random variable). Nevertheless, the relative importance of population structure and relative kinship depends on the actual genetic structure of the real population in question. It is likely that population structure can be more sensitive than relative kinship in some other cases.

In general, both the simulated and empirical data revealed that a robust relatedness estimate, particularly kinship estimate, requires an adequate number of background markers. Stabilized model-fitting statistics and variance component estimates can be used to assess how well the marker-based genetic relationships explain the phenotypic variation among individuals for multiple complex traits. For relative kinships, the pattern of changes in the likelihood values and variance ratio estimates was consistent across different data and analysis schemes: the simulated data with only kinship relationships, the maize data with a constant population structure, and the canine data. In practice, kinship construction with subsets of the whole marker panel and subsequent model testing could provide information on whether there is a sufficient number of background markers to quantify genetic relationships among individuals. Previous studies have developed a variety of estimators to quantify relatedness (Queller and Goodnight, 1989; Loiselle et al., 1995; Ritland, 1996b; Lynch and Ritland, 1999; Wang, 2002; Milligan, 2003). These estimators differ in their accuracy and precision (Blouin, 2003; Milligan, 2003). A comparison of different estimators with empirical data from multiple complex traits would be interesting but is beyond the scope of this study. We chose the kinship measure (Loiselle et al., 1995) because of its simultaneous estimation of inter- and intra-individual relationships with a symmetric matrix and it is free of assumption of Hardy–Weinberg equilibrium. While previous studies mainly used simulations to test the accuracy of marker-based relative kinship estimates to simulated true kinships, the current study is the first one, to our knowledge, in which multiple phenotypes were used to test the robustness of kinships based on different number of markers in the context of association mapping. For association mapping panels with diverse germplasm, our approach provided a more relevant test because it is impossible to obtain the true kinships among all individuals. The use of the ratio of the differences in the probabilities of identity in state was recently proposed to give a generic definition for inbreeding coefficient and relatedness (Rousset, 2002). It shifted from identity by descent for which a reference population is required, but extremely difficult to define in reality, to identity in state. Relatedness is, by definition, a relative

quantity in terms of time and population. For association analysis with diverse germplasm, this type of relative quantification serves well the purpose to account for the part of phenotypic variation that is stemmed from the hidden genetic factors in the given population.

A direct comparison of SSRs and SNPs in quantifying kinship in maize was not made because such a comparison is confounded by many factors that differ between the two types of markers, such as initial discovery and detection process, frequency distributions, and evolutionary history (Tenaillon et al., 2002). Instead, we focused on providing a parallel examination with available empirical data. As expected, with higher PIC, a much smaller number of SSRs provided equivalent relatedness estimates with respect to the fit of the phenotype model as do a large number of SNPs. Although the cost of generating SNP data through high-throughput genotyping techniques is much lower than that for the traditional SSRs, our results demonstrated that estimating relatedness using SSRs is feasible to initiate association mapping studies. As more random markers are used in population structure and kinship estimation, the dependency (i.e., correlated allele frequencies) among markers would increase. This has been considered in the latest method of inferring population structure by STRUCTURE (Falush et al., 2003). However, given the linkage disequilibrium decays within a short distance for most diverse germplasm examined so far (Flint-Garcia et al., 2003), this should not be a serious concern unless the marker number becomes much higher than the level we examined here. A detailed examination on marker dependency is desirable but beyond the scope of the current study.

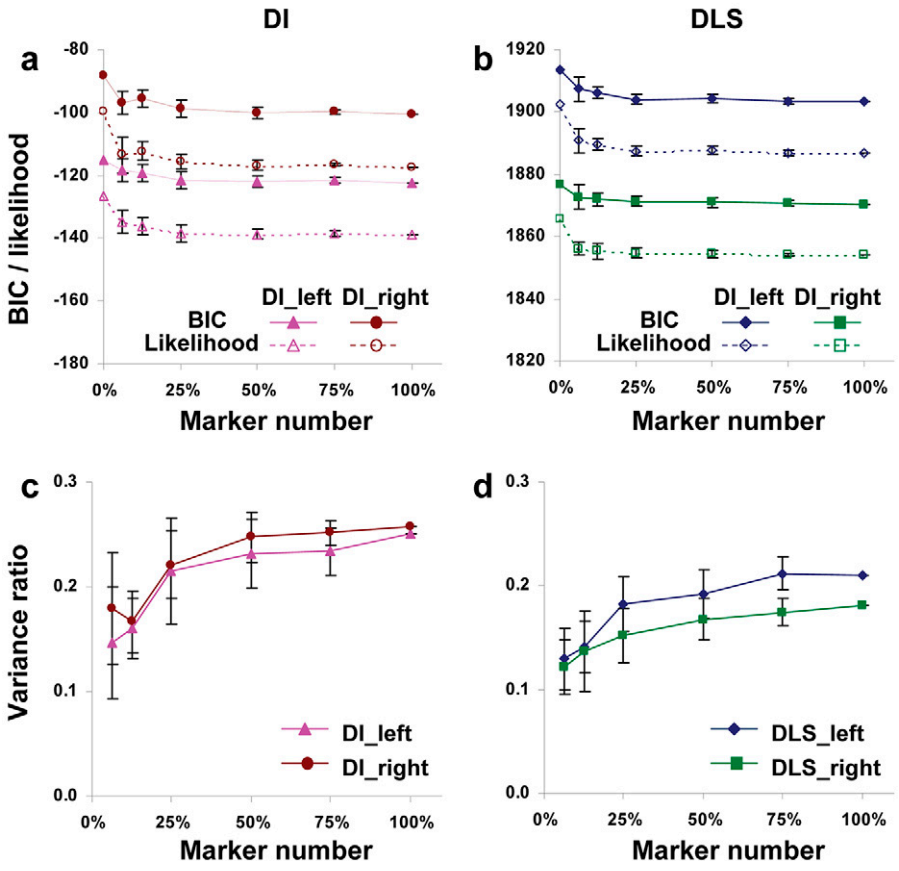

Figure 8. Model fitting for canine dysplastic traits. (a, b) Bayesian Information Criterion (BIC) and –2 residual log-likelihood values of mixed models with different relative kinship (K) matrices based on different proportion of 471 simple sequence repeats (SSRs). (c, d) Variance ratio [$V_g/(V_g + V_R)$] estimates from mixed models. Standard deviations are shown by vertical bars. DI, distraction index; DLS, dorsolateral subluxation.

In our simulation and analysis, the traditional assumption of the mixed model was kept throughout, that is, only the additive genetic effect was investigated. The importance of nonadditive effects (Lee and Van der Werf, 2006; Wardyn et al., 2007) may also affect the evaluation but is more relevant to the canine data (i.e., both simulated data and maize data were based on inbred lines with only additive-by-additive type of epistasis as the source of nonadditive effect). Moreover, the assumption of the background markers being selectively neutral may affect the relationship estimation and subsequent model testing. The approach we proposed in the current study, however, should generally apply since any inaccuracy in relationship estimation is expected to show in the process of linking these relationships to phenotypic variation observed. Although we only investigated biallelic markers in the computer simulations, the extension to multiallelic markers is straightforward as was demonstrated with SSR markers in maize and canine data.

While we observed a reduced variation in model-fitting statistics and variance ratios with an increased number of markers in the canine data, this trend was less obvious in maize. Presumably, besides the much higher genetic diversity in maize than in domestic dogs at the species level, the assembly of these worldwide maize inbred lines (Liu et al., 2003; Flint-Garcia et al., 2005) was much wider than that of the dogs, which had a limited number of founders from two different breeds (Todhunter et al., 2005). Indeed, the large number of SSRs scored for the canine panel was aimed at obtaining an initial genomewide scan rather than to provide background markers for quantifying genetic relationships (Todhunter et al., 2005). Among many other reasons, the scale and pattern differences in variance ratio estimates between maize and canine traits also may be due to differences in genetic architecture. While many empirical studies have suggested the quantitative nature of the aforementioned maize traits (Hallauer and Miranda Filho, 1988) and agreed with our results

of high variance ratios, studies for canine hip dysplasia suggested the possibility of fewer major loci (Todhunter et al., 2003a, 2003b, 2005), corroborating the low variance ratios found in this study. Variation in BIC values and variance ratios was not unexpected for two reasons. First, the kinship estimation procedure itself has certain bias and variances as shown in previous studies (Blouin, 2003; Milligan, 2003). Second, the variances of variance components by default are large and particularly so for small sample sizes (Lynch and Walsh, 1998). Because of this, we would recommend that multiple (e.g., 10 in our simulations) resampling repetitions at different subsets of markers are conducted in using the proposed method to avoid the detection of artifacts rather than a true plateau. Furthermore, caution must be taken in interpreting the results because our proposed approach starts with a full set of markers that presumably result in a close estimate of kinship to the true kinship.

In conclusion, we have shown that since both population structure and relative kinship are first estimated from markers and then fitted in the mixed model as covariates and variance–covariance matrix, the inaccuracy introduced in this estimation process decreases the maximum likelihood of the model explaining phenotypic variation. We have demonstrated with computer simulations and empirical data that a robust estimation of kinship for use in association mapping with diverse germplasm requires a certain amount of background markers (e.g., 300–600 biallelic markers for simulated pedigree materials, >1000 SNPs or 100 SSRs for the diverse maize panel, and about 100 SSRs for the canine panel). The robustness of relationship estimate can be tested by fitting multiple phenotypic traits with different estimates based on subsets of random markers. The number of markers required is much higher for biallelic SNPs than for multiallelic SSRs. Furthermore, the number of markers required for robust relationship estimation is not a statistical question by itself and needs to be addressed from both genetic and statistical perspectives. Future research should address the effect of relationship estimate on detection power and false discovery rate for different sizes of QTLs. As we gain a better understanding from many ongoing plant and animal association studies, a thorough theoretical examination with both population structure and relative kinship should provide more insights into the experimental design, genetic structure, and data analysis in association mapping. Fortunately, this process is being expedited with the continually decreasing cost of genotyping with rapidly evolving genomic technologies.

Acknowledgments

We thank two anonymous reviewers for their critical review of the manuscript. This project is supported by the National Research Initiative (NRI) Plant Genome Program of the USDA Cooperative State Research, Education and Extension Service (CSREES) to JY. We acknowledge funding support from Morris Animal Foundation (RST), Cornell Advanced Technology Biotechnology (RST), Master Foods Inc. (RST), USDA-ARS (ESB), and U.S. National Science Foundation (DBI-9872631 and DBI-0321467) (ESB and SK).

References

Anderson, J.A., G.A. Churchill, J.E. Autrique, S.D. Tanksley, and M.E. Sorrells. 1993. Optimizing parental selection for genetic-linkage maps. Genome 36:181–186.

Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. 2005. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 1:e60.

Blouin, M.S. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol. Evol. 18:503–510.

Breseghello, F., and M.E. Sorrells. 2006. Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177.

Camus-Kulandaivelu, L., J. Veyrieras, B. Gouesnard, A. Charcosset, and D. Manicacci. 2007. Evaluating the reliability of structure outputs in case of relatedness between individuals. Crop Sci. 47:887–890.

Camus-Kulandaivelu, L., J.B. Veyrieras, D. Madur, V. Combes, M. Fourmann, S. Barraud, P. Dubreuil, B. Gouesnard, D. Manicacci, and A. Charcosset. 2006. Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the *Dwarf8* gene. Genetics 172:2449–2463.

Casa, A.M., G. Pressoira, P.J. Brown, S.E. Mitchell, W.L. Rooney, M.R. Tuinstrac, C.D. Franks, and S. Kresovich. 2008. Community resources and strategies for association mapping in sorghum. Crop Sci. 48:30–40.

Darvasi, A., and S. Shifman. 2005. The beauty of admixture. Nat. Genet. 37:118–119.

Devlin, B., S.A. Bacanu, and K. Roeder. 2004. Genomic control to the extreme. Nat. Genet. 36:1129–1130.

Devlin, B., and K. Roeder. 1999. Genomic control for association studies. Biometrics 55:997–1004.

Devlin, B., K. Roeder, and L. Wasserman. 2001. Genomic control, a new approach to genetic-based association studies. Theor. Popul. Biol. 60:155–166.

Doerge, R.W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. 3:43–52.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. 14:2611–2620.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–374.

Flint-Garcia, S.A., A.C. Thuillet, J. Yu, G. Pressoir, S.M. Romero, S.E. Mitchell, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2005. Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant J. 44:1054–1064.

Hallauer, A.R., and J.B. Miranda Filho. 1988. Quantitative genetics in maize breeding. 2nd ed. Iowa State Univ. Press, Ames.

Hardy, O.J., and X. Vekemans. 2002. SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618–620.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

Henderson, C.R. 1984. Application of linear models in animal breeding. Univ. of Guelph, Guelph, ON.

Hey, J., and C.A. Machado. 2003. The study of structured populations: New hope for a difficult and divided science. Nat. Rev. Genet. 4:535–543.

Hirschhorn, J.N., and M.J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6:95–108.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

Lander, E.S., and N.J. Schork. 1994. Genetic dissection of complex traits. Science 265:2037–2048.

Lee, S.H., and J.H. Van der Werf. 2006. Using dominance relationship coefficients based on linkage disequilibrium and linkage with a general complex pedigree to increase mapping resolution. Genetics 174:1009–1016.

Lindblad-Toh, K., C.M. Wade, T.S. Mikkelsen, E.K. Karlsson, D.B. Jaffe, M. Kamal, M. Clamp, J.L. Chang, E.J. Kulbokas III, M.C. Zody, E. Mauceli, X. Xie, M. Breen, R.K. Wayne, E.A. Ostrander, C.P. Ponting, F. Galibert, D.R. Smith, P.J. DeJong, E. Kirkness, P. Alvarez, T. Biagi, W. Brockman, J. Butler, C.W. Chin, A. Cook, J. Cuff, M.J. Daly, D. DeCaprio, S. Gnerre, M. Grabherr, M. Kellis, M. Kleber, C. Bardeleben, L. Goodstadt, A. Heger, C. Hitte, L. Kim, K.P. Koepfli, H.G. Parker, J.P. Pollinger, S.M. Searle, N.B. Sutter, R. Thomas, C. Webber, J. Baldwin, A. Abebe, A. Abouelleil, L. Aftuck, M. Ait-Zahra, T. Aldredge, N. Allen, P. An, S. Anderson, C. Antoine, H. Arachchi, A. Aslam, L. Ayotte, P. Bachantsang, A. Barry, T. Bayul, M. Benamara, A. Berlin, D. Bessette, B. Blitshteyn, T. Bloom, J. Blye, L. Boguslavskiy, C. Bonnet, B. Boukhgalter, A. Brown, P. Cahill, N. Calixte, J. Camarata, Y. Cheshatsang, J. Chu, M. Citroen, A. Collymore, P. Cooke, T. Dawoe, R. Daza, K. Decktor, S. DeGray, N. Dhargay, K. Dooley, K. Dooley, P. Dorje, K. Dorjee, L. Dorris, N. Duffey, A. Dupes, O. Egbiremolen, R. Elong, J. Falk, A. Farina, S. Faro, D. Ferguson, P. Ferreira, S. Fisher, M. FitzGerald, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Littell, R.C., G.A. Milliken, W.W. Stroup, R.D. Wolfinger, and O. Schabenberger. 2006. SAS for mixed models. 2nd ed. SAS Inst., Cary, NC.

Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128.

Loiselle, B.A., V.L. Sork, J. Nason, and C. Graham. 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am. J. Bot. 82:1420–1425.

Lynch, M., and K. Ritland. 1999. Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766.

Lynch, M., and J.B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA.

Mackay, T.F. 2001. The genetic architecture of quantitative traits. Annu. Rev. Genet. 35:303–339.

Mateescu, R.G., Z. Zhang, K. Tsai, J. Phavaphutanon, N.I. Burton-Wurster, G. Lust, R. Quaas, K. Murphy, G.M. Acland, and R.J. Todhunter. 2005. Analysis of allele fidelity, polymorphic information content, and density of microsatellites in a genome-wide screening for hip dysplasia in a crossbreed pedigree. J. Hered. 96:847–853.

Milligan, B.G. 2003. Maximum-likelihood estimation of relatedness. Genetics 163:1153–1167.

Neale, D.B., and O. Savolainen. 2004. Association genetics of complex traits in conifers. Trends Plant Sci. 9:325–330.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.

Pritchard, J.K., and N.A. Rosenberg. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65:220–228.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Queller, D.C., and K.F. Goodnight. 1989. Estimating relatedness using genetic-markers. Evolution Int. J. Org. Evolution 43:258–275.

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

Ritland, K. 1996a. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. 67:175–186.

Ritland, K. 1996b. Marker-based method for inferences about quantitative inheritance in natural populations. Evolution Int. J. Org. Evolution 50:1062–1073.

Rousset, F. 2002. Inbreeding and relatedness coefficients: What do they measure? Heredity 88:371–380.

SAS Institute. 1999. SAS/STAT user's guide. Version 8. SAS Inst., Cary, NC.

Schwarz, G. 1978. Estimating dimension of a model. Ann. Stat. 6:461–464.

Senior, M.L., E.C.L. Chin, M. Lee, J.S.C. Smith, and C.W. Stuber. 1996. Simple sequence repeat markers developed from maize sequences found in the GENBANK database: Map construction. Crop Sci. 36:1676–1683.

Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732.

Sillanpaa, M.J., and J. Corander. 2002. Model choice in gene mapping: What and why. Trends Genet. 18:301–307.

Syvanen, A.C. 2005. Toward genome-wide SNP genotyping. Nat. Genet. 37(Suppl.):S5–S10.

Tenaillon, M.I., M.C. Sawkins, L.K. Anderson, S.M. Stack, J. Doebley, and B.S. Gaut. 2002. Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics 162:1401–1413.

Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. Nat. Genet. 28:286–289.

Todhunter, R.J., S.P. Bliss, G. Casella, R. Wu, G. Lust, N.I. Burton-Wurster, A.J. Williams, R.O. Gilbert, and G.M. Acland. 2003a. Genetic structure of susceptibility traits for hip dysplasia and microsatellite informativeness of an outcrossed canine pedigree. J. Hered. 94:39–48.

Todhunter, R.J., G. Casella, S.P. Bliss, G. Lust, A.J. Williams, S. Hamilton, N.L. Dykes, A.E. Yeager, R.O. Gilbert, N.I. Burton-Wurster, C.C. Mellersh, and G.M. Acland. 2003b. Power of a Labrador Retriever–Greyhound pedigree for linkage analysis of hip dysplasia and osteoarthritis. Am. J. Vet. Res. 64:418–424.

Todhunter, R.J., R. Mateescu, G. Lust, N.I. Burton-Wurster, N.L. Dykes, S.P. Bliss, A.J. Williams, M. Vernier-Singer, E. Corey, C. Harjes, R.L. Quaas, Z. Zhang, R.O. Gilbert, D. Volkman, G. Casella, R. Wu, and G.M. Acland. 2005. Quantitative trait loci for hip dysplasia in a cross-breed canine pedigree. Mamm. Genome 16:720–730.

Voight, B.F., and J.K. Pritchard. 2005. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1:e32.

Wang, J. 2002. An estimator for pairwise relatedness using molecular markers. Genetics 160:1203–1215.

Wang, W.Y., B.J. Barratt, D.G. Clayton, and J.A. Todd. 2005. Genome-wide association studies: Theoretical and practical concerns. Nat. Rev. Genet. 6:109–118.

Wardyn, B.M., J.W. Edwards, and K.R. Lamkey. 2007. The genetic structure of a maize population: The role of dominance. Crop Sci. 47:467–476.

Weir, B.S., A.D. Anderson, and A.B. Hepler. 2006. Genetic relatedness analysis: Modern data and new challenges. Nat. Rev. Genet. 7:771–780.

Wilson, L.M., S.R. Whitt, A.M. Ibanez, T.R. Rocheford, M.M. Goodman, and E.S. Buckler. 2004. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719–2733.

Yu, J., M. Arbelbide, and R. Bernardo. 2005. Power of in silico QTL mapping from phenotypic, pedigree, and marker data in a hybrid breeding program. Theor. Appl. Genet. 110:1061–1067.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208.

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An *Arabidopsis* example of association mapping in structured samples. PLoS Genet. 3:e4.