ORIGINAL PAPER

Mixed model approaches for the identification of QTLs within a maize hybrid breeding program

Fred A. van Eeuwijk · Martin Boer · L. Radu Totir · Marco Bink · Deanne Wright ·
Christopher R. Winkler · Dean Podlich · Keith Boldman · Andy Baumgarten ·
Matt Smalley · Martin Arbelbide · Cajo J. F. ter Braak · Mark Cooper

Abstract Two outlines for mixed model based approaches to quantitative trait locus (QTL) mapping in existing maize hybrid selection programs are presented: a restricted maximum likelihood (REML) and a Bayesian Markov Chain Monte Carlo (MCMC) approach. The methods use the in-silico-mapping procedure developed by Parisseaux and Bernardo ([2004]) as a starting point. The original single-point approach is extended to a multi-point approach that facilitates interval mapping procedures. For computational and conceptual reasons, we partition the full set of relationships from founders to parents of hybrids into two types of relations by defining so-called intermediate founders. QTL effects are defined in terms of those intermediate founders. Marker based identity by descent relationships between intermediate founders define structuring matrices for the QTL effects that change along the genome. The dimension of the vector of QTL effects is reduced by the fact that there are fewer intermediate founders than parents.

Furthermore, additional reduction in the number of QTL effects follows from the identification of founder groups by various algorithms. As a result, we obtain a powerful mixed model based statistical framework to identify QTLs in genetic backgrounds relevant to the elite germplasm of a commercial breeding program. The identification of such QTLs will provide the foundation for effective marker assisted and genome wide selection strategies. Analyses of an example data set show that QTLs are primarily identified in different heterotic groups and point to complementation of additive QTL effects as an important factor in hybrid performance.

Abbreviations
BLUE	Best linear unbiased estimator
GCA	General combining ability
GS	Genome wide selection
MAS	Marker assisted selection
MCMC	Markov chain Monte Carlo
QTL	Quantitative trait locus
REML	Restricted maximum likelihood
SCA	Specific combining ability
HP	Hybrid performance

F. A. van Eeuwijk (✉) · M. Boer · M. Bink · C. J. F. ter Braak
Biometris, Wageningen UR, P.O. Box 100,
6700 AC Wageningen, The Netherlands
e-mail: fred.vaneeuwijk@wur.nl
URL: www.biometris.wur.nl

L. R. Totir · D. Wright · C. R. Winkler · D. Podlich ·
K. Boldman · A. Baumgarten · M. Smalley · M. Arbelbide ·
M. Cooper
Pioneer Hi-Bred International, A DuPont Business,
Johnston, IA 50131, USA

Introduction

The transition from open-pollinated populations to double-cross and then single-cross hybrids in maize breeding was a major component of the long-term genetic gain for yield of maize in the US Corn Belt. It is tempting to think that increasing heterosis, i.e., an increasing difference between hybrid performance and average parental performance, was

an important driving force behind this gain. In a study of a time series of maize hybrids released between the early 1930s till 2001, Duvick et al. (2004) concluded that heterosis played a minor role in the improved performance of present day hybrids, the role of additive gene action was found far more important. Nevertheless, the phenomenon of heterosis remains interesting and papers keep appearing about the subject, where it is remarkable that hardly any consensus exists about the genetic mechanism(s) that may underlie the phenomenon: dominance, overdominance, pseudo-over-dominance, and epistasis, or combinations of these components. For example, Frascaroli et al. (2007) emphasized the role of dominance and overdominance, whereas they found little evidence for epistasis. In contrast, Bernardo (1996a, b), in line with Duvick et al. (2004), proposes to omit specific combining ability from the analysis of hybrid performance, implying a negligible role for dominance and overdominance. Similar conclusions on the low impact of specific combining ability (SCA) were recently reached by Fischer et al. (2008) and Schrag et al. (2009). Deviating from this trend, Melchinger et al. (2007) present an argument for the importance of epistasis in heterosis.

Although the importance and underlying mechanism for heterosis still elicit discussion, the heterosis question constitutes only a part of the more relevant and encompassing question concerning the prediction of hybrid performance (HP). Nowadays, the key question for HP prediction is whether marker information on the parents, or on related inbred lines suffices for HP prediction, thereby obviating field evaluations of the particular hybrids themselves. To enable HP prediction for a range of quantitative traits, we are interested in making use of pedigree relations (coancestry) between the lines in the parental generations, of phenotypic records on hybrid performance for hybrids other than the ones to be predicted, and of marker information from the parent generation.

Two types of approaches can be distinguished with respect to HP prediction. First, there is a class of approaches that can be characterized as distance methods: either HP or SCA, as part of HP, is regressed on marker information, whether in the form of a single predictor or a small set of predictors derived from operations on a matrix of similarity coefficients or coancestry coefficients (Charcosset et al. 1998) or in the form of a predictor set that consists of a subset of coded markers (Vuylsteke et al. 2000; Schrag et al. 2006, 2009). A second approach consists in a more classic elaboration of the quantitative genetic theory within the mixed model framework. This approach is advocated by Bernardo (1994, 1996a, b, 1999) and finds its culmination in Parisseaux and Bernardo (2004) and Yu et al. (2005). In the last two papers, the so-called method of in-silico-mapping is presented: the use of accumulated phenotypic data in public and private plant breeding programs

for quantitative trait locus (QTL) mapping. Four advantages are mentioned for in-silico-mapping over classical QTL mapping using designed crosses: (1) larger mapping populations are available; (2) evaluation takes place in multiple environments, so that results will be applicable across a wide range of future growing environments; (3) a wide sample of germplasm and genetic backgrounds is tested, so fewer problems will occur with respect to the validity of predictions for other genetic backgrounds; (4) field data are already available, so no extra costs need to be made to obtain the phenotypic data. In addition, new inbred lines are routinely genotyped for multiple purposes within commercial breeding programs.

The two classical stages of hybrid breeding programs are the development of promising inbred lines followed by the identification and selection of superior hybrids created from crosses between the inbred lines. Reliable prediction of HP on the basis of information produced in the second stage (hybrid selection) of on-going breeding programs would be extremely useful. A prerequisite for such a HP prediction strategy is the availability of advanced QTL mapping methodology, i.e., methodology that is able to accommodate the specifics of phenotypic, genotypic, and pedigree data represented in hybrid selection programs. In this paper, we propose a mixed model based statistical framework to map QTLs in hybrid selection programs. We use and extend the in-silico-mapping QTL model first described in Parisseaux and Bernardo (2004) in our development. The latter approach was a single point analysis restricted to evaluations at marker positions. Interval mapping, a multi-point analysis, which was indicated as computationally prohibitive by Parisseaux and Bernardo in their 2004 paper, is possible using our approach. Further differences are that we model QTL effects as random, whereas these effects were modeled fixed in Parisseaux and Bernardo (2004). The assumption of QTL effects being random allows us to impose structure on their variances and covariances. As a consequence, we are able to arrive at a QTL analysis that combines elements of linkage and linkage disequilibrium mapping, closely akin to the way in which Meuwissen et al. (2001) defined such a combined mapping strategy. This latter element is essential for any QTL mapping methodology adapted to the details of data generated from hybrid selection programs. The approach by Parisseaux and Bernardo (2004) focused on additive QTL effects. Our modeling framework can deal with both additive and dominant effects, although, in accordance with the remark by Bernardo (1999) about the negligibility of dominance effects in QTL mapping, we found a minor role for dominance in our illustration data and thus we will not further report on dominance aspects in this paper. Extensions to include epistasis in the models are currently under study.

Below we will first describe a restricted maximum likelihood (REML) based mixed model approach to QTL mapping in hybrid selection data, followed by a Bayesian Markov Chain Monte Carlo (MCMC) mixed model approach. The REML and Bayesian MCMC approaches share the same linear model structure and model terms, and they use equivalent definitions for the genetic relationships between the individuals in the pedigree. In the REML approach, the number of alleles per QTL locus is typically larger than two, whereas in the Bayesian approach this number is fixed at two. In the REML approach, a single QTL model is fitted at a grid of evaluation points across the genome. In the Bayesian approach, multi-QTL models with random numbers of causative loci located at random positions on the genome are fitted as competing models during the MCMC process. Following a description of the major theoretical aspects of the two mixed model approaches introduced above, we briefly illustrate their performance using data on ear height, a trait known to show intermediate heritability. The data stem from a maize hybrid selection program at Pioneer Hi-Bred International. The intention of this paper is to investigate the suitability of current mixed model methodology as run on standard PCs for mapping QTLs in hybrid selection programs; a feasibility study and not an exhaustive comparison of REML and Bayesian mixed models in a QTL mapping context (like, for example, Bauer et al. 2009).

Data

The maize hybrid selection data used for illustration of our models can be considered to represent a generic example data set. Our calculations are based on proprietary data of Pioneer Hi-Bred International. The phenotypic trait analyzed was ear height, originally measured in inches from the ground to the node from which the ear was attached, but below presented in centimeters. The estimated heritability of ear height was 0.36. Ear height was available for 1,700 hybrids produced from crosses between inbred parents that belong to two heterotic groups: 1 and 2. Hybrids were evaluated on average at 15 locations during two growing seasons, 2004 and 2005, in the US Corn Belt.

At the hybrid level, the phenotypic data points used in the QTL analyses came from a two-stage analysis (Smith et al. 2001). In the first stage, trials were analyzed by location to compute hybrid means (Best Linear Unbiased Estimates, or BLUEs), which were stored with their relative weights. In the second stage, these hybrid-by-location BLUEs were analyzed in a multi-location analysis using an additive mixed model with locations and hybrids as fixed effects. The resulting across-location hybrid BLUEs obtained in the second stage were used as phenotypic data

in the QTL analyses. We were interested in the average hybrid performance across locations and for that reason did not pursue analyses of genotype by environment interaction or QTL by environment interaction.

Our stage wise approach for calculating the hybrid means for use in the QTL analysis was dictated by the large computer requirements in our genetic (QTL) analysis; a one hit approach that would fit a genetic model starting from plot data was infeasible.

Heterotic group 1 consisted of 222 inbred parents, versus 213 inbred parents in group 2, making 435 parents in total. Pedigree data were available for the parents, where the pedigree was complete for three ancestral generations. Going back, three generations in the pedigree, 62 inbred founder lines were defined for group 1, versus 55 for group 2, making 117 founders in total. For all 435 parents and most ancestors, 768 SNP markers were scored. Furthermore, using a proprietary estimation method, pedigree relationships, and denser marker coverage than the 768 SNP markers mentioned above, genetic relationship matrices between the 117 founders were calculated at 1 centi-Morgan intervals along the full length of the genome.

REML approach

Structure of pedigree and nomenclature

Pedigree graphs show the genetic relationships between ancestors and descendents. Figure 1 gives a schematic pedigree graph for our maize hybrid breeding scheme, where the pedigree contains two heterotic groups (vertical split) each having a number of, mostly, group specific founders at the top, that transmit their alleles to the hybrids at the bottom. For computational and conceptual reasons, we partitioned the pedigree from true founder to parental lines (horizontal split) by defining an intermediate level of ancestral lines to which we will refer from now on as (intermediate) founders, bearing in mind that these lines represent not the ultimate founder lines. The exact structure of the pedigree beyond the (intermediate) founders is considered to be unknown, but quantitative summaries of those relationships are available in the form of a collection of symmetric relationship matrices containing marker based identity by descent (IBD) information (see below). From the (intermediate) founders to the parental lines of hybrids, we take pedigree relationships to be known and available.

Following Fig. 1, we now define the different types of individuals that will appear in our analyses. At the highest level in the (considered) known pedigree we find the (intermediate) founders (F), at the lowest levels the hybrids
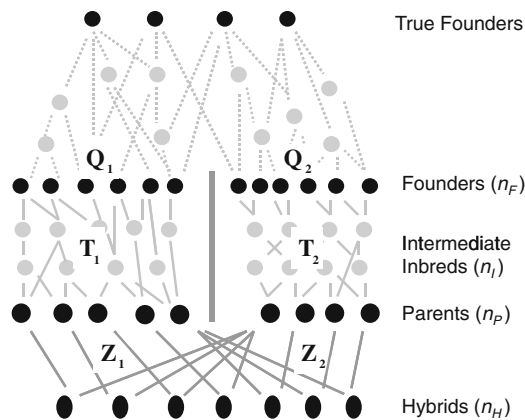
Fig. 1 Schematic pedigree relationships, composition of heterotic groups, and matrices that represent genetic relationships in mixed models

(H) with field trial data. As remarked above, additional quantitative relationship information was available at a grid along the genome beyond the level of the (intermediate) founders. Just above the hybrids, we see their parents (P), while inbetween parents and founders we have what we will call intermediate inbred lines (I). The numbers of founders, hybrids, parental lines and intermediate lines are given by, respectively, n_F, n_H, n_P, and n_I. To indicate the numbers of parents in heterotic group 1 and 2, we write n_{P1} and n_{P2}, and similarly n_{F1} and n_{F2} for the founders.

A reference mixed model for QTL mapping

Parisseaux and Bernardo (2004) developed a mixed model for QTL mapping in hybrid selection programs that served as a reference model for our QTL mapping approach. We give a description of the Parisseaux and Bernardo (2004) QTL model and introduce some notation. Their model reads as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}_1\boldsymbol{\alpha}_1 + \mathbf{M}_2\boldsymbol{\alpha}_2 + \mathbf{Z}_1\mathbf{g}_1 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e} \qquad (1)$$

with $\mathbf{y}$ the vector of phenotypic observations, $\mathbf{X}$ the design matrix for adjustment of non-genetic effects [in the Parisseaux and Bernardo (2004) paper this included corrections for multi-location trials], and $\boldsymbol{\beta}$ the corresponding vector of fixed effects. The vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ represented general combining abilities (GCA) associated with markers in group 1 and 2, and $\mathbf{g}_1$ and $\mathbf{g}_2$ represented GCA effects not related to markers in group 1 and 2. The dimensions of these vectors corresponded to the numbers of marker alleles in group 1 and 2 for one particular QTL, and the numbers of parent lines in groups 1 and 2 (n_{P1} and n_{P2}). Model (1) thus describes a model for a single QTL. The incidence matrices $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{Z}_1$, and $\mathbf{Z}_2$ allocated the corresponding effects to the hybrids. The marker effects $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ were treated as fixed, while $\mathbf{g}_1$ and $\mathbf{g}_2$ were treated as

random with a normal distribution and variance–covariance matrices $\mathbf{G}_1 V_{G1}$ and $\mathbf{G}_2 V_{G2}$, where $\mathbf{G}_1$ and $\mathbf{G}_2$ were the $n_{P1} \times n_{P1}$ and $n_{P2} \times n_{P2}$ coancestry matrices within groups 1 and 2, and V_{G1} and V_{G2} group specific variance components. The error term $\mathbf{e}$ had a diagonal variance–covariance matrix with the entries on the diagonal depending on the reciprocal of the number of replicates.

A mixed model for hybrid performance without marker information

The QTL model in Parisseaux and Bernardo (2004) described hybrid performance exclusively in terms of GCA, for both markers and polygenic background. Omitting the marker related terms from that model, gives a base line model for testing the effects of QTLs:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{g}_1 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e} \qquad (2)$$

The hybrid BLUEs in our data set were already adjusted for non-genetic effects in the analysis, so that the only parameter entering $\boldsymbol{\beta}$ was the general intercept, μ, while $\mathbf{X}$ consisted of the unity vector of length n_H. For each heterotic group, we calculated the coefficients of coancestry among inbred parents of hybrids, in terms of IBD probabilities, and formed the GCA variance–covariance structuring matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ mentioned above using a tabular method (Bernardo 2002). The matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ were used to structure GCA effects in both the REML and Bayesian approach. For the variance of $\mathbf{e}$, we simply took $\mathbf{I}V_e$, i.e., independent errors with constant variance, V_e.

Extending the reference mixed model for QTL mapping

We diverge from Parisseaux and Bernardo (2004) with respect to the exact form for introduction of marker and QTL information into the model. As alluded to above, instead of defining the QTLs at the level of the parents (P), we chose to define QTLs at the level of the (intermediate) founders (F). For this to be possible, the transition (descent) probabilities for alleles from founders to parents are needed. These probabilities were calculated at a 1 centi-Morgan grid along the genome using pedigree and marker data by a recursive or tabular method (see e.g., George et al. 2000; Bernardo 2002). The method is a top–down approach, starting with the (intermediate) founders and incorporates a Hidden Markov Model (HMM) approach (Lander and Green 1987).

For the first heterotic group, the $n_{P1} \times n_{F1}$ matrix of transition probabilities is called $\mathbf{T}_1$. When we want to test for a QTL being present in the first heterotic group at a particular position coinciding with the particular transition matrix, we can compare the following model with the GCA model (2) by a deviance test for a single variance component (Verbeke and Molenberghs 2000):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{T}_1\mathbf{a}_1 + \mathbf{Z}_1\mathbf{g}_1 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e} \qquad (3)$$

The vector $\mathbf{a}_1$ is a vector of random QTL allele effects corresponding to a QTL at the position of the $\mathbf{T}_1$ matrix of founder–parent transition probabilities in group 1. As mentioned, we can structure the variance–covariance matrix (VCOV) of the random QTL allele effects in $\mathbf{a}_1$ using pedigree and dense marker information on inbred lines preceding the (intermediate) founders. The structuring matrix for the VCOV of the intermediate founder alleles can be calculated at the same genomic grid as the matrix of transition probabilities, $\mathbf{T}_1$. For an overview of genetic relationships and their matrix representations in the mixed model, see Fig. 1.

The deviance test for variance components

Critical values for the deviance in the test for single variance components can be found from a two component, 0.5–0.5, mixture distribution consisting of a Chi-square distribution on zero degrees of freedom combined with a Chi-square distribution on one degree of freedom. However, as pointed out by a reviewer, this approximation is only valid under the assumption of independence between the hybrids. A more general test for single variance components, including the dependence configuration pertinent to our hybrid data, is presented by Greven et al. (2008), who show that the commonly used mixture of Chi-squares provides a conservative test. Because alternatives to standard Chi-square mixture approximations for deviance differences are still cumbersome, we will stick to these standard approximations, acknowledging that they produce conservative tests.

Linkage and linkage disequilibrium

We can interpret the structure of the VCOV of the founder marker allele effects as the addition of linkage disequilibrium information to our linkage analysis, in the spirit of Meuwissen et al. (2001). Effectively, the total of the pedigree and marker information on the inbred lines preceding the parents is split into two parts, where the split is determined by the choice of the (intermediate) founder individuals, in this case three generations before the parent lines. In the founder–parent part of the relationships, classical genetic relationships are assessed from pedigree and marker information using Hidden Markov Models. This information enters the model at the linkage level. For the relationships between inbred lines above the founder level, often characterized at high marker density, a summary is constructed in the form of a symmetric matrix of IBD probabilities. The latter information accounts for the covariance between the 'intermediate' founder level random terms and incorporates ancestral linkage disequilibrium information.

Decomposing founder effect structuring matrices

Define the structuring matrix for the variance of $\mathbf{a}_1$ in model (3) by $\mathbf{Q}_1$, then $\mathrm{VCOV}(\mathbf{a}_1) = \mathbf{Q}_1 V_{Q1}$, with V_{Q1} the corresponding variance component. The matrix $\mathbf{Q}_1$ may turn out to be non-positive definite, which will cause problems when fitting model (3). A solution is to use a spectral decomposition of $\mathbf{Q}_1 = \mathbf{U}_1\mathbf{U}_1^T$ and approximate $\mathbf{Q}_1$ by discarding the eigenvectors with negative eigenvalues (as long as these are small in absolute value): $\mathbf{Q}_1 = \mathbf{U}_1^* \mathbf{U}_1^{*T}$, where $\mathbf{U}_1^*$ represents the part of $\mathbf{U}$ with positive eigenvalues (Calinski et al. 2005; Piepho et al. 2008).

The spectral decomposition can also be used to replace the original founder structuring matrix $\mathbf{Q}_1$ by a low rank approximation that centers on founder groups instead of founders (both preceding the currently used 'intermediate' founders). The matrix $\mathbf{Q}_1$ is again decomposed as before: $\mathbf{Q}_1 = \mathbf{U}_1^* \mathbf{U}_1^{*T}$, but the number of columns in $\mathbf{U}_1^*$ is based on a small set of eigenvectors, in the order of 2–6, with the largest eigenvalues included. When we post-multiply $\mathbf{Z}_1\mathbf{T}_1$ by $\mathbf{U}_1^*$ and call this matrix $\mathbf{M}_1$; $\mathbf{M}_1 = \mathbf{Z}_1\mathbf{T}_1\mathbf{U}_1^*$, we can write (3) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{g}_1 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e} \qquad (4)$$

Model (4) has the same structure as model (1), but instead of testing for QTLs at marker positions exclusively, it can test for additive effect QTLs anywhere in the genome (on the 1 centi-Morgan grid, in our case). Further, when $\mathbf{U}_1^*$ consists of the first few eigenvectors, model (4) estimates QTL allele effects that are linear combinations of the (intermediate) founder QTL allele effects. The transformation of the initial founder effects $\mathbf{a}_1$ to the reduced set of effects $\mathbf{b}_1$ can be interpreted as the creation of approximating founder groups, going back to key individuals somewhere in the pedigree. Because of the linear transformation defined by $\mathbf{U}_1^*$, the effects in $\mathbf{b}_1$ are independent. The founder effects $\mathbf{a}_1$ can be found from those for $\mathbf{b}_1$ by $\mathbf{a}_1 = \mathbf{U}_1^*\mathbf{b}_1$.

An alternative to the spectral decomposition of $\mathbf{Q}_1$ is the factorization based on a latent class model as proposed by ter Braak et al. (2009): $\mathbf{Q}_1 = \mathbf{P}_1\mathbf{P}_1^T$, with the elements in $\mathbf{P}_1$ representing probabilities for individual founders to belong to a particular founder class. This latter factorization was especially useful in the Bayesian implementation of our approach.

Modeling two heterotic groups

It is straightforward to add to model (4) a QTL for the second heterotic group:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}_1\mathbf{b}_1 + \mathbf{M}_2\mathbf{b}_2 + \mathbf{Z}_1\mathbf{g}_1 + \mathbf{Z}_2\mathbf{g}_2 + \mathbf{e} \qquad (5)$$

Comparing model (5) to model (2) tests for QTLs in both heterotic groups, whereas comparing model (4) to model (2) tests only for a QTL in group 1.

As discussed above, the dimension of the QTL effect vectors in model (5) is relatively small. Therefore, it is not difficult to add dominance QTL effects, or even various forms of epistatic interactions, to the model. A dominance design matrix can be created from a multiplication of the columns of $\mathbf{M}_1$ by those of $\mathbf{M}_2$.

Single and multiple QTL models

The models described in this REML section are all single QTL models. To arrive at multi-QTL models, one can follow the standard practice of first performing a genome wide simple interval mapping scan using, for example, a comparison of model (5) with model (2) along the genome, and then retain a set of significant/interesting QTL. The set of QTLs identified in the simple interval mapping scan can be used to perform one or more rounds of composite interval mapping, or can immediately be used to carry out a backward selection procedure. Similar backward selection procedures can be started from the results of composite interval mapping scans. Forward selection procedures, give, of course, under certain conditions, another possible strategy for arriving at multi-QTL models (Bauer et al. 2009; Broman and Speed 2002). Alternatively, a one step whole genome selection approach of the type advocated by Meuwissen et al. (2001) can be considered.

Practicalities

We performed all REML computations using Genstat 12 (www.genstat.co.uk) and an Intel Core 2 Quad Q9550 processor with a clock rate of 2.83 GHz. The time required for a genome wide scan testing for additive QTLs in one or both heterotic groups was 4 h.

Bayesian approach

General description

The general mixed model formulation for the Bayesian MCMC approach is similar to the REML approach. Our Bayesian approach differs in two main respects: we use (1) a multi-QTL model in which the number of QTLs is a random variable, and (2) the assumption that the QTL effects are bi-allelic. One minor difference is that the GCA effects in both heterotic groups are assumed to have a common variance, implying a common GCA design matrix $\mathbf{Z}$ of dimensions $n_{\mathrm{H}} \times n_{\mathrm{P}}$.

The Bayesian MCMC equivalent of the REML based mixed model (2) for GCA effects, written as the data density (likelihood) for HP given non-genetic effects and GCA effects is

$$P(\mathbf{y}; \boldsymbol{\beta}, \mathbf{g}) = N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}, \mathbf{V}_{\mathrm{e}}) \qquad (6)$$

The distribution in (6) is normal with expectation $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}$ and variance $\mathbf{V}_{\mathrm{e}}$. The fixed effects have an uniform prior, the GCA effects have a normal prior, $P(\mathbf{g}) = N(0, \mathbf{G}V_{\mathrm{G}})$, with the coancestry matrix $\mathbf{G}$, as in the REML approach, being block diagonal as a consequence of the disconnectedness of the two heterotic groups.

The Bayesian multi-QTL model, closely following Bink et al. (2008), including residual GCA effects, has data density

$$\begin{aligned} P(\mathbf{y}|\boldsymbol{\beta}, &n_{\mathrm{QTL}}, \mathbf{W}_{1,\ldots,n_{\mathrm{QTL}}}, \mathbf{v}_{1,\ldots,n_{\mathrm{QTL}}}, \mathbf{g}) \\ &= N\left(\mathbf{X}\boldsymbol{\beta} + \sum_{k=1}^{n_{\mathrm{QTL}}} \mathbf{W}_k\mathbf{v}_k + \mathbf{Z}\mathbf{g}, \mathbf{V}_{\mathrm{e}}\right) \end{aligned} \qquad (7)$$

Similar to the REML based mixed model (5), the QTL effects in the Bayesian model (7), $\mathbf{v}_k$ ($k = 1,\ldots, n_{\mathrm{QTL}}$), are defined at the founder level. Unlike the REML mixed model (5), however, in the Bayesian model (7) the number of QTLs, n_{QTL}, is itself a random variable, following ideas in Heath (1997), Sillanpää and Arjas (1998), and Bink et al. (2002). The QTL effects can include both additive and dominance effects. With respect to the QTL effects, the assumption is made that these effects will be bi-allelic. When the QTL effects are chosen to be exclusively additive, the design matrices $\mathbf{W}_k$ will have dimension $n_{\mathrm{H}} \times 1$, whereas for inclusion of dominance effects they will have two columns, one for additive QTL effects and a second one for the dominance effects. The prior for the additive QTL effects is normal and similar to that in Bink et al. (2008). To decide upon the number of QTLs governing a particular trait, model selection procedures implementing the standard theory of Bayes factors (Kass and Raftery 1995) can be used. For illustrations of the use of Bayes factors in QTL mapping, see Bink et al. (2002, 2008).

Structuring VCOV of founder effects

The VCOV of the additive founder QTL effects is structured in the same way as in the REML mixed model by a combination of pedigree and marker information, where this structuring changes for each centi-Morgan. In the Bayesian approach, the VCOV of the additive effects was approximated by various methods of clustering the founders. For example, ancestral classes were constructed allocating individuals beyond a certain threshold value for

coancestry, 0.8 in our case, to the same founder class. The ancestral classes form an addition to the Bayesian framework described originally in Bink et al. (2002). An attractive alternative to the threshold algorithm is the use of latent class models as proposed by ter Braak et al. (2009). We are presently incorporating this latent class algorithm in our Bayesian QTL models.

Practicalities

Bayesian models (6) and (7) are analytically intractable and MCMC simulation is used to sample from the joint distribution of data and model parameters. For the Bayesian analyses, we used the FlexQTL[TM] software (http://www. flexqtl.nl), on a 64-bit Dual-core Opteron system with a clock rate of 2.2 GHz. The simulations were performed with chains of 500,000 iterations, while storing samples every 200th iteration to reduce auto-correlation among samples and to save disk memory. Visual inspection of trace plots of important parameters indicated an absence of burn-in periods and that these numbers of Markov iterations were sufficient for reliable inference. The required computation time for an analysis was up to 5 days, while the analysis of the single chromosome 10 required 14 h of CPU time. These numbers may be reduced by shorter Markov chains and optimization of computer coding or running parallel simulation chains. The stored samples were used for posterior inference on the parameters of interest, i.e., the posterior mean estimates were taken to summarize the accumulated knowledge.

Illustration of modeling approaches on maize ear height data

For the REML approach, Fig. 2 shows in the top panel the test statistic profiles for a genome wide scan for additive QTLs segregating in one or both heterotic groups, comparing model (5) with model (2), using the deviance difference between both models, assuming a mixture of Chi-square distributions (Verbeke and Molenberghs 2000), although this leads to a somewhat conservative test (Greven et al. 2008). To compensate, a rather liberal test level of 0.01 point wise was used, so that relatively many significant QTLs appear. The largest QTL was located on chromosome 10. The middle and bottom panel of Fig. 2 show test profiles for QTLs in the individual heterotic groups, comparing model (4) with model (2). It can be observed that the QTLs at chromosome 1, 3 and 5 principally segregate in the second heterotic group (bottom panel), whereas the QTLs at chromosomes 2, 4, and 6 segregate in the first group (middle panel). Figure 3 once again shows the complementarity of the QTLs segregating
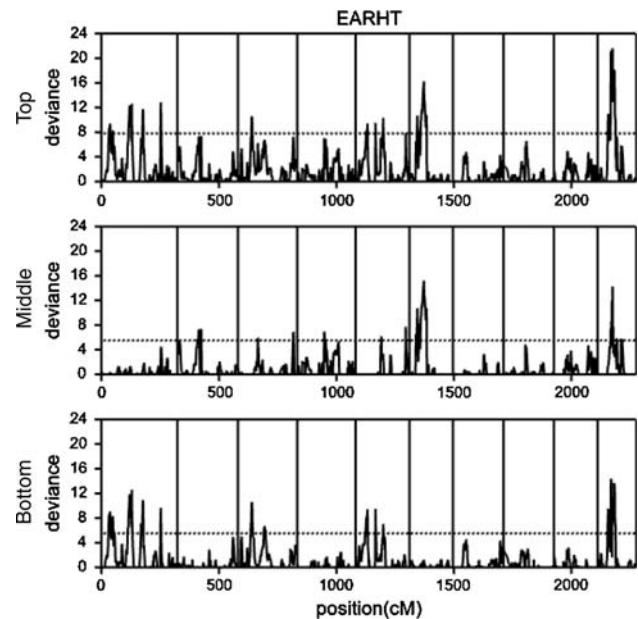

Fig. 2 Mixed model profiles for QTL effects for ear height across genome. *Upper panel* test statistic profile for existence of additive QTL effects in one or both heterotic groups. *Middle panel* test statistic profile for QTL in heterotic group 1. *Lower panel* test statistic profile for QTL in heterotic group 2. The *dashed horizontal lines* mark a point wise significance threshold at test level 0.01

in the different heterotic groups. Here we plot, on the left for ear height, for the full set of genomic evaluation points, the deviance statistics for the tests on a QTL in heterotic group 2 versus those for a QTL in heterotic group 1. It can be seen that most QTLs appear in only one group (points close to either the horizontal or vertical axis) rather than in both heterotic groups (points near the diagonal). To illustrate that this type of complementation was a general phenomenon and not trait specific, Fig. 3 (middle and right) gives the same plot for two other traits (although for those traits, the QTL analyses were performed on hybrids means that were calculated from a hybrid by location phenotypic analysis with hybrids taken random, see section on "Data"). The pattern of additive QTLs segregating in one or the other heterotic groups suggests that a substantial contribution to improved hybrid performance comes from the complementation of positive alleles in the hybrids. The large QTL at chromosome 10 segregates in both groups, while a second, smaller QTL at the same chromosome, segregates only in the first group (Fig. 2).

Figure 4 shows in more detail how both heterotic groups contribute to the detection of the QTL on chromosome 10 in the REML analysis. Figure 4 also emphasizes the strong influence of introducing structure on the founders. The left panel gives the test profiles for an analysis in which the founders are assumed to be unrelated, whereas the more irregular profiles of the right panel result from the analysis in which the VCOV of the founder alleles was structured
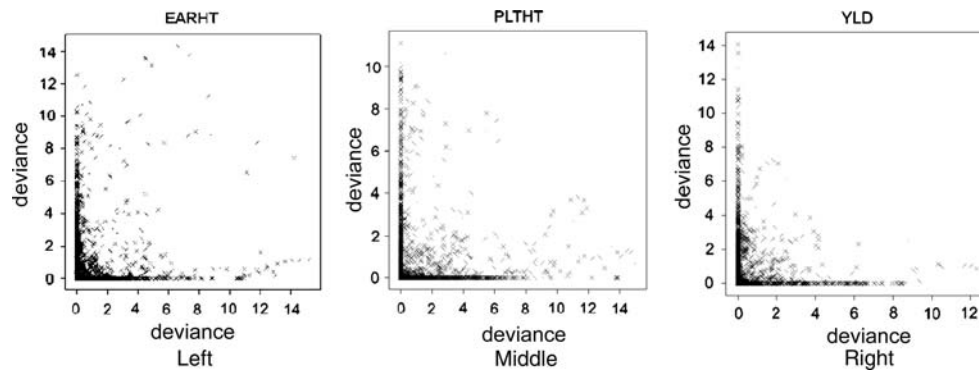
Fig. 3 Complementarity of QTL segregation patterns. Values of deviance statistics representing test on QTL in heterotic group 2 versus test in group 1 for all evaluation points in genome. *Left* for ear height, *middle* for plant height, *right* for yield. Concentration of points near *horizontal* and *vertical axes* together with relative absence of points close to diagonal forms an indication for complementarity of QTL segregation patterns in the two heterotic groups

Fig. 4 Mixed model QTL profile for ear height at chromosome 10. On the *left* without structuring VCOV of founders, i.e., assuming unrelated founders. On the *right* structuring VCOV using pedigree and marker information, i.e., introducing linkage disequilibrium information. The *upper panels* show the test for the detection of a putative QTL segregating in both heterotic groups. The *middle (lower) panels* show a test for a QTL effect in heterotic group 1 (2). The *dashed horizontal lines* mark a point wise significance threshold at test level 0.01
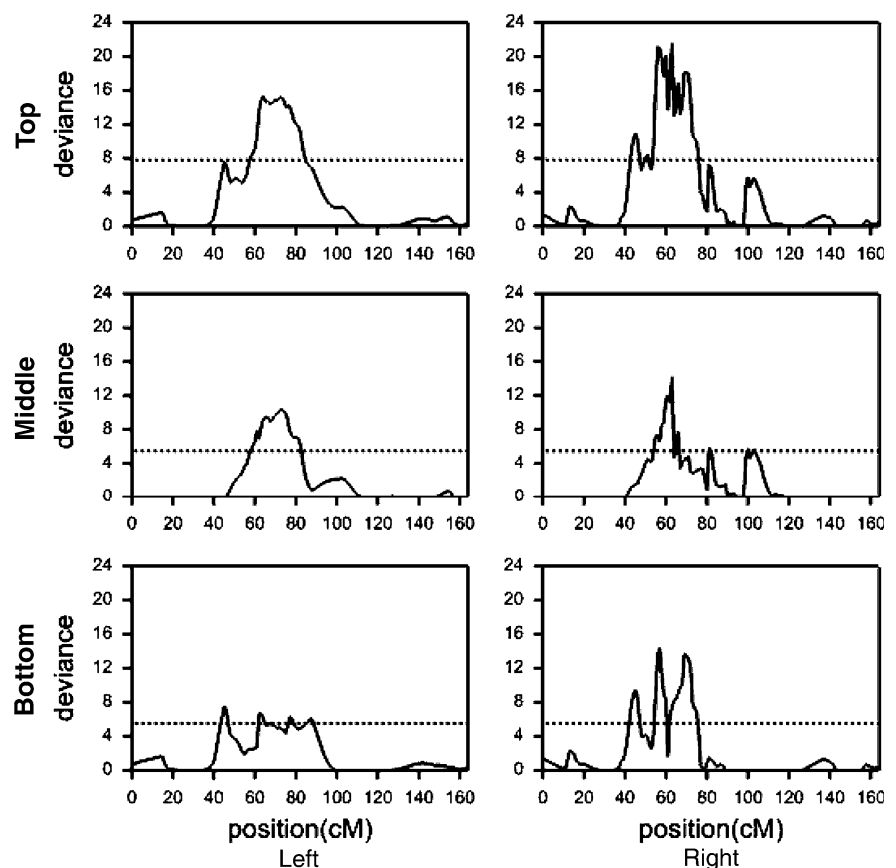
by pedigree and marker information. The irregularity of the test profiles when introducing founder relations follows from the amount of information on QTL effect contrasts changing rapidly across the chromosome. When in a particular chromosome region all founders contain equal or equivalent allelic information, no QTL information can be extracted. If a few centi-Morgans further more variation is present, the test profiles can make sudden jumps. Thus, ignoring the covariance between 'intermediate' founders

leads to overly optimistic test statistics in a number of chromosomal regions.

The distributions of QTL additive allelic effects in both heterotic groups, for the models with and without VCOV-structuring of QTL allele effects are given in Fig. 5 left and middle (REML approach). Again, clear differences between the effects in both models can be observed, where the difference between the heterotic groups is mainly observable in the model with VCOV structuring of allele
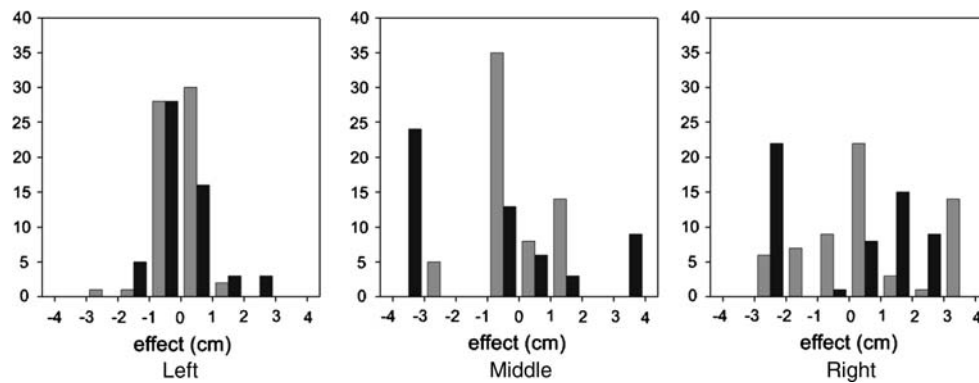
Fig. 5 Histogram for REML (*left* and *middle*) and Bayesian (*right*) mixed model founder effects at the peaks of the REML test profile and the Bayesian intensity profile at chromosome 10. In *gray* (*black*), the founder effects of heterotic group 1 (2). The *left panel* shows the estimated effects assuming that the founders are unrelated. In the *middle panel*, relations between founders are structured. In the *right panel*, Bayesian founder effects are given as obtained from the product of posterior founder QTL genotype probabilities and posterior QTL genotypic values
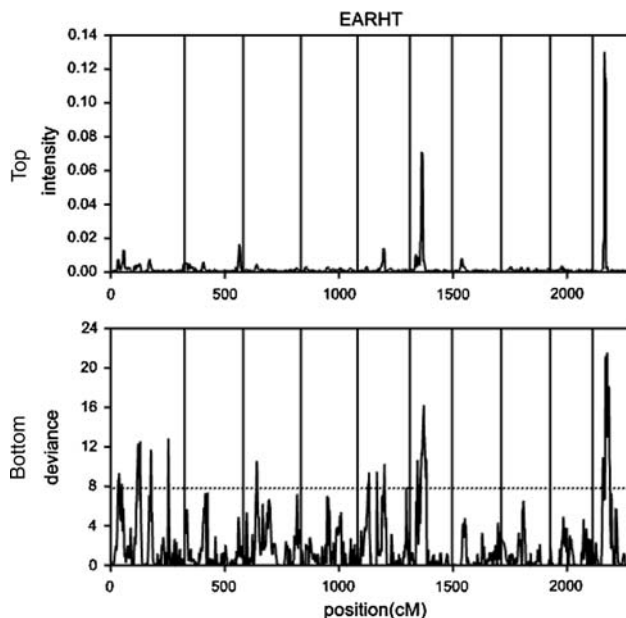

Fig. 6 *Top* Bayesian profile of posterior intensity of QTL position. *Bottom* REML deviance profile for test of QTL in either or both heterotic groups

effects. The distribution of the allelic effects in the second heterotic group (black) is less concentrated around zero than the distribution of the effects for the first heterotic group (gray). This larger variance of the effects in the second heterotic group leads to higher values in the corresponding test profile in Fig. 4: compare the bottom panel with the middle panel.

Figure 6 top provides genome wide information on the results of a Bayesian analysis in which QTLs were allowed to appear on any chromosome, while simultaneously allowing a residual GCA term (see model 7). The posterior mean estimate for the number of QTLs was equal to 4.7, where only chromosomes 6 and 10 had high Bayes Factors, i.e., 5.5

and 7.8, respectively, favoring a model with one QTL over a model with no QTLs for those chromosomes. The intensity profile of the top of Fig. 6 summarizes the posterior information for the local presence of QTLs. The larger a peak is, the stronger the presence of a QTL is indicated.

The QTL on chromosome 10 had most posterior evidence and explained 37 percent of the genetic variation by itself. At the place of maximum intensity on chromosome 10, we calculated the products of posterior founder QTL allele probabilities and QTL allele effects to arrive at quantities that closely resemble the predicted REML founder allele effects. The middle and right panels of Fig. 5 show that the Bayesian and REML models produce comparable estimated QTL allele effects.

For comparison, the bottom part of Fig. 6 repeats the REML profiles for a test in either or both of the heterotic groups. The REML and Bayesian analysis coincide with respect to the major QTLs, to be found on chromosomes 6 and 10. Further coincidences can be observed between the Bayesian intensity profile and the REML deviance profile, where it should be taken into account that the REML analysis is based on a series of single QTL models, whereas the Bayesian analysis is a multi-QTL analysis.

Discussion

Comparison of the REML and Bayesian MCMC mixed model based approaches

In this paper, we described a mixed model framework for the detection of QTLs in elite hybrid backgrounds. We used both REML and Bayesian implementations of this framework that are similar in the linear model structure for the response and the VCOV matrices for QTL and GCA

effects. One difference was that the Bayesian implementation required a bi-allelic QTL representation. Another difference consisted in the fact that the REML implementation was based on the fitting of single QTL models, whereas the Bayesian implementation contained the number of QTLs as a parameter inside the model.

Both approaches are computer-intensive, but do allow multi-point/interval mapping strategies to QTL mapping in a multi-population setting while accounting for pedigree relations between the parents of the offspring populations, something that was not possible so far for realistic hybrid selection data. The REML approach presented in this paper is an elaboration of mixed model QTL work in a REML context as presented earlier in Malosetti et al. (2004, 2007, 2008), Boer et al. (2007), and Paulo et al. (2008). An attractive point of this REML approach is that it allows QTL modeling that combines features of the phenotypic modeling of individual field trials (Boer et al. 2007), modeling of genotype by environment interaction and QTL by environment interactions across trials (Malosetti et al. 2004; Boer et al. 2007), and modeling of multiple traits (Malosetti et al. 2008). These features can be integrated with the approach to QTL mapping in a commercial maize breeding program that was developed here. For the current paper, we did use features on the modeling of multiple populations as hinted at by Paulo et al. (2008), as well as elaborations of the modeling of genetic relations via structuring of the VCOV matrix for the genetic effects in an association mapping context, as was described by Malosetti et al. (2007). Our REML mixed model approach to QTL mapping is thus flexible with respect to the spectrum of genetic phenomena it can handle.

A few choices keep the computing time and requirements acceptable. First, the splitting up of the pedigree information, i.e., the definition of a layer of intermediate founders in combination with a low rank approximation to the VCOV matrix of the random (intermediate) founder effects provides a powerful tool to efficiently manage computer requirements for large and complex data sets. Second, the essentially single QTL models within the REML approach reduce calculation requirements considerably. To arrive at multi-QTL models in the REML setting, one can use, for example, composite interval mapping or backward selection strategies in combination with the simple interval mapping strategy that was described.

The Bayesian mixed model approach has a more elegant solution to the questions on the number of QTLs and their location: it has the number of QTLs as a parameter with a prior to be specified and fits sequences of multi-QTL models with varying numbers of QTLs across the whole of the genome while running through its MCMC chains. The posterior distributions for the number of QTLs and their respective locations are an integration of the realizations of

the QTL models as occurring during the MCMC chain. The price for this elegance is that the computing requirements and times strongly increase, from 4 h for a REML analysis to 5 days for a Bayesian analysis, possibly up to a level that hinders extensions of the present Bayesian mixed models to multi-environment and multi-trait models. In contrast, when compared to the REML mixed model, the simplifying assumption of bi-allelic QTLs in the Bayesian implementation can provide greater versatility for defining epistatic interactions.

The Bayesian method can be extended to include the number of alleles at a QTL as a parameter in the model (Jannink and Wu 2003). However, Jannink and Wu (2003) concluded that for interconnected families there is insufficient information in DNA-marker and phenotypic data to determine with high probability the QTL allelic number unless each family contains many individuals (more than 100). Our own experience (unpublished results) is that simulations of tri-allelic QTLs with substantial allelic differences, the bi-allelic QTL approach fits two closely linked QTLs. In cases where the allelic effects did not differ substantially, the biallelic Bayesian method clustered the alleles into two allelic groups.

The REML and Bayesian approach could be used in a complementary way to combine the advantages of both approaches, i.e., the handling of a wide spectrum of genetic phenomena with the possibility of an elegant approach to the number and the location of QTLs. One could start with a genome wide scan using a REML implementation of our mixed model framework to get a first rough estimate of the number of QTLs, their locations and effects. Next, one could focus on specific genomic regions to run a more refined Bayesian analysis to obtain estimates for especially QTL location and to a lesser extent QTL number.

Approaches to hybrid prediction

When looking at the literature, various classes of approaches to predicting HP can be distinguished. On one extreme, we find the mixed model based approaches of which Parisseaux and Bernardo (2004) is the classical example. The modeling approaches developed in this paper are part of that tradition. The approach can be characterized by the fact that GCA and SCA effects are seen as random parameters whose VCOV matrix should be structured on the basis of pedigree and marker related information. QTLs then absorb part of the original GCA and SCA signal. On the other extreme, we find regression based approaches where HP and parts thereof, like GCA, SCA, mid parent performance, and mid parent heterosis are regressed on marker information or distance related information derived from markers (Vuylsteke et al. 2000; Schrag et al. 2006, 2009). Although QTLs are identified, their relationship with GCA and SCA is not direct, as

in the above-mentioned mixed models. A risky property of the latter approaches is that they typically estimate constitutive effects of HP (GCA, SCA, etc.) in a first linear model and then select markers for association with those effects in one or more further regression procedures, to finally put the pieces together again and to arrive at HP predictions that are composed of the estimates for the corresponding parameters (GCA, SCA, etc.), where these estimates come from different model fits and estimation procedures. For unbalanced data, as hybrid prediction data invariably are, interaction parameters as SCA (and GCA to a lesser extent) are not estimable within fixed linear models, which means that these parameters cannot be calculated as linear combinations of observations without imposing rather artificial constraints (Nelder 1994). When GCA and SCA are taken random, estimability of the parameters is less of a problem, but interpretation remains a problem. Parameters like SCA mainly indicate differences between model fits of increasing complexity, so these parameters can better be understood as residuals indicating lack of fit (from a main effects only model) than as genetic entities contributing to HP. Modeling of HP within one and the same modeling framework avoids all these interpretational problems and has our preference.

The problem of difficult to interpret genetic interaction parameters, like SCA and epistasis, in unbalanced data also affects a recent proposal of Melchinger et al. (2007) to construct NCIII designs to estimate dominance effects that are not contaminated by additive × additive epistatic interactions. Melchinger et al. (2007) see the unbiased estimation of dominance and additive × additive epistatic interactions as a condition for reliable HP prediction. QTL effect estimates would otherwise not carry over to other genetic backgrounds. Still, we feel that a mixed model QTL approach for use in an existing breeding population, as outlined in this paper, should always be a strong competitor to an approach requiring specifically designed breeding populations producing balanced data sets.

Hybrid performance and prediction

The identification of QTLs in biparental mating designs does not necessarily lead to a successful marker assisted selection strategy (MAS), because the sampled allelic diversity is (too) small and QTL × genetic background interactions are not taken into account. Furthermore, the statistical methods for QTL detection typically inadequately model polygenic traits with many small effects (Heffner et al. 2009). Recently, genome wide selection (GS) has been presented as an alternative to MAS in plant breeding (Bernardo and Yu 2007; Nordborg and Weigel 2008; Heffner et al. 2009; Piepho 2009). In MAS, a predictive model for genotypic performance is created on the basis of earlier identified QTLs and estimates for their allelic effects. In MAS, only the markers that are found to be significantly associated with traits are used. In GS, no prior testing of markers on significant association takes place, but all markers enter the predictive model. However, for GS, estimation methods are required that appropriately shrink QTL effects at markers close to minor QTLs, while leaving QTL effects at stronger QTLs relatively untouched. The idea is that in GS both oligogenic QTLs with strong effects and polygenic QTLs with small effects contribute to a genome wide breeding value estimate that can serve for genotypic prediction and selection.

Bernardo and Yu (2007) speculate that breeding value approaches as underlying GS will be useful for evaluating the combining ability of maize inbreds, but will be less useful for finding pairs of inbreds that perform well as single-cross hybrids. With regards to the latter aspect, we are less pessimistic. Our mixed model QTL approach applied within a breeding program should identify relevant QTLs and realistic estimates for their effects. We expect to have included the full allelic diversity. As mapping and prediction are performed with respect to the same breeding population, no QTL × genetic background interactions should show up. We expect to be able to benefit from the complimentary QTLs that we identified in our two heterotic groups by bringing them together in new hybrids using optimal combinations of inbred lines. Furthermore, using the Bayesian implementation of our mixed model framework, we can naturally combine the posterior intensities for QTLs with the QTL effects to arrive at a powerful index for genome wide selection. As a preliminary test of the mixed model methodology in this respect, we produced genome wide predictions from a REML and Bayesian analysis, using the 2004/2005 data. The validation data consisted of a set of 288 new hybrid combinations that were evaluated in the field in 2007/2008. Correlations between observations and predictions were 0.69 for the Bayesian analysis and 0.68 for the REML analysis. The correlation between the REML and Bayesian predictions was 0.89.

We have described two mixed model approaches to QTL mapping in existing hybrid selection programs. Both approaches produced encouraging results. We see this as a first step in the construction of a HP prediction protocol.

References

Bauer AM, Hoti F, von Korff M, Pillen K, Léon J, Sillanpää MJ (2009) Advanced backcross-QTL analysis in spring barley (*H. vulgare* ssp. spontaneum) comparing a REML versus a Bayesian model in multi-environmental field trials. Theor Appl Genet 119:105–123

Bernardo R (1994) Prediction of maize single cross performance using RFLPs and information from related hybrids. Crop Sci 34:20–25

Bernardo R (1996a) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:50–56

Bernardo R (1996b) Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Sci 36:872–876

Bernardo R (1999) Best linear unbiased prediction of single-cross performance. Crop Sci 39:1277–1282

Bernardo R (2002) Breeding for quantitative traits in plants. Stemma Press, Woodbury

Bernardo R, Yu JM (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Bink M, Uimari P, Sillanpää MJ, Janss LLG, Jansen RC (2002) Multiple QTL mapping in related plant populations via a pedigree-analysis approach. Theor Appl Genet 104:751–762

Bink M, Boer MP, ter Braak CJF, Jansen J, Voorrips RE, de Weg WEV (2008) Bayesian analysis of complex traits in pedigreed plant populations. Euphytica 161:85–96

Boer MP, Wright D, Feng LZ, Podlich DW, Luo L, Cooper M, van Eeuwijk FA (2007) A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics 177:1801–1813

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Stat Soc B 64:641–656

Calinski T, Czajka S, Kaczmarek Z, Krajewski P, Pilarczyk W (2005) Analyzing multi-environment variety trials using randomization-derived mixed models. Biometrics 61:448–455

Charcosset A, Bonnisseau B, Touchebeuf O, Burstin J, Dubreuil P, Barriere Y, Gallais A, Denis JB (1998) Prediction of maize hybrid silage performance using marker data: comparison of several models for specific combining ability. Crop Sci 38:38–44

Duvick DN, Smith JSC, Cooper M (2004) Long-term selection in a commercial hybrid breeding program. Plant Breed Rev 24:109–151

Fischer S, Mohring J, Schon CC, Piepho HP, Klein D, Schipprack W, Utz HF, Melchinger AE, Reif JC (2008) Trends in genetic variance components during 30 years of hybrid maize breeding at the University of Hohenheim. Plant Breed 127:446–451

Frascaroli E, Cane MA, Landi P, Pea G, Gianfranceschi L, Villa M, Morgante M, Pe ME (2007) Classical genetic and quantitative trait loci analyses of heterosis in a maize hybrid between two elite inbred lines. Genetics 176:625–644

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

Greven S, Crainiceanu CM, Küchenhoff H, Peters A (2008) Restricted likelihood ratio testing for zero variance components in linear mixed models. J Comput Graph Stat 17:870–891

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

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Jannink JL, Wu XL (2003) Estimating allelic number and identity in state of QTLs in interconnected families. Genet Res 81:133–144

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Lander ES, Green P (1987) Construction of multilocus genetic-linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Malosetti M, Voltas J, Romagosa I, Ullrich SE, van Eeuwijk FA (2004) Mixed models including environmental covariables for studying QTL by environment interaction. Euphytica 137:139–145

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175:879–889

Malosetti M, Ribaut JM, Vargas M, Crossa J, van Eeuwijk FA (2008) A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). Euphytica 161:241–257

Melchinger AE, Utz HF, Piepho HP, Zeng ZB, Schon CC (2007) The role of epistasis in the manifestation of heterosis: a systems-oriented approach. Genetics 177:1815–1825

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using dense marker maps. Genetics 157:1819–1829

Nelder J (1994) The statistics of linear models: back to basics. Stat Comput 4:221–234

Nordborg M, Weigel D (2008) Next-generation genetics in plants. Nature 456:720–723

Parisseaux B, Bernardo R (2004) In silico mapping of quantitative trait loci in maize. Theor Appl Genet 109:508–514

Paulo MJ, Boer M, Huang XQ, Koornneef M, van Eeuwijk F (2008) A mixed model QTL analysis for a complex cross population consisting of a half diallel of two-way hybrids in *Arabidopsis thaliana*: analysis of simulated data. Euphytica 161:107–114

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165–1176

Piepho HP, Mohring J, Melchinger AE, Buchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228

Schrag TA, Melchinger AE, Sorensen AP, Frisch M (2006) Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor Appl Genet 113:1037–1047

Schrag TA, Mohring J, Maurer HP, Dhillon BS, Melchinger AE, Piepho HP, Sorensen AP, Frisch M (2009) Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theor Appl Genet 118:741–751

Sillanpää MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373–1388

Smith A, Cullis B, Gilmour A (2001) The analysis of crop variety evaluation data in Australia. Aust N Z J Stat 43:129–145

ter Braak CJF, Kourmpetis Y, Kiers HAL, Bink MCAM (2009) Approximating a similarity matrix by a latent class model: a reappraisal of additive fuzzy clustering. Comput Stat Data Anal 53:3183–3193

Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York, 568 p

Vuylsteke M, Kuiper M, Stam P (2000) Chromosomal regions involved in hybrid performance and heterosis: their AFLP (R)-based identification and practical use in prediction models. Heredity 85:208–218

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