Theor Appl Genet (2009) 118:1181–1192
DOI 10.1007/s00122-009-0972-y

ORIGINAL PAPER

Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes

S. Maenhout · B. De Baets · G. Haesaert

Received: 20 June 2008 / Accepted: 15 January 2009 / Published online: 18 February 2009

Abstract Molecular markers allow to estimate the pairwise relatedness between the members of a breeding pool when their selection history is no longer available or has become too complex for a classical pedigree analysis. The field of population genetics has several estimation procedures at its disposal, but when the genotyped individuals are highly selected inbred lines, their application is not warranted as the theoretical assumptions on which these estimators were built, usually linkage equilibrium between marker loci or even Hardy–Weinberg equilibrium, are not met. An alternative approach requires the availability of a genotyped reference set of inbred lines, which allows to correct the observed marker similarities for their inherent upward bias when used as a coancestry measure. However, this approach does not guarantee that the resulting coancestry matrix is at least positive semi-definite (psd), a necessary condition for its use as a covariance matrix. In this paper we present the weighted alikeness in state (WAIS) estimator. This marker-based coancestry estimator is compared to several other commonly applied relatedness estimators under realistic hybrid breeding conditions in a number of simulations. We also fit a linear mixed model to phenotypical data from a commercial maize breeding programme and compare the likelihood of the different variance structures. WAIS is shown to be psd which makes it suitable for modelling the covariance between genetic components in linear mixed models involved in breeding value estimation or association studies. Results indicate that it generally produces a low root mean squared error under different breeding circumstances and provides a fit to the data that is comparable to that of several other marker-based alternatives. Recommendations for each of the examined coancestry measures are provided.

Communicated by J. Yu.

S. Maenhout (✉) · G. Haesaert
Department of Biosciences and Landscape Architecture,
University College Ghent, Voskenslaan 270,
9000 Ghent, Belgium
e-mail: Steven.Maenhout@hogent.be

B. De Baets
Department of Applied Mathematics,
Biometrics and Process Control, Ghent University,
Coupure links 653, 9000 Ghent, Belgium

Introduction

The coefficient of coancestry (CoC) between two individuals i and j is defined as the probability that at an allele drawn from both i and j at the same locus is identical by descent (ibd) from a recent common ancestor. This similarity measure is frequently used for modelling the covariance between the genetic background of plants involved in breeding programmes (Panter and Allen 1995a, b; Bernardo 1994, 1995, 1996a, b) or association studies (Jannink et al. 2001; Yu et al. 2006). Piepho et al. (2008) recapitulates the underlying quantitative genetic assumptions of incorporating a coancestry-based covariance matrix in these models, such as gametic-phase equilibrium of the base population and absence of epistasis, selection and drift. These assumptions are rarely or never honoured in a plant breeding context and even explicitly violated when the genotypes under study represent a set of highly selected inbred lines. However, in practice, despite the

numerous deviations from quantitative genetic theory, the CoC often results in an improved model fit compared to alternative methods for structuring the covariance between the genetic components of inbred lines.

If detailed pedigree information is available for all genotypes under study, one can calculate the CoC by means of the tabular method (Emik and Terrill 1949). The founding fathers of this pedigree are assumed to be unrelated and therefore set the reference of a zero CoC. Besides accurate pedigree information, the tabular method assumes an equal contribution of both parents to each offspring. The obtained estimators are therefore only valid when there is no selection or genetic drift in the population at hand. However, if inbred lines are obtained through iterative cycles of inbreeding and selection, by doubling haploids or the single seed descent method, the parental contributions are expected to deviate from their theoretical expectations. Molecular marker information allows to circumvent the assumption-burdened pedigree-based estimator, as the resulting allele identities reflect the unequal parental contributions caused by the breeding process. However, deducing the CoC from allele identities on marker loci results in an upwardly biased estimator, because an alikeness in state (AIS) of alleles in different genotypes does not guarantee a shared inheritance from a common ancestor (Cox et al. 1985; Lynch 1988). Bernardo (1993) shows how this bias can be reduced by taking into account the observed marker similarities between unrelated inbred lines. An alternative approach consists of using marker-based estimation procedures from population genetics, like the kinship coefficient of Loiselle et al. (1995) or the maximum likelihood estimator (MLE) described by Thompson (1975), to name but a few. These estimators have their foundations in population genetics but since none of the initial assumptions are met when the genotypes at hand are selected inbred lines, they reduce to the same level as Bernardo's ad-hoc method.

Irrespective of the estimation procedure used, the resulting pairwise CoC values are often arranged in a symmetric relationship matrix $\boldsymbol{A}$ which is then used to model the covariance structure between specific components involved in a linear mixed model analysis of genetic evaluation data. This matrix should therefore be at least positive semi-definite (psd) which implies that all eigenvalues of the matrix are greater than or equal to zero, or equivalently that

$$\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} \geq 0, \quad \forall \boldsymbol{v} \neq \boldsymbol{0}.$$

If we were to model the variance of a vector of random additive genetic effects $\boldsymbol{u}$ as $2\sigma^2_{\text{gca}}\boldsymbol{A}$ (Lynch and Walsh 1998), $\boldsymbol{A}$ would have to be psd, as the variance of any linear combination of the additive effects Var $(\boldsymbol{v}'\boldsymbol{u}) = 2\sigma^2_{\text{gca}}\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}$ must be positive or 0. A marker-based CoC estimation procedure should therefore guarantee that any derived relationship matrix is psd. Unfortunately, most published estimation procedures can result in a non-psd $\boldsymbol{A}$ matrix, while for those who seem empirically psd, a formal proof of this property has not been established. Trying to fit a non-psd covariance structure in a linear mixed model is however not without consequence. Most linear mixed model packages use the psd property to decompose the variance matrix of the model by means of a Cholesky decomposition. If the variance matrix of the linear mixed model is not psd, the linear mixed model package either quits with an error message referring to a problem in the initial likelihood calculation (SAS PROC MIXED, Wombat) or gives a warning message and continues the analysis (ASReml). In the latter case, convergence problems of the REML algorithm are frequently observed and the resulting BLUPs should be interpreted with caution as the estimation procedure can now force certain BLUPs to expand away from zero instead of shrinking them.

Several estimation procedures can possibly result in estimated CoC values that are greater than 1 or smaller than 0. From a sheer model fit perspective, a negative covariance between certain genetic components might be justifiable, but when a biological interpretation of the estimated variance components or BLUPs according to Stuber and Cockerham (1966) is needed, the CoC should be a probability and thus bounded by zero and one. To accommodate an interpretation of the CoC estimator according to its original definition, Bernardo (1993) proposes to truncate the out of bound values at the boundaries of the parameter space. As a consequence, even if the used CoC estimation procedure is proven to always generate a psd relationship matrix, the possibility of a post-hoc truncation of the out of bound values results in a loss of this mathematical property.

If a non-psd coancestry matrix should arise for whatever reason, it can always be bent towards the closest psd matrix. The term matrix bending was first coined by Hayes and Hill (1981) for describing a procedure which shrinks the range of eigenvalues of a matrix involved in selection index calculations. The authors indicate, rather as a side-effect, that this procedure allows to make a non-psd, genotypic or phenotypic variance matrix psd. More than 20 years later, Sørensen et al. (2002) used this procedure for bending estimated CoC matrices and compared its performance to two other procedures based on spectral decomposition. Unfortunately, all three described procedures allow to obtain CoC values outside the parameter space. Henshall and Meyer (2002) published two programs which focus on bending non-psd covariance matrices which might arise in multi-trait genetic evaluations. The iterative matrix bender described by Jorjani et al. (2003) focuses on the same problem and allows to give different

weights to each entry in the covariance matrix, depending on its reliability. The described algorithm even allows to incorporate the restrictions specific to correlation matrices but these obviously differ from coancestry matrices.

The main objective of our research was to develop a new marker-based CoC estimation procedure for specific use in hybrid breeding programmes. This procedure should therefore allow for a mix of heterozygous and inbred genotypes. All pairwise CoC values should be interpretable as a probability and therefore lie in the unit interval [0,1]. Any resulting relationship matrix should be guaranteed to be psd which avoids the need for any bending procedure. In the next section we derive this new estimation procedure and give a formal proof of its psd property. In the two following sections we compare its behaviour to other CoC estimation procedures by means of simulations and an application to actual maize breeding data. We conclude by presenting the results of these calculations and a general discussion.

Materials and methods

WAIS

A codominant molecular fingerprint of a diploid genotype i can be represented as an integer row vector $\boldsymbol{x}_i$. Each position in this vector represents an allele at a certain locus that is represented in the genotyped breeding pool. The vector position is set to 2 if the matching allele is homozygous for genotype i, 1 in case the allele is present at only one of the two homologous chromosomes and 0 in case of absence. $\boldsymbol{x}_i$ therefore has length $p = \sum_{k=1}^{l} n_k$ where l is the number of genotyped loci and n_k is the number of alleles observed in the collection of genotypes at locus k. Using these vectors we can calculate f_{ij}^{AIS} between two genotypes i and j as

$$f_{ij}^{\mathrm{AIS}} = \frac{1}{4l}\boldsymbol{x}_i\boldsymbol{x}_j'.$$

If we arrange the row vectors $\boldsymbol{x}_i$ of length p for all m genotyped individuals in an $m \times p$ matrix $\boldsymbol{X}$ we can calculate the symmetric AIS matrix as

$$\boldsymbol{A}^{\mathrm{AIS}} = \frac{1}{4l}\boldsymbol{X}\boldsymbol{X}'.$$

$\boldsymbol{A}^{\mathrm{AIS}}$ can be shown to be at least psd (Gower 1971) as

$$\boldsymbol{v}'\boldsymbol{A}^{\mathrm{AIS}}\boldsymbol{v} = \frac{1}{4l}\boldsymbol{v}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{v} = \frac{1}{4l}(\boldsymbol{X}'\boldsymbol{v})'(\boldsymbol{X}'\boldsymbol{v}) = \frac{1}{4l}\sum_{z=1}^{p} u_z^2 \geq 0,$$

for all m-sized vectors $\boldsymbol{v} \neq \boldsymbol{0}$ where $(u_1, u_2, \ldots, u_p)$ is the transpose of the column vector $\boldsymbol{X}'\boldsymbol{v}$.

Despite being psd, AIS is upwardly biased and therefore not the preferred similarity measure for linear mixed modelling of breeding data. Therefore, we want to incorporate a correction factor without loosing the psd property. To calculate this correction factor, we start from a normal hybrid breeding scenario which assumes that the inbred lines, for which we want to estimate the pairwise relationships, all belong to the same heterotic group. We also assume that we have a complementary heterotic group of genotyped inbred lines at our disposal. All inbred lines from the first heterotic group are assumed to be completely unrelated to the lines belonging to the second heterotic group. We are now able to define several probabilities that are needed to introduce the correction factor. Imagine we draw a random allele from individuals i and j, at the same locus and both alleles α_i and α_j turn out to be allele z. We define the conditional probability ω_z for two random individuals as

$$\begin{aligned}\omega_z &= P(\alpha_i \overset{\text{ibd}}{=} \alpha_j \mid \alpha_i = z, \alpha_j = z) \\ &= \frac{P(\alpha_i = z, \alpha_j = z) - P(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\text{ibd}}{=} \alpha_j)}{P(\alpha_i = z, \alpha_j = z)}, \end{aligned} \quad (1)$$

where $P(\alpha_i = z, \alpha_j = z)$ is the probability that the two alleles, drawn from two random individuals i and j of the same heterotic group at the locus to which z belongs, are equal to z and therefore AIS. $P(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\text{ibd}}{=} \alpha_j)$ is the same probability but with the additional constraint that the AIS is not caused by a shared inheritance from a nearby ancestor (i.e. an ancestor that is still unrelated to all lines in the complementary heterotic group). $P(\alpha_i = z, \alpha_j = z)$ can be estimated from the $m(m-1)/2$ possible pairs of genotyped members of the heterotic group as

$$P(\alpha_i = z, \alpha_j = z) = \frac{\sum_{i=1}^{m}\sum_{j>i}^{m} x_{(i,z)}x_{(j,z)}}{2m(m-1)}, \quad (2)$$

where $x_{(i,z)}$ and $x_{(j,z)}$ represent the corresponding entries in matrix $\boldsymbol{X}$ for genotypes i and j and the column corresponding to allele z. If we now assume that individual i belongs to one heterotic group and j to another, we can estimate the probability of obtaining an AIS for allele z that did not originate from a shared inheritance from a nearby ancestor. If we define m_1 and m_2 as the number of genotyped members in the first and second heterotic group, respectively, then we can estimate $P(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\text{ibd}}{=} \alpha_j)$ for both groups as

$$P(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\text{ibd}}{=} \alpha_j) = \frac{\sum_{i=1}^{m_1}\sum_{j=m_1+1}^{m_1+m_2} x_{(i,z)}x_{(j,z)}}{4m_1m_2}, \quad (3)$$

where i and j now index over individuals from the first and second heterotic group, respectively. Due to small sample size effects, it is possible that the estimator for $P(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\text{ibd}}{=} \alpha_j) > P(\alpha_i = z, \alpha_j = z)$ in which case the conditional probability ω_z should be set to 0. For rare alleles $P(\alpha_i = z, \alpha_j = z)$ might be 0 but in those cases the

conditional probability is not needed for the calculation of the coancestry. If we now arrange the conditional probabilities ω_z from Eq. 1 for each allele z on the diagonal of an all zero square matrix $\boldsymbol{W}$ of size p, we can calculate f_{ij}^{WAIS} for two individuals i and j belonging to the same heterotic group as

$$f_{ij}^{\mathrm{WAIS}} = \frac{1}{4l}\boldsymbol{x}_i\boldsymbol{W}\boldsymbol{x}_j', \tag{4}$$

where the index WAIS is shorthand for weighted alikeness in state (WAIS). The procedure thus far has assumed that i and j are different genotypes belonging to the same heterotic group. For the calculation of the symmetric matrix $\boldsymbol{A}^{\mathrm{WAIS}}$ we also need to calculate f_{ii}^{WAIS} for each of the m individuals in the heterotic group. In this case, the conditional probability of Eq. 1 underestimates the actual ibd probability and this is even more the case when genotype i has been inbred for g_i generations as is common in hybrid breeding. If we draw two alleles α_{i1} and α_{i2} at the same locus of inbred line i, the conditional probability of Eq. 1 should be corrected to

$$\begin{aligned} y'_{i,z} &= P\left(\alpha_{i1} \overset{\mathrm{ibd}}{=} \alpha_{i2} \mid \alpha_{i1} = z, \alpha_{i2} = z\right) \\ &= \frac{1}{2} + \frac{1}{2}\left[1 - \left(\frac{1}{2}\right)^{g_i} + \left(\frac{1}{2}\right)^{g_i}\omega_z\right] \\ &= \left[1 - \left(\frac{1}{2}\right)^{(g_i+1)}\right] + \omega_z\left(\frac{1}{2}\right)^{(g_i+1)}, \end{aligned} \tag{5}$$

where ω_z is the entry in the diagonal of $\boldsymbol{W}$ corresponding to allele z. If we define

$$\begin{aligned} y_{i,z} &= y'_{i,z} - \omega_z \\ &= \left[1 - \left(\frac{1}{2}\right)^{(g_i+1)}\right](1 - \omega_z), \end{aligned}$$

we can see that $y_{i,z}$ can never be negative. In case all genotyped individuals i in the heterotic group have the same level of inbreeding g we can drop the index i in this last equation and use the same value y_z for all individuals. For each of the m genotyped individuals in the heterotic group we calculate

$$q_i = \sum_{z=1}^{p} x_{(i,z)}^2 y_{i,z},$$

and arrange these values on the diagonal of an all zero square matrix $\boldsymbol{Q}$ of size m. We can now calculate the WAIS coancestry matrix as

$$\boldsymbol{A}^{\mathrm{WAIS}} = \frac{1}{4l}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}' + \boldsymbol{Q}). \tag{6}$$

The estimated matrix $\boldsymbol{A}^{\mathrm{WAIS}}$ is guaranteed to be psd as the sum of two psd matrices $\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'$ and $\boldsymbol{Q}$ is always psd. It is easy to show that $\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'$ is psd as for any m-sized vector $\boldsymbol{v}$

$$\boldsymbol{v}'\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'\boldsymbol{v} = (\boldsymbol{X}'\boldsymbol{v})'\boldsymbol{W}(\boldsymbol{X}'\boldsymbol{v}) = \sum_{z=1}^{p} u_z^2\omega_z \geq 0,$$

where the last inequality follows from the fact that for all alleles z, ω_z is always greater than or equal to zero. Also matrix $\boldsymbol{Q}$ is psd as it is a diagonal matrix and all entries q_i are greater than or equal to zero.

Simulations

In population genetics, the statistical behaviour of marker-based coancestry estimators is usually determined by repeatedly simulating pairs of genotypes for which the true relatedness belongs to a discrete number of predefined classes (Ritland 1996; Lynch and Ritland 1999; Van de Casteele et al. 2001; Milligan 2002). The mean, standard error, bias and possibly other statistical features are examined with loci number, allele number and allele frequency distributions as variables. All of the previously mentioned studies focus on natural populations and therefore assume linkage equilibrium throughout the genome. However, Stich et al. (2005, 2007) show the presence of significant linkage disequilibrium (LD) between SSR marker loci of elite, European and US maize germplasm. In a later study, Stich et al. (2007) demonstrate, by means of simulation studies, that selection and drift are the major forces generating this LD. As we want to study the behaviour of different relatedness estimators under realistic breeding circumstances, we must incorporate LD between marker loci. Therefore, each simulation tracks selection by means of several breeding cycles from open-pollinated varieties (OPV) towards elite inbred lines.

The simulations used in this study follow the approach of Stich et al. (2007) and therefore indirectly mimic the breeding scheme of the University of Hohenheim. We assume that the inbred lines are genotyped with 101 microsatellite loci, which are evenly distributed over the maize genome according to a proprietary linkage map of the breeding company RAGT R2n. We also generate 250 QTL loci of the selection trait (e.g. yield) which are randomly positioned on the genetic map. The QTL effects and resulting phenotypic values for line per se and testcross performance were calculated according to Stich et al. (2007). An important difference in the presented simulations is the determination of the number of alleles and the allele frequency distribution of all loci on the map. Stich et al. (2007) use SSR allele frequencies obtained from five Central European OPVs and copy these on the simulated QTLs. Other studies assume identical allele frequency distributions across loci (Ritland 1996; Lynch and Ritland 1999) or allow independent draws from a Dirichlet distribution for each locus (Milligan 2002). We follow the latter approach but also allow the number of alleles to differ

between loci. We obtain the number of alleles for each locus as an independent draw from a Poisson distribution plus two, where the parameter λ varies between 0 and 12. This last upper bound was determined by observing little change in the behaviour of the different CoC estimators at higher values of λ.

Each simulation starts by generating an initial base population in Hardy–Weinberg equilibrium. Allele frequencies of each locus are drawn from a Dirichlet distribution with all parameters set to one. From this base generation, we generate the allele frequencies of two subpopulations which have diverged because of artificial selection or geographical differentiation. We assume that on average individuals within each subpopulation share more ancestry compared to individuals belonging to different subpopulations. Wright's Fst value (Wright 1943, 1951) is a measure for this population stratification and we assume this value to be constant over all loci. The allele frequencies in the subpopulations for locus k are drawn from a Dirichlet distribution with parameters $\theta \boldsymbol{p}_k$ where $\boldsymbol{p}_k$ is the vector of allele frequencies at locus k in the base population and $\text{Fst} = 1/1 + \theta$ (Balding 2003). A total of 50 individuals are randomly drawn from each of the two populations as an entry point for the first breeding cycle. Each breeding cycle consists of 6 generations of inbreeding and subsequent phenotypical selection based on line per se or testcross performance as described by Stich et al. (2007). This results in 28 almost homozygous inbred lines within each heterotic group (former subpopulation) which are either intercrossed to produce 50 new genotypes for the next breeding cycle or used to compare the different relatedness estimators. For each allele in the breeding pool we keep track of the original founder allele from which it originated. This allows us to calculate the true pairwise CoC values between pairs of inbred lines as an average of the actual ibd relationships over all genotyped loci. We also calculate the pedigree-based coefficient of coancestry (PED), AIS, WAIS and the estimators described by Bernardo (1993) (BNO), Thompson (1975) (MLE) and Loiselle et al. (1995) (LOI). Some BNO values are negative while LOI admits to values smaller than 0 or greater than 1. These values are consequently truncated to either 0 or 1 to obtain estimators within the biologically meaningful parameter space.

Maize breeding data

Besides simulations we use the described relatedness estimators to determine the CoC of a set of selected inbred lines from the maize breeding programme of the private company RAGT R2n. This data set, described in earlier studies (Maenhout et al. 2007, 2008), contains 40,432 phenotypic measurements on 2,367 hybrids originating from 92 Iodent and 105 Iowa Stiff Stalk Synthetic (ISSS) lines. These hybrids are tested in 1280 multi-environment trials (METs). A Met always takes place during one growing season and is spread out over an average of 3.6 locations in Europe. As the total number of locations is limited, it is quite common that the trials belonging to different METs take place at the same location. Within one location, the environmental conditions such as the level of irrigation and fertilisation, or sowing and harvesting dates, can vary between trials. As a consequence, one MET can contain two trials at the same location, where each trial receives a different treatment. In 67% of all trials over all METs and locations, there is only one replication, while the plots at the remaining trials are laid out in a randomised complete block design. The data is severely unbalanced as, on average, a hybrid is tested in only 2.6 METs. All locations are however connected through the measurements on 3,022 check varieties for which no parental marker or pedigree information is available. We consider all environmental factors as fixed, while the genotypical components and G × E interactions are considered as random effects. The full model for the mean of the vector of phenotypical measurements $\boldsymbol{y}$ can be represented as

$$E[\boldsymbol{y}] = \mu + \boldsymbol{X}_{(g)}\boldsymbol{g} + \boldsymbol{X}_{(l)}\boldsymbol{l} + \boldsymbol{X}_{(g.l)}\boldsymbol{g.l} + \boldsymbol{X}_{(m)}\boldsymbol{m} + \boldsymbol{X}_{(m.l)}\boldsymbol{m.l} + \boldsymbol{X}_{(m.l.t)}\boldsymbol{m.l.t} + \boldsymbol{X}_{(m.l.t.b)}\boldsymbol{m.l.t.b}. \quad (7)$$

Here μ represents the global phenotypical mean, while $\boldsymbol{g}$, $\boldsymbol{l}$, $\boldsymbol{m}$, $\boldsymbol{t}$, $\boldsymbol{b}$ represent vectors containing the effects for growing seasons, locations, METs, trials and blocks respectively. The interaction terms in the model are represented as a listing of the appropriate vector symbols, separated by a dot. The $\boldsymbol{X}_{(*)}$ matrices link the effects in each vector to the phenotypical measurements in vector $\boldsymbol{y}$. The effects in vector $\boldsymbol{m}$ are nested within growing seasons, but the METs have received a unique identifier and therefore the notation $\boldsymbol{g.m}$ has been replaced by $\boldsymbol{m}$. We were not able to fit model terms containing treatment effects as we have no information about the specific treatment (irrigation, fertilisation,…) applied in each trial. The main effects for year and location are removed from the model for the mean as all their levels are confounded with those of higher level interaction terms. The term $\boldsymbol{X}_{(m.l)}$ $\boldsymbol{m.l}$ was also dropped from Eq. 7 as 98% of the location/growing season combinations contain only trials belonging to separate METs. Most of the effects in vector $\boldsymbol{m.l}$ are therefore confounded with the effects in the higher interaction term $\boldsymbol{m.l.t}$. Furthermore, the data contains little or no information for the remaining effects in $\boldsymbol{m.l}$, as different treatments were applied in the few cases where two trials of the same MET were placed within the same location/growing season combination.

The main effects of the random part of the mixed model can be represented as

$$\boldsymbol{Z}_{(c)}\boldsymbol{c} + \boldsymbol{Z}_{(I_1)}\boldsymbol{I}_1 + \boldsymbol{Z}_{(I_2)}\boldsymbol{I}_2 + \boldsymbol{Z}_{(d)}\boldsymbol{d} + \boldsymbol{e}. \quad (8)$$

Vector $\boldsymbol{c}$ contains the total genotypical effects for all checks, and $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ are vectors containing GCA effects for the inbred lines belonging to the ISSS and Iodent heterotic groups, respectively. Vector $\boldsymbol{d}$ contains the SCA effects for each of the 2,367 hybrids and $\boldsymbol{e}$ contains a residual error for each phenotypical measurement in $\boldsymbol{y}$. The rows of the matrix $\boldsymbol{Z}_{(c)}$ corresponding to measurements on genotyped hybrids are set to 0, while all rows of the remaining $\boldsymbol{Z}$-matrices are set to 0 when their corresponding entries in vector $\boldsymbol{y}$ pertain to check varieties.

Random G × E interaction terms are introduced in the full model for the variance by pairwise interacting the first four model terms in Eq. 8 with all the model terms in Eq. 7 except ***m.l*** and ***m.l.t.b***. Due to a software restriction in the maximum number of unknown variance parameters and the prohibitively large computer memory requirements, the possibly improved model fit of factor analytic and reduced rank variance structures for the G × E interaction terms can not be verified. For the same reasons, heterogeneous residual variances can not be fitted. Akaike's information criterion is used to identify the important variance components. At this stage, AIS is used to model the covariance between the general and specific combining abilities of the hybrids according to Stuber and Cockerham (1966). The other random effects are assumed to have a diagonal variance matrix. The described model selection procedure is repeated for the traits grain yield (q/ha at 15% moisture), grain moisture content and days until flowering. The logit transformation is applied to the measurements of grain moisture content as to reduce the skewness in the distribution of the residuals. To avoid convergence problems during REML iterations, these transformed measurements are multiplied with a scaling factor of 100. For both yield and grain moisture content, Akaike's information criterion indicates that the full model for the variance, containing 25 variance parameters, is to be preferred. For days until flowering on the other hand, the three interactions between the SCA effects and ***l***, ***m*** and ***m.l.t*** are dropped from the model for the variance. This reduces the number of variance parameters for this trait to 22.

All variance components are estimated through REML optimisation by means of the Average Information algorithm as implemented in the software tool ASReml (Gilmour et al. 2002). The model fit of the different CoC matrices, obtained by applying each of the examined procedures, is determined by replacing them for the AIS-based matrices in the covariance models of the vectors $\boldsymbol{I}_1$, $\boldsymbol{I}_2$ and $\boldsymbol{d}$ in Eq. 8 and evaluating the resulting restricted log-likelihood at the end of the REML iteration. Both BNO and LOI produce CoC values that are outside the biologically meaningful parameter space ($0 \leq f_{ij} \leq 1$ for all genotypes i and j) and these values are therefore truncated at the boundaries. For both heterotic groups the MLE and the bounded LOI numerator matrices are non-psd and therefore need bending towards the nearest psd matrix.

Bending procedures

The first examined bending procedure applies a spectral decomposition of the non-psd matrix and replaces all negative eigenvalues with a small positive value as described in Sørensen et al. (2002) and Jorjani et al. (2003). This procedure does however not constrain the elements of the bended matrix within the unit interval such that new boundary infringements might arise during bending. To enforce these boundary constraints we implemented an MCMC procedure, inspired by FLBEND (Henshall and Meyer 2002), to transform non-psd coancestry matrices towards the closest psd matrix within the parameter space. The idea behind FLBEND is to generate a symmetric matrix $\boldsymbol{B}$ by means of an iterative Monte Carlo procedure such that the distance between the psd matrix product $\boldsymbol{BB}'$ and the non-psd input matrix $\boldsymbol{A}$ is minimised. Perturbations in $\boldsymbol{B}$ that increase this distance are accepted at reduced probability. Our modified algorithm rejects alterations in $\boldsymbol{B}$ that allow the elements of $\boldsymbol{BB}'$ to stray outside the unit interval. To allow for a faster convergence under this restricted setting, we continuously update the variance of new perturbations by means of a Metropolis-Hastings step. We also allow the matrix $\boldsymbol{B}$ to be non-symmetrical as this results in a better approximation of the input matrix, at the cost of a higher computational demand.

Results

Simulated breeding populations

The first breeding cycle in each simulation produces 28 unrelated inbred lines. The selective pairwise mating of these inbred lines produces 50 hybrids, which represent the starting point for the next selection cycle. At the end of each breeding cycle we can calculate the actual CoC between all pairs of inbred lines by averaging over the true ibd relationships at the SSR marker loci. Figure 1 depicts this average CoC at each breeding cycle.

At the end of each breeding cycle, only the best performing inbred lines are retained, regardless of their pairwise relatedness. This behaviour mimics a real hybrid breeding programme where decisions are based on phenotypical performance data. Unfortunately this implies that it is not possible to control the pairwise CoC between inbred lines at each breeding cycle which would allow to quantify the standard error of the different estimators at

predefined levels of coancestry (parent-offspring, half-sibs,…). Instead, we determine for the 378 pairwise combinations of the 28 selected inbred lines within a heterotic group the actual CoC based on the average ibd relationships at the SSR loci. This allows us to determine the average bias and root mean squared error (RMSE) of each CoC estimator. Figure 2 visualises for each estimator this RMSE at different values of the Fst between the initial OPVs and different values of λ. The presented RMSE values are averaged over 100 independent iterations of the simulation routine.

The AIS, WAIS and PED estimators are guaranteed to produce a psd coancestry matrix, while the other three estimators (BNO, MLE and LOI) are not. For every CoC estimator the proportion of non-psd numerator relationship matrices was determined by means of an eigenvalue analysis. During simulations, BNO never resulted in a non-psd $\boldsymbol{A}^{\mathrm{BNO}}$ matrix despite the fact that several small truncations were necessary to confine the estimator within the biologically meaningful parameter space. MLE and LOI are more likely to produce a non-psd coancestry matrix as can be seen from Fig. 3.

Maize breeding data

AIS, PED, MLE and WAIS all produce CoC estimators within the unit interval. BNO on the other hand can produce negative CoC values and the same holds for LOI which also allows to obtain CoC values greater than one. Figure 4 shows the range of pairwise CoC values between genotypes belonging to the same heterotic group for all examined estimation procedures.

The covariance structure of the GCA and SCA effects of Eq. 8 is modelled by means of the six examined coancestry estimators. We can compare the goodness-of-fit of these coancestry estimators by means of the restricted log-likelihood at the final REML iteration, as the fixed effects structure and the number of estimated variance components for each model are constant. To allow for a fair comparison between estimators we restrict all CoC values to lie within the unit interval. This decision only has a minor effect on the model fit as the difference in restricted log-likelihoods between the bounded and unbounded variants of BNO and LOI is negligible for all three traits under study. MLE is bounded by nature but results in non-psd CoC matrices for both the Iodent and the ISSS heterotic groups and so does the bounded LOI variant. The MCMC bending procedure results in a smaller distance between the original non-psd matrix and the bended output matrix compared to the spectral decomposition approach. For the MCMC bending procedure the maximum element-wise average distance is only 0.00075, while it is 0.0015 for the spectral decomposition approach. This superiority is however barely reflected in an improved model fit as the restricted log-likelihoods of the LOI and MLE CoC matrices bended with the MCMC procedure are usually identical or slightly higher than those bended with the spectral decomposition approach. Table 1 gives an overview of the restricted log-likelihoods for each of the examined CoC estimators and for each of the three traits under study.

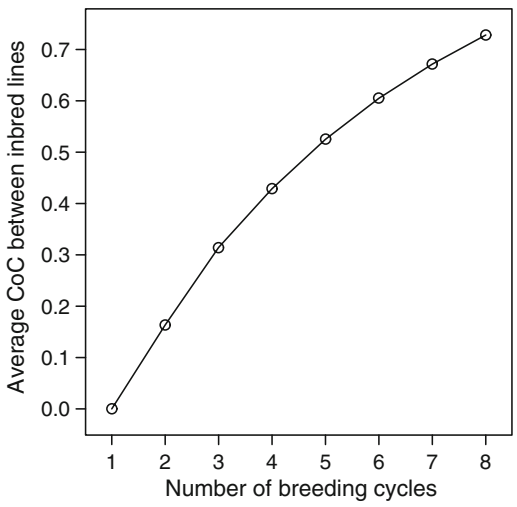

Fig. 1 Average coefficient of coancestry between inbred lines at each breeding cycle

Discussion

The CoC is often used to model the covariance between genetic components of genotypes under selection, despite the inherent conflicts with the underlying quantitative genetic theory. In hybrid breeding programmes and certain association studies the genotypes at hand are highly selected inbred lines with little or no information concerning their selection history. Analysing phenotypical data originating from such inbred lines or their pairwise matings by means of a linear mixed model which uses a CoC estimator to model the covariance between GCA or SCA components, should be considered as an approximation, since the resulting variance components and BLUPs are biased. Nevertheless, good results have been obtained in practice using different CoC estimators based on pedigree or molecular marker information.

In this paper we present a psd, codominant marker-based relatedness estimator called the weighted alikeness in state or WAIS estimator. This estimator is only applicable in the specific case that a reference set of genotyped individuals, unrelated to the genotypes in the sample, is available. As hybrid breeders make extensive use of unrelated heterotic groups, this estimator is particularly suited for this type of selection. It should be clear that WAIS is not claimed to be

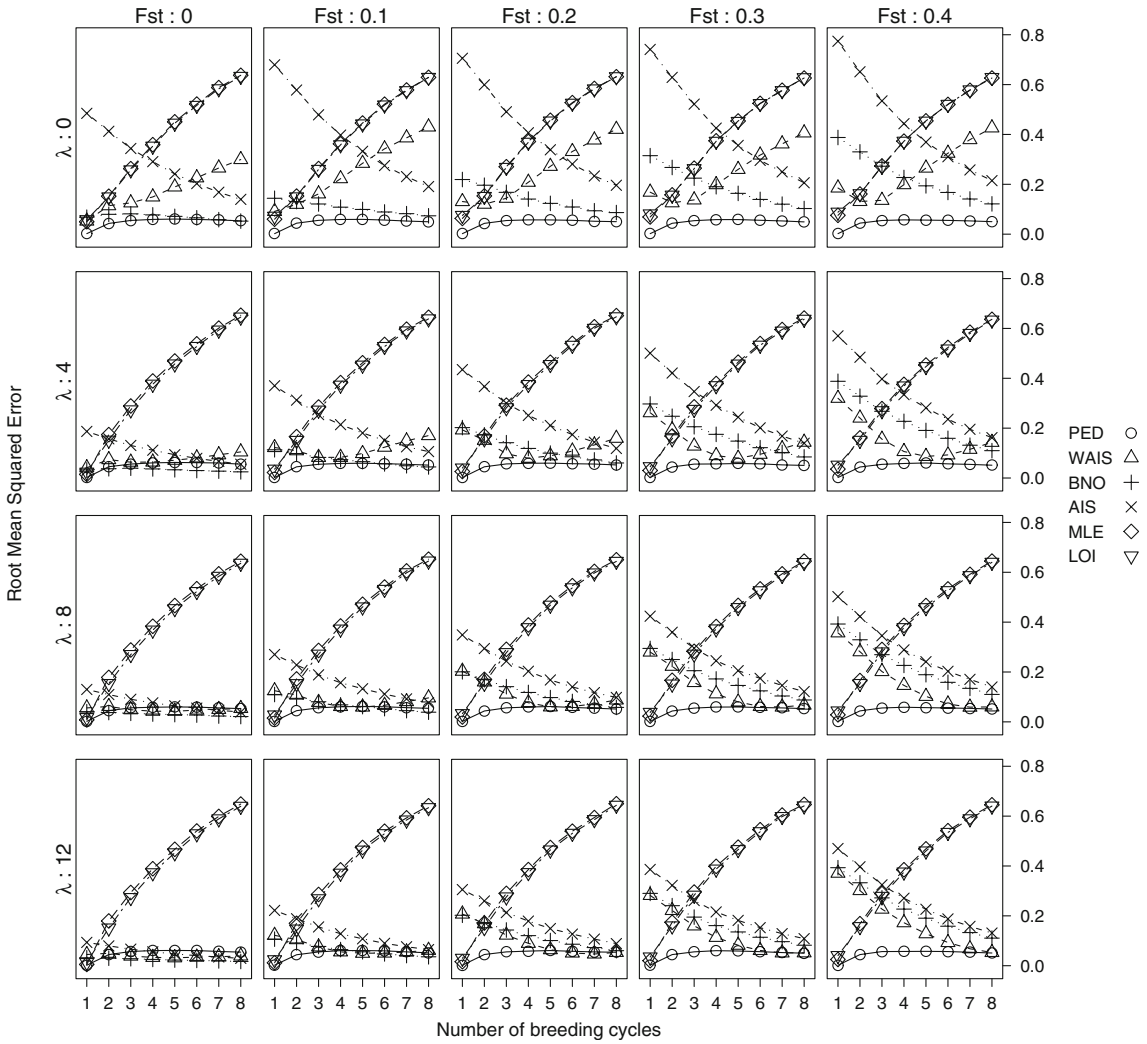

Fig. 2 Root mean squared error of each CoC estimator at the different stages of a hybrid breeding programme. Panels are sorted according to the Fst value of the two initial OPVs from which the selection routine started and λ, the expected value of the Poisson distribution which was used to draw the number of alleles at each locus. RMSE values are averaged over 100 iterations of the simulation routine

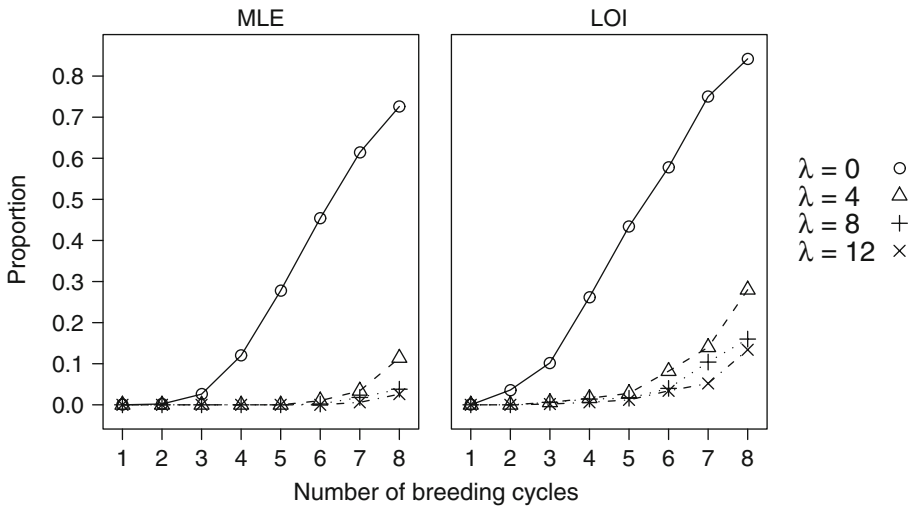

Fig. 3 Proportion of non-psd coancestry matrices for MLE and LOI

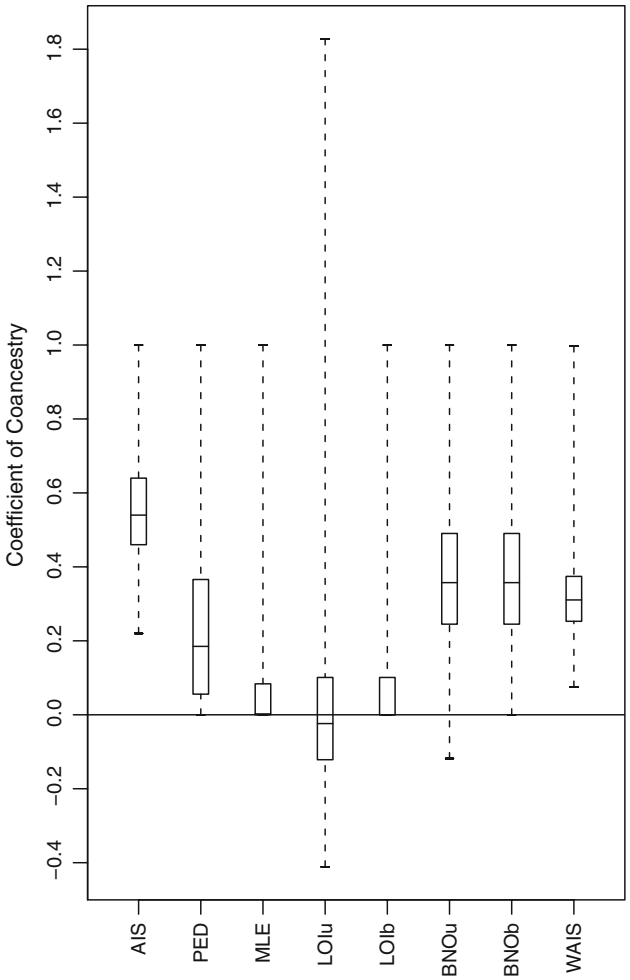

Fig. 4 Pairwise CoC values between members of the same heterotic group, estimated by means of the six examined procedures. For LOI and BNO both the unbounded (suffix u) and bounded (suffix b) ranges are presented

Table 1 Restricted log-likelihoods for each of the six coancestry estimators that were used to model the covariance for GCA and SCA effects in Eq. 8 for the traits yield, grain moisture content and days until flowering

	Yield	Moisture %	Flowering
PED	−222740.1 (4)	−194696.6 (1)	−55339.6 (1)
AIS	−222734.8 (2)	−194710.8 (2)	−55343.8 (2)
BNO	−222734.8 (1)	−194712.9 (3)	−55344.1 (3)
WAIS	−222739.2 (3)	−194715.3 (4)	−55347.7 (4)
MLE	−222743.2 (6)	−194716.2 (5)	−55357.0 (5)
LOI	−222741.0 (5)	−194725.6 (6)	−55361.9 (6)

The number between brackets represents the relative ordering of the estimators when sorted according to decreasing restricted likelihood. BNO and LOI values were bounded within the unit interval. MLE and the bounded LOI matrices were bended towards the closest psd matrix using the MCMC algorithm

assumption free as it for example relies on gametic phase equilibrium. This assumption is surely not met in advanced breeding pools, so we study the behaviour of WAIS and other CoC estimators under a typical hybrid breeding selection scheme by means of simulations and actual breeding data.

Marker-based estimators

Bernardo (1993) uses the observed marker similarities between unrelated lines to correct the AIS-based estimator for lines belonging to the same heterotic group. Besides the often violated assumption of gametic-phase equilibrium between loci, there is also the problem of obtaining negative values for BNO when the correction factor exceeds AIS.

Thompson (1975) demonstrates how the pairwise relationship between non-inbred individuals can be estimated by means of a likelihood function that incorporates the three possible identity by descent probabilities (Jacquard 1974). Milligan (2002) compares the behaviour of this estimator to 5 prominent, non-likelihood estimators (Queller and Goodnight 1989; Li et al. 1993; Ritland 1996; Lynch and Ritland 1999; Wang 2002). He concludes that under all simulated scenarios, MLE exhibits a lower variation compared to the other estimators. However, this reduction in standard error comes at a price, as the likelihood estimator shows considerably more bias, especially at the boundary of the parameter space. A second advantage lies in the fact that the likelihood maximisation procedure is constrained to produce biologically meaningful results $(0 \leq \text{MLE} \leq 1)$, but this property could in fact be enforced on the other estimators as well, again at the cost of increasing the bias. Nevertheless, we consider the MLE to be the most appropriate candidate for use in breeding pools as it explicitly handles inbred individuals. Other implicit assumptions like linkage equilibrium between marker loci and exact knowledge of population allele frequencies are most likely to be violated when the fingerprinted genotypes are all inbred lines but this is the case for all other estimators as well. Anderson and Weir (2007) extended the maximum likelihood approach for the case where the examined genotypes belong to subpopulations of a population with known allele frequencies. However, we did not adopt this approach as its resulting coancestry measures refer to the ancestral population, while all other examined estimators refer to the subpopulation itself.

The problem of finding the most likely ibd relationship between two genotypes can be formulated as the maximisation of a function over a vector $\boldsymbol{\Delta}$ containing nine single-locus, identity by descent modes (Jacquard 1974; Thompson 1975). As the simulations in Milligan (2002) assume large, non-inbred populations, the parameter space can be reduced

to having 2 dimensions. Hepler (2005) explores the possibility of inbred individuals which expands the parameter space to eight dimensions. Both Milligan (2002) and Hepler (2005) use the downhill simplex method (Nelder and Mead 1965), a heuristic optimisation technique, because an algebraic solution of the maximisation problem is not feasible. The original version of this heuristic neither allows the incorporation of the boundary constraints $(0 \leq \Delta_i \leq 1)$ nor the linear constraint $\left(\sum_{i=1}^{9} \Delta_i = 1\right)$ Hepler (2005) introduces these constraints by rejecting solutions outside the parameter space during the optimisation process. This results in numerous lost iterations, especially when certain values of $\boldsymbol{\Delta}$ are near the boundary of the parameter space. To allow for simulations to be performed in an acceptable time frame, we use a quasi-Newton nonlinear interior-point method (Meza et al. 2007) for the maximisation of $L(\boldsymbol{\Delta})$. This approach reduces the needed processor time per genotype pair drastically, while the resulting estimators of $\boldsymbol{\Delta}$ are always nearly identical compared to those of the constrained simplex algorithm. The resulting matrix $\boldsymbol{A}^{\mathrm{MLE}}$, containing all pairwise estimates of MLE, is not guaranteed to be psd which limits its use in a mixed model setting. If the $\boldsymbol{A}^{\mathrm{MLE}}$ happens to be non-psd, the nearest psd matrix should be used instead.

Loiselle et al. (1995) describe a marker-based coancestry estimator which quantifies the correlation in allele frequencies between two individuals belonging to a population in Hardy-Weinberg equilibrium. Despite the obvious violations of underlying theoretical assumptions, this marker-based estimator is sometimes used to model the covariance between genotypes originating from breeding programmes (Yu et al. 2006; Zhang et al. 2007; Casa et al. 2008). LOI is not guaranteed to lie within the parameter space so truncations are often necessary at the boundaries. The resulting coancestry matrix $\boldsymbol{A}^{\mathrm{LOI}}$ is not guaranteed to be psd.

Simulations

The simulated selection scheme follows the maize breeding programme of the University of Hohenheim (Stich et al. 2007). Each breeding cycle consists of 6 generations of inbreeding and selection after which the best performing inbred lines are mated to provide the initial population for the next breeding cycle. The actual CoC, obtained as an average over SSR loci, gradually increases as the number of subsequent breeding cycles rises. The trend observed in Fig. 1 is independent from the initial parameter selection (Fst, λ), which indicates that the number of breeding cycles can be used as an indirect measure for the average relatedness within the heterotic groups.

From Fig. 2 we can see that the pedigree-based estimator outperforms all marker-based estimators by producing the lowest RMSE under all parameter settings. The bias introduced by unequal parental contributions as a consequence of the selection process seems to be negligible compared to the bias of the marker-based estimators. The advantage of the pedigree-based estimator might not be so apparent under practical breeding circumstances, as detailed and accurate pedigree records, tracing back to the initial OPVs, are usually not available. Looking at the marker-based estimators we see that the behaviour of MLE and that of LOI are nearly identical under all simulated scenarios. The performance of these estimators deteriorates as the number of breeding cycles increases which is probably caused by the increasing deviations from population genetics assumptions on which they rely. AIS shows a rather reversed picture as it tends to become more accurate as the number of breeding cycles increases. The overestimation of AIS at low levels of selection is more pronounced when the expected number of distinct alleles at each locus (λ) is small or the differentiation between the populations from which the heterotic groups are developed (Fst) is large. The influence of the value of the Fst is rather surprising as AIS makes no use of a reference population. A possible explanation might lie in the constraints that are imposed on the allele frequencies as a consequence of fixing the Fst value. This might have the same effect as lowering the effective number of distinct alleles at each locus.

The RMSE of WAIS and BNO is usually at a considerably lower level compared to the other marker-based estimators. When $\lambda = 0$, which is equivalent to fixing the number of distinct alleles of each SSR or QTL locus at 2, WAIS has a higher RMSE compared to BNO. This rather unrealistic scenario allows AIS to outperform WAIS when the number of breeding cycles is high. As soon as λ increases to a more realistic setting, WAIS outperforms AIS and can compete with BNO. Ho et al. (2005) estimate the Fst between Corn Belt dent populations to be 0.142 which is somewhat similar to the 0.15 found earlier by Labate et al. (2003). At this level of differentiation WAIS and BNO perform at a comparable level, although WAIS performs slightly better when allelic diversity is high. WAIS also outperforms BNO when the Fst value increases, except when λ is small and the number of breeding cycles is high.

WAIS is specifically designed to guarantee a psd coancestry matrix. This property is necessary when this matrix is used to model the covariance between genetic components in a linear mixed model. BNO, despite not being a psd estimator, always produced a psd coancestry matrix for all simulated populations and the real hybrid maize data set. Several truncations towards 0 were necessary but these were small in absolute value. These arguments allow to conclude that BNO is a stable estimator

which produces natural coancestry measures under variable circumstances. This cannot be said for MLE and LOI which both produced non-psd coancestry matrices for a rather large proportion of the simulated heterotic groups. This proportion is highly dependent on the number of distinct alleles at each locus where a value of λ of 0 and 8 consecutive breeding cycles results in a very high probability of obtaining a non-psd matrix. LOI performs slightly worse than MLE, but both estimators generally exhibit the same increase in proportion of non-psd matrices when the allelic diversity decreases.

Maize breeding data

Figure 4 shows that BNO and LOI both produce negative CoC estimates and that LOI also allows the estimators to become greater than one. The infractions of BNO on the lower bound are rather limited in frequency as well as in size as only 37 of all 9,843 estimates are smaller than zero with a mean negative deviation of 0.04. Truncation of BNO at the lower bound therefore has little impact on the model fit. LOI on the other hand, ranges from -0.41 to 1.83 and only 43% of the estimated CoC values fall within the biologically meaningful parameter space. Truncation of LOI at the boundaries therefore cripples the distribution of CoC values as more than half of the estimates are set to 0 or 1. The bounded LOI distribution looks very similar to that of MLE, the other estimator from population genetics. This is to be expected as MLE forces all estimates to lie within the unit interval by means of the constrained optimisation algorithm. The other estimators produce more natural looking distributions where AIS is generally at a higher level than PED and both BNO and WAIS take more intermediate positions. The unbounded BNO and LOI result in psd CoC matrices for both the ISSS and Iodent heterotic groups while MLE produces non-psd matrices. After bounding of BNO and LOI only LOI results in non-psd CoC matrices for both heterotic groups such that bending needs to be applied.

For the non-psd matrices produced by MLE and the bounded variant of LOI, two matrix bending procedures were examined. The spectral decomposition approach is computationally quite fast but does not allow to constrain the elements within the unit interval. The application of this bending procedure to the bounded LOI estimator results for example in 2,770 new boundary infringements, though it should be noted that these are rather small in absolute value. The MCMC procedure is computationally quite demanding but allows to constrain all CoC values within the aforementioned range and produces a psd matrix that is closer to the original input matrix than the matrix resulting from the spectral decomposition approach. This difference between both bending procedures is however negligible when comparing restricted log-likelihoods of linear mixed models in which the bended CoC matrices are used to model covariances between random GCA and SCA effects.

In Table 1 we can see that PED results in the highest restricted log-likelihood at the end of the REML iterations for the traits grain moisture content and days until flowering, while BNO, AIS and WAIS outperform the pedigree estimator for the trait yield. If we focus on the marker-based estimators, we see that the uncorrected AIS results in the highest restricted log-likelihood for the traits grain moisture content and days until flowering while for yield the difference with BNO is negligible. Although surprising at first, this behaviour of AIS is consistent with the simulations as it was shown that the RMSE of AIS decreases to that of BNO and WAIS when the number of consecutive breeding cycles is high. Taking into account that AIS always results in a psd coancestry matrix, this estimator deserves a revaluation when applied to highly selected breeding material. When summing over rank scores, BNO takes third position while WAIS takes fourth. Constraining the resulting coancestry matrix to be psd comes at the price of a slightly reduced model fit. MLE and LOI, both originating from population genetics, give the lowest log-likelihoods for all three traits under study.

Results from this study indicate that the pedigree-based CoC estimator is superior to the available marker-based alternatives when accurate and complete pedigree information is available for a set of highly selected inbred lines. Comparisons between marker-based CoC estimation procedures, for the specific case that the inbred lines are subdivided in unrelated heterotic groups, indicate that procedures from population genetics like MLE or LOI should generally be avoided as a considerable deviation from the actual ibd relationship can be observed when the inbred lines have a long breeding history. Results also indicate that in this specific case, the observed allele identities need little correction and therefore AIS results in a good approximation of the true CoC. However, if the breeding history is not that long or unknown, BNO and WAIS should be used. BNO generally results in a slightly better model fit, but when the psd property of the resulting CoC matrix needs to be guaranteed, for example when used in a linear mixed model for breeding value estimation or an association study, the new relatedness estimator WAIS should be preferred.

Acknowledgments The authors would like to thank the people from RAGT R2n for their unreserved and open-minded scientific contribution to this research. We would also like to thank the two anonymous reviewers for their helpful comments and suggestions during the review process of this paper.

References

Anderson AD, Weir BS (2007) A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics 176:421–440

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol 63:221–230

Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet 85:1055–1062

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

Bernardo R (1995) Genetic models for predicting maize single-cross performance in unbalanced yield trial data. Crop Sci 35:141–147

Bernardo R (1996a) Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Sci 36:50–56

Bernardo R (1996b) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:872–876

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30–40

Cox TS, Kiang YT, Gorman MB, Rodgers DM (1985) Relationship between coefficient of parentage and genetic similarity indices in soybean. Crop Sci 25:529–532

Emik L, Terrill C (1949) Systematic procedures for calculating inbreeding coefficients. J Hered 40:51–55

Gilmour A, Gogel B, Cullis B, Welham S, Thompson R (2002) ASREML User Guide Release 1.0. VSN International Ltd

Gower JC (1971) A general coefficient of similariy and some of its properties. Biometrics 27:857–874

Hayes JF, Hill WG (1981) Modification of estimates of parameters in the construction of genetic selection indices ('Bending'). Biometrics 37:483–493

Henshall JM, Meyer K (2002) PDMATRIX–Programs to make matrices positive definite. In: Proceedings of the 7th world congress on genetics applied to livestock production, vol 33. Communication No. 28–12, pp 753–754

Hepler AB (2005) Improving forensic identification using Bayesian networks and relatedness estimation. Dissertation, North Carolina State University, Raleigh, NC

Ho JC, Kresovich S, Lamkey KR (2005) Extent and distribution of genetic variation in U.S. Maize: historically important lines and their open-pollinated dent and flint progenitors. Crop Sci 45:1891–1900

Jacquard A (1974) The genetic structure of populations. Springer, New York

Jannink JL, Bink MCAM, Jansen RC (2001) Using complex plant pedigrees to map valuable genes. Trend Plant Sci 6:337–342

Jorjani H, Klei L, Emanuelson U (2003) A simple method for weighted bending of genetic (co)variance matrices. J Dairy Sci 86:677–679

Labate JA, Lamkey KR, Mitchell SE, Kresovich S, Sullivan H, Smith SC (2003) Molecular and historical aspects of corn belt dent diversity. Crop Sci 43:80–91

Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. Hum Hered 43:45–52

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). Am J Bot 82:1420–1425

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. Mol Biol Evol 5:584–599

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Inc, Sunderland

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

Maenhout S, De Baets B, Haesaert G, Van Bockstaele E (2007) Support vector machine regression for the prediction of maize hybrid performance. Theor Appl Genet 115:1003–1013

Maenhout S, De Baets B, Haesaert G, Van Bockstaele E (2008) Marker-based screening of maize inbred lines using support vector machine regression. Euphytica 161:123–131

Meza JC, Oliva RA, Hough PD, Williams PJ (2007) OPT++: an object-oriented toolkit for nonlinear optimization. ACM (TOMS) 33(2), article 12

Milligan BG (2002) Maximum-likelihood estimation of relatedness. Genetics 163:1153–1167

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7(4):308–313

Panter DM, Allen FL (1995a) Using best linear unbiased predictions to enhance breeding for yield in soybean. 1. Choosing parents. Crop Sci 35:397–405

Panter DM, Allen FL (1995b) Using best linear unbiased predictions to enhance breeding for yield in soybean. 2. Selection of superior crosses from a limited number of yield trials. Crop Sci 35:405–410

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

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. Evolution 43:258–275

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res 67:175–185

Sørensen AC, Pong-Wong R, Windig JJ, Woolliams JA (2002) Precision of methods for calculating identity-by-descent matrices using multiple markers. Genet Sel Evol 34:557–579

Stich B, Melchinger AE, Frisch M, Maurer HP, Heckenberger M, Reif JC (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111:723–730

Stich B, Melchinger AE, Piepho HP, Hamrit S, Schipprack W, Maurer HP, Reif JC (2007) Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. Theor Appl Genet 115:529–536

Stuber C, Cockerham C (1966) Gene effects and variances in hybrid populations. Genetics 54:1279–1286

Thompson EA (1975) The estimation of pairwise relationships. Ann Hum Genet 39:173–188

Van de Casteele T, Galbusera P, Matthysen E (2001) A comparison of microsatellite-based pairwise relatedness estimators. Mol Ecol 10:1539–1549

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

Wright S (1943) Isolation by distance. Genetics 28:114–138

Wright S (1951) The genetical structure of populations. Ann Eugen 15:323–354

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhang Z, Todhunter RJ, Buckler ES, Van Vleck LD (2007) Technical note: use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. J Anim Sci 85:881–885