2008

Comparison of Marker-based Pairwise Relatedness Estimators on a Pedigreed Plant Population

Amy D. Anderson
Western Washington University, amy.anderson@wwu.edu

Marco C.A.M. Bink
Wageningen UR

W. Eric van de Weg
Wageningen UR

Elizabeth A. Thompson
University of Washington

Recommended Citation

Anderson, Amy D.; Bink, Marco C.A.M.; van de Weg, W. Eric; and Thompson, Elizabeth A., "Comparison of Marker-based Pairwise Relatedness Estimators on a Pedigreed Plant Population" (2008). *Mathematics*. 37.
https://cedar.wwu.edu/math_facpubs/37

Theor Appl Genet (2008) 117:843–855
DOI 10.1007/s00122-008-0824-1

ORIGINAL PAPER

Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population

Marco C. A. M. Bink · Amy D. Anderson ·
W. Eric van de Weg · Elizabeth A. Thompson

Received: 20 September 2007 / Accepted: 8 June 2008 / Published online: 1 July 2008

Abstract Several estimators have been proposed that use molecular marker data to infer the degree of relatedness for pairs of individuals. The objective of this study was to evaluate the performance of seven estimators when applied to marker data of a set of 33 key individuals from a large complex apple pedigree. The evaluation considered different scenarios of allele frequencies and different numbers of marker loci. The method of moments estimators were Similarity, Queller-Goodknight, Lynch-Ritland and Wang. The maximum likelihood estimators were Thompson, Anderson-Weir and Jacquard. The pedigree-based coancestry coefficients were taken as the point of reference in calculating correlations and root mean square error (RMSE). The marker data comprised 86 multi-allelic SSR markers on 17 linkage groups, covering 11 Morgans. Additionally, we simulated 10 datasets conditional on the real pedigree to support the results on the real dataset. None of the estimators outperformed the others. Knowledge of allele frequencies appeared to be the most influential, i.e., the highest correlations and lowest RMSE were found when frequencies from the founder population were available. When equal allele frequencies were used, all estimators resulted in very similar, but on average lower, correlations. The use of allele frequencies estimated from the set of 33 individuals gave, on average, the poorest results. The maximum likelihood estimators and the Lynch-Ritland estimator were the most sensitive to allele frequencies. The results from the simulation study fully supported the trends in results of the real dataset. This study indicated that high correlations (up to 0.90) and small RMSE (below 0.03), may be obtained when population allelic frequencies are available. In this scenario, the performances of the various estimators were similar, but seemed to favor the maximum likelihood estimators. In the absence of reliable allele frequencies the method of moments estimators were shown to be more robust. The number of marker loci influenced the average performance of the estimators; however, the ranking was not affected. Correlations up to 0.80 were obtained when two markers per chromosome and appropriate allele frequencies were available. Adding more markers to the current dataset may lead to marginal improvements.

Communicated by M. Sillanpää.

M. C. A. M. Bink (✉)
Biometris, Wageningen University and Research Centre,
Wageningen, PO Box 100, 6700, The Netherlands
e-mail: marco.bink@wur.nl

A. D. Anderson
Department of Biostatistics, University of Washington,
Seattle, WA 98195-7232, USA

W. E. van de Weg
Department of Plant Breeding, Wageningen University
and Research Centre, Wageningen, The Netherlands

E. A. Thompson
Department of Statistics, University of Washington,
Seattle, WA 98195-4322, USA

Present Address:
A. D. Anderson
Department of Mathematics, Western Washington University,
Bellingham, WA 98225-9063, USA

Introduction

Molecular marker data have been used to infer the degree of relatedness between two individuals in a variety of contexts (Weir et al. 2006). The primary interest is often in the relatedness estimates themselves. Alternatively, it may lie in using the estimates in a subsequent analysis in quantitative genetics (Lynch and Walsh 1998). Marker-based estimates have been combined with phenotypic observations to obtain heritability estimates for natural populations (Ritland 1996b; Thomas and Hill 2000). More recently, Yu et al. (2006) incorporated a marker-based relationship matrix into a mixed model approach for association mapping. The marker-based relatedness among individuals effectively accounted for hidden levels of relatedness that otherwise resulted in spurious associations between genetic markers and quantitative traits (Yu et al. 2006). In addition, including known relationships between parents of multiple mapping populations generally increases the power to detect and map QTL in linkage analysis (Bink et al. 2002). Traditionally, relationships are calculated from a known pedigree (Crow and Kimura 1970; Lynch and Walsh 1998). When pedigree records are missing or unreliable, the relationships may be obtained by marker-based estimators.

In the estimation of pairwise relatedness in plant breeding populations, the following two issues seem relevant. First, effective population sizes are often small and therefore, the presence of recent inbreeding needs to be taken into account. Second, accurate estimates for population allelic frequencies of markers are not available for many plant species.

The quantification of relatedness can also be done through probabilities of identity by descent (IBD). For non-inbred individuals, these are the probability that two individuals share two alleles IBD and the probability that they share one allele IBD. A commonly used measure of relatedness is the coancestry coefficient (Wright 1922). Many estimators for pairwise relatedness have been developed for the case where the loci are unlinked and the individuals are not inbred, for example the maximum likelihood (ML) estimator of Thompson (1975), and a variety of method of moments (MOM) estimators (Li et al. 1993; Lynch and Ritland 1999; Queller and Goodnight 1989; Ritland 1996a; Wang 2002). Milligan (2003) compared the statistical properties of these estimators and concluded that the ML estimator exhibits a smaller standard error but is more biased than the others are. However, these differences became smaller when marker information increased (Milligan 2003). All these estimators rely heavily on accurate knowledge of the marker allelic frequencies, although one of the MOM estimators was claimed to be more robust to misspecification (Wang 2002). Recently, Thompson's ML estimator has been extended by incorporating a population structure parameter to examine individuals belonging to a subpopulation that have diverged from the original population to which known allele frequencies are applied (Anderson and Weir 2007). The potential of this latter approach for situations where the allele frequencies of the original population are unknown is yet to be explored. The ML estimator of Thompson (1975) can be generalized to a less parsimonious estimator that handles inbred individuals (Jacquard 1972).

In this study, we evaluate the performance of multiple pairwise relatedness estimators on a set of 33 key individuals from a dataset consisting of genotypes for several related apple cultivars as well as simulated datasets reflecting the properties of this dataset. The dataset is available from the EU project HiDRAS. The pedigree of this population is known and consistent with available marker data. Therefore, the pedigree-based coancestry coefficients are taken as the point of reference.

Materials and methods

Prologue

Two alleles are said to be identical by state (IBS) if they share the same allelic type or value. In the case of a multi-allelic locus, nine IBS modes are possible for a pair of genotypes, see Tables 1 in either Milligan (2003) or Anderson and Weir (2007). Two alleles are said to be identical by descent (IBD) if they are identical copies of an allele segregating from a common ancestor within the defined pedigree. Compared to unrelated individuals, individuals that are related are more likely to have similar genotypes as they have an increased probability of sharing alleles IBD from a recent common ancestor.

One commonly used parameter in relationship estimation is Wright's coancestry coefficient θ_{XY}, which represents the probability that a randomly chosen allele from individual X is IBD to an allele randomly chosen from individual Y. For non-inbred individuals, Wright's relatedness coefficient equals twice the coancestry coefficient, i.e., $r_{XY} = 2\theta_{XY}$. The inbreeding coefficient is the probability that the two alleles at any locus in an individual are IBD (Malécot 1948). Jacquard (1972) described a set of nine identity states that gives a full description of the IBD relationships between the set of four alleles from a pair of possibly inbred individuals. In the absence of inbreeding, the number of possible IBD modes reduces to three (Thompson 1975). We will not consider more elaborate parameterizations such as those presented by Cockerham (1954), Kempthorne(1954), and Harris (1964).

Estimators

We used the same set of estimators for pairwise coancestry coefficients that were used by Anderson and Weir (2007), plus the Jacquard estimator. The estimators may be grouped into a set of MOM estimators (based on matching the *sample* moments with the corresponding *distribution* moments) and a set of ML estimators (based on determining the parameters that maximize the probability of the sample data).

Method of moments (MOM) estimators

SI The similarity index (SI), denoted S_{xy} (Li et al. 1993) is here defined as the average fraction of alleles at a locus in a reference individual *X* for which there is another allele in the proband *Y* that is IBS (Lynch and Ritland 1999). Let *x* and *y* be the genotypes of individuals *X* and *Y*, respectively, and let *a*, *b*, *c*, and *d* denote alternative alleles. The frequency of the allele *i* is given as p_i. The values that S_{xy} can take are equal to 1.0 (*x* = *aa* and *y* = *aa*, or *x* = *ab* and *y* = *ab*), 0.75 (*x* = *aa* and *y* = *ab*; or *x* = *ab* and *y* = *aa*), 0.5 (*x* = *ab* and *y* = *ac*), and 0 (*x* = *ab* and *y* = *cd*) A single locus estimator for θ_{xy} can be written as $\hat{\theta}_{xy} = 0.5\hat{r}_{xy}$, with $\hat{r}_{xy} = \left(S_{xy} - S_0\right)/(1 - S_0)$, where $S_0 = \sum_{i=1}^{n} p_i^2 (2 - p_i)$ is the expected value of *S* for unrelated individuals in a random-mating population. The multi-locus estimator is taken as the non-weighted average across all loci, i.e., $\hat{\theta}_{XY} = (1/L)\sum_L \hat{\theta}_{xy}$, where *L* is the number of loci.

QG This MOM estimator was primarily proposed for estimating the average degree of relatedness within groups of individuals (Queller and Goodnight 1989). However, Lynch and Ritland (1999) expressed the estimator in terms that can be used to estimate pairwise relatedness for individuals *X* and *Y*. Let the reference individual *X* with genotype *x* have alleles *a* and *b* and the proband individual *Y* with genotype *y* have alleles *c* and *d*. Then let S_{ab} denote the indicator variable for sharing of pairs of alleles, and this variable equals 1 or 0 if the reference individual is homozygous or heterozygous, respectively. Likewise for the indicators S_{ac}, S_{ad}, S_{bc}, S_{bd}, and S_{cd}, the estimator for the coancestry coefficient is $\hat{\theta}_{xy} = 0.5 \times \hat{r}_{xy}$, with $\hat{r}_{xy} = (0.5(S_{ac} + S_{ad} + S_{bc} + S_{bd}) - p_a - p_b)/(1 + S_{ab} - p_a - p_b)$, i.e., half the relatedness coefficient. Note that this estimator fails in case of heterozygous individuals at diallelic loci since $S_{ab} = 0$ and $\hat{r}_{xy}$ is not defined. Similarly, to the SI estimator, the multi-locus estimator is calculated as the non-weighted average across all loci.

LR Lynch and Ritland (1999) proposed the single locus estimator as $\hat{r}_{xy} = \left(p_a(S_{bc} + S_{bd}) + p_b(S_{ac} + S_{ad}) - 4p_ap_b\right)/((1 + S_{ab})(p_a + p_b) - 4p_ap_b)$ and used $w_{r,x} = ((1 + S_{ab})(p_a + p_b) - 4p_ap_b)/(2p_ap_b)$ as a locus-specific weight. Then, the multilocus estimator is $\hat{r}_{XY} = \left(\sum_L w_{r,x}\hat{r}_{xy}\right)/\sum_L w_{r,x}$, where $w_{r,x}\hat{r}_{xy}$ reduces to $0.5((S_{bc} + S_{bd})/p_b + (S_{ac} + S_{ad})/p_a - 4)$. The estimator for the coancestry coefficient is calculated as $\hat{\theta}_{XY} = 0.5 \times \hat{r}_{XY}$.

W The estimator proposed by Wang (2002) has some similarity to the LR estimator as it also starts with the equation $r = \phi/2 + \Delta$ where the weighted least-squares estimators of ϕ and Δ were obtained by solving Eqs. 9 and 10 in Wang (2002). However, this estimator was devised to handle better the uncertainty in allele frequency estimates (Wang 2002). Again, the estimator for coancestry coefficient becomes $\hat{\theta}_{XY} = 0.5 \times \hat{r}_{XY}$.

Note that both the QG and LR estimators are asymmetric with respect to the two individuals, that is $\hat{\theta}_{xy} \neq \hat{\theta}_{yx}$, and we use the average of the estimates taken from the different orderings of the two individuals (Anderson and Weir 2007). Furthermore, not all MOM estimators restrict the estimates of coancestry coefficients to be non-negative. However, since we define a coancestry coefficient as IBD probability, we can adjust negative estimates by setting them equal to zero, and we denote the adjusted estimators as **QGa**, **SIa**, **LRa**, and **Wa**, respectively.

Maximum likelihood (ML) estimators

The single locus likelihood of a relationship specified by Δ may be given by $L(\Delta) = \Pr(IBS_i|\Delta) = \sum_j \Pr\left(IBS_i|IBD_j\right)\Delta_j$, where $\Pr\left(IBS_i|IBD_j\right)$ is the probability of the *i*th IBS mode, given the *j*th IBD mode. The probability that a pair of individuals will be in the *j*th IBD mode is denoted Δ_j. For the Jacquard and Thompson estimators, these probabilities are given in Table 7, p. 43 of Thompson (1986) and also in Table 1 of both Milligan (2003) and Anderson and Weir (2007). The probabilities for the Anderson-Weir estimator are in Table 2 of Anderson and Weir (2007). Note that these two tables are identical when the population structure parameter, F_{ST}, equals zero. For unlinked marker loci, the multiple-loci likelihood is the product of all single locus likelihoods.

Th This estimator assumes no inbreeding (Milligan 2003; Thompson 1975) with a parameter space constrained by $\Delta_1 = \ldots = \Delta_6 = 0$, reducing the likelihood to a function of parameters $k_2(\equiv \Delta_7)$, $k_1(\equiv \Delta_8)$, and k_0 $(\equiv \Delta_9)$, with $0 \leq k_i \leq 1$ (*i* = 0, 1, 2) and $k_0 + k_1 + k_2 = 1$. The coancestry coefficient is estimated as $\hat{\theta}_{xy} = 0.5\left(\hat{k}_2\right) + 0.25\left(\hat{k}_1\right)$.

AW This estimator assumes some degree of population structure, that is specified by the population structure parameter F_{ST} (Anderson and Weir 2007). We examined the following values for F_{ST}: 0.01, 0.03, 0.05 and 0.10, denoted as AW01, AW03, AW05, and AW10, respectively. The same constraint was put on the parameter space as for Thompson's estimator.

J The estimator with the nine Jacquard coefficients (Hepler 2005). This is the only estimator that does not assume non-inbred individuals. The Jacquard estimator thus provides estimates for the inbreeding coefficients of the individuals. Obeying laws of probability, we apply the constraints $0 \le \Delta_j \le 1$ (j = 1, …, 9) and $\sum_{j=1}^{9} \Delta_j = 1.0$, where the coancestry coefficient is estimated as $\hat{\theta}_{xy} = 1.0 \times \left(\hat{\Delta}_1\right) + 0.5 \times \left(\hat{\Delta}_3 + \hat{\Delta}_5 + \hat{\Delta}_7\right) + 0.25 \times \left(\hat{\Delta}_8\right)$.

We found that the simplex method (Press et al. 2002) to obtain ML estimates, as used by Anderson and Weir (2007), did not work well for the Jacquard estimator. We developed an EM algorithm, originally proposed by Dempster et al. (1977), to obtain ML estimators for the model parameters (unpublished work, Hepler et al.). The EM algorithm begins with a starting estimate of Δ_1,…, Δ_9. Given these values, we use Bayes theorem to calculate the probability that the pair is in each of states S1,…, S9 at each marker. The values of Δ_1,…, Δ_9 are then updated by calculating the expected number of markers at which the pair is in states S1,…, S9 (given that we know the probability of the pair being in each state at each marker). In this way, we iterate between updating Δ_1,…, Δ_9 and updating the probability of the pair being in each IBD state at each marker until convergence is reached. For all ML estimators, we applied multiple starting configurations and we used the value of 1.0e-4 as a convergence criterion.

HiDRAS data

The pairwise relatedness estimation was performed on a set of 33 parents of multiple mapping populations that were the resource for a large QTL mapping study for apples in the EU project HiDRAS (Gianfranceschi and Soglio 2004) (http://www.hidras.unimi.it/). As this project was still ongoing and both marker and pedigree data may be added or modified, we took the dataset available on 25 June 2007 (Table 1). The dataset contained 135 founder individuals, i.e., individuals whose parents were unknown. The 135 founder individuals are taken as the founder population to maximize the allele counts used to calculate allele frequencies. Only 38 of these founders were true ancestors of the 33 parents. The other founders were included in the HiDRAS project to study allelic diversity and were helpful in the estimation of allele frequencies in the founder population. Let n_i denote the number of individuals, then the number of pairwise coancestry coefficients is $n_i(n_i + 1)/2$ and equals 561 for the 33 parents.

Table 1 The pedigree composition of the HiDRAS apple data

	HiDRAS		Simulation	
		Parents		*Parents*
Founders	135	*1*	38	*1*
Non-founders	250	*32*	70	*32*
Total	385	*33*	108	*33*

The core pedigree (n = 108) pertains to the set of 33 parents and their ancestors and was used in the simulation datasets

Release date of this dataset was 25 June 2007

The 33 parents were crossed to produce the HiDRAS mapping populations

The set of 33 parents and all their ancestors form the core pedigree of 108 individuals (Table 1). This core pedigree will be used in the simulation datasets (see later). The pedigree records of the HiDRAS population were available for up to six ancestral generations and pedigree structure was complex as generations were overlapping and more importantly, many marriage loops were present that induced inbreeding (see Fig. 3 in Bink et al. (2007)). The complexity of the pedigree structure caused considerable variation in the pedigree-based coancestry coefficients among the 33 parents (Fig. 1). The pedigree-based distribution of coancestry coefficients was continuous instead of a distinct clustering at 0 (not related), 0.0625 (e.g., second cousins), 0.125 (half sibs), and 0.25 (e.g., full sibs or parent-offspring). Close to 50 percent of the coancestry coefficients were zero. In the set of 33 parents, 9 were inbred with their inbreeding coefficients equal to 0.01, 0.02, 0.03, 0.05, 0.06 (3×) and 0.19 (2×). Note also that several coancestry coefficients exceeded 0.5 (Fig. 1). In all our figures, we will depict the coancestry coefficients for pairs involving one or two inbred individuals by a different color and symbol, i.e., red triangle-up versus black triangle-down to visualize the possible influence of inbreeding.

The marker data comprised 86 SSR markers on 17 linkage groups, covering 11 Morgans with the number of markers per linkage group ranging from 2 to 10 (Fig. 2) (Patocchi et al. 2007). The average marker distance was 0.16 Morgan and we treat the markers as unlinked in this study. The number of alleles per marker ranged from 2 to 19 in the full HiDRAS population. For most marker loci, the fraction of typed individuals among the 135 founders was close to 0.3 (Fig. 2).

Some of the estimators use an estimate of the population structure parameter F_{ST} (Wright 1951) as presented in equation 5.2 in Weir (1996), for different populations. This

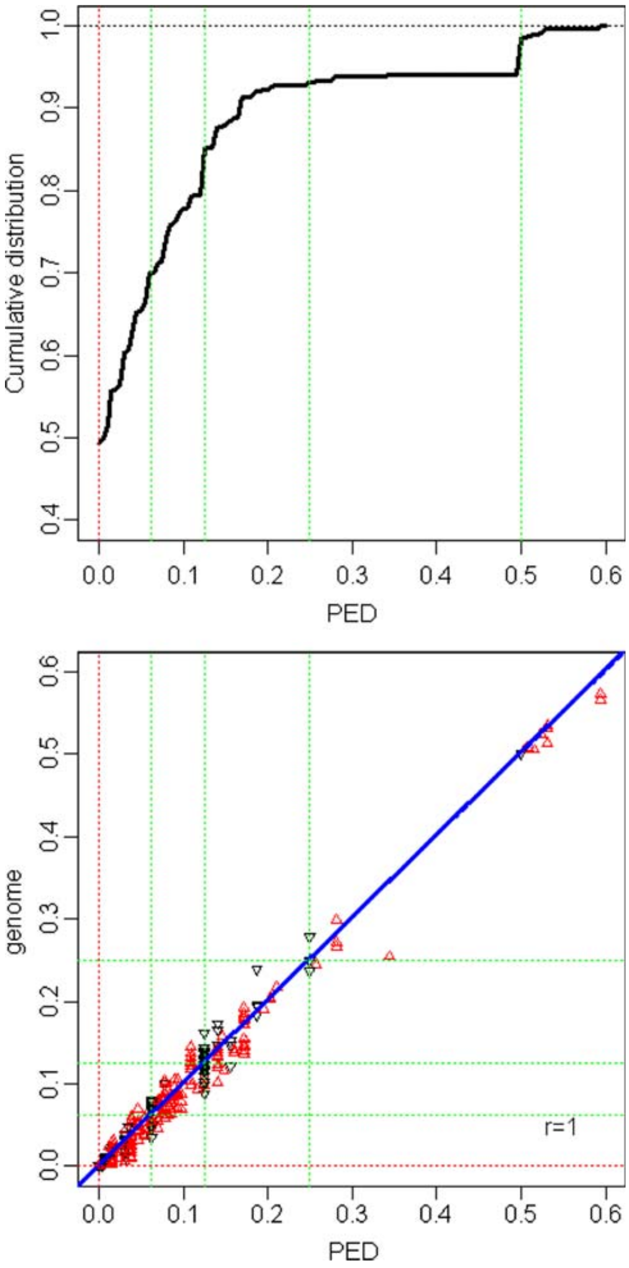

Fig. 1 Cumulative distribution of pedigree-based pairwise coancestry coefficients among the 33 HiDRAS parents (*top panel*), and the relationship between pedigree-based coefficients and realized genome-wide IBD proportions among these individuals for the first replicate of simulation (*bottom panel*)

was achieved by considering the set of 33 parents and the set of 138 founders as separate populations. Note that the individuals in the set of parents are not unrelated which hampers the estimation of the allele frequencies. To account for multiple alleles, let v denote the number of alleles at a particular locus; then we estimate F_{ST} for each marker as $\hat{F}_{ST} = \sum_{u=1}^{v} s_u^2 / \sum_{u=1}^{v} \tilde{p}_{u.}(1 - \tilde{p}_{u.})$, where $s_u^2 = \frac{1}{(r-1)\bar{n}} \sum_{i=1}^{r} n_i \left(\tilde{p}_{u,i} - \tilde{p}_{u.}\right)^2$ with $\bar{n} = \sum_{i=1}^{r} n_i / r$, where r equals 2 as we compare the set of founders with the set of parents; $\tilde{p}_{u,i}$ and $\tilde{p}_{u.}$ refer to the frequency of allele u as estimated from the ith population and to the average allele frequency across populations, respectively. The estimates for F_{ST} for all 86 loci were mostly below 0.03, however, some loci had estimates greater than 0.05 (Fig. 2). The higher F_{ST} values may be random noise (Lewontin and Krakauer 1973) or the corresponding genomic regions may have been under strong selection pressure as parents of the QTL mapping populations are usually consciously chosen, i.e., having a specific highly desired trait and/or frequently used in breeding programs. The estimates for Nei's genetic distance (Nei 1972) were strongly correlated with the estimates for F_{ST} (results not shown).

Marker allele frequencies

We consider three scenarios of marker allele frequencies, two (**D** and **E**) pertaining to absence of knowledge about population allelic frequencies and one (**F**) where we have access to unrelated founder individuals from the population.

Dataset-derived (**D**) We estimate the allele frequencies from the dataset itself, i.e., the 33 HiDRAS parents. We note that the unrelatedness assumption is not met in this scenario. The major pitfall here is that the frequency of a rare allele shared by related individuals in the dataset is overestimated, thus not being recognized as being rare. Consequently, the coancestry coefficients for these pairs are underestimated.

Equal (**E**) We assume a discrete uniform distribution, i.e., for a particular marker locus each allele is equally probable. That is, if n alleles were observed in the sampled dataset (Fig. 2), all alleles occur at frequency equal to $1/n$. The major pitfall will be that alleles at high frequency are not recognized as such and the coancestry coefficient of pairs sharing these alleles will be overestimated.

Founder population (**F**) As pedigree records were available and we had access to the original founders of a pedigreed population, we estimated the allele frequencies from the set of founders. Here, we had DNA for marker genotyping available of a fraction (close to 0.3) of the 135 founder individuals (Fig. 2). Some alleles were not present in the typed founders but were present in the set of 33 parents. To prevent estimates being equal to zero, we added 1 copy to every allelic class when estimating the allele frequencies in the founder population.

We note that both the allele frequencies in scenarios D and F are essentially estimated from the same HiDRAS pedigreed population dataset. However, different sets of individuals were used leading to different estimates of allele frequencies, as is illustrated by the estimated F_{ST} values (Fig. 2). Furthermore, it should be noted that we aim to measure pairwise relatedness relative to the founder

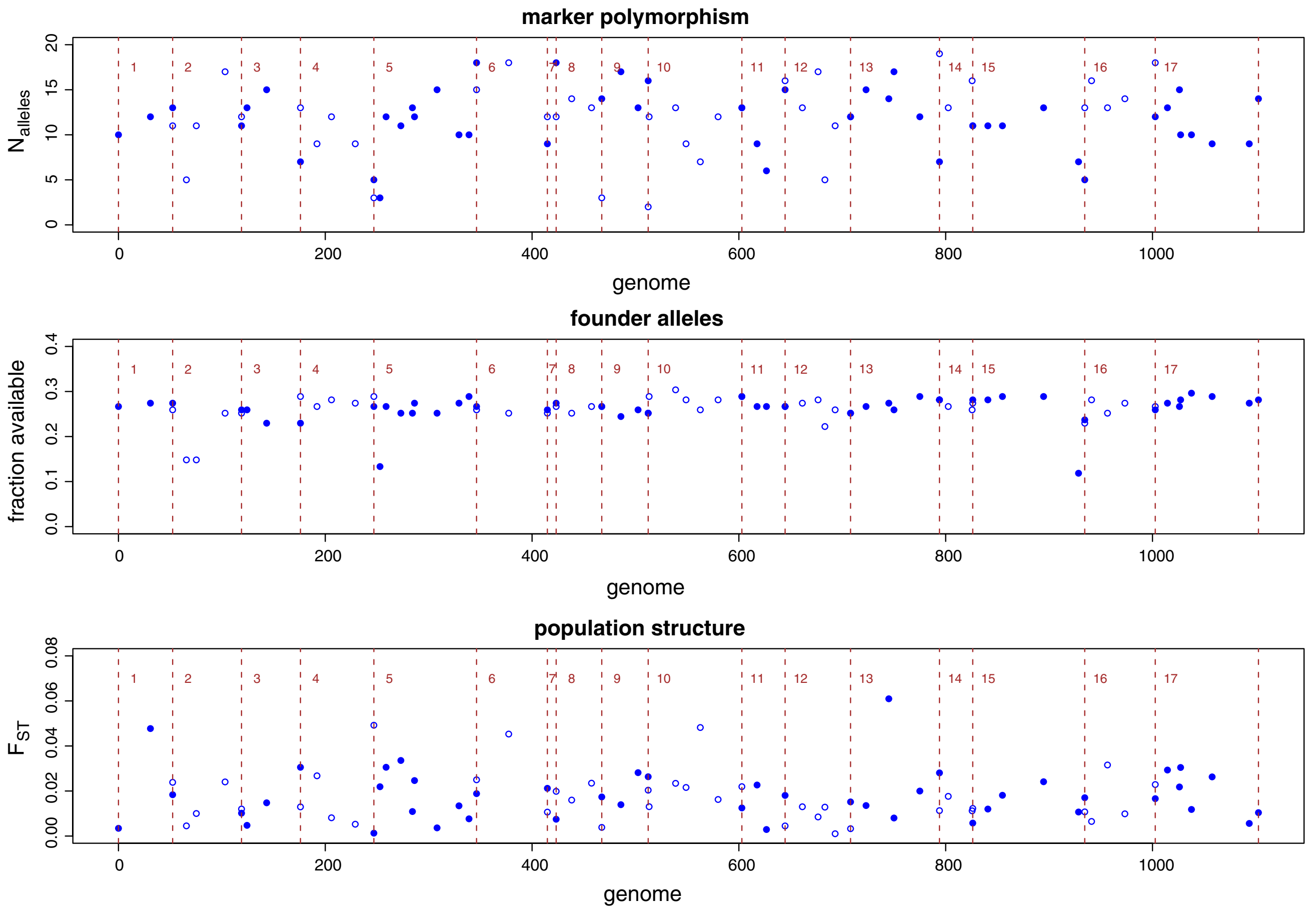

Fig. 2 Number of alleles per marker locus in the HiDRAS population (*top panel*), the fraction of typed individuals among 135 HiDRAS founders (*middle panel*), and population structure (F_{ST}) between the set of 135 founders and the set of 33 parents (*lower panel*). The genetic linkage map holds 17 groups comprising 86 SSR markers and a total genetic distance of approximately 11 Morgans

population, but all methods actually measure relatedness with respect to some (possibly fictional) generation in which the allele frequencies we are using first began to hold.

Influence of the number of marker loci

We also investigated the influence of the number of markers on the performance of the various estimators by selecting 17, 34, and 50 marker loci. These numbers were obtained by picking at every chromosome the first; the first and final; and the first, final and most intermediate loci (except chromosome 7 which only had 2 loci), respectively.

Simulated data

To study the performance of the various pairwise relatedness estimators further, we simulated 10 replicated datasets conforming to the pedigree relationships of the core of the real data set (Table 1). Thus, we simulated 38 independent founders and their descendants according to the known pedigree structure. A gene-dropping method (Maccluer et al. 1986) was used to simulate Mendelian inheritance of marker alleles from parents to offspring while Haldane's mapping function (Haldane 1919), was used to transform linkage distances into recombination fractions. The genetic linkage map differed from the real data set as markers were placed equidistantly along the genome. That is, using Haldane's mapping function we simulated 17 linkage groups each with length equal to 0.8 Morgan and with markers every 0.2 Morgan, resulting in 85 marker loci. For every marker locus, four alleles were simulated with their frequencies equal to 0.125, 0.125, 0.25, and 0.50, respectively. These allele frequencies were taken to mimic the inequality in frequencies observed in the real data. The marker genotypes were simulated to be in gametic and Hardy Weinberg equilibrium. We emphasize that in the simulation datasets the pedigree-based coancestry coefficients truly reflect the known pedigree, i.e., the founders are independent and unrelated, which was unlikely to be true in the real data.

In the simulation, the transmission of founder genomic segments was known with a resolution of 0.01 Morgan. In other words, we traced the origin of genomic segments back to the founder chromosomes at 0.01 Morgan-bins. From this, we calculated the realized genome-wide pairwise IBD proportions for the set of 33 parents. The relationship between the pedigree-based coefficients and realized proportions was highly accurate (Fig. 1). For the 10 replicated datasets, the averaged correlation and RMSE (see later) were close to 1.0 and below 0.01, respectively (not shown). The variation in the meiosis process and finite length of the genome caused sampling variation as shown in Fig. 1. For example, the pedigree-based coancestry coefficient 0.125 corresponds to a range in genome-based proportions (0.09–0.15). In the simulated datasets, 76 (38 × 2) alleles were available to estimate the allele frequencies in founder population.

Criteria for comparing performance of estimators

In this study, we choose to have the known pedigree of the 33 parents as the point of reference to assess the performance of the marker-based estimators of coancestry coefficients. We plot the pedigree-based versus the marker-based estimates of coancestry coefficients and calculate the Pearson correlation coefficients and the linear regression lines. The pedigree-based estimates are taken as the predictor variable (plotted along the *x*-axis) and the marker-based estimates are used as response variable (plotted along the *y*-axis). The correlations and regression lines are based on pairs of distinct individuals, i.e., excluding the coancestry coefficients of each individual with itself. In the visualization, we make a distinction between pairs that include one or two inbred individuals (*red triangle-up*) and pairs with no inbred individuals (*black triangle-down*). Omitting the within-individual coancestry coefficients, the number of pairwise coancestry coefficients is $n_{\mathrm{i}}(n_{\mathrm{i}}-1)/2$ and equals 528 for the 33 parents.

The root mean square error (RMSE) comprises both bias and standard error and was here quantified as $\mathrm{RMSE}(\theta) = \sqrt{\left(1/n_{\{\mathrm{i,j}\}}\right)\sum_{k=1}^{n_{\{\mathrm{i,j}\}}}\left(\hat{\theta}_k-\tilde{\theta}_k\right)^2}$ where $n_{\{\mathrm{i,j}\}}$ is the number of pairs of individuals, $\hat{\theta}_k$ and $\tilde{\theta}_k$ are the estimated coancestry coefficient for the *k*th pair from a marker-based and a pedigree-based estimator, respectively. We calculated the RMSE across all pairs of individuals and across all pairs of unrelated individuals ($\tilde{\theta}_k = 0$), respectively (Table 2).

Results

HiDRAS results

Data derived (D) The scenario of estimating the allele frequencies from the dataset itself resulted in considerable variation in estimated correlations (Figs. 3, 4). The highest correlation was obtained for the Wang estimator after adjustment of negative estimates. In this scenario, all MOM estimators had a major proportion of negative estimates for coancestry coefficients. The adjustment of negative values resulted in very different regression lines and, except for the LR estimator, it resulted always in higher correlations compared to the unadjusted estimates (Fig. 3, 4). The adjustment of negative values also improved the RMSE values (Table 2) of the MOM estimators. The LR estimator performed relatively poorly and the adjustment of negative estimates resulted in a even lower correlation and a too steep slope of the regression line (Fig. 3). The correlations of the ML estimators were in general lower, especially for high values of F_{ST} in the AW estimator (Fig. 4). Apparently, these higher values for F_{ST} caused a (too) strong regression of the estimated coancestry coefficients towards zero. This is also seen in the estimated RMSE as the RMSE across all pairs increases when F_{ST} increases. However, the RMSE of the ML estimators were lower than those of the unadjusted MOM estimators (Table 2).

Equal (E) This scenario resulted in a general over-estimation of the pairwise coancestry coefficients (Fig. 3). That is, the estimated regression lines were all biased to higher values. The magnitude of the bias was largest for the Jacquard estimator and smallest for the Wang estimator. All the correlation estimates were very similar, ranging from 0.79 to 0.81 (Fig. 4). In this scenario, there were no negative estimates for coancestry coefficients so the adjusted estimators might have been omitted here. The severe bias is also reflected in the estimated RMSE (Table 2), where the effect of the population parameter F_{ST} is also clearly visible. The AW10 had the lowest estimate as the higher F_{ST} values cause a stronger regression of the coancestry coefficient towards zero and thereby counteracting the effect of assuming equal allele frequencies. The high RMSE for the Jacquard estimator may be due to the less accurate estimation of the large number of parameters of this estimator.

Founder population (F) The use of founder allele frequencies resulted, on average, in the highest estimated correlations (Fig. 3, 4) with the highest value (0.89) for the AW03 estimator. This AW03 estimator gave very low estimates for RMSE as well (Table 2). However, higher values for F_{ST} resulted in an undesired regression towards zero of the estimated coancestry coefficients, which resulted in an increased RMSE for all pairs. The MOM estimators gave negative estimates and the adjustment of these negative values resulted in higher correlation and lower RMSE estimates. However, these were never better than the ML estimators (except AW10). The LR estimator had comparable high correlations and low RMSE estimates.

Table 2 Root mean square error (RMSE) of pairwise relatedness estimates $(\theta_k)^2$ from MOM and ML estimators to HiDRAS and simulation data

	Data derived (D)				*Equal (E)*				*Founder population (F)*			
	HiDRAS		Simulation		HiDRAS		Simulation		HiDRAS		Simulation	
	θ_k	$\theta_k = 0$	θ_k	$\theta_k = 0$	θ_k	$\theta_k = 0$	θ_k	$\theta_k = 0$	θ_k	$\theta_k = 0$	θ_k	$\theta_k = 0$
SI	6	7	7	7	12	13	9	10	7	7	4	4
QG	8	8	7	7	13	14	11	11	6	6	4	4
LR	7	4	7	5	13	13	11	12	3	3	4	3
W	6	6	6	6	12	13	10	11	6	6	4	4
SIa	4	2	3	1	12	13	9	10	7	7	3	3
QGa	4	1	4	1	13	14	11	11	5	6	3	3
LRa	6	1	5	1	13	13	11	12	3	3	3	2
Wa	3	1	3	1	12	13	10	11	6	6	3	3
Th	5	1	5	1	11	12	9	11	4	3	3	2
AW01	5	1	5	1	11	12	9	10	3	3	3	1
AW03	5	0	5	1	1	0	8	9	3	2	3	1
AW05	6	0	6	0	9	9	7	7	3	1	4	1
AW10	6	0	6	0	6	6	4	4	4	1	5	0
J	5	1	5	1	15	16	14	14	4	4	3	2

Marker allele frequencies were estimated from the data ($n = 33$), taken equal, or estimated from the founder population. For the simulation, the averages of 10 replicates are presented

Values have been multiplied by 100 to improve readability of Table

θ_k is relatedness of pair k; $\theta_k = 0$ pertains to those pairs that involve two individuals not related by pedigree

Method of Moments estimators: *SI* similarity index, *QG* Queller-Goodknight, *LR* Lynch-Ritland; and *W* Wang. QGa, SIa, LRa, and Wa refer to the estimates adjusted for negative values (set equal to zero)

Maximum Likelihood estimators: *Th* Thompson, *AW* Anderson-Weir and *J* Jacquard. The substrings 01, 03, 05, and 10 refer to $F_{ST} = 0.01$, 0.03, 0.05, and 0.10, respectively

Influence of the number of marker loci We also investigated the influence of the number of markers on the performance of the various estimators (Fig. 5) by selecting 17, 34, and 50 marker loci. The increase in correlation was the largest when doubling the number of loci from 17 to 34 for all estimators. Correlations up to 0.80 were obtained when two markers per chromosome and appropriate allele frequencies were available. Several differences between the performance of MOM estimators (excluding the LR estimator) and the LR and ML estimators were observable. In the scenario of data derived frequencies, the MOM estimators show an additional increase in correlation when going from 50 to 86 which was not observed for the other estimators.

Simulation results

The estimated correlations and RMSE, mean and standard deviations across ten replicates, for the simulation datasets showed almost identical patterns to those observed in the HiDRAS dataset (Fig. 4; Table 2). This may indicate that the relatedness through unknown common ancestors of the founders did not add substantially to the ancestral relatedness in the known part of the pedigree.

Equal (E) In case of the equal allele frequencies the estimated correlations were somewhat higher than those for the HiDRAS dataset, however, the uniformity in results across estimators is remarkable. The AW10 estimator appeared to be superior to the other estimators, which may be the result of by allowing deviations in allele frequencies via the population structure parameter.

Data derived (D) The variation in mean estimates of correlations was highest for the scenario of estimating the allele frequencies from the dataset itself. The MOM estimators, except the LR estimator, resulted in the highest correlations, especially after correction for the negative values for estimated coancestry coefficients. The influence of increasing the value of the population structure parameter F_{ST} was clearly visible in this scenario as higher values for F_{ST} resulted in much lower correlations and higher the RMSE estimates (Table 2). The AW estimators performed relatively poorly as the regression towards zero caused lower correlation and higher RMSE estimates among pairs with non-zero coancestry coefficients.

Founder population (F) The simulation results supported the results on the HiDRAS data fully. The ML estimators performed very well as long as the value for F_{ST} was not too large, i.e., up to 0.05. The LR estimator

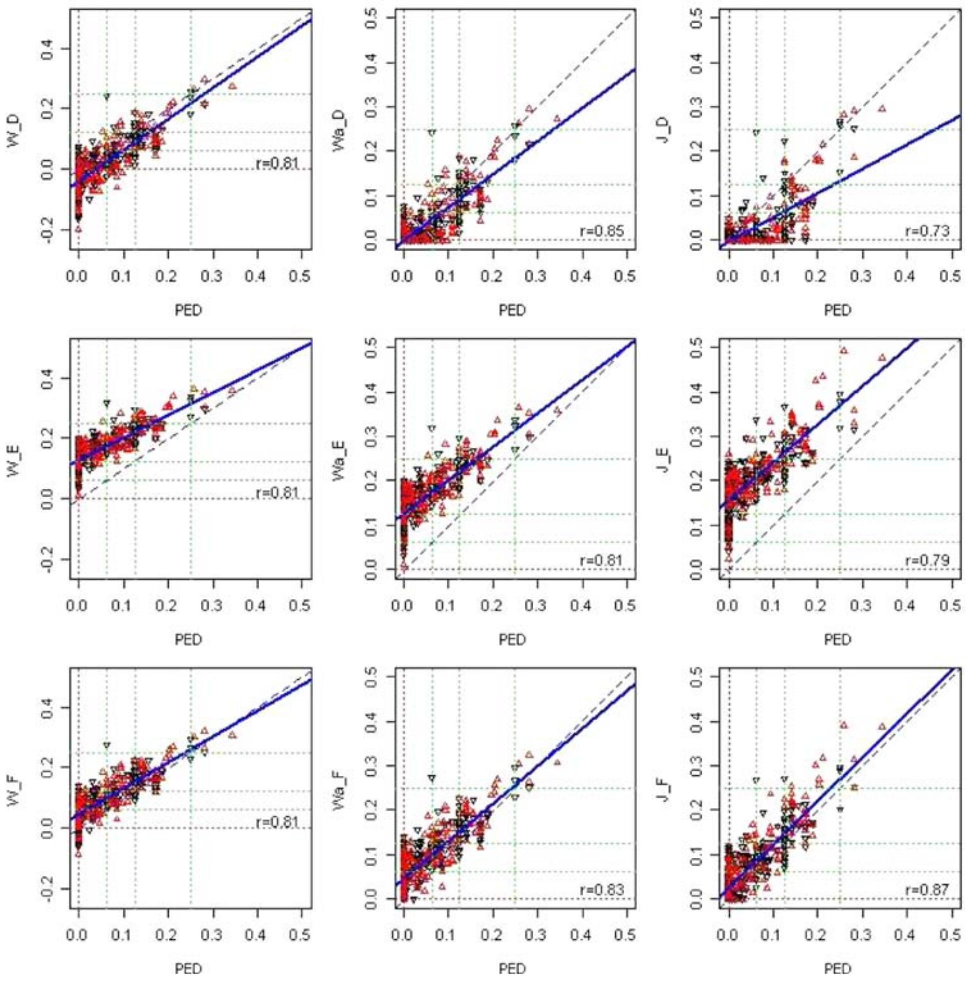

Fig. 3 Correlation (*r*) among estimates for pairwise coancestry coefficients among 33 HiDRAS individuals from the **W**, **Wa**, and **J** estimators, taking data derived (**D**); equal (**E**); or founder population (**F**) allele frequencies. The coefficients involving at least one inbred individual is indicated by *red triangles* (*black otherwise*). The *dotted lines* indicate coefficients equal to 0.0 (*red*) 0.0625, 0.125 and 0.25 (*green*). The estimated and unity regression is indicated by the *blue solid* and *dashed lines*, respectively

performed as well as the ML estimators and the other MOM estimators were relatively close after adjustment of negative estimates, both with respect to correlation and RMSE.

Discussion

We have compared several marker-based estimators of pairwise coancestry coefficients by analyzing real and simulated datasets. The pedigree-based coancestry coefficients were taken as the standard for comparison. The results of the analysis of the real data were strongly supported by those from the simulated datasets. The availability of (founder) population allele frequencies resulted in high correlations, in the simulated data as high as 0.90 in the simulated data. The performance of ML estimators and MOM estimators, especially the LR estimator, were basically the same in the real and simulated data. However, when allele frequency information is absent the use of estimates from the sampled dataset itself or naively taking equal allele frequencies resulted in lower correlations and higher RMSE values. Furthermore, the MOM estimators, with the exception of the LR estimator, performed relatively better as they seemed to be more robust to having incorrect specification of allele frequencies. This might especially pertain to the Wang estimator (Fig. 4) which is claimed to be robust for having unknown relatives being included in samples for estimating allele frequencies (Wang 2002). Without adjustment for negative estimates, the MOM estimators always gave higher RMSE values than the ML estimators, but this trend was reversed when this adjustment was made.

Influence of the number of marker loci

Several differences between the performance of MOM estimators (excluding the LR estimator) and the LR and ML estimators were observable. In the scenario of data derived frequencies, the MOM estimators show an additional increase in correlation when going from 50 to 86 loci, which was not observed for the other estimators. The correlations of the MOM estimators (except LR) only varied by a small amount across the different allele frequency scenarios. For

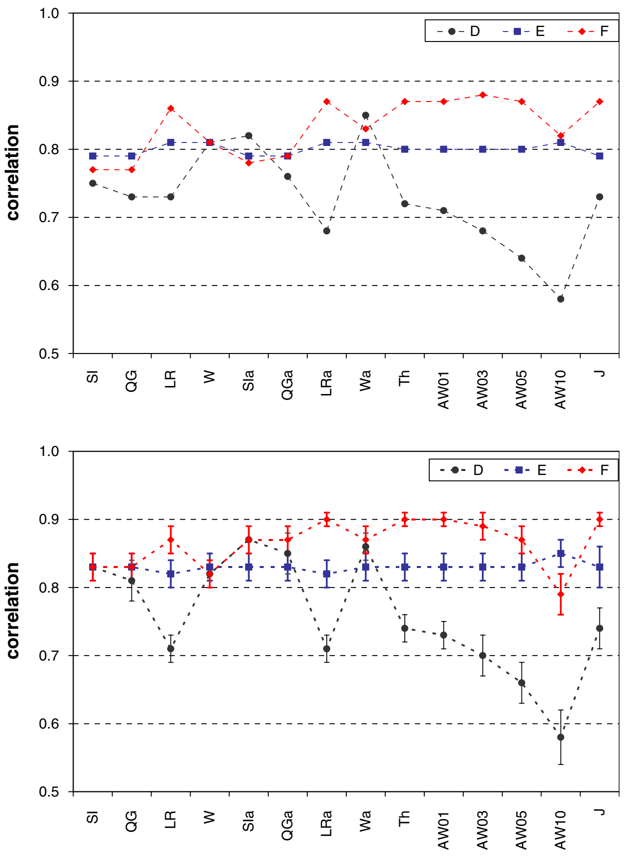

Fig. 4 Correlation of pairwise relatedness estimates from MOM1 and ML2 estimators to HiDRAS (*top panel*) and simulation data (*bottom panel*). Marker allele frequencies were taken estimated from the data ($n = 33$) (**D**), equal (**E**), or estimated from the founder population (**F**). The *y-error bars* represent the standard deviations for 10 simulated datasets

the other estimators, the correlations were highest across all numbers of marker loci in the scenario of founder allele frequencies. Adding more markers to the current dataset may only marginally increase the correlations. Figure 5 also shows that the SIa, QGa, and Wa estimators were insensitive to allele frequencies as correlation estimates were highly similar. The influence of the number of markers on the RMSE estimates was small, i.e., the RMSE only decreased mildly when increasing the number of marker loci. The RMSE estimates in the equal frequencies scenario were always distinctively highest for all estimators. The ranking in performance of the various estimators was not affected by the number of loci, which was consistent to the results of Oliehoek et al. (2006).

Influence of linked loci

All estimators in this study were based on the assumption of unlinked markers. The removal of 12 relatively closely linked markers from the set of 86 loci (distance less than 0.09 Morgan) resulted into estimated coancestry coefficients that were highly similar to those obtained from the full set of markers (results not shown). In addition, the increasing availability of (SNP) markers will increase marker density and linkage between markers will affect the variance of the estimators, although it will not change the expected values. Methods that consider linkage tend to have long running times as the dimensionality of IBD parameters increases (McPeek and Sun 2000; Sieberts et al. 2002) and a Baum algorithm (Baum et al. 1970) is needed to compute the likelihood for a given IBD configuration. This type of approach was not explored in this study.

Influence of inbreeding

The dataset contained several inbred individuals (see “Materials and methods”) and most estimators have been devised for outbred populations. However, the results did not point to a dramatic influence of inbreeding in the performance of the various estimators. In all figures, we used different colors for estimates between pairs that contained one or two inbred individuals versus pairs without inbred individuals (Fig. 3) and no clustering could be observed. The estimates of correlation and RMSE for these two groups were also similar (results not shown). Furthermore, the Jacquard estimator did not perform better than the Thompson estimator even though the first allows for inbreeding and the second does not. This lack of improvement might have two explanations; either the inbreeding levels were too low to influence the Thompson estimator (only 2 out of 9 non-zero coefficients larger than 0.06, see “Materials and methods”), or the accuracy of the Jacquard estimator is decreased because it uses more parameters.

Influence of population structure parameter

The allele frequencies in the sampled set of 33 parents were different from those in the founder population, as seen from the estimates of F_{ST} (Fig. 2). Our expectation was that accommodating deviations in allele frequencies into the estimator might improve the estimates of coancestry coefficients. The AW estimators with small values for F_{ST} (0.01, 0.03) performed as well as the original Thompson estimator whereas the estimators with higher values for F_{ST} (0.05, and 0.10) seemed to suffer from allowing deviations in the allele frequencies for the data derived allele frequencies. In retrospect, this should not have been surprising as the AW estimator was designed to accommodate situations where allele frequencies are available from a more distantly related population, not the sampled population itself. This also explains why this estimator, with higher values, performed well in the scenario of equal allele frequencies, i.e., these frequencies deviated from both those in

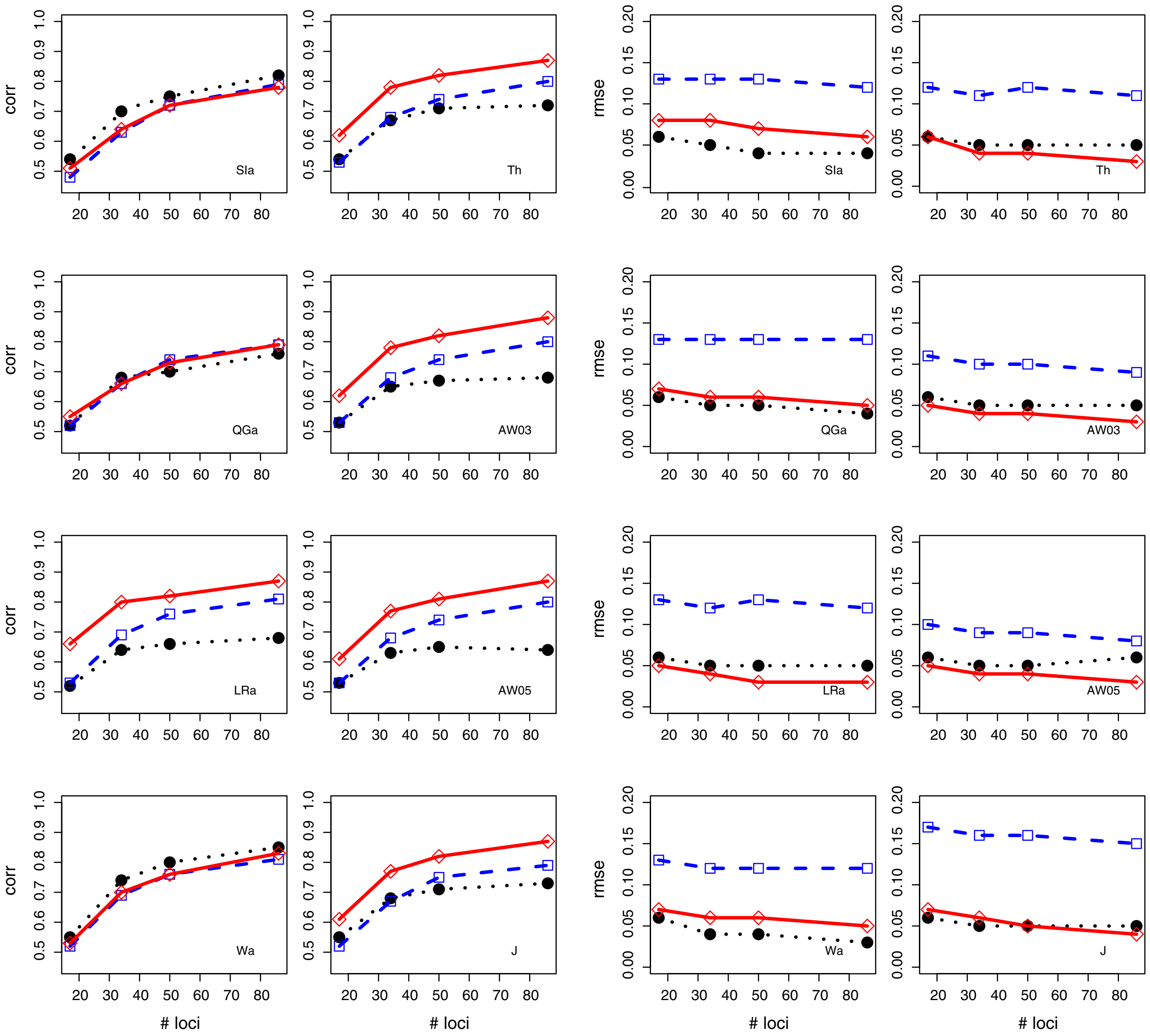

Fig. 5 Correlation (*corr*) and Root Mean Square Error (*rmse*) among estimates for pairwise coancestry coefficients as a function of the number of markers. Allele frequencies were data derived (*black circle*)(**D**), equal (*blue square*)(**E**), and founder (*red diamond*)(**F**), respectively

the founder population and those in the sampled population. In addition, when knowledge about recent population history and population substructures is available, other approaches may be considered (Gasbarra et al. 2007; Meuwissen and Goddard 2001).

Subsequent analysis

When pedigree records are inaccurate or missing, the estimation of coancestry coefficients might be a first step in estimation of genetic parameters of quantitative traits. Thomas (2005) reviewed different approaches to relationship estimation with particular attention on optimizing the use of this relationship information in subsequent heritability estimation. A real data example in Soay sheep (Thomas et al. 2002) using a set of 12 markers, did not result in reliable heritability estimates probably because of the low average relatedness in the population (Thomas et al. 2002). Our results indicate that 12 markers do not suffice to obtain accurate relationship estimates. The reliability of heritability estimates is expected to increase when higher numbers of markers are available and we plan to verify this in a separate study.

Practical guidelines

The results of this study did not identify one estimator for pairwise coancestry coefficients that outperformed other estimators. However, the allele frequencies used were critical. Taking equal allele frequencies generally leads to an overestimation of relatedness. This may be neutralized by taking a somewhat too high a value for F_{ST} in the AW estimator. Using dataset-derived allelic frequencies leads to underestimation, especially for the ML and LR estimators, as rare alleles shared by relatives are not recognized as such. We are currently investigating the possibility to estimate allele frequency and relatedness jointly. Nevertheless, access to reliable allele frequencies seems most beneficial since highest correlation and lowest RMSE estimates were obtained. In that scenario, the performances of the various estimators were not so different but seem to favor the ML estimators. In this study, a set of 34 polymorphic loci seemed to be a good balance between performance of estimators and marker genotyping costs as the increase in performance became marginal for higher number of loci. This optimal number will likely depend on the research question, the desired accuracy, and the number of chromosomes and total length of the genome.

Acknowledgments This study was performed while the first author was a visiting scholar to the Department of Statistics, University of Washington and he was supported in part by the Netherlands Science Organization (NWO) grant 050-72-429. Fruitful discussions with Bruce Weir are gratefully acknowledged. We gratefully acknowledge the constructive comments of the editor and two anonymous reviewers on earlier versions of the manuscript.

The genetic marker data used in this project were generated by partners (Plant Pathology, Institute of Integrative Biology (IBZ), ETH Zurich, CH-8092 Zurich, Switzerland; East Malling Research (EMR), East Malling, England; Bundesanstalt für Züchtungsforschung an Kulturpflanzen, Institut für Obstzüchtung, Dresden, Germany; Laboratory of Basic Research in Horticulture, Faculty of Horticulture and Landscape Architecture, Warsaw Agricultural University (WAU), Poland; Genetics and Horticulture (GenHort), National Institute for Agronomical Research (INRA), Beaucouzé Cedex, France; Istituto Agrario di San Michele all'Adige, Trento, Italy), Department of Plant Breeding, Plant Research International b.v., Wageningen UR, Wageningen, The Netherlands of the European Community research program "Quality of Life and Management of Living Resource", QLK5-2002-01492 "High quality Disease Resistant Apple for a Sustainable Agriculture", coordinated by L. Gianfranceschi from the University of Milan (It.).

References

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

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 41:164

Bink MCAM, Boer MP, ter Braak CJF, Jansen J, Voorrips RE, van de Weg WE (2007) Bayesian analysis of complex traits in pedigreed plant populations. Euphytica. doi:10.1007/s10681-10007-19516-10681

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

Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39:859–882

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via em algorithm. Journal of the Royal Statistical Society Series B-Methodological 39:1–38

Gasbarra D, Pirinen M, Sillanpaa MJ, Salmela E, Arjas E (2007) Estimating genealogies from unlinked marker data: a Bayesian approach. Theor Popul Biol 72:305–322

Gianfranceschi L, Soglio V (2004) The European project HiDRAS: Innovative multidisciplinary approaches to breeding high quality disease resistant apples. Acta Horticulturae 663:327–330

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. Journal of Genetics 8:299–309

Harris DL (1964) Genotypic covariances between inbred relatives. Genetics 50:1319

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

Jacquard A (1972) Genetic information given by a relative. Biometrics 28:1101–1114

Kempthorne O (1954) The correlation between relatives in a random mating population. Proceedings of the Royal Society of London Series B-Biological Sciences 143:103–113

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. Genetics 74:175–195

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

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

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

Maccluer JW, Vandeberg JL, Read B, Ryder OA (1986) Pedigree analysis by computer-simulation. Zoo Biology 5:147–160

Malécot G (1948) Les mathématiques de l'hérédité; préf. de L. Blaringhem. Masson, Paris

McPeek MS, Sun L (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet 66:1076–1094

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

Milligan BG (2003) Maximum-likelihood estimation of relatedness. Genetics 163:1153–1167

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Oliehoek PA, Windig JJ, van Arendonk JAM, Bijma P (2006) Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics 173:483–496

Patocchi A, Fernández-Fernández F, Evans K, Silfverberg-Dilworth E, Matasci CL, Gobbin D, Rezzonico F, Boudichevskaia A, Dunemann F, Stankiewicz-Kosyl M, Mathis F, Durel CE, Soglio V, Gianfranceschi L, Costa F, Toller C, Cova V, Mott D, Komjanc M, Barbaro E, Voorrips RE, Rikkerink E, Yamamoto

T, Cevik V, Gessler C, van de Weg WE (2007) Development of a set of apple SSRs markers spanning the apple genome, genotyping of HiDRAS plant material and validation of genotypic data. Acta Horticulturae

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) Numerical recipes in C++. The art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

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

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

Ritland K (1996b) Marker-based method for inferences about quantitative inheritance in natural populations. Evolution 50:1062–1073

Sieberts SK, Wijsman EM, Thompson EA (2002) Relationship inference from trios of individuals, in the presence of typing error. Am J Hum Genet 70:170–180

Thomas SC (2005) The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. Philosophical Transactions of the Royal Society B-Biological Sciences 360:1457–1467

Thomas SC, Coltman DW, Pemberton JM (2002) The use of marker-based relationship information to estimate the heritability of body weight in a natural population: a cautionary tale. J Evol Biol 15:92–99

Thomas SC, Hill WG (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics 155:1961–1972

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

Thompson EA (1986) Pedigree analysis in human genetics. Johns Hopkins University Press, Baltimore

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

Weir BS (1996) Genetic data analysis II: methods for discrete population genetic data/Bruce S. Weir. Sinauer Associates, Sunderland, Mass

Weir BS, Anderson AD, Hepler AB (2006) Genetic relatedness analysis: modern data and new challenges. Nature Reviews Genetics 7:771–780

Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56:330–338

Wright S (1951) The genetical structure of populations. Annals of Eugenics 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