Estimating inbreeding using dense panels of biallelic markers and pedigree information

A. Bouquet[1,*], M.J. Sillanpää[1,2], J. Juga[1]

[1] *Department of Agricultural Sciences, P.O. Box 27, FIN-00014 University of Helsinki*
[2] *Department of Mathematics and Statistics, P.O. Box 68, FIN-00014 University of Helsinki*

(*Corresponding author, e-mail: alban.bouquet@helsinki.fi)

Abstract:
The aim of this simulation study was to compare performances of inbreeding (F) estimators using dense panels of biallelic markers. Two types of population structures were derived for 10 discrete generations starting from an ancestral population at mutation drift equilibrium simulated with an effective size of 1000 and a mutation rate (μ=5.10^{-4}). Subpopulations differed by the level of selection applied both on males and females: no selection or a structure close to a breeding program with selection of the best 40 males and 500 females on EBV with accuracy of 0.85 and 0.71, respectively, on a trait with heritability of 0.3. Marker panels were made up of 36 000 biallelic markers (18 per cM) and were available for animals in the last 4 generations. Pedigrees were recorded on the last 8 generations. For each scenario, 30 replicates were carried out. Analysed estimators were the correlation (VR1) and regression (VR3) estimators described by VanRaden in 2008 to build the genomic relationship matrices. Other estimators included the weighted corrected similarity (WCS) estimator published by Ritland in1996 and a modified WCS estimator accounting for pedigree information (WPCS). Marker-based estimators were also compared to the pedigree estimator (PED). F estimates were correlated and regressed to the true simulated values of inbreeding to assess the precision and bias of estimators, respectively. Main results show that use of dense marker information improves the estimation of F, whatever the scenario. The accuracy of F estimates and the bias were increased in presence of selection, except for PED. Across scenarios, VR3, WCS and WPCS were the most correlated with true F values. In the situation where pedigree was exhaustive, VR3 performed as well as WCS and WPCS but had a larger variability over replicates. Although less biased on average, VR1 was less accurate than other estimators especially when allele frequencies were not properly defined. Accounting for pedigree information into WCS did not increase its accuracy and did not reduce bias in the tested scenarios. Thus, the results indicate that WCS, which can be also used with multiallelic markers, is a promising estimator both to build the genomic relationship matrix for genomic evaluations and to better assess genetic diversity in selected populations.

Keywords: genomic inbreeding, genomic relationship matrix, single nucleotide polymorphism

I- Introduction

Following recent advances of high-throughput genotyping technologies, the breeding industry has been largely adopting genetic markers for selection purposes. Based on the use of dense marker panels, genomic selection offers promising opportunities to improve rates of genetic gain using genomically enhanced evaluations (Schaeffer, 2006) but also the management of genetic diversity of livestock populations (Daetwyler et al., 2007). Compared to conventional BLUP evaluations, accounting for genomic information improves the accuracy of estimated breeding values because it allows estimating the mendelian sampling term, even early in life. Hence, genomic selection makes it possible to reduce inbreeding rates due to increased emphasis on own rather than family information (Daetwyler et al., 2007). Besides, molecular markers can be used instead of or coupled with pedigree information to improve the estimation the level and evolution of inbreeding coefficients (**F**). Therefore, genomic prediction and careful monitoring of genetic resources will be efficient provided accurate F estimators. So far, most marker-based estimators were conceived to be applied with sparse marker maps, in which markers were assumed to be independent, and with unselected and large populations in which existing levels of inbreeding could be ignored. Violation of both assumptions is obvious when estimating F coefficients in highly selected populations using dense marker maps. The objective of this study was to assess the performance of different F estimators for different types of population structures using simulated data. A marker-based estimator accounting for pedigree information was also derived and compared here to existing pedigree- and marker-based estimators.

II- Material and Methods

Description of simulated data. Performance of inbreeding estimators was assessed based on simulated datasets which were generated with QMSim program (Sargolzaei and Schenkel, 2009). Two different population structures were simulated which differed by the occurrence of selection. To simulate genotypic data with realistic linkage disequilibrium (**LD**) structures, both populations were derived from a common ancestral population at mutation-drift equilibrium (**MDE**). This population had an effective size of 1000 individuals and was simulated by randomly mating 500 males and 500 females for 5000 discrete generations. Two offspring were produced per mating. In generation 5000, individuals were considered as base founders for deriving the 2 populations used for analyses. Each subpopulation diverged independently for 10 generations and all genotypes, phenotypes and pedigrees were recorded for the last 8 generations.

In the first subpopulation, all demographic and reproductive parameters were the same as in the ancestral population to generate a panmictic and unselected population at MDE. The second subpopulation was simulated to mimic a strongly selected population on a trait with heritability of 0.3. Therefore, the best 40 males and 500 females were kept for reproduction based on estimated breeding values whose accuracies were set to 0.85 and 0.70 for males and females, respectively. To make selection effective on the female side, the number of females was increased by distorting the proportion of born female calves to 66%, thus avoiding increasing the population size.

Individuals had diploid genomes comprising twenty 1 M long chromosomes, each bearing 100 quantitative trait loci (**QTL**) with 2 equifrequent alleles in the first ancestral generation. A mutation rate of 2.5 10^{-5} per haploid site affected QTL so that new alleles could emerge during the simulation. QTL allelic effects were sampled from a gamma distribution with scale and shape parameters equal to 5.40 and 0.42, respectively (Hayes and Goddard, 2001). Neither dominance nor epistatic effects were simulated. QTL effects explained 75% of the genetic variance of the trait, the other 25% being attributable to polygenes. Each chromosome was also bearing 4000 evenly spaced markers with 2 equifrequent alleles in the first ancestral generations. In the first ancestral generation, all alleles were tagged with a unique label and transmission of ancestral alleles was followed over generations. A mutation rate of 5.10^{-4} per haploid site was applied to marker loci and kept constant over generations. When mutated, a new label was created to clearly distinguish identical by state (**IBS**) and identical by descent (**IBD**) alleles. The SNP allele value was then sampled from a Bernoulli distribution with probability 0.5. Hence, mutation was not recurrent. A panel of 36 000 markers was constituted by sampling at random among loci with minor allele frequencies (**MAF**) higher than 5%. Recombination was modelled with Haldane´s mapping function, assuming a mean number of 1 crossing-over per M and no interference.

Inbreeding coefficient estimators

Considering a single point in the genome, F can be defined as the probability that the two homologous alleles within an individual are IBD with respect to a defined based population. Considering the genome as a whole, the achieved inbreeding coefficient of an individual is the proportion of its genome which is IBD. Variation in homozygosity by descent (**HBD**) depends then both on pedigree and the extent to which alleles at different loci are jointly IBD (Hill and Weir, 2011). In this study, true F coefficients were directly computed by counting the genome-wide (**GW**) proportion of HBD loci over the 80 000 simulated markers. This was facilitated by the use of unique labels defined for each ancestral allele. To be comparable in magnitude with other inbreeding estimates, true coefficients were expressed relative to the mean HBD of the base population comprising individuals recorded as founders in the pedigree file using conventional change of base population (Powell et al., 2010).

Pedigree-based estimator. Pedigree-based F coefficients ($\mathbf{F_{ped}}$) were estimated with Relax2 (Strandén and Vuori, 2006) using algorithm by Meuwissen and Luo (1992) and exhaustive pedigree information over the last 8 generations.

Regression of genomic over pedigree-based coefficients. VanRaden (2008) proposed an estimator based on the linear regression of marker genotype sharing over the pedigree-based relationship matrix (A) adjusting for the mean homozygosity of the population:

MM' = $g_0$11' + g_1A + E, where g_0 and g_1 are the intercept and slope of the regression model, respectively. Matrix E includes differences of true from expected fractions of DNA in common plus measurement error. Then, the genomic relationship matrix can be obtained by reversing the calculations using g_0 and g_1 estimated in the first step as described in Van Raden (2008).

Genome-wide covariance of minor allele counts. This approach initially introduced by Li and Horwitz (1953), was recently used by VanRaden (2008) to efficiently build the genomic relationship matrix used for genomic evaluations. Assuming Hardy-Weinberg equilibrium (**HW**) and linkage equilibrium (**LE**), F can be derived as the correlation between gametes constituting an individual (Powell et al., 2010). As described by VanRaden (2008), locus specific F estimates can be efficiently obtained from biallelic marker data as GW-homozygosity by state (**HBS**) corrected for mean HBS in the base population divided by the variance in homozygosity expected at this locus under HW-LE. Locus specific estimates were averaged over loci to obtain a GW estimate. This estimator was denoted VR1.

Weighted corrected similarity. Ritland (1996) extended the preceding approach to any kind of codominant markers by applying a 3-step procedure to compute at first allele-specific estimators which are combined into locus specific estimators and again combined into GW estimates. Briefly, the probability s_{ikl} to sample 2 homozygous alleles of value l at a marker k in an individual i can be partitioned as $s_{ikl} = F_{ikl}*p_{kl}+ (1-F_{ikl})*p_{kl}^2(1)$ where F_{ikl} is the F coefficient of individual i estimated using allele l at marker locus k and p_{kl} denoting allele frequency of allele l at locus k. By reversing (1), allele specific F_{ikl} coefficients can be estimated as $F_{ikl} = (s_{ikl}- p_{kl}^2)/(p_{kl}*(1- p_{kl}))$. In this moment estimator, expectation of allelic similarity s_{ikl} was replaced by its observed value for allele l at locus k $S_{xykl} = 0.25*(I_{ackl} + I_{adkl} + I_{bckl} + I_{bdkl})$ with I_{xykl} an indicator variable equal to 0 if paternal allele x (a or b) and maternal allele y (c or d) are homozygous.

At a locus, since all allelic types are not equally informative depending on their allele frequencies, it is desirable to find an optimal linear combination of allele specific estimators: $F_{ik} = \sum_l w_l F_{ikl}$ which maximizes accuracy and minimizes bias of the locus specific estimator. The vector of optimal weights can be derived with a Lagrangian optimization procedure under constraint of minimal variance of the estimator by minimizing the derivative of F_{ikl} under the constraint no bias. Using Lagrangian multipliers, it can be shown that w_{kl} is equal to $w_{kl}=\mathbf{V}^{-1}.\mathbf{1}/(\mathbf{1}'.\mathbf{V}^{-1}.\mathbf{1})$ where **V** denotes the matrix of variance-covariance between allele specific estimators at a locus as described by Ritland (1996) and **1** is a vector of 1. Similarly, a second optimization procedure canbe carried out to optimally combine locus-specific coefficients accounting for differences in informativeness between loci arising from differences in MAF across loci and statistical dependencies between markers. In the original estimator, Ritland (1996) set F_{ijk} to 0 and supposed that markers were independent. Under those conditions, optimal locus specific weights were equal to the inverse of the locus specific estimator variance. This estimator was denoted WCS.

Inclusion of pedigree information into Ritland´s estimator. Ignoring prior knowledge of pedigree to derive optimal weights may bias GW estimates, especially in intensively selected or small populations. Therefore, F_{ped} estimates were included in the computation of weights to construct the V matrix. In this study, locus specific weights were also derived assuming independence between markers. This estimator including pedigree information was denoted WPCS.

Comparison criteria of estimators. F coefficients were estimated using allele frequencies observed either in the population of pedigree founders or in the genotyped population (4 last generations). Performance of estimators was assessed based on i) the average bias (B), ii) the correlation coefficient between true HBD and each estimator (ρ) and finally iii) the linear regression slope (α) of true HBD on each estimator.

III. Results and Discussion.

Distribution of allele frequencies. The distribution of observed allele frequencies followed a uniform distribution (results not shown) in unselected as well as in selected populations. Over replicates, the mean correlations between observed allele frequencies in the genotyped and the pedigree founder populations were equal to 0.99 and 0.97 for unselected and selected populations, respectively.

Table 1. Mean bias (B), correlation coefficient with true HBD (ρ) and regression coefficient (α) of true HBD on each estimator along with the standard errors of the mean (SEM) obtained for unselected populations over 30 replicates

Estimator	Mean B (SEM)	Mean ρ (SEM)	Mean α (SEM)
PED	0.00 (6.82E-04)	0.72 (0.11)*	1.02 (0.12)
VR1	0.00 (7.67E-04)	0.91 (0.03)*	0.79 (0.04)*
VR3	0.00 (1.41E-03)	0.97 (0.01)	0.88 (0.02)*
WCS	0.00 (5.90E-04)	0.97 (0.01)	0.90 (0.02)*
WPCS	0.00 (6.17E-04)	0.96 (0.01)	0.90 (0.02)*

*Significantly different from 0 (or 1, accordingly) at a 5% error level

Performance of estimators in unselected populations. Over replicates, mean true HBD for individuals born in the last 4 generations was equal to 0.16% and mean standard deviation of HBD was 1.78%. Average biases of each estimator are presented in Table 1 along with correlation coefficients (ρ) and regression coefficients (α) of true HBD over inbreeding estimates. In unselected populations, no estimator was significantly biased (Table1). Pedigree estimator was less variable than marker-based estimators whereas variability of marker-based estimates was similar to the one of true HBD.

Correlation coefficients of analyzed F estimators with true HBD presented larger range of values. Pedigree estimator was the least correlated with true HBD (0.72) although the coefficient obtained in the present study was much higher than in other studies (Keller et al., 2011). This is mainly due to the effects of scaling down the size of genomes in simulation studies which increases the variability of true HBD whereas Fped assumes an infinitely large number of independent chromosome segments. VR1 was strongly correlated with true HBD (0.91). VR3, WCS and WPCS were even more strongly correlated with true HBD (0.97) confirming that use of marker information clearly improves the estimation of achieved inbreeding. Although no bias was detected on average for estimators in unselected populations, analysis of regression coefficients revealed occurrence of value-dependent bias. Average regression coefficient was close to 1 for F_{ped} indicating that bias was not increasing with increasing values of true HBD. However, the variability of α was larger for F_{ped} than for marker-based estimators. Regression coefficients for WCS and WPCS estimators were lower than 1 indicating that estimation bias tended to increase for animals with the most extreme values of HBD. This trend was even clearer for VR3 and VR1 (Table 1). In unselected populations, similar results were obtained whatever the definition of the base population since changes in allele frequencies were tiny between the pedigree founders and genotyped populations.

Performance of estimators in selected populations. Over replicates, mean true HBD for individuals born in the last 4 generations increased to 2.61% and average standard deviation of HBD was 3.60%. Comparison criteria were summarized in Table 2 and were computed using allele frequencies observed in the pedigree founder population. Pedigree estimates were significantly lower than true HBD (B= -0.009) and were also less variable (0.024). VR1 estimator was not significantly biased whereas WCS and WPCS significantly underestimated true HBD (Table 2). VR3 estimator tended to underestimate true genomic inbreeding but average bias did not significantly deviate from zero due to large variability of B over replicates.

Table 2. Mean bias (B), correlation coefficient with true HBD values (ρ) and regression coefficient (α) of true HBD on each estimator along with standard errors of the mean (SEM) obtained for selected populations over 30 replicates.

Estimator	Mean B (SEM)	Mean ρ (SEM)	Mean α (SEM)
Ped	-0.01 (0.003)*	0.67 (0.024)*	1.01 (0.043)
VR1	0.00 (0.003)	0.96 (0.004)*	0.90 (0.010)*
VR3	-0.01 (0.010)	0.98 (0.002)*	0.97 (0.012)*
WCS	-0.01 (0.002)*	0.99 (0.001)*	0.97 (0.004)*
WPCS	-0.01 (0.002)*	0.98 (0.002)*	0.97 (0.004)*

*Significantly different from 0 (or 1, accordingly) at a 5% error level

In selected populations, the correlation coefficient between true HBD and F_{ped} slightly decreased, indicating decreased estimation accuracy. However, α coefficient indicated that bias was still independent of HBD values. In selected populations, accuracy of marker-based estimators (ρ>0.91) increased. Indeed, since increases in HBS were supposed to arise mainly from increase in HBD, marker-based estimators were more able to explain variability in F between individuals based on their levels of HBS. Accuracy of VR1 was significantly lower than the one of other marker-based estimators (Table 2). Although no average bias was detected for VR1, regression coefficient α indicated that a value-dependent bias occurred. Thus, with VR1, no bias was observed at the population level whereas individual biases could happen, especially for individuals with most extreme HBD values. Other marker-based estimators also suffered, although to a smaller extent, from value-dependent bias. At this point, it should be noted that including pedigree into WCS did not improve estimation accuracy which was already high. However, it should be noted that proportion of animals with large F values was very small. When considering only animals with true HBD larger than 6.25%, including pedigree information reduced B and made α closer to 1, although those changes were not significant based on 30 replicates.

As expected, when using allele frequencies observed in the current genotyped population, estimated mean F levels were close to 0 because this defined the genotyped population as the base population. All marker-based estimators requiring the use of allele frequencies (VR1, WCS and WPCS) were equally biased in this scenario (-0.027 ± 0.003). Interestingly, VR1 appeared to be more sensitive to the definition of allele frequencies than WCS and WPCS both in terms of ρ (0.87 ± 0.028) and α (0.81± 0.033) coefficients, although changes in allele frequencies were small between pedigree founder and genotyped populations. This meant that variance in true HBD explained by VR1 was strongly reduced and that value-dependent bias also increased. Therefore, it is clear that the allele frequencies to be used to construct genomic relationship matrices for genomic evaluations should be the ones estimated in the pedigree founder population and not the ones observed in the genotyped population (VanRaden, 2008). WCS estimator appeared to be robust to the different definitions of allele frequencies both in terms of accuracy (ρ=0.98 ± 0.003) and value-dependent bias (α=0.98± 0.007). On the contrary, estimation accuracy of WPCS was slightly reduced (ρ=0.95 ±0.009) and value-dependent bias increased (α=0.93±0.008) when considering genotyped individuals as the base population. This meant that more weight was given to allele frequencies in the weighting procedure when accounting for pedigree information leading to sub-optimality of weights, especially for individuals with the largest HBD values.

IV. Conclusion

Results of this study showed that using dense panels of markers could significantly improve the accuracy of estimation of genome-wide inbreeding coefficients. Accuracy of marker-based estimators was improved for populations which have already been selected for many generations. In such selected populations, defining pedigree founders as base populations clearly reduced the estimation bias. Including pedigree information into WCS seemed to be useless given the high accuracy of this estimator and the low mean level of inbreeding in simulated populations. Besides, accounting for pedigree information into WPCS rendered the estimator more sensitive to the allele frequency definition.

Ackowlegements. This work was funded by FABA through TEKES grant 433/31/09. Computational facilities were provided by CSC-IT, Espoo (Finland).

References

Daetwyler, H. D., Villanueva, B., Bijma, P., Woolliams, J.A., 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet., 124:369-376

Hayes, B.J., Goddard, M.E. 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33:209-229

Hill and Weir, 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res., 93:47-64.

Keller, M.C., Visscher, P.M., Goddard, M.E. 2011. Quantification of inbreeding due to distant ancestors and its detection using dense SNP data. Genetics, published on-line on June 24th, 2011.

Li, C.C., Horvitz, D.G. 1953. Some methods of estimating the inbreeding coefficients. Am. J. Hum. Genet. 5:107-117.

Meuwissen, T.H.E., Luo, Z. 1992. Computing inbreeding coefficients in large populations. Genet. Sel. Evol. 24:305–313.

Powell, J.E., Visscher, P.M., Goddard, M.E. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800-805

Ritland, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res., 67:175-185.

Sargolzaei, M., Schenkel, F.S. 2009. QMSim: a large scale genome simulator for livestock. Bioinformatics, 25:680-681.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet., 123, 218–223.

Strandén, I., Vuori, K. 2006. Relax2 : pedigree analyses program. In : Proc. of the 8th WCGALP, 13-18 Aug. 2006, Belo Horizonte, MG, Brazil.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.