Commission on Animal Genetics **Session G3.5**

55th Annual Meeting of the European Association for Animal Production
Bled, 5 - 8 September 2004

Investigations on the Selection of Breeds for Conservation of genetic Variance

Jörn Bennewitz*[a], and Theo H. E. Meuwissen[b]

[a]Institute of Animal Breeding and Husbandry, Christian-Albrechts-Universität, D-24098 Kiel, Germany

[b]Institute of Animal and Aquacultural Sciences, Agriculture University of Norway, Box 5052, 1432 Aas, Norway

INTRODUCTION:

On a world-wide level there are roughly 6000 breeds of 30 domestic mammalian and bird species. Around 35% of them are classified as having a high risk of extinction (FAO 2000). A key question is the choice of breeds for conservation programs. Ruane (1999) reported several criteria for this selection, one is the maintenance of the genetic variance, an aspect upon which we will focus exclusively throughout this paper. Pairwise genetic distances and their graphical representation in distance trees are common tools used to assess the genetic uniqueness of a particular breed within a set of breeds. Weitzman (1992) developed an algorithm for the estimation of the genetic diversity using genetic distances. However, as pointed out by Eding & Meuwissen (2001) and Caballero & Toro (2002), the use of this method on a within-species breed level might be inappropriate. Instead of genetic distances, Eding & Meuwissen (2001) and Caballero & Toro (2002) used average kinships between and within breeds for the description of the genetic diversity. Kinship is defined as the probability that two gametes randomly drawn from a population are identical by descent. Following this, the average kinship between two breeds is an estimate of the fraction of alleles that these breeds have in common. Eding & Meuwissen (2001, 2003) described a method that estimates average kinships between breeds using similarities of genetic marker alleles. To prioritise breeds for conservation, Eding et al. (2002) defined a core set that is built by relative contributions of the breeds under consideration in order to minimise the mean kinship in this core set. The core set maximises the variance of a hypothetical quantitative trait that can be found in a hypothetical population obtained from interbreeding the conserved breeds (Eding et al. 2002). The breeds are ranked for their priority for conservation according to their relative contribution to the core set. By way of simulations, Eding & Meuwissen (2003) demonstrated that the accuracy of the estimation of the breed's contribution to the core set from marker estimated kinships is only moderate, indicating that there is scope for improvement.

The aim of this paper is to put forward a conservation criterion that values the differences between breeds more than the core set method of Eding et al. (2002) does. For this purpose an algorithm is introduced that estimates the contribution of breeds to a core set in order to maximise the total additive genetic variance of a hypothetical quantitative trait. The method was compared to the core set method of Eding et al. (2002) using simulated and real data. Additionally, a bootstrap based method is introduced that improves the accuracy of contribution estimates substantially.

MATERIALS AND METHODS

Assume a set S of N breeds with a known average kinship matrix, $\mathbf{M}$, of dimension N x N. The off-diagonal elements of $\mathbf{M}$ are the inbreeding coefficient of putative offspring from the corresponding between breed mating and the diagonal elements of $\mathbf{M}$ are the inbreeding coefficients of putative offspring from within breed mating (Falconer and MacKay 1996). The additive genetic variance of a hypothetical quantitative trait within set S can then be described by

$$\mathrm{var}(u_S) = \frac{1}{N}\sum_{i=1}^{N}\left[(1+\mathrm{M}_i) - 2\overline{\mathrm{M}}\right]$$

where M_i is the within breed kinships of breed i obtained from the diagonal elements of $\mathbf{M}$ and $\overline{\mathrm{M}}$ denotes the mean of all the elements of $\mathbf{M}$. A one is added to M_i to make it comparable to the diagonal elements of the numerator relationship matrix. In order to maximise the total genetic variance, a core set S_{mvt} is formed by relative contributions of the N breeds of S. This core set is termed maximum variance total (MVT) core set method in the following. The contributions of the breeds to the MVT core set are stored in a vector, $\mathbf{c}_{\mathbf{mvt}}$, of dimension N. The total genetic variance conserved by the MVT core set, $\mathrm{Var}_{\mathrm{total}}(\mathrm{S}_{\mathrm{mvt}})$, is then calculated by:

$$\mathrm{Var}_{\mathrm{total}}(\mathrm{S}_{\mathrm{mvt}}) = 1 + \mathbf{c}'_{\mathbf{mvt}}\mathbf{F} - 2\mathbf{c}'_{\mathbf{mvt}}\mathbf{M}\mathbf{c}_{\mathbf{mvt}}$$

with $\mathbf{F}$ being a vector of dimension N that contains the within breed kinship, i.e $\mathbf{F} = \mathrm{diag}(\mathbf{M})$. The solutions for $\mathbf{c}_{\mathbf{mvt}}$ under the restriction $c_{mvt}(i) \geq 0$ and $\sum_{i=1}^{N} c_{mvt}(i) = 1$ is (the one is omitted):

$$\mathbf{c}_{\mathbf{mvt}} = \tfrac{1}{4}\left(\mathbf{M}^{-1}\mathbf{F} - \frac{\mathbf{1}'_{\mathbf{N}}\mathbf{M}^{-1}\mathbf{F} - 4}{\mathbf{1}'_{\mathbf{N}}\mathbf{M}^{-1}\mathbf{1}_{\mathbf{N}}} \cdot \mathbf{M}^{-1}\mathbf{1}_{\mathbf{N}}\right)$$

This core set method is compared to the core set method of Eding et al. (2002), which maximises the variance in a hypothetical population that could be bred from the conserved breeds, i.e. the offspring variance. The Eding et al. (2002) core set method will be termed the maximum variance offspring core set method (MVO). The MVO obtains the relative contribution vector ($\mathbf{c}_{\mathbf{mvo}}$) by minimising the average kinship within the core set, $\bar{f}_{\min}(\mathrm{S}_{\mathrm{mvo}})$, see Eding et al. (2002) for a full description.

Both core set methods are based on average kinships between and within population, stored in the kinship matrix $\mathbf{M}$. Ideally, $\mathbf{M}$ is calculated using pedigree information, but often no appropriate pedigree information is available. In these cases $\mathbf{M}$ can be estimated using molecular marker information (Eding & Meuwissen 2001, 2003, Caballero & Toro 2002). Here a method of Eding & Meuwissen (2003) was used. In brief, an average similarity index between and within breeds was calculated using the marker genotype information. This similarity index is a function of the Alike in State (AIS) probabilities and the Identical by Descent (IBD) probabilities. A weighted log-linear model was set up to separate the IBD probability from the AIS probability, the estimated IBD probabilities were transferred into the kinship matrix $\hat{\mathbf{M}}$. See Eding & Meuwissen (2003) for a full description.

The breed contribution vectors to the MVT and MVO core sets (i.e. $\hat{\mathbf{c}}_{\mathbf{mvt}}$ and $\hat{\mathbf{c}}_{\mathbf{mvo}}$), respectively, were calculated from $\hat{\mathbf{M}}$ and $\hat{\mathbf{F}}$. If breeds showed negative contributions, the most negative contribution was set to zero and the contribution vector was recalculated without the corresponding breed. This was repeated until no further negative contribution estimates were observed. We will refer to this method of contribution vector estimation as the Eding et al. method (EEA), because it is analogous to the idea of Eding et al. (2002). As stated in the introduction, the accuracy of $\hat{\mathbf{c}}_{\mathbf{mvo}}$ is only moderate when estimated by EEA (Eding & Meuwissen 2003). To obtain a higher accuracy of the contribution vectors for both core set methods ($\hat{\mathbf{c}}_{\mathbf{mvt}}$ and $\hat{\mathbf{c}}_{\mathbf{mvo}}$, respectively), a non-parametric bootstrap method was tested as described in the following. A bootstrap sample was generated by performing two sampling steps. In one step, N individuals were sampled with replacement out of the pool of N original individuals within breed. In the second step, L marker loci were sampled with replacement out of the pool of L original marker loci across breeds. Hence, a bootstrap sample consisted of sampled individuals and sampled marker loci. Throughout the study 100 bootstrap samples were generated ($B = 100$). For each bootstrap sample, the average similarities were calculated and the weighted log-linear model was set up to obtain the kinship of the bootstrap sample. The weights were obtained empirically using a bootstrap approach (Efron & Tibshirani 1993). This method of contribution vector estimation is denoted as the weighted bootstrap method (WBM).

To test and to compare the methods outlined above, two series of Monte Carlo simulations were carried out, the number of replicates always being 10. In the first series a phylogenetic situation was simulated over 50 generations. A base breed was simulated that consisted of 50 individuals, the size was kept constant throughout the simulation. For each individual a number of L unlinked genetic marker loci was assumed (L = 10 and 20, respectively) and the alleles were randomly assigned to the individuals. Each next generation was generated by randomly assigning sires and dams from the current generation as parents of the individuals of the next generation. New breeds were randomly generated by base breed fission between generations 10 and 49 by sampling sires and dams randomly as parents of the individuals of the new breed. The effective size of the new breed was randomly chosen from the interval 24/76 and was kept constant throughout the simulation. In total, 15 breeds were simulated for each replicate, one base breed and 14 breeds that were formed by fission from the base breed. During the second series of simulations the phylogenetic structure was fixed. A base breed of size 50 was bred for 50 generations as described above (*L* = 20). From this, four new breeds were generated at generation 10 by fission as outlined above. To obtain different within breed kinships after 50 generations, the effective size of the breed was chosen to be 10, 20, 30 and 40, respectively. In generation 50 the genotypic data of the breeds were used for the estimation of the breed contribution vectors for both core sets by use of the EEA and WBM methods as outlined above. Further, the full pedigree information was recorded during the simulation and this was used to estimate the true average kinships between the breeds using path analysis. From the true average kinship matrix, the true breed contributions were obtained by the two core set algorithm.

The correlation between the estimated and the true contribution vectors and the mean square error of the estimated contributions was calculated, serving as an empirical measure of the ability of EEA and WBM to estimate correct breed core set contributions. For the second series of simulations the individual contributions of the five breeds to the core set, averaged over all replicates, were recorded. This revealed the differences between the MVT and the MVO core set methods regarding the breed contributions as a function of their within breed kinships and regarding the variance conserved, respectively. To compare the two core set methods using a real case, the field data of Eding & Meuwissen (2003) were reanalysed.

RESULTS:

Results from the simulations: The accuracy of the WBM contribution vector is highest in both core set methods. These contribution estimates showed the highest average correlations with the true contributions and on average the lowest mean square error (Tab. I and II). Most total variance in the MVT core set and most offspring variance in the MVO core set is conserved, respectively, when the contribution vectors were estimated by WBM. All breeds contributed to the MVO core set (results from the true contributions, Tab. I), but when the contributions were estimated with EEA, many breeds erroneously showed a zero contribution estimate. This was also reported by Eding et al. (2002). This problem did not occur when the WBM method was applied. In contrast, a substantial fraction of breeds did not contribute to the MVT core set (results from the true contributions, Tab. I), but nearly all breeds showed a contribution estimate greater than zero, when using WBM. However, these estimates were in general only very small.

Table I: Correlation (Corr) between the estimated and the true core set contributions, mean square error (MSE) of contribution estimates and actual conserved total genetic variance (Var_{total}) and offspring genetic variance ($\mathrm{Var}_{offspring}$) as a function of the core set, of the number of simulated loci (*L*) and of the method of contribution estimation.

Core set	*L*	Method	MSE	Corr	n_{zero}[1]	Var_{total}[2]	$\mathrm{Var}_{offspring}$[2]
MVT	10	TRUE	0	1	5.0	1.197	0.926
		EEA	0.057	0.618	7.4	1.176	0.917
		WBM	0.035	0.803	0.1	1.185	0.924
	20	TRUE	0	1	3.9	1.188	0.925
		EEA	0.049	0.641	7.2	1.172	0.918
		WBM	0.032	0.767	0.2	1.179	0.922
MVO	10	TRUE	0	1	0	1.185	0.931
		EEA	0.066	0.589	6.4	1.151	0.916
		WBM	0.038	0.754	0.1	1.169	0.923
	20	TRUE	0	1	0	1.178	0.930
		EEA	0.057	0.653	5.6	1.152	0.918
		WBM	0.036	0.835	0	1.169	0.924

Results from the first series of simulation, averaged over all replicates. [1]n_{zero} number of breeds that showed zero contribution to the core set; [2]Var_{total} and $\mathrm{Var}_{offspring}$, actual conserved total genetic variance and offspring genetic variance estimated by the use of the contribution vectors and the true kinship matrix.

In Table II, the within breed kinships of the five simulated breeds (results from the second series of simulation, fixed phylogenetic situation and fixed and unequal population size) are given as an average over all replicates. The breeds with low kinships contributed more to the MVO core set. This was expected, because this core set method aims to keep the average kinship of the set as small as possible (Eding et al. 2002). In contrast, the true contributions of the breeds with a higher kinship are slightly larger in the MVT core set. This could not be observed from the estimated contributions, probably due to sampling errors.

Table II. Mean square error (MSE) of contribution estimates, actual conserved total genetic variance (Var$_{\text{total}}$) and offspring genetic variance (Var$_{\text{offspring}}$) as functions of the core set and of the method for contribution estimation, and contribution estimates as a function of the within breed kinship (f_i).

Core set	Method	MSE	Var$_{\text{total}}$^{1}	Var$_{\text{offspring}}$^{1}	Core set contributions of breeds of size n				
					Breed 1	Breed 2	Breed 3	Breed 4	Breed 5
					$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
					$f_i = 0.86$	$f_i = 0.64$	$f_i = 0.48$	$f_i = 0.40$	$f_i = 0.33$
MVT	TRUE	0	1.323	0.881	0.221	0.211	0.201	0.190	0.176
	EEA	0.089	1.277	0.873	0.175	0.200	0.173	0.207	0.245
	WBM	0.047	1.312	0.888	0.178	0.205	0.193	0.207	0.217
MVO	TRUE	0	1.288	0.899	0.112	0.152	0.197	0.243	0.296
	EEA	0.090	1.206	0.878	0.070	0.088	0.187	0.277	0.378
	WBM	0.066	1.247	0.889	0.088	0.126	0.182	0.268	0.326

Results from the second series of simulation, averaged over all replicates. ^{1}Var$_{\text{total}}$ and Var$_{\text{offspring}}$, actual conserved total genetic variance and offspring genetic variance estimated by the use of the contribution vectors and the true kinship matrix.

Results from the field data: From the estimated kinship matrix from the field data (Tab. III) it can be seen that the Heck breed shows the highest level of within breed kinship followed by Galloway, Dutch Black Belted and Improved Red Pied. The Holstein Friesian and Limousine breeds contributed considerably to the MVT core set, despite their low within breed kinships. This is due to their comparatively low mean kinship, and consequently these two breeds are also the major contributors to the MVO core set (Tab. III and IV). The third breed of the top three contributors was different in the two core sets. The Heck breed showed the largest contribution to the MVT core set but only a small contribution to the MVO core set. The high MVT contribution of the Heck breed is due to its high within breed kinship and its low mean kinship. The opposite was true for the Dutch Red Pied (low contribution to MVT and high contribution to MVO core set).

Table III. Estimated kinships within and between the 10 cattle breeds using a weighted log-linear model.

Breed	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Belgian Blue	**0.104**									
(2) Dutch Red Pied	0.055	**0.088**								
(3) Dutch Black Belted	0.076	0.050	**0.179**							
(4) Limousine	0.027	0.026	0.036	**0.089**						
(5) Holstein Friesian	0.049	0.027	0.048	0.001	**0.102**					
(6) Galloway	0.038	0.030	0.088	0.045	0.041	**0.213**				
(7) Dutch Friesian	0.062	0.040	0.086	0.027	0.044	0.061	**0.099**			
(8) Improved Red Pied	0.049	0.076	0.063	0.033	0.041	0.058	0.052	**0.172**		
(9) Blonde d'Aquitaine	0.043	0.038	0.038	0.052	0.025	0.036	0.042	0.046	**0.085**	
(10) Heck	0.059	0.032	0.055	0.022	0.000	0.029	0.056	0.006	0.059	**0.310**
Mean (within and between)	**0.056**	**0.046**	**0.072**	**0.036**	**0.038**	**0.064**	**0.057**	**0.060**	**0.046**	**0.063**

DISCUSSION:

This study introduces the MVT core set method for prioritising breeds for conservation in order to optimally maintain total additive genetic variance of a hypothetical quantitative trait. The differences between the two core sets become obvious when focussing on the individual breed contributions. From the optimisation term it follows that the MVT core set prioritises breeds with a high within breed kinship and a low average between breed kinship. In contrast, the MVO core set method favours breeds that show a low average kinship both within and between breeds. This was shown by the results of the second series of simulations (Tab. II). Considering the breed contributions to the two core sets from all simulated configurations and all replicates, the Spearman rank correlation was around 0.7 (averaged over all replicates), indicating that the two core sets prioritise different breeds only to some extent.

Table IV. MVT and MVO core set contribution vectors for the 10 cattle breeds.

	MVT core set		MVO core set	
	EEA	Bootstrap	EEA	Bootstrap
Belgian Blue	0	0.002	0	0.032
Dutch Red Pied	0	0.028	0.207	0.115
Dutch Black Belted	0.011	0.049	0	0.021
Limousine	0.186	0.169	0.359	0.304
Holstein Friesian	0.242	0.209	0.321	0.229
Galloway	0.135	0.139	0	0.051
Dutch Friesian	0	0.004	0.037	0.086
Improved Red Pied	0.169	0.133	0	0.049
Blonde d'Aquitaine	0	0	0	0.039
Heck	0.257	0.267	0.075	0.072

The differences between the two core set methods are demonstrated by the following simple hypothetical example. Consider two unrelated breeds A and B, and a breed AB are available for conservation, where AB is obtained by crossbreeding of A and B. The MVO method will not give extra value to A and B when AB is already conserved, whilst the MVT method would give extra value to A and B. The latter is because, although A and B contain no genes that are not present in AB, they may contain combinations of genes that are not present in AB, because in AB the genes of A and B are mixed up. Thus, it may be easier to find particular combinations of genes when A and B are conserved instead of AB, i.e. the genetic variance is more accessible in a set containing A and b compared to a set containing only AB. In terms of allele frequencies, the MVO core set avoids extreme frequencies and therefore, maximises the possible directions of selection. The MVT core set prioritises breeds with more extreme frequencies and thus prioritises breeds that show different combinations of genotypes. Hence, the MVT core set method aims to conserve breeds that show large differences in the respective population mean of a hypothetical quantitative trait. This makes the MVT core set method attractive, because the efficiency of upgrading a breed by introducing genetics from another breed is a function of the difference in the respective population means. The MVT core set thus enables a faster reaction on putative changed conditions compared to the MVO core set.

The conceptual differences between the two core set methods on the one hand and the Weitzman algorithm on the other hand is that both core set methods do account for the within breed variance and do account for migration between breeds. The ignorance of these two aspects was seen as the main point of criticism of the application of the Weitzman algorithm on a within-species breed level (Eding & Meuwissen 2001, 2003, Caballero & Toro 2002), but in the conservation of farm animal genetic resources we are acting on this level. Nevertheless, Eding et al. (2002) showed that the nice mathematical and biological properties of the Weitzman algorithm (Thaon d' Arnoldi et al. 1998) also hold in the MVO core set. The MVT core set method fulfils also the Weitzman criteria.

CONCLUSION:

The introduced core set method (MVT) prioritises breeds for conservation by maximising the total genetic variance for a hypothetical quantitative trait in the core set. It was shown that the numeric results of this core set approach and of the MVO core set approach of Eding et al. (2002) are to some extent similar. The differences were most clearly shown by the results of the field data analysis. The MVT core set approach suggests the conservation of breeds that show comparatively large differences in the respective population mean of a hypothetical quantitative trait. This maximises the speed of achieving selection response for a putative changed breeding objective, which makes the MVT core set method attractive. For the estimation of the core set contribution vectors from molecular marker data we recommend the use of the weighted bootstrap approach (WBM), because this method produced the most accurate estimates, regardless of the core set method.

ACKNOWLEDGEMENTS:

The authors thank the Van Haeringen Laboratory in Wageningen, the Netherlands, for the permission to use the field data and Herwin Eding for preparing them. This study was prepared while J. B. was on leave at the Agriculture University of Norway in Aas. He was supported by a grant from the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD).

REFERENCES:

Caballero A. & Toro M. A. (2002). Analysis of genetic diversity for the management of conserved subdivided populations, Conserv. Genet. **3,** 289-299.

Eding J. H. & Meuwissen T. H. E. (2001). Marker based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics **118,** 141-159.

Eding J. H., Crooijmans R. P. M. A., Groenen M. A. M. & Meuwissen T. H. E. (2002). Assessing the contribution of breeds to genetic diversity in conservation schemes. Genetics, Selection, Evolution **34,** 613-633.

Eding J. H. & Meuwissen T. H. E. (2003). Linear methods to estimate kinships from genetic marker data for the construction of core sets in genetic conservation schemes. Journal of Animal Breeding and Genetics **120,** 289-302.

Efron B. & Tibshirani R. B. (1993). An introduction to the bootstrap. Chapman and Hall, New York.

Falconer D. S. & MacKay T. F. C. (1996). Introduction to quantitative genetics, Longman, London.

FAO (2000). World watch list for domestic animal diversity, third edition, Rome, http://www.fao.org/dad-is/.

Ruane J. (1999). A critical review of the value of genetic distances in conservation of animal genetic resources. Journal of Animal Breeding and Genetics **116,** 317-323.

Thaon d'Arnoldi C., Foulley J.-L. & Ollivier L. (1998). An overview of the Weitzman approach to diversity. Genetics, Selection, Evolution **30,** 149-161.

Weitzman M. L. (1992). On diversity. The Quarterly Journal of Economics **107,** 363-405.