STELLINGEN

1. De genetische diversiteit aanwezig in een set van populaties, kan bepaald worden met behulp van de gemiddelde bloedverwantschappen tussen en binnen populaties (*Dit proefschrift*).

2. Vanwege de versnelde erosie van de genetische diversiteit binnen een bedreigde populatie als gevolg van de kleine populatieomvang, leidt het verlies van een bedreigd ras over het algemeen tot een gering verlies aan genetische diversiteit (*dit proefschrift*).

3. Het gebruik van genetische afstanden als maat voor genetische diversiteit leidt tot behoud van de meest ingeteelde rassen (*dit proefschrift*).

4. Een ras is een ras als genoeg mensen zeggen dat het een ras is (*Hammond, pers. med.*).

5. Een van de meest waardevolle inzichten uit diverse genoomprojecten (*C. Elegans, Drosophila* of het Human Genome Project) is dat zelfs moleculaire genetici niet onder de toepassing van wiskundige statistiek uitkomen (*B Walsh, 2001*)

6. Als de MKZ-crisis iets heeft aangetoond, dan is het dit: Een model is ook maar een standpunt.

7. Een zekere mate van gene flow vermindert de effecten van inteelt. In dat opzicht is de achterdocht voor 'import' in plaatselijke dorpsgemeenschappen contraproductief.

8. *Music calms the savage mind.* Het nummer 'Break stuff' van Limp Bizkit heeft definitief afgerekend met dit misverstand.

Stellingen behorend bij het proefschrift
Conservation of genetic diversity: assessing genetic variation using marker estimated kinships.
Herwin Eding
Wageningen Universiteit, 18 januari 2002

PROPOSITIONS

1. The genetic diversity present in a set of populations can be assessed using mean kinships between and within populations (*this thesis*)

2. The small population size of a population at risk causes accelerated erosion of the genetic diversity within a population at risk of extinction. Hence, the loss of a population at risk usually leads to a small loss in genetic diversity (*this thesis*).

3. Using genetic distances as a measure of genetic diversity leads to the conservation of the most inbred populations (*this thesis*).

4. A breed is a breed if enough people say it is (*Hammond, pers. comm..*).

5. One of the most valuable insights form the various genome projects (*C. Elegans*, *Drosophila* or the Human Genome Project) is that even molecular geneticists will have to start to use statistical mathematics (*Walsh, 2001*).

6. If the Foot and Mouth crisis in the Netherlands has shown one thing, it is this: A model is just another opinion.

7. A certain amount of gene flow lessens the effects of inbreeding. In that light the suspicion with which newcomers are regarded in small village communities is counter-productive.

8. *Music calms the savage mind.* The song 'Break Stuff' by Limp Bizkit has put a definitive end to this misunderstanding.

Propositions accompanying the doctoral thesis
Conservation of genetic diversity: assessing genetic variation using marker estimated kinships.
Herwin Eding
Wageningen University, 18 January 2002

Conservation of Genetic Diversity

Assessing Genetic Variation Using Marker Estimated Kinships

Promotor: Prof. dr. ir. E.W Brascamp
Hoogleraar Fokkerij en Toegepaste Genetica
Wageningen Universiteit

Co-promotor: Dr. ir. T.H.E. Meuwissen
Senior Onderzoeker divisie Dier en Omgeving
ID-Lelystad

Samenstelling promotiecommissie:
Dr. J.A. Woolliams (Roslin Institute, Engeland)
Prof. dr. R.F. Hoekstra (Wageningen Universiteit)
Dr. ir. L. Visser (Plant Research International, Wageningen)
Prof. dr. H.H.T. Prins (Wageningen Universiteit)

Conservation of Genetic Diversity

Assessing Genetic Variation Using Marker Estimated Kinships

Herwin Eding

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
prof. dr. ir. L. Speelman
in het openbaar te verdedigen
op vrijdag 18 januari 2002
des middags te half twee in de Aula

Conservation of genetic diversity: assessing genetic variation using marker estimated kinships.

Eding, Herwin

ISBN 90-5808-556-2

Eding, H. Conservation of genetic diversity: assessing genetic variation using marker estimated kinships. This dissertation focuses on assessing genetic diversity in a quantitative way through the use of Malecots coefficients of kinship. Kinships between and within populations and individuals can be estimated using microsatellite marker genes that are assumed to be selectively neutral.

Genetic diversity is estimated from such Marker Estimated Kinships (MEK) by (1 – average (MEK)), where genetic diversity of a set of breeds is defined as the maximum genetic variation in a population that can be bred from this set of breeds. The concept of core sets is applied to livestock genetic diversity and a new measure of genetic diversity present in a set of breeds, based on the mean kinship within a core set, is developed. Log-linear (mixed) models can be used to simultaneously estimate kinships and the probability for alleles alike in state (AIS). Error variance of the kinship estimates may lead to populations that have incorrectly received a null-contribution. An analysis of a data set concerning African cattle populations, using the developed methods is described. Effects of conservation by breed type or regional versus continental conservation are examined, in terms of efficiency of conservation and changes of priorities of breeds.

Ph.D. thesis, Animal Breeding and Genetics Group, Department of Animal Sciences, Wageningen University, P.O. Box 338, 6700 AH Wageningen, The Netherlands.

VOORWOORD

Dit proefschrift is the neerslag van vier jaar onderzoek gedaan bij het instituut ID-Lelystad en de Vakgroep Fokkerij en Genetica van de Landbouw Universiteit Wageningen. 'Een proefschrift schrijf je niet alleen', zo luidt het cliche. Vandaar dat ik op deze plaats toch een aantal personen die, materieel en immaterieel, hebben bijgedragen aan het tot stand komen van dit proefschrift.

Theo, jouw bijdrage aan dit proefschrift is onschatbaar van waarde. Zonder jouw inzicht, begeleiding, geduld en vertrouwen was dit proefschrift er niet geweest.

John, even though we met only occasionally, the discussions we had regarding the whole concept of livestock core sets proved to be invaluable, as did your encouragement for the approach we'd taken.

Professor en Ab, jullie kritische en verstandige beoordeling van de artikelen en het manuscript hebben dit proefschrift zoveel beter gemaakt.

Ed and Olivier, thank you for sharing data and knowledge, as well as the hospitality you and everybody at ILRI have shown me. Especially your enthusiasm regarding the first results from analysis of the 'Africa data' I showed you both convinced me we were on to something after all.

Everybody participating in the EU concerted action workshops in Lelystad, Toulouse and Edinburgh. You all provided me with an excellent start of my project. I learned a lot of all the differing views you expressed.

Pap en Mam, jullie grenzeloze liefde en vertrouwen in mij hebben mij op moeilijke moementen op de been gehouden. Dit proefschrift is voor jullie.

Jos en Rik, jullie vermogen tot relativeren hebben mij met de voeten stevig aan de grond gehouden. Zeker op die momenten dat ik mezelf te serieus dreigde te nemen.

Anna en Jack, jullie waren geweldige kamergenoten. Altijd bereid voor een praatje of een goede discussie. En erg geduldig of maandagochtenden.

And last but by no means least, Wilma en aanhang. Y'all know why.

Lelystad, november 2001

sDg

CONTENTS

Chapter 1

GENERAL INTRODUCTION

Genetic variation in nature can be observed in the existence of different species of plant and animals. Within species populations are generally be divided in breeds. Genetic variation in livestock breeds is most obvious in phenotypic differences. These differences range from coat colour and conformation traits to production traits and adaptation to the environment in which breeds are kept. Genetic variation between breeds can have a number of causes. Adaptation to local circumstances are the main force behind breed differentiation in for instance Africa, whereas the differences between breeds in the Western world is also caused by herd books and specific selection of animals with respect to breeding goals associated with herd books (Oldenbroek, 1999).

Genetic diversity can be observed both within and between breeds or populations. However, there is a trend that high producing breeds or strains are replacing indigenous, locally adapted breeds, which subsequently decline in numbers and sometimes become extinct. In the third edition of the World Watch List, FAO states that '32% of the recorded animal genetic resources globally are at high risk of loss' (Scherf, 2000). As a consequence the between breed variation decreases and traits and genotypes, possibly of use now or in the future, are lost.

The loss of genetic variation within and between breeds is a negative trend, not only from the perspective of culture, but also with regard to utility. Traits, genotypes and alleles with possible economic interest risk being lost. Within breeds high rates of loss of genetic variation leads to decreased fitness through inbreeding depression. Furthermore, breeds are exposed to a greater loss of alleles and haplotypes, as a consequence of small effective population sizes or, equivalently, high rates of inbreeding (Falconer and MacKay, 1996). Continued loss of within breed genetic variation also diminishes the possibility of genetic improvement of breeds.

Diversity in Animal Genetic Resources

Within species of livestock genetic diversity is most obvious in differences between breeds. Breeds are defined as populations within a species of which the members can be determined by a set of characteristics particular to the breed (FAO, 1998). This definition assumes that there is a clear boundary between expression of characteristics, or traits, between populations or breeds. In Europe a situation of (relative) isolation of breeds from others exists only after the establishment of herd books, some 200 years ago (Ruane, 1999). In other regions, on the African continent for instance, such a clear definition of breeds is not always possible, due to widespread crossing between populations. Assigning animals to breeds in these regions is subjective and often questionable (Scherf, 2000).

Definition of genetic diversity in terms variation in traits or genotypes removes the need for clearly defined breeds. Populations, whether they are clearly defined breeds or sub-populations of a less clearly defined livestock population, can be assessed more objectively with respect to the variation in genotypes and traits.

Assigning priorities

Conservation efforts should be as efficient as possible, securing a maximum amount of genetic diversity given limited resources. To this end, breeds at risk need to be evaluated in terms of the amount of genetic diversity they contribute. The manner in which this is evaluated, is very much dependent on the rationale for conservation (Ruane, 1999). The most obvious criterion is the degree of endangerment of a breed. The priority given to a breed at risk can be based on several additional criteria: 1) adaptation to specific environments, 2) possession of traits of current or future economic importance, 3) possession of unique traits, that may be of scientific interest, 4) genetic uniqueness and 5) cultural or historic value. Note that all of these criteria, except 5) are based on genetic considerations, although cultural or historic value could be a result of considerations falling under criteria 1) to 4).

With the availability of relatively easy to use molecular genetic techniques, such as genotyping of microsatellite marker genes and in the absence of reliable information on relations between breeds (such as pedigree records), overall genetic diversity between breeds is mostly studied using genetic distances (Ruane, 1999). Genetic distances express the differences between populations either in terms of numbers of mutations or in terms of differences in allele frequencies or genetic drift. Breed formation occurred rather recent on the evolutionary scale. For this reason genetic diversity between populations is usually quantified using genetic distances based on genetic drift only, ignoring the effect of mutation. Within a breed diversity is usually expressed in terms directly related to the (rate of) inbreeding within the breed, such as heterozygosity, effective population size, effective number of alleles per locus or Wright's F-statistics, usually also calculated from allele-frequencies of microsatellite marker genes.

To evaluate a breed correctly with respect to genetic diversity, both the within and between breed genetic diversity need to be accounted for. Otherwise, the use of genetic distance measures to assess genetic diversity can lead to undesirable results, as we will argue in Chapter 2.

Conservation Methods

Conservation efforts are generally divided into two classes: *In-situ* and *ex-situ* conservation. *In-situ* conservation is the conservation of a breed in its region of origin and kept in a production system for which the breed was developed; *ex-situ* conservation conserves a breed outside of its production system of origin.

It is generally accepted that *in-situ* conservation is the most viable option in the long term. When a breed of livestock is productive economically, farmers will be more interested in keeping that breed. Therefore, *in-situ* conservation often involves a scheme of niche marketing of specialised products for which the breed in question supplies the raw material. This conservation strategy has been applied successfully in a number of cases (Gandini and Oldenbroek, 1999). For instance the Reggiana breed in Italy is used for the production of a brand of Parmigiano Reggiano cheese, marketed as the 'original' Parmigiano Reggiano cheese, which is sold at a higher price then the common Parmigiano Reggiano. The recovery of the Reggiano breed (from 500 in the 1980's to 1200 in 1998) is attributed to this operation. *In-situ* conservation must be regarded as the preferred situation. The breed is kept in its natural environment to which it is adapted and continues to evolve. Even when a breed is conserved *ex-situ*, attempts should be made to establish a breed *in-situ*, such that the breed keeps evolving. However, the current status of a breed, in terms of numbers of breeding animals can be such that *in-situ* conservation is not (yet) an option, because the breed might be vulnerable to the effects of random drift and inbreeding.

Ex-situ conservation means keeping conserved breeds outside their native environment in protected surroundings, for instance in zoos or museum farms. However, there is a more extreme form of *ex-situ* conservation: gene banks. In gene banks genetic material is stored in cryogenic conditions. The genetic material is usually semen, but also embryos, ovae or somatic cells can be stored in gene banks.

Core Sets

The concept of core sets was first introduced in plant breeding (Frankel and Brown, 1984). In its original form a core set is a sub-set of breeds or strains in a gene bank, chosen in such a way that the amount of genetic 'overlap' is minimised. This set is the 'core' of the gene bank, representing the genetic diversity contained in a gene bank in an efficient number of breeds or strains. In Chapters 3, 4 and 5 the concept of core sets is developed and applied to livestock populations in an attempt to categorise populations according to their importance for genetic diversity.

The defining character of a core set is the minimisation of genetic overlap, or the maximisation of genetic diversity in the core set. The genetic overlap, or genetic similarity between individuals or populations, can be described using a coefficient of kinship. Malecot (1948) defined a coefficient of kinship f between individuals as the probability that two randomly drawn alleles from two individuals are identical by descent. The coefficient of kinship describes genetic diversity both in terms of alleles (Caballero and Toro, 2000) and quantitative genetic variation in a general way, without requiring detailed knowledge on the genetics involved or the mean and variances for any trait that is to be conserved. The genetic variance in a random breeding population is proportional to $(1-\bar{f})$ (Falconer and Mackay, 1996), hence if we minimise the average kinship in the core set, $\bar{f}$, we will maximise the genetic diversity of a population that we breed from the core set.

If the average kinships between and within populations are known, we can calculate the average kinship in a core set given the contribution (as fractions of total resources) of each breed to the core set. In the case of a core set, we have to choose these contributions such that the average kinship in the core set is minimised. Thus, the question is not whether or not a breed is included in the core set, but how much it contributes to the core set. By calculating these theoretical contributions to a core set, the populations under study can be ranked according to their genetic uniqueness, which may help in identifying breeds or populations at risk as being important to the conservation of genetic variation.

In this thesis, we develop a method that is capable of ranking populations according to their contribution to overall genetic diversity. Both the concept of core sets and Malécots coefficient of kinship are central to this method. This method will be able to account for both within and between population (or individual) genetic diversity. Furthermore, we propose a definition of overall genetic diversity, which is the maximum quantitative genetic variance present in a population in Hardy-Weinberg equilibrium derived from the populations present in the core set.

Outline of the thesis

Chapter 2 deals with estimating kinships between and within populations and individuals using microsatellite marker genes that are assumed to be selectively neutral. The argument for the use of kinships in genetic diversity studies as opposed to the use of genetic distances and related measures is also developed in this chapter. **Chapter 3** introduces the concept of core sets applied to livestock genetic diversity and a new measure of genetic diversity present in a set of breeds, based on the mean kinship within a core set, is developed. This is subsequently demonstrated in an example of Dutch poultry populations. **Chapter 4** compares a number of methods to simultaneously estimate kinships and the probability for alleles alike in state (AIS) with regard to their accuracy and robustness, especially when the kinship matrix is not properly constructed due to error variance of the kinship estimates. The latter leads to populations that have incorrectly received a null-contribution. The methods are compared using simulated data and illustrated using a small example involving Dutch populations of cattle. **Chapter 5** describes the analysis of a data set concerning African cattle populations, using the methods developed in the previous chapters. Effects of conservation by breed type or regional versus continental conservation are examined, in terms of efficiency of conservation and changes of priorities of breeds. Finally in **Chapter 6** the results described in the previous chapters are discussed. Special attention is paid to the relevance of the core set method in planning conservation efforts.

Chapter 2

MARKER BASED ESTIMATES OF BETWEEN AND WITHIN POPULATION KINSHIPS FOR THE CONSERVATION OF GENETIC DIVERSITY

Eding, Herwin[*)] and Theo H.E. Meuwissen

Institute for Animal Science and Health, Box 65, 8200 AB Lelystad, The Netherlands

Journal of Animal Breeding and Genetics **118**, 141-159

[*)] Institute for Animal Science and Health

Box 65, 8200 AB Lelystad, The Netherlands

Tel: +31-(0)320-238238; Fax +31-(0)320-238050

E-mail: j.h.eding@id.wag-ur.nl

ABSTRACT

In this paper coefficients of kinship between and within populations are proposed as a tool to assess genetic diversity for conservation of genetic variation. However, pedigree based kinships are often not available, especially between populations. In this paper a method of estimation of kinship from genetic marker data is applied to simulated data from random breeding populations to study the suitability of this method for livestock conservation plans. Average coefficients of kinship between populations can be estimated with low Mean Square Error of Prediction, although a bias will occur from alleles alike in state in the founder population. The bias is similar for all populations, so the ranking of populations will not be affected. Possible ways of diminishing this bias are discussed. The estimation of kinships between individuals is imprecise unless the number of marker loci is large (>200). However, it allows distinction between highly related animals (fullsibs, halfsibs and equivalent relations) and animals that are not directly related if about 30 - 50 polymorphic marker genes are used. The marker based estimates of kinship coefficients yielded higher correlations than genetic distance measures with pedigree based kinships and thus to this measure of genetic diversity, although correlations were high overall. The relation between coefficients of kinship and genetic distances are discussed. Kinship based diversity measures conserve the founder population allele frequencies, whereas genetic distances will conserve populations with extreme allele frequencies. Marker based kinship estimates can be used for the selection of breeds and individuals as contributors to a genetic conservation program.

Key words: Genetic diversity, Kinship, Coancestry, Genetic Distance, Genetic markers

Chapter 2

INTRODUCTION

The importance of conservation of genetic diversity in livestock has received widespread attention in recent years. Food security (Hammond, 1994) and sustainable livestock production (de Wit *et al.*, 1995) are the main reasons. A major problem with regard to conservation efforts is the assessment of genetic diversity within and between populations.

Many studies have described genetic diversity of several populations within species based on genetic distances (Eding and Laval, 1999; Moazami-Goudarzi *et al.*, 1997; Ruane, 1999; Thaon d'Arnoldi *et al.*, 1998). On the other hand are measures, which are based on some form of genetic similarity index (Lynch 1988). These similarity indices can be adjusted to estimate relatedness between individuals within a population (Li *et al.*, 1993; Lynch and Ritland, 1999).

As a third option, minimizing the mean kinship between animals within a population selected for conservation purposes has been suggested as a general approach to conservation of genetic diversity (Frankham, 1994; Haig *et al.*, 1990; Johnston and Lacy, 1995; Toro *et al.*, 1998; Zheng *et al.* 1997). The coefficient of kinship is defined as the probability that two alleles randomly sampled from the same locus in two individuals are Identical By Descent (IBD, Malecot, 1948). Therefore, if we minimize the mean kinship in a set of individuals we will minimize duplicates of alleles descending from the same ancestor. Furthermore, this parameter is on average valid for the entire genome and is not limited to the loci under study. Kinships are calculated from pedigree records using for instance path analysis (Falconer and MacKay, 1996). The need for pedigree records means that in situations where they do not exist (poor administration or between breed analysis), pedigree based kinships can not be used as a measure of genetic diversity. In plant breeding a method was developed to estimate kinship between individuals and populations using marker gene data (Bernardo, 1993). This method consists of a similarity index S between individuals based on the concept of identity by descent.

The main focus of this paper will be the question to what extent missing pedigree data can be substituted by kinship estimates based on marker information in conservation decision making. First, we will study the behaviour of kinship (actual pedigree based and estimated from a similarity index) between and within (sub) populations over time. Next we will investigate by simulation how well kinships can be predicted by a similarity index using marker gene information. As a secondary aim we will investigate the relationship between coefficients of kinship and marker based estimates of genetic diversity, specifically genetic distances and similarity indices. We will argue that the similarity index used in this paper has the most consistent relation with both actual kinship coefficients and genetic diversity.

METHODS

Similarity index

The similarity index that is used is based on the concept of identity by descent (IBD, Lynch, 1988; Jacquard, 1983). The scoring rules can be written mathematically as:

(1) $$S_{xy,l} = \tfrac{1}{4}\left[I_{11} + I_{12} + I_{21} + I_{22}\right]$$

where I_{ij} is an indicator variable which is 1 when allele i on locus *l* in the first individual and allele j on the same locus in the second individual are identical, otherwise it is 0. Note that $S_{xy,l}$ can have four possible values: 1, ½ and ¼ and 0. When three indicators have value 1 the fourth will necessarily be 1 also, eliminating the possibility of a value of ¾. Under the assumption of founder alleles, S_{xy} averaged over multiple loci is an estimator of the coefficient of kinship f_{xy} (*i.e.* probability of IBD). Using Jacquards (1974) identity coefficients, Appendix 2.A shows S_{xy} is an unbiased estimator of kinship when founder alleles are unique.

When founder alleles are not unique, the pairwise similarity between two individuals is determined not only by the probability that two randomly sampled alleles are IBD, but also by the probability that they are alike in state (AIS). Let f_{ij} be the probability two alleles are IBD and *s* the probability that two alleles are AIS. Then the expected value of the similarity score for a locus *l* between two individuals *i* and *j* becomes (Lynch, 1988):

(2) $$\mathrm{E}(S_{ij}) = f_{ij} + (1 - f_{ij})s,$$

i.e. S is upwardly biased by *s*. We assume there is a founder population from which all populations descend. All population are therefore related at least through this founder population. We further assume all relations in the founder population are zero, i.e.

f_{ij}=f_{ff}=0 The probability of two alleles being AIS, but not IBD is: $s = S_{ff} = \sum q_k^2$, where S_{ff} is the similarity in the founder population and q_k is the frequency of the k-th allele in the founder population. Note that s is only defined by the founder population, in which all relations are assumed to be zero.

Rearrangement of equation (2) gives:

$$\text{(3)} \qquad \hat{f}_{ij} = \frac{S_{ij} - s}{1 - s} \qquad \text{(Lynch, 1988)}$$

where s can be of assumed value or be estimated per locus from founder population data.

The estimate of f_{ij} between two individuals i and j can be obtained through averaging over L analysed loci. If however the probability s differs per locus, we may use the inverse of the variance of the estimate of f_{ij} as weights (see Appendix 2.B for derivation):

$$\text{(4)} \qquad \hat{f}_{ij} = \frac{\sum_{l=1}^{L} \hat{f}_{ij,l} \left(\frac{1 - s_l}{s_l + f_{ij,l}(1 - 2s_l) - f_{ij,l}^2 (1 - s_l)} \right)}{\sum_{l=1}^{L} \left(\frac{1 - s_l}{s_l + f_{ij,l}(1 - 2s_l) - f_{ij,l}^2 (1 - s_l)} \right)}$$

Average similarities between and within populations

On the level of populations the average pairwise similarity between population x and y for a locus with K alleles can be expressed in terms of allele frequencies as:

$$\text{(5)} \qquad \bar{S}_{xy} = \sum_k p_{xk} p_{yk}$$

where p_{xk} is the frequency of the k-th allele in population x. This expression has been used many times in the field of conservation genetics. Applied within a population (x=y) it expresses homozygosity under Hardy-Weinberg equilibrium. Its complement, heterozygosity has been used as a measure of genetic diversity (Toro *et al.*, 1998). Moreover, the coefficient of inbreeding has been proposed as a measure of genetic diversity (notably F_{ST}) and is defined as the excess of homozygosity relative to Hardy-Weinberg equilibrium genotype frequencies. The reciprocal of expression (5) was used by Kimura (Crow and Kimura, 1970) to estimate effective number of alleles and in Nei's standard distance D expression (5) appears in the numerator of the coefficient of identity.

Simulation

The behaviour of similarity index S and the estimates of f_{ij} were tested by simulation. A base population was simulated, which developed into 5 separate populations according to the phylogeny given in Figure 2.1. Divergence was obtained by doubling the number of offspring in the generation in which fission occurred to avoid bottleneck effects. The population of each line consisted of 50 individuals with equal numbers of males and females. Each round of mating produced again 25 males and 25 females. Parents of each offspring were sampled at random from the preceding generation. Generations were discrete. For each individual a genome was simulated consisting of 200 autosomal, unlinked selectively neutral loci. Every generation information on all alleles of every individual was recorded. Simultaneously a pedigree file was written containing all pedigree information. For reasons of simplicity, linkage was ignored in this study, as were selection, mutation and migration, such that the relationship between the similarity and the actual kinship was not affected by these effects.

The size of each population was limited to a maximum 50 breeding individuals, to save on computer time. The length and structure of the history was variable. In this paper results will be presented as a function of t/N_e, since genetic drift depends on t/N_e rather than only N_e or time t (Crow and Kimura, 1970).

The simulation was run for founder alleles (all founder animals have a unique set of alleles per locus) and for founder populations with a limited number of alleles per locus (2, 5, 10 and 20, resp.), with approximately equal allele frequencies in the founder population. Before the first population fission, the founder population was allowed to breed for a number of generations to generate a realistic distribution of frequencies.

Over generations a number of statistics were calculated: average pairwise f between and within populations calculated from the full pedigree (f_{ij}, this statistic was taken to be the 'true' value of genetic similarity and was used to test the other statistics against), Marker Estimated Kinships (MEK) from average pairwise similarities (S_{ij}) and average population similarities from allele frequencies (S_{xy}), Nei's standard distance D (Nei, 1972), Reynold's distance D_R (Reynolds, 1983) and F_{ST} based on marker gene information (Nagylaki, 1998).

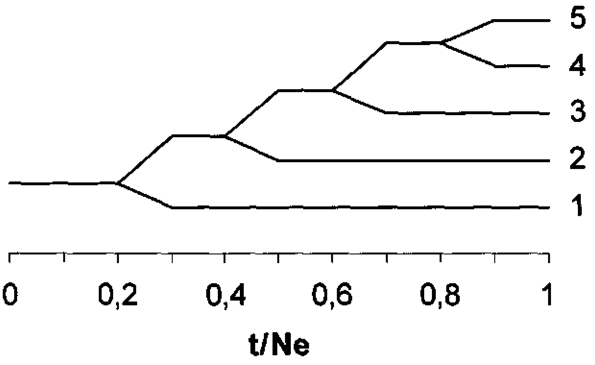


Figure 2.1 General structure of the phylogenetic tree used in the simulation for the case of 5 populations.

RESULTS

Actual average kinships between populations

Figure 2.2 shows scatter plots of the development of the average actual kinship between and within populations for a single replicate. Figure 2.2a shows *f* calculated from the recorded pedigree and Figure 2.2b MEK from the 200 loci, where the number of alleles per locus was 2 ('worst case'). Correction for alleles AIS, was done by setting *s* to 0.5, the expected probability of AIS. Data on all 200 loci was used to eliminate random drift effects. This was done to verify MEK does behave according to actual kinships. The population has a phylogeny as given in Figure 2.1. In the figure we can distinguish a main line (×), increasing with time. This line corresponds to the within population average actual kinship. At intervals of 0.2N$_e$ generations a horizontal line separates from the main line. These lines (□, Δ, ◊, o) show the average actual kinship between one population and the cluster of populations that are the descendants of this population, and their value is equal to the average population kinship within the population just prior to fission. The lowermost of these lines in the figure (at f_{ij} = 0.098; □) corresponds to the kinship between population 1 (the oldest population) and the cluster of populations (2, 3, 4, 5). The next line (at f_{ij} = 0.189; Δ) depicts the kinship between population 2 and the cluster (3,4,5), the third line (◊) corresponds to the kinship between 3 and (4,5) and the last line (o) is the average actual kinship between populations 4

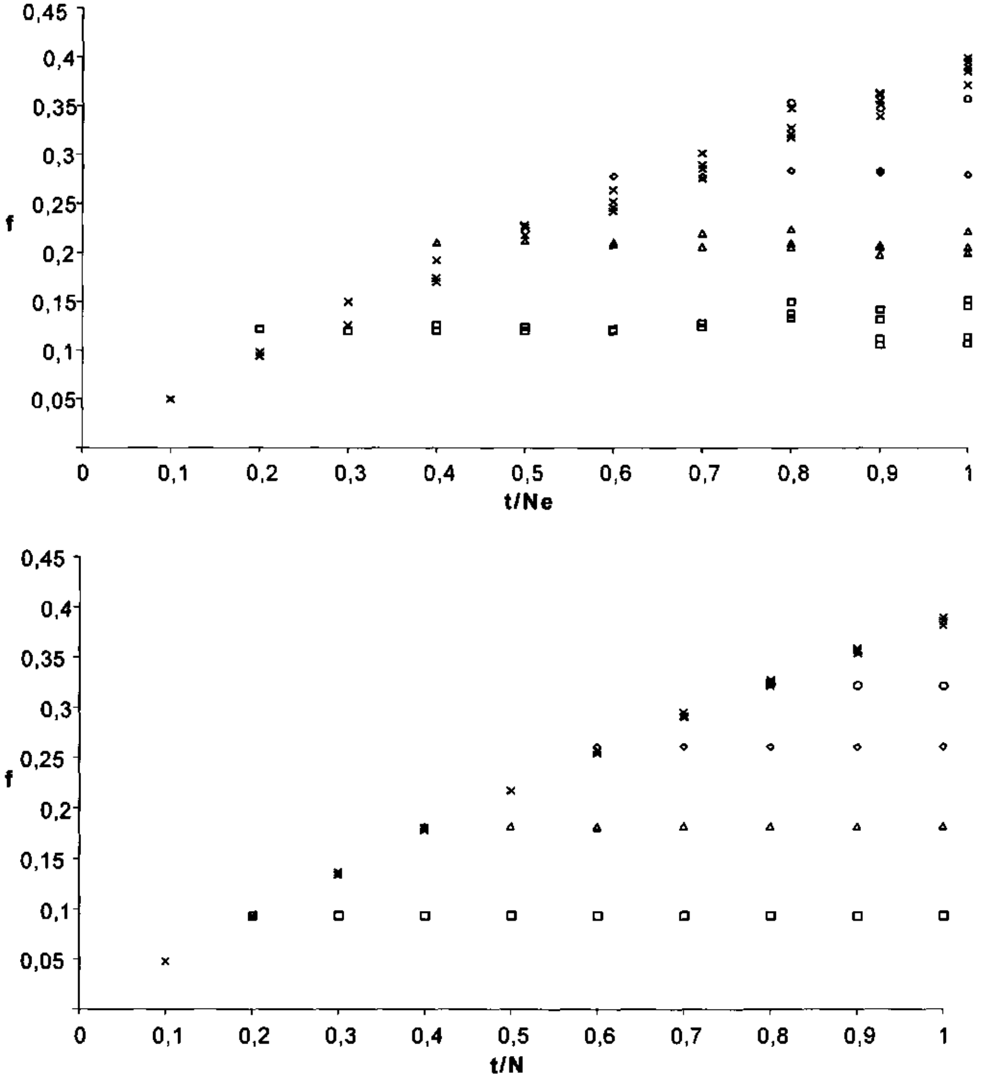


Figure 2.2 Scatterplot of the *actual* coefficient of kinship *f* (calculated from pedigree) versus t/N$_e$ (*above*) and estimated *f* using markers with two alleles per locus in the founder population (*below*) versus t/N$_e$ for a single replicate. Five populations were simulated. The populations have a phylogeny as given in Figure 2.1. (×) corresponds to the within population average actual kinship. (□) corresponds to the kinship between population 1 (the oldest population) and the cluster of populations (2, 3, 4, 5). (Δ) depicts the kinship between population 2 and the cluster (3,4,5), (◊) corresponds to the kinship between 3 and (4,5) and (o) is the average actual kinship between populations 4 and 5.

and 5. Note that after splitting the average kinship between populations remains constant in both 2a and 2b, even though genetic distances between populations would increase over time (see Discussion). Although some sampling deviations occur, Figure 2.2b generally depicts the same trend as Figure 2.2a.

Estimation of average kinships

In Table 2.1 the regression factor and the Mean Square Error of Prediction (MSEP), calculated as the square root of $\sum_{ij}\left(\hat{f}_{ij} - f_{ij}\right)^2 / n$, of average population *f* are given for a relatively short (t/N_e=0.4) and a relatively long (t/N_e=1) period of time. The case with M=200 refers to the full genetic model with which the simulation was done and is included for reference. In the upper half of the table founder alleles were assumed.

Table 2.1 Regression coefficients b, of the regression of the population averages of $\hat{f}_{ij}$ on f_{ij} and the square root of the Mean Square Error of Prediction (MSEP)[1]. Values of b and the MSEP were calculated over 20 replicates.

	t/N_e= 0.4		**t/N_e= 1.0**	
	b	MSEP	b	MSEP
No. of markers	founder alleles			
10	0.972	0.058	1.020	0.079
20	0.986	0.034	1.002	0.068
30	0.998	0.025	1.000	0.058
50	0.999	0.021	0.998	0.041
200	1.010	0.007	1.008	0.012
No. of alleles	200 markers			
2	0.852	0.020	0.940	0.028
5	0.970	0.009	0.992	0.018
10	1.000	0.009	1.003	0.015
20	0.998	0.008	1.001	0.013

[1] MSEP = $\sqrt{\sum_{ij}\left(\hat{f}_{ij} - f_{ij}\right)^2 / n}$, where n = 20 replicates

Table 2.2 Mean Square Errors of Prediction (MSEP) of estimated kinship $\hat{f}$ per pair of animals within a population. The probability of alleles being alike in state, but not identical by descent s had a value based on the distribution of alleles in the founder population. **t/N$_e$** is the time since establishment of the founder population. The regression estimates were taken from data over the entire history. Regression factors are from the regression $\hat{f} = b_0 + b_1 f + error$.

		# markers used							**Regression**		
# alleles	**t/N$_e$**	5	10	15	20	30	50	200	b_0	b_1	**MSEP$^{*)}$**
2	0.4	0.260	0.179	0.147	0.130	0.108	0.086	0.050	-0.007	1.084	0.042
	1.0	0.289	0.207	0.166	0.153	0.123	0.097	0.047			
5	0.4	0.154	0.109	0.089	0.077	0.065	0.054	0.037	0.002	0.980	0.026
	1.0	0.177	0.122	0.101	0.088	0.073	0.056	0.032			
10	0.4	0.123	0.089	0.076	0.067	0.058	0.048	0.035	0.002	0.999	0.023
	1.0	0.145	0.104	0.087	0.074	0.059	0.048	0.028			
20	0.4	0.107	0.077	0.067	0.059	0.051	0.043	0.034	0.005	0.992	0.021
	1.0	0.129	0.091	0.076	0.067	0.054	0.042	0.025			
Founder	0.4	0.094	0.069	0.059	0.053	0.047	0.040	0.033	0.002	0.992	0.019
	1.0	0.115	0.082	0.068	0.060	0.049	0.039	0.023			

*) Number of alleles per locus in the founder population. Alleles were assigned randomly with probability 1/(# alleles), except in the case of founder alleles, where each individual received a unique pair of allele.

The lower half of Table 2.1 gives the regression factors and MSEP of $\hat{f}$ with increasing numbers of alleles per locus at time $t/N_e = 0.4$ and 1, respectively. Regression coefficients between f and $\hat{f}$ were close to 1, indicating the estimator was approximately unbiased. The MSEP approached that of founder alleles. The estimation of $\hat{f}$ for non-founder alleles was by expression (5) and assumed known s.

Within populations estimates of kinship

The regression of the pairwise MEKs on the actual kinships was 1 and had relatively small MSEP. The right hand portion of Table 2.2 shows that the regression factors, b_0 and $b_{1,}$ are close to 0 and 1, respectively, which indicates an approximately unbiased estimation of f_{ij}.

For the left hand portion of Table 2.2 two situations were compared: one with a relatively short history ($t/N_e = 0.4$) and another with relatively long history ($t/N_e = 1$). Numbers of loci used were varied as was the number of alleles per locus in the founder population.

The general trend is a decreasing MSEP with increasing numbers of loci and increasing number of alleles per locus in the founder population. There is not a clear distinction in the importance between number of loci used and the number of alleles per locus. If the number of alleles per locus is low, extra alleles are more informative than extra loci.
MSEP was overall rather large. Especially when looking at scenarios that presently are used in the studies of genetic diversity with 10-15 loci, we see that it is virtually impossible to distinguish even full sibs from half sibs. To be able to accurately distinguish between non-inbred full sibs and half sibs ($p<0.05$) the results suggest that at moderate numbers of alleles per locus (5 -10) at least 30 to 50 unlinked markers have to be used, which confirms observations in similar studies of marker based relationship estimates (Lynch and Ritland, 1999).

Estimates of kinship and genetic distances

In Table 2.3 the proportion of variance explained by regression of genetic distances and similarity parameters on kinship, R^2, at time $t/N_e = 1$ are given for cases with different numbers of alleles in the founder population. All measures have an apparently strong relationship with kinship. Only F_{ST} shows a very weak relation with kinship when the number of alleles is 2. This might be due to the combination of relatively large variance on the

Table 2.3 Proportion of variance explained by the regression of average pairwise similarity S_{xy}, population similarity S_{ij}, Nei's standard distance D, Reynolds distance D_R or F_{ST} (from allele frequencies) at $t/N_e=1$ on actual average kinship (calculated from pedigree), R^2. Estimates of the parameters were based on full genetic information (i.e. 200 markers).

	Parameter				
# alleles/locus	S_{xy}	S_{ij}	$D^{*)}$	$D_R^{*)}$	F_{ST}
2	0.944	0.959	0.881	0.870	0.041
5	0.979	0.983	0.917	0.954	0.831
10	0.984	0.987	0.905	0.965	0.899
20	0.984	0.987	0.905	0.965	0.915
Founder	0.990	0.992	0.863	0.971	0.967

*) Genetic distances were calculated between populations only.

estimator and low estimates of F_{ST} due to the number of allels per locus. Although these strong relationships can be explained by the fact that all populations evolved similarly (constant and equal N_e) it illustrates that genetic distance measures have a tendency to be highly related (Hedrick, 1974; Takezaki and Nei, 1996).

R^2 of both measures of S with kinship is consistently higher than those of genetic distances. Note that the correlation of Nei's distance with kinship is reduced when founder alleles are used. This is due to the non-linearity with t/N_e of Nei's distance.

Looking over time the relationships between kinship and genetic distance becomes more complicated. In Figures 3a and b scatter plots are given of S and Nei's standard distance respectively versus the true kinship. S was calculated in two alternative ways: averaging all pairwise similarities, S_{xy} and estimation from allele frequencies, S_{ij}. Results were very similar so they are not presented separately. Both S_{ij} and S_{xy} were calculated from founder alleles, so $S = \hat{f}$. The points in the scatter plots represent kinships and the statistics mentioned above between populations at 10 intervals in time between $t/N_e=0$ and $t/N_e=1$ for 20 replicates. The four groups of data points in Figure 2.3a and 3b (from left to right) correspond to the kinship/distance of population 1 and the cluster of populations (2,3,4,5), populations 2 and (3,4,5), 3 and (4,5) and the kinship distance between populations 4 and 5. In Figure 2.3b, each group of data points starts on the x-axis (distance =0), as this is the moment where population

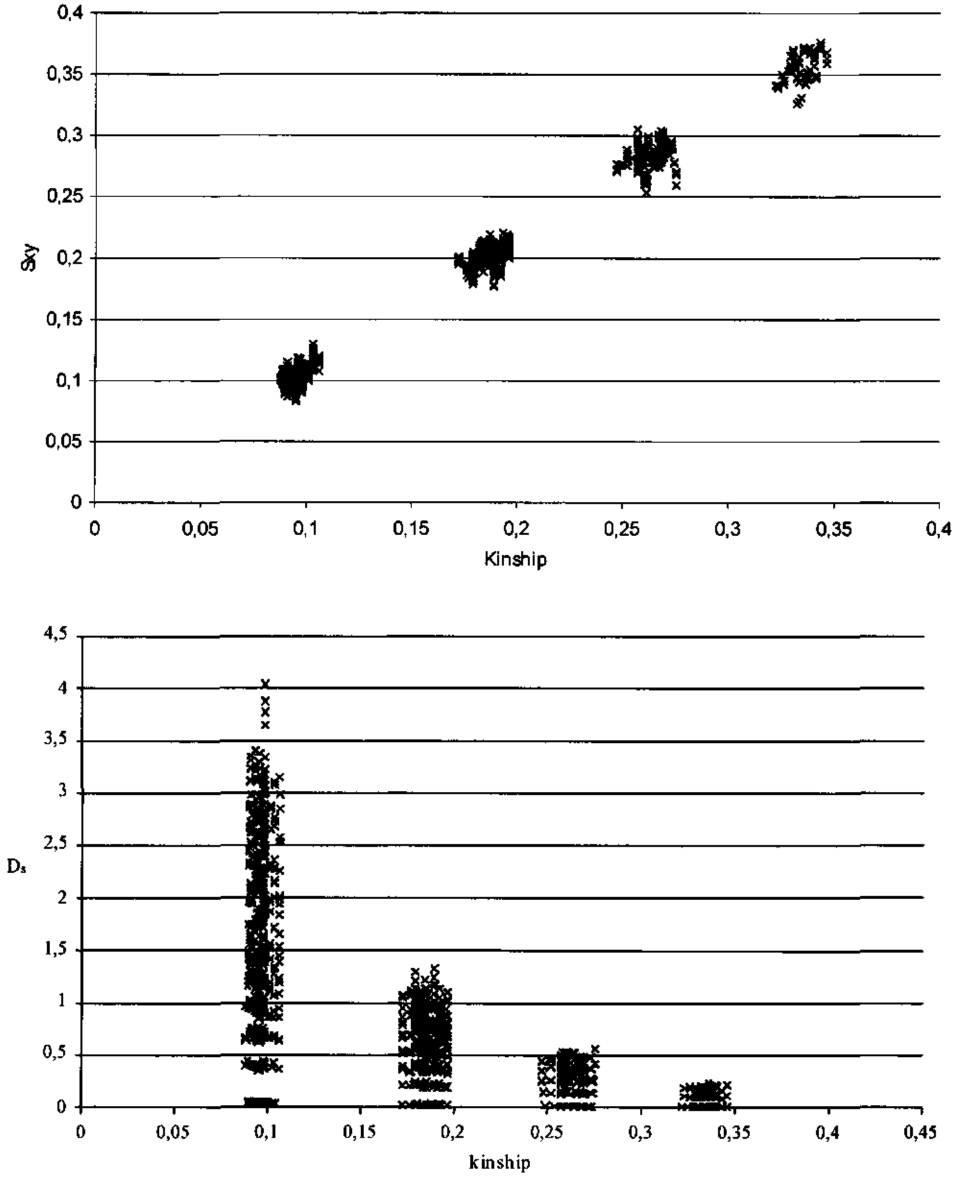


Figure 2.3 Scatter plots of between population diversity estimators versus the true kinship. Five populations were simulated according to Figure 2.1. All information (all individuals and all 200 loci) was included. For all measures founder alleles were assumed. *Above*) $\hat{f}$ based on *S*, *below*) Nei's standard genetic distance.

fission took place (D=0). Over the next time interval, the distances increase. The kinship between populations remains the same however, resulting in a cloud of points directly above the previous ones. Looking at Figure 2.3b, it is clear a distance measure can be associated with any number of combinations of kinship coefficients, making the interpretation of genetic distances in terms of genetic diversity ambiguous. Figure 2.3b shows this relationship for Nei's standard distance, but was similar for Reynold's distance and F_{ST}.

The average kinship f_{xy} between two populations x and y is an estimate of the time, or rather t/N_e between establishment of the founder populations and the time of divergence of the two populations. It is approximately equal to inbreeding in the parent population at time of divergence. After population fission f_{xy} will remain constant, while x and y will drift further apart, resulting in increasing distance estimates between population x and y, which explains the differences between kinship and distance measures in Figure 2.3.

DISCUSSION

Kinship/similarity as measure for genetic diversity

In this paper we argue that average kinship is a good measure of genetic diversity. Moreover, as can be seen from expression (5) most of the distance and diversity measures involve terms that estimate kinship. Kinship or similarity indices can be used to assess genetic diversity within and between populations. For conservation purposes kinship as a measure of diversity has some properties with intuitive appeal:

1) Within populations, kinships can generally only increase while diversity can only decrease over time (ignoring mutation).

2) After population fission kinship between populations becomes constant very quickly causing between population diversity to remain constant. For example, even after two descendant populations have become fully inbred there will be a fraction of loci at which the same allele has been fixed in both populations. Assuming founder alleles this fraction will be equal to the mean kinship in the parent population just prior to population fission, hence, the constant between population kinship. Because some of the fixed alleles in fully inbred populations will differ, some genetic diversity remains as predicted from a kinship coefficient smaller than 1. If the founder allele assumption is relaxed the fraction of alleles fixed in both populations (i.e. the similarity) will be larger than the average kinship between these populations. However, both s and f between two populations are defined by preceding

generations and therefore not subject to change. The expectation of the per locus similarity score will therefore also stay constant.

3) The definition of the coefficient of kinship as the probability that two randomly sampled alleles drawn from two individuals are identical by descent *f*, which implies that (1-*f*) is the probability they are not identical by descent and can therefore be interpreted as an upper limit for genetic diversity.

4) The coefficient of kinship is also involved in the variance of quantitative traits. In Appendix 2.C we show how the minimization of kinship will lead to conservation of variance of quantitative traits.

Between populations the marker-based estimates of *f* (including between a population with itself) show relatively low MSEP (Table 2.1), and are useful as genetic diversity measures. Between individuals the estimates of *f* suffer from relatively high MSEP (Table 2.2). Using a reasonable number of marker alleles (30-50) which are relatively polymorphic (5-10 alleles per locus) it is possible to distinguish animals with low kinship from pairs of animals with a high degree of kinship. Estimating between individual kinships based on marker estimation, even with a low number of marker loci, is useful however. Use of these estimates to calculate between population kinships introduces less assumptions about the population structure and implicitly accounts for structures within a population (herds, for instance).

Estimates of relations between individuals have been developed by many authors (Thompson, 1975; Lynch, 1988; Li *et al.* 1993; Lynch and Ritland, 1999). Each of these estimates has its merits but is not entirely suitable for the purposes we describe in this paper. Either they are not linear with Malecot's coefficient of kinship (Lynch, 1988) or can realistically only be applied within a population. Lynch and Ritland (1999) state that there are problems with the sampling error of the similarity index used in this paper. However, the case cited in Lynch and Ritland corrects for alleles alike in state by replacing *s* in Equation (3) by J_0, the expected homozygosity under Hardy-Weinberg equilibrium. While this is a good approximation for estimations of first and second order relationships, it should be clear that this is not the desired method when assessing genetic diversity. Using the expected homozygosity of a population spanning multiple generations defines the founder population somewhere between the oldest and the youngest generation in the population. When J_0 is used within populations a problem occurs in that populations cannot be compared for their genetic diversity content. Furthermore, inbreeding is not accounted for, while this is an important part of genetic diversity within a population. In practice, the use of J_0 as the probability of AIS leads to negative estimates of the kinship coefficient in cases where the common ancestor(s) is (are) a member of the oldest generations and is not a matter of sampling error alone.

All of the above authors and many others have concluded that it requires a large amount of genetic marker data to obtain reliable estimates of between individual coefficients of kinship. If there exists pedigree information other than from genetic marker data (i.e. herd books) it seems advisable that once populations have been identified for conservation, the existing pedigree information is incorporated to facilitate selection of individual contributors to a conservation plan or gene bank. This might be done by using Wright's (1968) F-statistics:

$$(1-F_{IT})=(1-F_{IS})(1-F_{ST})$$

where F_{IT} is defined as the total kinship between two individuals within a population. F_{IS} is the kinship between two individuals relative to the present population and can be extracted from the (limited) pedigree information. Then for F_{ST} we substitute the average kinship within the population under study estimated from genetic marker data (*i.e.* MEK). This method removes a large part of the error of the estimates of kinships between individuals based on marker data only.If pedigree information does not exist the Marker Estimated Kinships can still be used to avoid selection of full sibs or half sibs as contributors.

The strength of the presented method is that the same method is being applied on the level of breeds, populations, herds down to individuals which, as shown above can relatively easy incorporate existing pedigree information. Both Marker Estimated Kinships and pedigree information are tranferred to kinship coefficients and are therefore easily combined. The result is a comprehensive approach to assessing the genetic diversity that is maintained in a gene bank and thus can be used to prioritise breeds or populations for genetic conservation.

In this study a genome was simulated consisting of a maximum of 200 autosomal, unlinked loci. In nature, linkage does occur of course and will have an influence on the accuracy with which f is estimated. Accounting for linkage however is complicated and lies beyond the scope of this paper.

Weitzman (1992) developed criteria which have to be fulfiled by proper measures of diversity (Thaon d'Arnoldi *et al.*, 1998). These criteria are:

1) The 'twin property', which means that the inclusion of a population identical to a population already in a set of conserved populations must not increase the diversity in the set. In the case of kinship inclusion of such a population would increase the average kinship, i.e. diversity would be decreased.

2) The total amount of diversity in a set of populations cannot increase when a population is removed from the set. It can be shown that the average kinship can decrease, i.e. diversity can increase, when a population is removed from the set. However, this can only happen when the between population kinships are (almost) as large the within population kinships. The latter is not likely to occur in practice.

3) Continuity in distance: If distances are slightly modified, the change in diversity is slight too. Average kinship is a continuous function, so any small change leads to a small difference in average kinship.

4) Monotonicity in distance: If distances increase, diversity should increase also: If the kinship between two population decreases, diversity will increase.

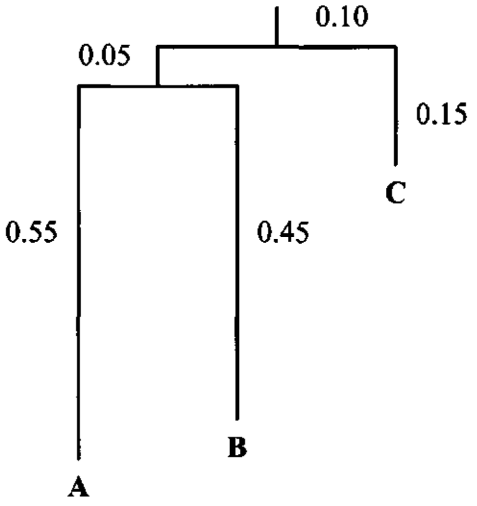
Thus the average kinship as a measure of diversity has some problems with the comparison of sets of unequal sizes, i.e. Weitzman's criteria 1 and 2. These problems do not seem to be very important in practical situations, where the number of populations in the genebank will often be limited and thus constant. We are in the process of modifying the average kinship criterion to a weighted average kinship, which should fulfill all of Weitzman's criteria.

Kinship and genetic distances

Being proportional to time since divergence, genetic distances create the impression of increasing diversity between two populations, even when there is no change in the actual genetic diversity in terms of allelic diversity or coefficient of kinships. The average kinship within a population can be written as:

$$f_x = f_{xy} + \Delta f_x$$

That is: the within population kinship is the sum of the between population kinship (*i.e.* the kinship within the population just prior to fission, f_{xy}) and the increase in within population kinship since fission (Δf_x).



distance	B	C
A	1.00 (0.64)	0.75 (0.62)
B		0.65 (0.59)

Kinship	A	B	C
A	0.70	0.15	0.10
B		0.60	0.10
C			0.25

Figure 2.4 Hypothetical phylogenetic tree of three breeds. The numbers in the figure refer to the increase in average coefficient of kinship within the line. The table in the figure gives the genetic distances between the breeds in a general form ($d(x,y)=f_x+f_y-2f_{xy}=\Delta f_x+\Delta f_y$) and Nei's standard genetic distance D (in parentheses) assuming founder alleles (i.e. D = -log(I), with $I=f_{xy}\big/\sqrt{f_x f_y}$) and kinships. From the table can be seen that even though the pair (A,B) has less diversity (higher between and within population coefficients of kinship), the distance between A and B is larger then the distances between them and C.

In terms of coefficients of kinship, a generic distance between populations x and y can be written as:

$$\begin{aligned} d(x,y) &= f_x + f_y - 2f_{xy} \\ &= \Delta f_x + \Delta f_y \end{aligned}$$

This expression explains the relation between genetic distances and kinship. Although f_{xy} stays constant over time, f_x and f_y increase over time and this results in an increase of the distance between x and y for the same value of f_{xy}.

Suppose we have a phylogenetic tree as given in Figure 2.4. In this figure the lengths of the branches are given in terms of *f*. The distances between (A,B), (A,C) and (B,C) in terms of average kinship are given in the table in Figure 2.4. In parentheses Nei's distances are given, assuming founder alleles.

If two populations were chosen for conservation based on these distances, the choice would be the pair (A,B) since they have the largest distance between them and seem the furthest apart. However, both the within and between population kinship is smaller (and consequently the conserved diversity larger), when the pair (A,C) or (B,C) is chosen for conservation instead of (A,B). The robust method of Weitzman results in population C being the link element in the diversity tree, which implies that the loss of population C is less consequential for the diversity than any other element. Clearly, the loss of population C in our example would yield the highest loss of diversity. Genetic distances are useful to picture genetic diversity, e.g. in the form of phylogenetic trees. However, genetic distances increase with increasing levels of inbreeding of the populations, and thus diversity decreases. In general genetic distances will conserve the more extreme genotypes and allele frequencies by placing more emphasis on differences between populations, while minimizing kinships attempts to conserve the founder population allele frequencies.

Correction for alleles being alike in state

Estimation of kinships with genetic marker data is easiest under the assumption of founder alleles somewhere in the history of the population. Toro *et al.* (1998) have used this assumption in their study of the use of marker information in a live conservation of a single breed. If the assumption of founder alleles is relaxed the estimate of kinship needs to be corrected for the probability two alleles are alike in state, *s*. When kinship or numbers of alleles per locus are relatively small, the influence of the distribution of alleles in the founder population is considerable (Table 2.2). There is an advantage in using estimates of *s* in that it makes weighing over loci possible which reduces the variance of the estimator (Equation (4)). Note that since we assume a single founding population, *s* will be of equal value for all populations and individuals and the ranking of pairs of individuals or populations is not affected by the assumed value of *s*.

In a set of populations we can assume *s* to be the value of the between population similarity of the populations descending from the oldest fission (i.e. *s* equals the smallest between population kinship). In the populations structure used in this study this would mean taking the average value of the between population similarity of population 1 and the cluster (2,3,4,5) (see Figure 2.1). This defines the generations with parents of 1 and 2 as the base population. This method requires the least amount of assumptions about the character of the founder population: information on the founder population can be inferred from the between population similarity of the two oldest populations or clusters. This seems to be the best approach to the question of founder population definition. It should be noted that the definition of a founder population is artificial. It is a convenient entity to specify more precisely what the relationships are and to minimize the prediction error of kinships estimates using equation (4). For conservation purposes the estimate of s need not be accurate, because the MEK will still be proportional to the true f. This will leave the outcome of a selection procedure of animals for a genebank unaffected, which. has been verified in an example (results not shown).

In this study mutation was not accounted for. Mutation will bias information about kinships between and within populations and individuals. However, studies of the effect of mutation on genetic distances generally indicate that these effects will not disturb estimates very much, unless the number of generations and the population size are very large (Slatkin, 1995; Nauta and Weissing, 1996). In studies of breed formation, both the population size and the time since divergence are expected to be relatively small on an evolutionary scale and therefore the influence of mutations is not expected to be of great importance.

Generally, when using marker information, it is recommended to use markers that are as polymorphic as possible (Bretting and Widerlechner, 1995). The panel of microsatellite markers proposed by FAO in the study of genetic diversity in European cattle (as part of the MoDAD project) was chosen on the basis that the markers had to have at least 4 different alleles per locus (FAO Primary Guidelines, 1998). Selection of highly polymorphic markers is equal to selection of markers with small s. Since the method presented in this paper includes a correction for s, this selection of highly polymorphic markers is not expected to bias the kinship estimates. Marker loci used should however display more then two alleles per locus. Writing the estimate of the coefficient of kinship in Jacquards notation for a locus with only two alleles in the founder population shows that this situation is no longer yielding an estimate of Malecot's kinship coefficient. This explains the poorer performance of the diversity measures in this paper for situations in which only two alleles per locus were used.

Conclusion

Kinship coefficients appear to be of central importance in the definition and measurement of genetic diversity. As the results show, it is possible to obtain estimates of between population kinship with acceptably low MSEP. These estimates may be biased by the unknown *s* (the probability two alleles are alike in state, but not identical by descent). However, since it is expected that this bias is equal for all populations (*s* being a function of the homozygosity in the founder population; see before) it will not affect the selection of populations for genetic conservation. The Marker Estimated Kinships will allow us to identify those populations and individuals that have the least kinship and will therefore help to make optimal use of limited resources for genetic conservation. However, the MSEP of the between individual estimates are such that it is advisable to use existing pedigree information for the selection of individuals of a population that is to be conserved.

ACKNOWLEDGEMENT

We would like to thank Pim Brascamp, Ab Groen and Kor Oldenbroek for their useful comments on the manuscript.

APPENDIX 2.A

The 15 states of identity defined by Jacquard are given in Figure 2.A1 condensed in 9 condensed coefficients of identity (Taken from Lynch and Walsh, 1999). Note that these states of identity presuppose the existence of more than two alleles for a locus.

Ignoring alleles alike in state (AIS) Malecot's coefficient of kinship can be written in these condensed identity coefficients as (Lynch and Walsh, 1998):

$$f_{xy} = \Delta_1 + \tfrac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \tfrac{1}{4}\Delta_8$$

The similarity index S_{xy} is defined as given in Table 2.A1 with the corresponding condensed identity coefficients. Assuming founder alleles and summing over all four possible values we get:

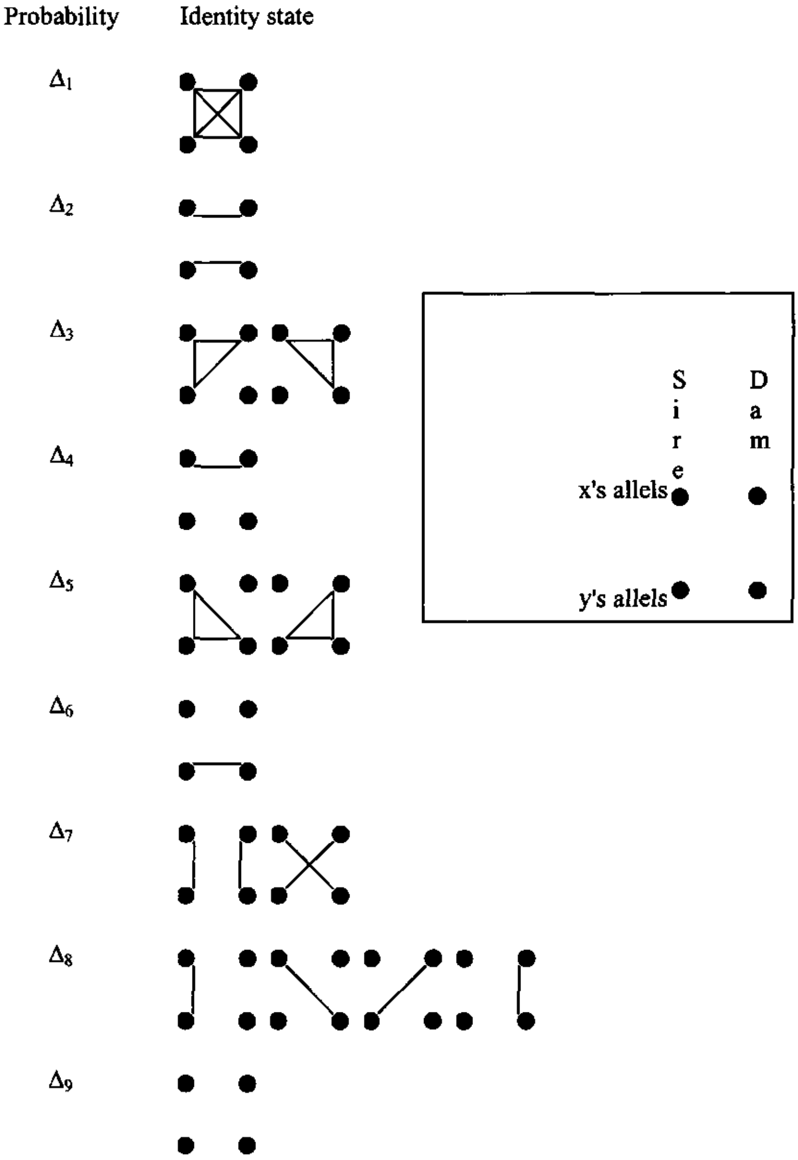
$$S_{xy} = \Delta_1 + \tfrac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \tfrac{1}{4}\Delta_8 = f_{xy}$$

i.e., assuming founder alleles S_{xy} is an unbiased estimator of $f_{xy.}$ Moreover, S_{xy} will be linear with f_{xy} as long as the number of alleles per locus is larger than two. When only two alleles per locus are assumed Δ_8 is undefined and S_{xy} is no longer strictly linear with f_{xy}. Note that this situation is different from the situation where Δ_8 equals 0, i.e. more than two alleles were present in the founder population. In the latter case S_{xy} is still linear with f_{xy}.

Table 2.A1 The four possible values of the similarity index and their corresponding condensed coefficients of identity.

Similarity	**value**	**Identity coefficient**
AA - AA	1	Δ_1
AA - AB	1/2	$\Delta_3 + \Delta_5$
AB - AB	1/2	Δ_7
AB - BC	1/4	Δ_8
Total		$\Delta_1 + \tfrac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \tfrac{1}{4}\Delta_8$

Figure 2.A1 The nine condensed coefficients of identity for a locus in two individuals. Alleles that are identical by descent are connected by lines (Taken from Lynch and Walsh, 1998).



Lynch and Ritland (1999) define a coefficient of relatedness, which should estimate twice the kinship coefficient of Malecot:

$$r_{xy} = \frac{\phi_{xy}}{2} + \Delta_{xy}$$

Where ϕ_{xy} is the probability that one allele in x is IBD with one allele in y, and Δ_{xy} is the probability that both alleles in x are IBD with alleles in y. Lynch and Ritland do not account for inbreeding. This removes the probability of individuals being homozygous for alleles IBD. If we rewrite f_{xy} and r_{xy} under these terms we get:

$$f_{xy} = \tfrac{1}{2}\Delta_7 + \tfrac{1}{4}\Delta_8$$

and

$$r_{xy} = \Delta_7 + \tfrac{1}{2}\Delta_8$$

As can be seen from the above: The estimator of Lynch and Ritland agrees with Malecots coefficient of kinship if inbreeding is non-existent. However, if individuals are allowed to be homozygous for alleles IBD, i.e. inbreeding does occur the estimator presented by Lynch and Ritland can be expressed as:

$$r_{xy} = \Delta_1 + \Delta_3 + \Delta_7 + \tfrac{1}{2}(\Delta_5 + \Delta_8)$$

which is no longer agrees with Malecots coefficient of kinship.

APPENDIX 2.B

As stated in the main text, the relation between S and the kinship f_{ij} between i and j can be written as:

$$\begin{aligned} \mathrm{E}(S_l) &= p_{ij,l} \\ &= f_{ij} + (1-f_{ij})s_l \\ &= s_l + (1-s_l)f_{ij} \end{aligned} \qquad (1)$$

where $S_{ij,1}$ is the similarity between two individuals for locus l and s_l is the probability of alleles of locus l being alike.

This result leads to the variance of $\hat{f}$ in that

$$\operatorname{var}\left(\hat{f}_{ij}\right)=\frac{1}{\left(1-s_l\right)^2}\operatorname{var}\left(S_{ij,l}\right) \tag{2}$$

Since S is the probability that two random alleles drawn from two individuals are alike, the distribution of S is binomial. The variance of S between two individuals i and j for a locus l is given as:

$$\operatorname{var}(S_{ij,l})=p_{ij,l}\left(1-p_{ij,l}\right) \tag{3}$$

Filling in (1) in (3) yields:

$$\begin{aligned}\operatorname{var}\left(S_{ij,l}\right)&=f_{ij}\left(1-s_l\right)+s_l-\left[f_{ij}^2\left(1-s_l\right)^2+2f_{ij}s_l+s_l^2\right]\\&=f_{ij}\left(1-s_l\right)\left(1-2s_l\right)+s_l\left(1-s_l\right)-f_{ij}^2\left(1-s_l\right)^2\end{aligned} \tag{4}$$

Substitution of (5) in (2) gives:

$$\begin{aligned}\operatorname{var}\left(\hat{f}_{ij}\right)&=\frac{f_{ij}\left(1-s_l\right)\left(1-2s_l\right)+s_l\left(1-s_l\right)-f_{ij}^2\left(1-s_l\right)^2}{\left(1-s_l\right)^2}\\&=\frac{s_l+f_{ij}\left(1-2s_l\right)-f_{ij}^2\left(1-s_l\right)}{1-s_l}\end{aligned} \tag{5}$$

APPENDIX 2.C

Suppose an animal i has a breeding value u$_i$ for an (unspecified) trait. The total variance of breeding value u$_i$ equals the variance of the mean plus the variance of deviations within the population:

$$\operatorname{var}\left(u_i\right)=\operatorname{var}\left(\bar{u}\right)+\operatorname{var}\left(u_i-\bar{u}\right)\Rightarrow$$

$$\operatorname{var}\left(u_i-\bar{u}\right)=\operatorname{var}\left(u_i\right)-\operatorname{var}\left(\bar{u}\right)$$

The total amount of genetic diversity in a population is described by $\operatorname{var}\left(u_i-\bar{u}\right)$ and it is this quantity we want maximized. The total variance of the breeding value, var(u$_i$), is fixed and unknown and thus cannot be maximized. Therefore a conservation plan can only affect $\operatorname{var}\left(\bar{u}\right)$. This last factor can be interpreted as the variance of the average breeding value of all possible genebanks assembled from the population under study.

In matrix notation $\mathrm{var}(\bar{u})$ equals $\mathrm{var}(\mathbf{c}'\mathbf{u}/\mathbf{c}'\mathbf{c})$, where $\mathbf{u}$ is an n x 1 vector containing the breeding values of the animals in the population and **c** denotes a vector of ones and zeros indicating which individuals in the total population are selected for conservation.

Now,

$$\mathrm{var}(\mathbf{c}'\mathbf{u}/n_{gb}) = \mathbf{c}'\,\mathrm{var}(\mathbf{u})\mathbf{c}/n_{gb}^2 = \mathbf{c}'\left[\sigma_u^2\mathbf{A}\right]\mathbf{c}/n_{gb}^2$$

where **A** is the relationship matrix and n_{gb} = **c'c** is the number of individuals in the genebank. Elements a_{ij} of **A** are the additive genetic relationships between individuals i and j and Malecot's coefficient of kinship is $f_{ij} = 0.5(a_{ij})$. We can see that $\mathrm{var}(\bar{u})$ is proportional to $\mathbf{A}/n^2$, hence it follows that maximization of genetic diversity in any quantitative trait implies minimization of average kinship.

Chapter 3

ASSESSING THE CONTRIBUTION OF BREEDS TO GENETIC DIVERSITY IN CONSERVATION SCHEMES

Eding, Herwin [*,1)] , Richard P.M.A. Crooijmans[2], Martien A.M Groenen[2] and Theo H.E. Meuwissen[1]

[1] Institute for Animal Science and Health, Box 65, 8200 AB Lelystad, The Netherlands

[2] Animal Breeding and Genetics group, Wageningen Institute for Animal Science, Wageningen University, Box 338, 6700 AH Wageningen, The Netherlands

Genetics Selection Evolution (Submitted)

[*)] Institute for Animal Science and Health
Box 65, 8200 AB Lelystad, The Netherlands
Tel: +31-(0)320-238238; Fax +31-(0)320-238050
E-mail: j.h.eding@id.wag-ur.nl

ABSTRACT

Quantitative assessment of genetic diversity within and between populations is important for decision making in genetic conservation plans. In this paper we introduce a definition of genetic diversity that is based on Marker Estimated Kinships. First we calculate the relative contribution of populations to a core set of populations in which overlap of genetic diversity is minimised. The total genetic diversity in a set of populations is defined as the average kinship in this core set. This definition satisfies the Weitzman criteria for a measure of genetic diversity. The application of the method is illustrated by an example involving 45 Dutch poultry breeds. The calculations used are easy to implement and not computer intensive. The method gives a ranking of breeds according to their contributions to genetic diversity. Losses in genetic diversity ranged from 2.1% to 4.5% for different subsets relative to the entire set of breeds, while the loss of founder genome equivalents ranged from 22.9% to 39.3%

INTRODUCTION

In conservation genetics of livestock the question of which breeds to conserve is important. Decisions on which breeds to conserve can be based on a number of different considerations, degree of endangerment being the most important (Oldenbroek, 1999). Forced by limited resources to concentrate efforts on only a few populations under threat, we need insight into the genetic variation present in each population. Quantitative assessment of genetic diversity within and between populations is a tool for decision making in genetic conservation plans. Weitzman proposed a method to quantify diversity in a set of populations (Weitzman, 1992), which was based on pairwise genetic distances between the populations. In the same paper, Weitzman put forth a number of criteria (see METHODS section for further details), to which a meaningful measure of diversity should adhere. Thaon d'Arnoldi *et al.* demonstrated this method in a set of cattle breeds (Thaon d'Arnoldi *et al.*, 1998). They noted that because of the recursive nature of Weitzman's method, the algorithm to calculate the total diversity in a set of breeds and the loss of genetic diversity when a breed is excluded from the set is complex and computer intensive, limiting its use to sets of 25 populations or less. A simpler method, which does not have these limitations, would be advantageous.

In this paper we develop such a method based on Marker Estimated Kinships (MEK). Eding and Meuwissen proposed the use of MEK to asses genetic diversity (Eding and Meuwissen, 2001), a measure which expresses genetic diversity in terms of average (estimated) kinships between (and within) populations using genetic marker genes. In contrast, the Weitzman method expresses only between population diversity. Furthermore, kinships have a direct relationship with other well-known indicators of genetic diversity (Caballero and Toro, 2000). A population that is the result of random mating within and between populations of a conserved set will show the conserved genetic variance which is: $\sigma_w^2 = (1-\bar{f})\sigma_a^2$, where σ_a^2 is the total original genetic variance and $\bar{f}$ is the average kinship within the set of populations (Falconer and Mackay, 1996) (page 265; their term 'line' refers to the conserved set here).

From the former, it follows that a kinship based method of assessing genetic diversity is essentially based on genetic variance. Thaon d'Arnoldi *et al.* observe that variance based estimates do not necessarily comply with Weitzman's criteria. For instance, it is possible that the removal of a population from the set leads to an increase in diversity (Thaon d'Arnoldi *et al.*, 1998).

In this note we propose a MEK based definition of total genetic diversity in a set of populations, which is consistent with Weitzman's criteria. The calculations used are non-recursive and therefore easier to implement and less computer intensive then the Weitzman approach. Moreover, this method accounts for both within and between population diversity simultaneously. The method relies on estimation of the contribution of each breed to a Core Set (Core set). These estimated contributions provide a way of ranking breeds according to their importance with regard to genetic diversity.

METHOD

As an example, consider 3 populations, where population 2 and 3 are identical, while population 1 is unrelated to both 2 and 3. The kinship matrix is:

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

The average kinship in **M** is 5/9 (5 ones over 9 elements). Removal of population 3 from **M** leads to

$$\mathbf{M}^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and the average kinship has decreased to 2/4, which implies an increase of genetic diversity. According to the Weitzman criteria, the removal of a population should have either a negative or zero effect on the measure of diversity.

The decrease in average kinship that occurred with the removal of population 3 from the set, occurred because populations 3 and 2 are the same population. There is one population that contributes twice to the average kinship of set *S*. However, the diversity in set *S* depends only on whether a population is represented in *S* or not, not how often. This problem is avoided by basing the diversity contained in set *S* on the average kinship of a core set of set *S*, where the core set is a mixture of populations such that 'genetic overlap' within the core set is minimized (Frankel and Brown, 1984). This is done by minimising the average kinship within a core set

The coefficient of kinship is defined as the probability that two randomly drawn alleles from two individuals are identical by descent. Thus the average coefficient of kinship between two populations indicates the fraction of alleles two populations have in common through common ancestors. To eliminate as much genetic overlap as possible, the average coefficient of kinship in the core set of S should be minimised. In the case of the former example the solution would be removal of population 3 (or equivalently, removal of population 2). This removal does not affect the diversity contained in the core set, which seems intuitively correct.

Optimal contributions to a core set

Consider an n×n matrix **M** containing within and between population kinships for n populations in set S. Also define an n-dimensional vector **c** that will contain the relative contribution of each population to the gene bank, such that the elements of **c** sum up to one. We can calculate the average kinship in the set, given **c**, as:

$$(1) \qquad f(S) = \mathbf{c}'\mathbf{M}\mathbf{c}$$

For the construction of the Core Set we must find contributions in **c** such that the average kinship in the core set is minimal. To this end we introduce a Lagrangian multiplier λ that restricts the **c** vector such that the elements of **c** sum up to 1, leading to the Lagrangian equation:

$$(2) \qquad L(S) = \mathbf{c}'\mathbf{M}\mathbf{c} - \lambda(\mathbf{c}'\mathbf{1}_n - 1)$$

Where $L(S)$ is the average kinship in the core set. $\mathbf{1}_n$ is a n dimensional vector of ones.

Setting the first derivative of (2) with respect to **c** to zero we get:

$$\frac{\partial L(S)_{\min}}{\partial \mathbf{c}} = 2\mathbf{Mc} - \lambda \mathbf{1}_n = 0$$

(3)

$$\mathbf{Mc} = \tfrac{1}{2}\lambda \mathbf{1}_n$$

$$\mathbf{c} = \tfrac{1}{2}\lambda \mathbf{M}^{-1}\mathbf{1}_n$$

And since $\mathbf{c}'\mathbf{1}_n = 1$

(4)

$$\mathbf{c}'\mathbf{1} = \tfrac{1}{2}\lambda \mathbf{1}_n{}'\mathbf{M}^{-1}\mathbf{1}_n = 1$$

$$\lambda = \frac{2}{\mathbf{1}_n{}'\mathbf{M}^{-1}\mathbf{1}_n}$$

Substituting this results in (3) we obtain:

(5) $$\mathbf{c}_{\min} = \frac{\mathbf{M}^{-1}\mathbf{1}_n}{\mathbf{1}_n{}'\mathbf{M}^{-1}\mathbf{1}_n}$$

The value of $f(S)_{\min}$ can be obtained from

(6)

$$\mathbf{c}_{\min}{}'\mathbf{Mc}_{\min} = \frac{1}{\left(\mathbf{1}_n{}'\mathbf{M}^{-1}\mathbf{1}_n\right)^2} \cdot \mathbf{1}_n{}'\mathbf{M}^{-1}\mathbf{1}_n$$

$$= \frac{1}{\mathbf{1}_n{}'\mathbf{M}^{-1}\mathbf{1}_n}$$

Because the genetic variance contained within set S is proportional to ($1 - f(S)_{\min}$), the genetic diversity Div(S) in set S is defined as Div(S) = $1 - f(S)_{\min}$.

<u>*The Weitzman criteria*</u>

Weitzman (1992) defined four criteria for a proper measure of diversity:

Criterion 1) *The link property.* The total amount of diversity in a set of populations should not increase when a population is removed from the set.

Criterion 2) *The twin property.* The addition of an element identical to an element already in the set should not change the diversity content in a set of populations.

Criterion 3) *Continuity in distance.* A small change in distance measures should not results in large changes in the diversity measure.

Criterion 4) *Monotonicity in distance.* The diversity contained in a pair of populations should increase if the distance between these populations increases.

With regard to the first criterion: Since kinship is essentially a measure of variance it is possible that the estimated genetic diversity in terms of kinship increases when a population is removed from the set (Thaon d'Arnoldi *et al.*, 1998). However, when the contribution of each population is optimised, the average kinship is at a minimum. Removal of a breed from the set will give a solution away from the minimum average kinship if the contribution of this breed is non-zero and genetic diversity will decrease. In the case a population is identical to another population in the set (or a inbred sub-population of another population) its contribution is zero and can be excluded from the set without affecting the diversity, which satisfies criterion 2.

With regard to criteria 3): The measure of genetic diversity in a set of breeds as presented above is a continuous function of the (estimated) average kinships between and within breeds. Hence, the measure of genetic diversity presented here changes only slightly, when distances change slightly.

With regard to criterion (4), in the short term an increase in distances is not necessarily equal to an increase in genetic diversity, because an increase in genetic distance can be achieved in two ways (in terms of kinships): An increase in the within population average kinships, or a decrease in the between population average kinship (Eding and Meuwissen, 2001). In the former case genetic diversity decreases, in the latter case diversity increases.

Application to real marker data

As an illustration of the use of the MEK/Core set method, we present here the results from a data set containing microsatellite data from 46 lines of poultry. DNA was isolated of pooled blood samples (approximately 50 animals per line) as described by Crooijmans *et al.*(1996). In case of the Sumatra's only 10 animals were present in the pool. These 46 lines were genotyped for 17 microsatellites. Within the lines three major groups can be distinguished: Commercial layer lines (N_l= 9) which can be subdivided into brown layers (25, 26, 27, 29 and 57) and white layers (17, 18, 20, 56), commercial broiler lines (N_b= 17) and non-commercial breeds of poultry (N_h=20). The latter included indigenous Dutch breeds, which are mainly kept and bred as fancy breeds, and the *Bankiva* and *Sumatra* breed. The data is summarised in Table 3.1.

Table 3.1 Summary of the data on poultry lines and genetic markers used in the application of the Marker Estimated Kinship/ Core Set method.

Indigenous populations	**Commercial lines**	**Markers Used**	**# alleles**
Assendelft fowl	*Broiler CD*	ADL0112	5
Bankiva	*Broiler CG*	ADL0114	8
Barnevelder A	*Broiler CH*	ADL0268	8
Barnevelder B	*Broiler CK*	ADL0278	7
Bearded Polish	*Broiler CO*		
Brabanter	*Broiler CP*	LEI0166	5
Breda fowl	*Broiler CQ*	LEI0228	26
Drents fowl	*Broiler CR*		
Dutch Bantam	*Broiler CT*	MCW0111	6
Dutch booted bantam	*Broiler CV*	MCW0014	11
Dutch Owl-bearded	*Broiler CZ*	MCW0150	8
Frisian fowl	*Broiler DA*	MCW0183	13
Groninger Mew	*Broiler DB*	MCW0248	10
Hamburgh	*Broiler DD*	MCW0295	8
Kraienkoppe	*Broiler DE*	MCW0330	6
Lakenvelder	*Broiler EE*	MCW0004	17
Non-bearded Polish	*Broiler GB*	MCW0067	7
Noord Hollands hoen		MCW0078	8
Sumatra	*Layer 17 (white)*	MCW0081	11
Welsummer	*Layer 18 (white)*		
	Layer 20 (white)		
	Layer 56 (white)		
	Layer 25 (brown)		
	Layer 26 (brown)		
	Layer 27 (brown)		
	Layer 29 (brown)		
	Layer 57 (brown)		

Per locus similarity scores were calculated from the allele frequencies (Eding and Meuwissen, 2001). We defined the population that existed just before this first fission as the founder population, in which all animals are unrelated. Analysis of the similarity scores indicated that the earliest detectable population fission was between the *Bankiva* and the cluster of broiler lines. The per locus average similarity between the *Bankiva* and the broiler cluster were assumed to be s (i.e. the probability of alleles Alike In State). MEKs between and within populations were calculated as the weighted average of kinship estimates per locus, where the standard errors of the estimates are used for weighing (Eding and Meuwissen, 2001).

RESULTS

Figure 3.1 is a contour plot of the 46 x 46 **M** matrix containing the MEKs. A schematic representation of the relations is given as a Neighbour-Joining tree in Figure 3.2. The tree was constructed using the Phylip package (Felsenstein, 1995). For the construction of this tree kinship estimates had to be converted to 'kinship distances' by:

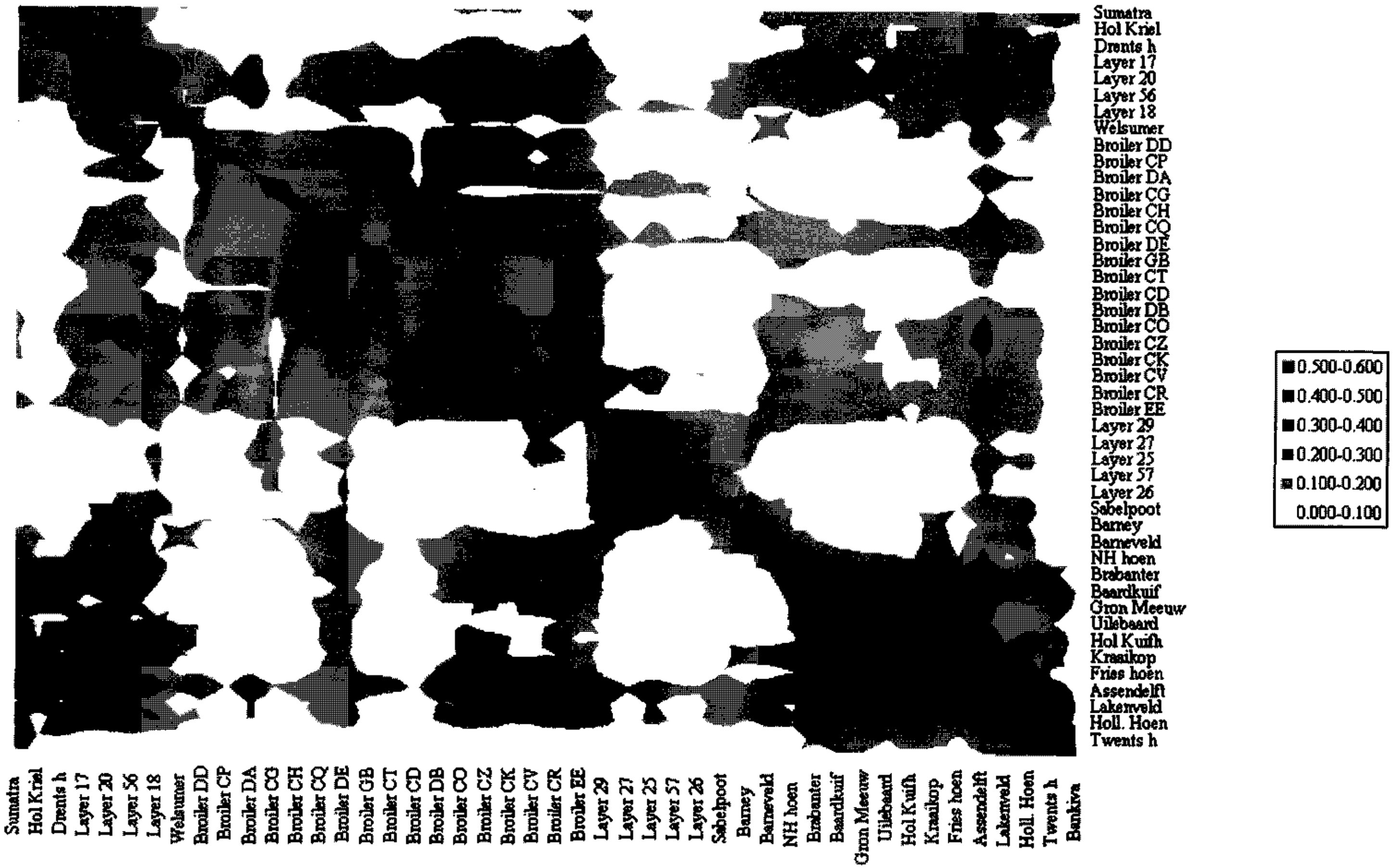
$$d(i,j) = \hat{f}_{ii} + \hat{f}_{jj} - 2\hat{f}_{ij}$$

Note that this distance is analogous to Nei's minimum distance. In the contour plot of Figure 3.1 the populations are ranked according to the dendrogram of Figure 3.2

The dendrogram resulting from the kinship distances shows three main clusters. The *Bankiva* breed, generally considered to be closely related to the ancestral population of all poultry breeds, constitutes one cluster, the *Sumatra* another. All the old Dutch fancy breeds and commercial lines are clustered together in what could be termed a 'Western cluster'. Within the Western cluster we see two separated clusters of layer lines and two closely related clusters of broiler lines. The distinction between the two clusters of broiler lines can be seen from the contour plot.

Figure 3.1, next page Contour plot of the estimated average population relation between 46 populations of poultry. Populations are ranked according to the clustering in the dendrogram in Figure 3.2. Shading is dependent on the value of the MEK.

Core sets


Sumatra
Hol Kriel
Drents h
Layer 17
Layer 20
Layer 56
Layer 18
Welsumer
Broiler DD
Broiler CP
Broiler DA
Broiler CG
Broiler CH
Broiler CQ
Broiler DE
Broiler GB
Broiler CT
Broiler CD
Broiler DB
Broiler CO
Broiler CZ
Broiler CK
Broiler CV
Broiler CR
Broiler EE
Layer 29
Layer 27
Layer 25
Layer 57
Layer 26
Sabelpoot
Barney
Barneveld
NH hoen
Brabanter
Baardkuif
Gron Meeuw
Uilebaard
Hol Kuifh
Kraaikop
Fries hoen
Assendelft
Lakenveld
Holl. Hoen
Twents h
Bankiva
■ 0.500-0.600
■ 0.400-0.500
■ 0.300-0.400
■ 0.200-0.300
▨ 0.100-0.200
0.000-0.100

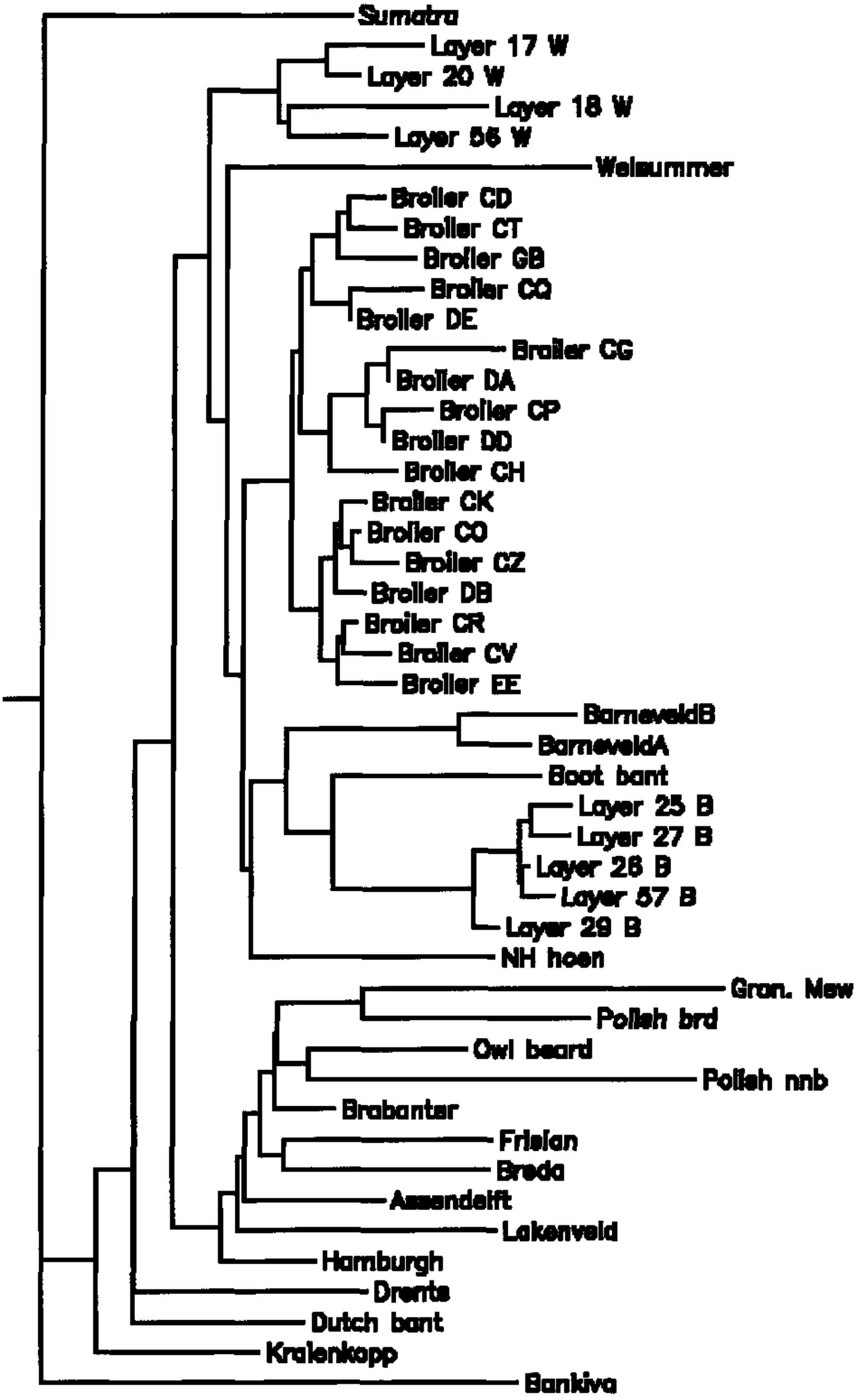

Sumatra
Layer 17 W
Layer 20 W
Layer 18 W
Layer 56 W
Welsummer
Broiler CD
Broiler CT
Broiler GB
Broiler CQ
Broiler DE
Broiler CG
Broiler DA
Broiler CP
Broiler DD
Broiler CH
Broiler CK
Broiler CO
Broiler CZ
Broiler DB
Broiler CR
Broiler CV
Broiler EE
BarneveldB
BarneveldA
Boot bant
Layer 25 B
Layer 27 B
Layer 26 B
Layer 57 B
Layer 29 B
NH hoen
Gron. Mew
Polish brd
Owl beard
Polish nnb
Brabanter
Frisian
Breda
Assendelft
Lakenveld
Hamburgh
Drente
Dutch bant
Kraienkopp
Bankiva

Figure 3.2, previous page Neighbour-Joining tree representation of relationships between 46 populations of poultry

The first cluster, comprised of broiler lines CD through CH, has a generally low kinship with the other populations in the set, whereas the second cluster (broiler lines CK to EE) is related not only to the first broiler cluster, but also to a cluster of Layer lines (Layer 17,20, 56 and 18) and a number of indigenous breeds. A similar pattern can be observed in the two clusters of layer lines. The cluster consisting of Layer lines 25, 26, 27, 29 and 57 (the brown layer lines) lines 17, 20, 56 and 18 (the white layer lines) are related to a cluster of indigenous breeds (the cluster beginning with *Groninger Mew* and ending with *Hamburgh*), apart from the relation with the aforementioned cluster of broiler lines.

Considering that the length of the branches correspond to extent of inbreeding, we can see from the tree representation, as well as in the contour plot, that there are a number of indigenous poultry breeds (e.g. *Welsummer, Noord Hollands hoen, Groninger Mew, Non-bearded Polish fowl, Assendelft*), that seem to suffer from higher levels of inbreeding than commercial lines. The within population MEK ranged from 0.17 to 0.28 for broiler lines, 0.29 to 0.42 for layer lines and 0.26 to 0.65 for Dutch indigenous breeds, averaging 0.24, 0.36 and 0.41 for broilers, layers and indigenous populations respectively.

There were a number of negative estimates of MEK, most notably for the *Bankiva* (MEKs with broiler lines), *Drents fowl* and *Welsummer* (both for MEKs with the brown layer lines). These negative estimates ranged from –0.01 to –0.06 and are caused by sampling errors on the kinships estimates (Appendix 3.B). Note that in the case of the *Bankiva* and broiler lines the between population similarity was used to estimate *s*, implying that their expected kinship is zero.

Results of the Core set method are given in Table 3.2. In the uncorrected solution we saw negative contributions. These arise when the **M**-matrix is not consistent, for instance when some within population kinship estimate is lower then any between population average kinship with the same population (see Appendix 3.B for a derivation of this result). To correct for these inconsistencies we iteratively removed the breed with the most negative contribution from the Core Set setting its contribution to zero, until all contributions were equal or greater then zero. This procedure results in the solution under $\mathbf{c}_{cor}$ (Table 3.2). Only populations with non-zero contributions are given.

Fourteen of the 46 populations received a contribution greater then zero. Six of these were commercial lines, while 7 Dutch indigenous breeds and the *Bankiva* also contributed to the core set. Contributions of commercial lines totalled 51%, while indigenous breeds contributed 37%. The broiler lines with non-zero contributions all stem from one of the two clusters of broilers, namely the cluster of broilers that is relatively isolated (see before). The layer lines with non-zero contributions also stem from one cluster: the brown layer cluster [25, 26, 27, 29 and 57], which was relatively more isolated.

Table 3.2 Optimal contributions to a core set of Dutch poultry populations $\mathbf{c}_{cor}$. **Div(M)** is the genetic diversity captured and is calculated as $1 - f_{cs}$, where f_{cs} is the average kinship in the core set.

Breed	$\mathbf{c}_{cor}$
Broiler CD	0.177
Broiler CP	0.167
Drents fowl	0.130
Bankiva	0.122
Layer 57	0.094
Dutch Bantam	0.094
Welsummer	0.066
Owl-bearded	0.056
Layer 26	0.043
Layer 27	0.024
Barneveld B	0.020
Booted bantam	0.005
Kraienkoppe	0.002
Broiler CH	0.001
Div(M)	**0.935**

Following Thaon d'Arnoldi *et al.* (1998) we defined a set of breeds that are not likely to become extinct (the **Safe** set, consisting of all commercial lines) and compare the diversity lost by only retaining this **Safe** set to the safe set plus one other breed (**Safe** + 1). This was done by comparing the diversity of the core set constructed from the **Safe** set with the diversity of the core set created from the **Safe** +1 set. Results are shown in Table 3.3. Genetic diversity was calculated in two ways: Div(M)= 1-f_{cs}, where f_{cs} is the average estimated kinship in the core set, and $N_{ge}= (2\,f_{cs})^{-1}$, where N_{ge} is the number of founder genome equivalents (Lacy, 1989) represented in the core set. Changes in Div(M) are directly related to changes in genetic variation of quantitative traits. Changes in

N_{ge} indicate the loss of founders represented in the core set, i.e. the potential loss of alleles and/or haplotypes.

In terms of Div(M) the loss in genetic diversity by keeping only the **Safe** set compared to keeping the entire set of populations is rather small: 4.5% (Table 3.3). The loss in founder genome equivalents is substantially higher: 39.3%. This pattern remains throughout the different **Safe** + 1 sets.

Of the populations not in the **Safe** set only the *Assendelft* showed a contribution of zero. This can be attributed to the relatively high estimated kinships with all other populations in the whole set (Figure 3.1). All other populations contributed moderately to substantially when added to the **Safe** set (Table 3.3). The contributions of breeds to the core set are not very closely related to the loss due to exclusion of the breed. For instance, inclusion of the *Hamburgh* incurs the same increase in diversity as inclusion of the *Barnevelder A*. However, its contribution is 33% higher: 0.121 for the *Hamburgh* versus 0.091 for the *Barnevelder A*.

From Table 3.3 the first four breeds (*Drents fowl, Dutch bantam, Bankiva* and *Kraienkoppe*) have large contributions to genetic diversity, both in terms of their relative contributions (c_{S+1}) and added genetic diversity, Div(M). Further down the list the contributions are markedly lower and the % losses markedly higher. Looking at Figure 3.2 we see that these four breeds have a distinct position in the dendrogram. They form clusters only with themselves and the average kinships with the other populations indicate these breeds are relatively older and/or more isolated.

Comparing the results from Table 3.2 with the results from Table 3.3, we see the top indigenous contributors are the same, although some reranking has occurred. However, in Table 3.2 both the *Barnevelder B* and the *Dutch Booted bantam* receive non-zero contributions, while in Table 3.3 they rank among the lowest in diversity contributed to the **Safe**+1 set.

Table 3.3 Relative loss in genetic diversity, when only a fixed set of breeds is kept (**Safe,** consisting of commercial broiler and layer lines) or the **Safe** set plus one other population. Div(M) is the genetic diversity and N_{ge} is the number of founder genome equivalents(Lacy, 1989) in the core set constructed from the populations in the indicated set. **Whole** is the entire set of 46 populations. Losses are calculated relative to either the genetic diversity or N_{ge} of the **Whole** set. $\mathbf{c_{s+1}}$ is the contribution of a population to the core set constructed from the appropriate **Safe** + 1 set.

Set	c_{S+1}	Div(M)	% loss	N_{ge}	% loss
Whole		0.935		7.69	
Safe only		0.893	4.49	4.67	39.25
Safe +1 set:					
Drents fowl	0.247	0.916	2.06	5.93	22.89
Dutch bantam	0.269	0.915	2.12	5.90	23.35
Bankiva	0.180	0.914	2.29	5.79	24.77
Kraienkoppe	0.241	0.911	2.60	5.60	27.21
Dutch Owl-bearded	0.168	0.902	3.49	5.12	33.40
Welsummer	0.157	0.902	3.57	5.08	33.94
Brabanter	0.167	0.900	3.70	5.02	34.74
Frisian fowl	0.132	0.900	3.72	5.01	34.87
Breda fowl	0.138	0.900	3.72	5.01	34.87
Polish bearded	0.115	0.899	3.82	4.97	35.45
Sumatra	0.106	0.899	3.88	4.94	35.83
Polish non-bearded	0.100	0.898	3.91	4.92	36.02
Groninger Mew	0.079	0.897	4.05	4.86	36.83
Lakenvelder	0.109	0.897	4.05	4.86	36.83
Hamburgh	0.121	0.895	4.24	4.78	37.86
Barnevelder A	0.091	0.895	4.24	4.78	37.86
Booted bantam	0.098	0.895	4.26	4.77	37.98
Barnevelder B	0.067	0.894	4.35	4.73	38.51
Noord-Hollands hoen	0.051	0.894	4.44	4.69	38.97
Assendelft	0.000	0.893	4.49	4.67	39.25

DISCUSSION

In principle the Core set method offers an alternative to the Weitzman approach in quantifying genetic diversity and support of decision making in conservation genetics. The Core set method has a number of advantages over the Weitzman method.

First, it is easy to use. Calculations in the Weitzman method are complex and time consuming, because of the recursive nature of the Weitzman method. The Core set method is a straightforward optimization procedure requiring less programming and computations. Additional programming is required if negative contributions need to be eliminated. But even then the calculations needed are a fraction of the calculations needed for the Weitzman approach. Also, the MEK/Core set method could be applied at the level of individuals, optimising the individual contributions to a conservation scheme. In contrast, the number of calculations needed in the Weitzman method limit the amount of data that can be used as input, thus preventing the Weitzman method from being used in larger conservation problems (Thaon d'Arnoldi *et al.*, 1998). The MEK/Core set method could also be extended to incorporate additional data, such as the economic valuation of genetic diversity, or data on additional considerations for conservation, such as socio-economic and traditional reasons. Alternatively, by using weights per marker locus one could place emphasis on the importance of certain genomic regions.

Second, the Core set method uses between and within breed diversity simultaneously. Within and between population diversity are measured in the same units (kinship) and the within breed diversity is weighed against the between breed diversity. This means that a relatively inbred population will receive a non-zero contribution if it is distant from all other populations in the set (i.e. low average between breed kinships). In the Weitzman method some additional weighing is needed to account for within breed diversity. Following Weitzman (1992) Thaon d'Arnoldi *et al.* (1998) suggest weighing with expected probabilities of extinction of each breed in the set. However, this suggestion could lead to results opposite from the Core set method. A highly inbred breed will receive a lower contribution in the Core set method. Because of the higher risk of extinction, following the suggestion by Thaon d'Arnoldi *et al.* such a breed would get a higher weight, increasing its priority in conservation decisions. Extinction risk could be accomodated in the Core set method by calculating the expectation of Div(M), where the expectation is taken over a vector **I** of indicator variables that indicates whether population i is expected to become extinct or not (I_i= 0 means population i will become extinct).

Third, using average population kinships is a natural way for measuring genetic diversity, because of its relationship with genetic variation. Average population kinships are closely related to well-know concepts as effective population sizes and inbreeding (Caballero and Toro, 2000). Most genetic distances used in the analysis of microsatellite data can be written in terms of kinships between and within population kinships (Eding and Meuwissen, 2001). Additionally, the MEK/Core set method closely links genetic diversity to variation in quantitative traits, putting less emphasis on the conservation of rare alleles and more on conservation of a wide range of genotypes.

Due to the nature of the kind of optimisation algorithm used in this study, relationships need only be known proportionally. Different definitions of the founder population (which is a major factor determining the values of Marker Estimated (Eding and Meuwissen, 2001)) will have no effect on the solution to the $\mathbf{c}_{\mathbf{min}}$ vector, which means that the composition of the core set does not change if the definition of the founder population changes (Appendix 3.A).

The tree representation in this paper was constructed using the Neighbor Joining method on 'kinship-distances' (which essentially is twice Nei's minimum distance corrected for allele-frequencies in the founder population). Generally this approach seems to give results that correlate well with the actual estimates of the average kinship coefficients (Figure 3.1). However, tree representations as in Figure 3.2 assume population fission and subsequent isolation and therefore do not show migration or crossbreeding patterns. A contour plot as given in Figure 3.1 is able to show patterns of gene flow. The combination of dendrogram and contour plot, where the dendrogram is used to determine the sorting order of the populations in the contour plot seems to give a clear image of both relatedness and gene flow between (clusters of) populations.

Overall, the kinship estimates and more specifically the low within breed kinship estimates (relative to the between breed estimates) suggest that migration between populations is quite large. In such situations the MEK/Core set method would seem to be preferable to other methods, since complete isolation of populations after fission is not assumed. Between populations kinships may be increased due to migration and the Core set method will account for the migration.

Sampling errors on the MEKs causes the **M**-matrix to contain inconsistencies that lead to negative contributions (Table 3.3; Appendix 3.B). In this paper we simply restricted populations with negative contributions to zero and recalculated the **c**-vector. We also used a genetic algorithm (Holland, 1975) to find a **c**-vector that minimized the kinship under the constraint that all contributions are equal or greater then zero. The resulting **c**-vectors were similar to the results from the culling procedure, where the culling procedure seemed to reach maximum diversity. The genetic algorithm, although always near maximum, did not achieve true maximum diversity (results not shown). The similarity between **c**-vectors from both methods seems to suggest that given the data the culling process gives the maximum diversity. Even so, the MEK/Core set method seems to be sensitive to inconsistencies in the tree, which lead to negative contributions. As these inconsistencies are caused by error variance on the MEKs (see Appendix 3.B), methods are needed that account for the error variance, making the method more robust and eliminating negative contributions (e.g. methods similar to the Bending method of Hayes and Hill, 1981). One straightforward solution is increasing the number of markers used in the estimation of kinship coefficients, reducing the variance of the estimates.

The per locus average similarity between the *Bankiva* and the broiler cluster were assumed to be s, because the genetic similarities between the *Bankiva* and the broiler clusters were lowest, indicating the oldest population fission. From Figure 3.2 we can see that this actually indicated the first population fission resulting in the *Bankiva* line and a line that was the ancestor to all 'Western' lines. This definition of s is somewhat ad-hoc. Other, more formal methods for the simultaneous estimation of f and s will be described in a subsequent paper.

The base population is assumed to be the population that might have existed at the time the population first split into two separate populations. The Core set method weighs the contributions of each breed in such a way that the genetic diversity in the base population is recovered as fully as possible. In the different sets for which solutions were calculated, genetic diversities ranged from 0.935 (full set) to 0.893 ('safe' set; see Table 3.3). The MEK/Core set method implicitly assumes a base population in which all individuals are unrelated and therefore Div(Base)=1.00. This suggests that the solutions to the **c**-vector conserve approximately 90% or more of the genetic variation of the hypothetical founder population. It may be noted that exclusion of a breed causes an adjustment of the contributions of the remaining populations in such a way that the loss in diversity is minimised. This readjustment uses the overlap in genetic diversity between breeds, increasing weights of breeds that are genetically related to the removed breed.

However, there remains a rather large loss in founder genome equivalence (23 – 39%) while the loss in genetic variation is small (2.0 – 4.5%). Given that (1-f) represents genetic variation that is conserved, it seems that even a substantial loss in founder genome equivalents (or related measures) has little effect on the amount of genetic variation retained in the present population. The actual relation between founder genome equivalents and genetic diversity is not clear and therefore it is difficult to indicate how much diversity is lost when 39% of the founder genome equivalents are lost.

The results from the MEK/Core set method seem promising. Application of the method is flexible and not computer intensive. According to the results presented in this paper it is possible to conserve most of the genetic diversity originally found in the founder population. The definition of total genetic diversity as the complement of the minimum average kinship in a set of breeds obeys the criteria set by Weitzman (1992). The MEK/Core set method employed in this paper provides a clear ranking of breeds according to their 'diversity content', both relative to the entire set and relative to alternative sets (in this study the **Safe** set).

The **c**-vector could also be used to allocate resources to a gene bank. But such an approach carries the risk that some breeds will be allocated insufficient resources to maintain it as an independent, viable population. In those cases crossbreeding might be used to conserve the diversity of breeds. However, this could mean the loss of valuable genotype and allele combinations that need to be conserved. Ultimately, the decision to conserve a breed is dependent on a number of considerations of which genetic diversity in the terms presented in this paper is only one (Oldenbroek, 1999; Ruane, 1999).

APPENDIX 3.A

Invariance of the contributions vector to probability of alleles Alike In State (AIS).

We consider the set M of m populations. Suppose **A** is an m×m matrix containing the actual (unknown) kinships between populations. The vector containing optimal contribution to the core set should be calculated through:

$$(1) \qquad \mathbf{c}_{min} = \frac{\mathbf{A}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{A}^{-1}\mathbf{1}}$$

Where **1** is an m-dimensional vector of ones.

However, for a locus L where alleles can be alike in state without being identical by descent the similarity matrix $\mathbf{M}_L$ will be of the form:

$$(2) \qquad \mathbf{M}_L = (1 - s_L)\mathbf{A} + \mathbf{11}'s_L$$

Where s_L is the probability of alleles being alike in state but not identical by descent (Eding and Meuwissen, 2001). Substituting the similarity matrix $\mathbf{M}_L$ for **A**, expression (1) changes into:

$$(3) \qquad \hat{\mathbf{c}}_{min} = \frac{\mathbf{M}_L^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{M}_L^{-1}\mathbf{1}}$$

For the calculation of the estimate of $\mathbf{c}_{min}$ we need the inverse of $\mathbf{M}_L$. Setting $\mathbf{M} = (1 - s_L)\mathbf{A}$, we get:

$$(4) \qquad \begin{aligned} \mathbf{M}_L^{-1} &= \left[\mathbf{M} + \mathbf{11}'s_L\right]^{-1} \\ &= \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{1}\left[\mathbf{1}'\mathbf{M}^{-1}\mathbf{1} + s_L^{-1}\right]^{-1}\mathbf{1}'\mathbf{M}^{-1} \end{aligned}$$

Multiplication by **1** gives:

$$(5) \qquad \begin{aligned} \mathbf{M}_L^{-1}\mathbf{1} &= \mathbf{M}^{-1}\mathbf{1} - \mathbf{M}^{-1}\mathbf{1}\left[\mathbf{1}'\mathbf{M}^{-1}\mathbf{1} + s_L^{-1}\right]^{-1}\mathbf{1}'\mathbf{M}^{-1}\mathbf{1} \\ &= \mathbf{M}^{-1}\mathbf{1}\left[1 - \frac{\mathbf{1}'\mathbf{M}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{M}^{-1}\mathbf{1} + s_L^{-1}}\right] \end{aligned}$$

Substituting (5) in (3) and substituting $\mathbf{M} = (1 - s_L)\mathbf{A}$ we see that

$$(6) \qquad \hat{\mathbf{c}}_{min} = \frac{\mathbf{M}_L^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{M}_L^{-1}\mathbf{1}} = \frac{\mathbf{A}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{A}^{-1}\mathbf{1}} = \mathbf{c}_{min}$$

The vector $\mathbf{c}_{min}$ is insensitive to the probability of alleles AIS, provided this probability is equal for all populations in M. This holds true for probabilities of alleles AIS in general. If estimates of f_{ij} are made, correction will take place for the probabilities of alleles being alike in state at different loci. However, inherently there will be some probability of alleles AIS left because we implicitly assume a founder population, where the relations among animals and inbreeding are zero. The above shows that the choice of founder population will not affect the contributions of populations to the core set.

APPENDIX 3.B

Negative contributions of breeds to the Core set occur as a result of the quality of the data used in the analysis. Large variation on the Marker Estimated Kinships (MEKs) may lead to inconsistencies in the resulting **M** matrix, such as a population having a within population kinship that is smaller than their kinships with other populations.

For a MEK matrix **M** the contribution vector $\mathbf{c}_{\mathbf{min}}$ that minimises the average kinship in a set of N populations is:

$$\mathbf{c}_{\min} = \frac{\mathbf{M}^{-1}\mathbf{1}_{n}}{\mathbf{1}_{n}'\mathbf{M}^{-1}\mathbf{1}_{n}}$$

where $\mathbf{1}_{\mathbf{n}}$ is a vector whose N elements equal one.

Suppose population P has a negative contribution. The inverse of the **M** matrix may be partitioned as:

$$\mathbf{M}^{-1} = \left[\begin{array}{cc:c} & & \\ & \mathbf{M}_{11} & \mathbf{M}_{1P} \\ \cdots & \cdots & \cdots \\ & \mathbf{M}_{P1} & \mathbf{M}_{PP} \end{array}\right]^{-1}$$

$$= \left[\begin{array}{cc:c} & & \\ & & \\ \cdots & \cdots & \cdots \\ & -\mathbf{Q}^{-1}\mathbf{M}_{P1}^{-1}\mathbf{M}_{11}^{-1} & \mathbf{Q}^{-1} \end{array}\right]$$

and

$$\mathbf{Q}^{-1} = \mathbf{M}_{PP} - \mathbf{M}_{P1}\mathbf{M}_{11}^{-1}\mathbf{M}_{1P}$$

where $\mathbf{M}_{11}$ is the (N-1) x (N-1) partition of matrix of **M** in which population P is excluded. $\mathbf{M}_{1P}$ is a vector containing the MEKs between P and all other populations. $\mathbf{M}_{PP}$ is the within population kinship estimate of P.

The contribution of P follows from formula (5) in the main text. The contribution of population P is negative, when the P-th element of $\mathbf{M}^{-1}\mathbf{1}<0$.

$$\begin{aligned}
\left[\mathbf{M}^{-1}\mathbf{1}\right]_P &= -\mathbf{Q}^{-1}\mathbf{M}_{P1}\mathbf{M}_{11}^{-1}\mathbf{1}+\mathbf{Q}^{-1}\mathbf{1} \\
&= \mathbf{Q}^{-1}\left(\mathbf{1}-\mathbf{M}_{P1}\mathbf{M}_{11}^{-1}\mathbf{1}\right) \\
&= \mathbf{Q}^{-1}\left(\mathbf{1}-\mathbf{M}_{P1}\cdot\frac{\mathbf{M}_{11}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{M}_{11}^{-1}\mathbf{1}}\cdot\mathbf{1}'\mathbf{M}_{11}^{-1}\mathbf{1}\right) \\
&= \mathbf{Q}^{-1}\left(\mathbf{1}-\frac{\mathbf{M}_{P1}\mathbf{c}_{\min,11}}{f_{\min,11}}\right)
\end{aligned}$$

Where $\mathbf{c}_{\mathrm{min},11}=\dfrac{\mathbf{M}_{11}^{-1}\mathbf{1}_n}{\mathbf{1}_n'\mathbf{M}_{11}^{-1}\mathbf{1}_n}$ is the optimum contributions vector of populations 1 to (N-1) to their core set (A) and $f_{\min,11}=\dfrac{1}{\mathbf{1}'\mathbf{M}_{11}^{-1}\mathbf{1}}$ is the average kinship within the core set A

From this we see that the contribution of population P is smaller then zero if

$$\mathbf{Q}^{-1}<0 \quad \text{or} \quad \mathbf{M}_{P1}\mathbf{c}_{\mathrm{min},11}>f_{\mathrm{min},11}$$

where $\mathbf{Q}^{-1}$ is a diagonal element of $\mathbf{M}^{-1}$. Since **M** is a relationship matrix and therefore a variance/covariance matrix, **M** must be positive definite. $\mathbf{Q}^{-1}<0$ indicates that **M** is not positive definite and thus not a proper relationship.

The scalar $f_{\mathrm{min},11}$ is the average minimal kinship within core set A, composed of all populations except population P. $\mathbf{M}_{P1}\mathbf{c}_{\mathrm{min},11}$ is the minimal average kinship between population P and the composite population A. A negative contribution therefore occurs when the kinship between populations P and A is greater then the kinship within population A. Since the true within population kinship is always greater or equal then the true between population kinship, the occurrence of negative contributions is due to sampling errors on the MEKs in **M**.

Chapter 4

ESTIMATION OF MARKER BASED KINSHIPS TO CONSTRUCT CORE SETS FOR GENE BANKS

Eding, Herwin[*)] and Theo H.E. Meuwissen

Institute for Animal Science and Health, Box 65, 8200 AB Lelystad, The Netherlands

Genetics Selection Evolution (Submitted)

[*)] Institute for Animal Science and Health

Box 65, 8200 AB Lelystad, The Netherlands

Tel: +31-(0)320-238238; Fax +31-(0)320-238050

E-mail: j.h.eding@id.wag-ur.nl

ABSTRACT

Three log-linear models were developed to improve the estimates of kinships between breeds (MEK) and of Alike In State probabilities (AIS) using all marker data and all pairs of animals simultaneously. These models were developed to 1) increase accuracy of MEK, 2) improve AIS estimates (especially compared to methods that simply take average allele ferquencies) and 2) to reduce the number of zero contributions of breeds with actual contributions larger than zero. The models are: Unweighted Log-linear Model (ULM), Weighted Log-linear Model (WLM), where marker data is weighted to account for the amount of information per locus and Weighted Log-linear Mixed Model (WLMM), where the solution is restricted such that a maximum of one zero-contribution remains. These models were tested using simulated data and compared to the results from the Weighted Least Similarity, where the per locus probabilities of alleles Alike in State (AIS) are taken from the similarities between the pair of populations with the minimum average similarity. An example using field data on 10 cattle populations in the Netherlands is discussed. Differences in accuracy between the four methods were small, although substantial differences in contribution of breeds to the core set were found. In terms of conserved variation WLM was the most efficient, followed by WLMM. WLMM yielded the smallest number of zero contributions of breeds and provides a more conservative solution (i.e. fewer breeds will be erroneously excluded).

INTRODUCTION

Genetic diversity is considered to be important for the survival of species. With the continued expansion of only a few breeds per species of livestock, methods to measure and conserve genetic variation in livestock have received considerable attention. Average kinships describe genetic diversity in terms of probabilities that alleles are Identical By Descent (Malécot, 1948) and thus indicate 'genetic variability' or allelic richness (Caballero and Toro, 2000). Kinships also have a linear relationship with genetic variation in quantitative traits (Falconer and Mackay, 1996). Average kinships can be calculated between and within populations. Coefficients of kinship are typically calculated from pedigree records. However, lack of pedigree records, especially those that describe breed formation, necessitates retrieving kinship estimates from other sources of information, which often results in the use of genetic marker information.

Eding and Meuwissen presented a method of estimating average kinships between populations from similarities of genetic marker alleles (Eding and Meuwissen, 2001a). Unbiased estimation of kinship from marker data depends on an accurate correction for probabilities of alleles alike in state (AIS) for each locus, i.e. the allele frequencies in the founder population. Eding and Meuwissen suggested setting AIS equal to the per locus similarity indices of the pair of populations with the lowest average genetic similarity, which is equal to the within population similarity of the parent population just prior to fission. The latter is because expectation of the similarity between two populations remains unchanged after population fission. Hence, if we assume a founder population from which all populations in a set descended, the lowest average similarity is an estimator of the within population similarity at the time of the first population fission. Note that only two populations are used to estimate the AIS probability, namely those with the lowest marker similarity. Also, simply calculating AIS from the average of the allele frequencies of all populations (Slatkin, 1995) will yield a biased estimate when some populations are more related than others.

Eding *et al.* (Eding *et al.*, 2001) presented a method of ranking populations according to their contributions to a core set, in which the overlap in genetic variation is minimised. To this end a matrix **M** is constructed containing the average Marker Estimated Kinship (MEK) between and within populations. Subsequently, relative contributions of each population are calculated in such a way, that the average Marker Estimated Kinship in the core set is minimal.

However, due to sampling errors on the marker similarities, it is possible that the matrix **M** is not a positive definite relationship matrix, causing zero or negative contributions. Eding *et al.* showed that all contributions to a core set will be positive or zero when a proper relationship matrix is used and the set of populations does not contain genetic influences from populations not represented in the set. When negative contributions were encountered, contributions of populations that were most negative were set to zero and optimal contributions recalculated, until all contributions were equal to or greater then zero. Nevertheless, many zero contributions of breeds indicates sampling error on the matrix of kinships, and implies erroneously that many breeds do not contain any genetic diversity.

A simultaneous estimation of Marker Estimated Kinships and probabilities of alleles AIS from all data available could give more accurate estimates on both MEK and probabilities of alleles AIS. Increased accuracy in the MEKs could also reduce or even eliminate the occurrence of zero contributions to a core set. We will explore methods of simultaneous estimation of MEKs and probabilities of alleles AIS. A log-linear transformation of the similarity index yields a linear model with which the data can be analysed using standard linear model techniques. Implementation of such techniques has the added advantage of making use of all available data to infer the homozygosity in the founder population, i.e. the probabilities of alleles AIS.

Introducing weights in the linear model takes account of the information content of the loci. Furthermore, a mixed model method for estimating kinships is proposed, which is designed to further reduce the number of zero contributions. We will demonstrate the method using simulated data, which makes comparison to the actual kinships and AIS possible and using a field data set.

The aim of this study is to find a method that can minimises the errors on Marker Estimated Kinships and thus give more accurate kinship estimates, which in turn should lead to a more efficient conservation of genetic variation.

MATERIAL AND METHODS

The similarity index used is based on the probability that two marker alleles drawn from two individuals of populations i and j are identical. On the level of populations, the similarity index (i.e. the average similarity between populations) for a locus *l* can be expressed as (Lynch, 1988; Eding and Meuwissen, 2001a):

$$\text{(1a)} \qquad S_{ij,l} = \frac{\sum_{k=1}^{n_{ij}} S_{ij,l}(k)}{n_{ij}}$$

where $S_{ij,l}$ is the similarity between populations *i* and *j* for locus *l*, and n_{ij} is the number of pairs of individuals and $S_{ij,l}(k) = \frac{1}{4}\left[I_{11}(k) + I_{12}(k) + I_{21}(k) + I_{22}(k)\right]$, where I_{xy} is an indicator variable that is 1 when allele *x* in one individual and allele *y* in a second individual are identical and 0 otherwise (Eding and Meuwissen, 2001a).

For a pair of populations i and j the expected similarity between populations *i* and *j* for a locus *l* is (Bernardo, 1993):

$$\text{(1b)} \qquad \mathrm{E}(S_{ij,l}) = f_{ij} + (1 - f_{ij})s_l$$

where f_{ij} is the average kinship between *i* and *j* and s_l is the probability of alleles AIS for locus *l*. Note that the kinship between population *i* and *j* is expected to be equal for all loci and that the probability of alleles AIS for locus *l* is expected to be equal for all pairs of populations.

Weighted Least Similarity

The Weighted Least Similarity is the model in which the per locus similarities of the least similar pair of populations are taken as estimates of s_l (the probability of alleles AIS) (Eding and Meuwissen, 2001a). Kinships are then obtained from the weighted mean of similarities for a pair of populations after correction for s_l. This model will be referred to as the Weighted Least Similarity model or WLS model.

Log-linear regression

Given the relation in formula (1b) we can construct a log-linear regression model. Subtracting both the left and the right-hand side of (1b) from 1 and applying a logarithmic transformation yields:

(2)
$$\begin{aligned}\ln\left(1-S_{ij,l}\right) &= \ln\left[\left(1-f_{ij}\right)\left(1-s_l\right)\right]+error_{ij,l}\\ &= \ln\left(1-f_{ij}\right)+\ln\left(1-s_l\right)+error_{ij,l} \quad \Leftrightarrow\\ y_{ij,l} &= a_{ij}+b_l+error_{ij,l}\end{aligned}$$

where $y_{ij,l}$ = ln(1-$S_{ij,l}$); a_{ij} = ln(1-f_{ij}) and b_l = ln(1 – s_l). In matrix notation and including the similarity data of ½N(N+1) breed combinations for L marker loci, formula (2) becomes:

(3)
$$\mathbf{y}=\mathbf{X_a a}+\mathbf{X_b b}\ +\mathbf{e}$$

where **y** is a vector of ½N(N + 1)L elements containing the ln(1 – $S_{ij,l}$) per combination of populations per locus, the vectors **a** and **b** contain the effects of ln(1-f_{ij}) and ln(1 – s_l) respectively. $\mathbf{X_a}$ and $\mathbf{X_b}$ are design matrices for the **a** and **b** estimates.

Hence the estimates of **a** and **b** can simultaneously be calculated from

(4)
$$\begin{bmatrix}\hat{\mathbf{a}}\\ \hat{\mathbf{b}}\end{bmatrix}=\begin{bmatrix}\mathbf{X_a'X_a} & \mathbf{X_a'X_b}\\ \mathbf{X_b'X_a} & \mathbf{X_b'X_b}\end{bmatrix}^{-}\begin{bmatrix}\mathbf{X_a'y}\\ \mathbf{X_b'y}\end{bmatrix}$$

Because of dependencies in the design matrices, a generalised inverse of the coefficient matrix is used to get solutions for **a** and **b**, which are consequently not unique. The solution to **a** can be restricted in such a way that the smallest between population MEK is set to zero (Eding *et al.*, 2001).

We can rewrite (3) as

(5)
$$\begin{aligned}y_{ijk} &= a_{ij}+b_k+error_{ijk}\\ &= \left(a_{ij}-\alpha\right)+\left(b_k+\alpha\right)+error_{ijk}\end{aligned}$$

Since a_{ij} = ln (1 - f_{ij}), α is set to MAX(a_{ij}). The maximum value of a_{ij} yields the highest ln(1 – f_{ij}) and thus the smallest f_{ij}, which is set to zero by subtracting α from the a_{ij} values. We will refer to this model as the Unweighted Log-linear Model (ULM).

Weighted log-linear regression

Estimation errors of the solutions may be reduced by accounting for the error variances of the data y_{ijk} by weighing each similarity score per locus with the expected variance on the similarity score (Eding and Meuwissen, 2001a). This results in more informative markers having a larger influence on the solutions of both *f* and *s*, while less informative markers have less influence. The model equations are:

$$(6) \qquad \begin{bmatrix} \mathbf{X}_a'\mathbf{W}^{-1}\mathbf{y} \\ \mathbf{X}_b'\mathbf{W}^{-1}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_a'\mathbf{W}^{-1}\mathbf{X}_a & \mathbf{X}_a'\mathbf{W}^{-1}\mathbf{X}_b \\ \mathbf{X}_b'\mathbf{W}^{-1}\mathbf{X}_a & \mathbf{X}_b'\mathbf{W}^{-1}\mathbf{X}_b \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}$$

where all observations are weighted by the matrix $\mathbf{W}^{-1}$, where the diagonals of $\mathbf{W}$ contain the error variances of each observation. Appendix 4.A shows that

$$(7) \qquad \text{var}\left[\ln\left(1 - S_{ij,l}\right)\right] \approx \frac{\hat{f}_{ij} + (1 - \hat{f}_{ij})\hat{s}_l}{4n_{ij}(1 - \hat{f}_{ij})(1 - \hat{s}_l)}$$

Since all observations are assumed to be independent, the off-diagonals are zero. Note that the weights are determined by the estimates. Hence, the equations in (6) need to be solved iteratively, until convergence is reached. Since only the relative values of the weights are important, the factor $4n_{ij}$ may be removed when all pairs of populations contribute an equal number of observations n_{ij}. We will refer to this model as the Weighted Log-linear Model (WLM).

A mixed model method to obtain positive contributions

Here, the within and between population MEK are regressed towards their respective means, using Mixed Model methods. To this end the means of the between and within population MEK are introduced as an extra term in (3).

$$(8) \qquad \mathbf{y} = \mathbf{Z}\boldsymbol{\mu} + \mathbf{X}_a\mathbf{a}^* + \mathbf{X}_b\mathbf{b} + \mathbf{e}$$

where μ is a 2 x1 vector of estimates of the between and within population MEK on the log scale; $\mathbf{Z}$ is a ½ N(N+1) x 2 design matrix for the means μ. The vectors $\mathbf{a}^*$ contains deviations from the mean such that $\mathbf{a} = \mathbf{Z}\boldsymbol{\mu} + \mathbf{a}^*$.

The model equations are:

$$(9) \qquad \begin{bmatrix} \mathbf{Z}'\mathbf{W}^{-1}\mathbf{y} \\ \mathbf{X}_a'\mathbf{W}^{-1}\mathbf{y} \\ \mathbf{X}_b'\mathbf{W}^{-1}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{W}^{-1}\mathbf{Z} & \mathbf{Z}'\mathbf{W}^{-1}\mathbf{X}_a & \mathbf{Z}'\mathbf{W}^{-1}\mathbf{X}_b \\ \mathbf{X}_a'\mathbf{W}^{-1}\mathbf{Z} & \mathbf{X}_a'\mathbf{W}^{-1}\mathbf{X}_a + \lambda\mathbf{I} & \mathbf{X}_a'\mathbf{W}^{-1}\mathbf{X}_b \\ \mathbf{X}_b'\mathbf{W}^{-1}\mathbf{Z} & \mathbf{X}_b'\mathbf{W}^{-1}\mathbf{X}_a & \mathbf{X}_b'\mathbf{W}^{-1}\mathbf{X}_b \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{a}^* \\ \mathbf{b} \end{bmatrix}$$

where the term $\lambda\mathbf{I}$ is included to regress the deviations from the means, $\mathbf{a}^*$, back to zero, such that $\mathbf{a}$ is regressed back to its mean. The extent to which $\mathbf{a}^*$ is regressed to zero depends on λ. As with WLM, this system of equations is solved iteratively, in order to find the weights. The factor λ is chosen such that the smallest contribution to the core set equals zero, i.e. all contributions are larger or equal to zero. We will refer to this model as the Weighted Log Linear Mixed Model (WLMM).

After obtaining solutions to the equations kinships were transformed back through

(10) $$\mathbf{f} = 1 - e^{\mathbf{a}}$$

where **f** contains the (½)N(N+1) kinship estimates. These were subsequently used to construct the estimated kinship matrix $\hat{\mathbf{F}}$. Vectors containing the contributions to the core set were then calculated through (Eding *et al.*, 2001):

(11) $$\mathbf{c} = \frac{\mathbf{1}'\hat{\mathbf{F}}^{-1}}{\mathbf{1}'\hat{\mathbf{F}}^{-1}\mathbf{1}}$$

Simulated data

Per replicate a phylogeny was simulated over 50 generations during which 10 population formed from a single base population. The base population consisted of 50 individuals. For each individual 20 autosomal, unlinked loci were simulated. At the start each locus was randomly assigned a number of alleles ranging from 10 to 50. The base population was then allowed to breed undisturbed for at least 10 generations, in order to create more realistic allele frequency distributions. The number of alleles chosen for the initial (pre-founder) generations was chosen in such a way, that after 50 generations a reasonable number of alleles still segregated in the population, although fixation of alleles could not be prevented without resorting to very unrealistic numbers of alleles per locus.

New populations were formed by fission from a randomly drawn existing population at a randomly drawn time between generation 10 and 50. The size of each population was kept constant and was allowed to range from 20 to 100 random breeding individuals, half of which were male and half were female. Each next generation was generated by randomly assigning a sire and a dam from the previous generation as parents of each member of the new generations. In the case of population fission, parents of individuals in the new population were drawn from the parent population.

Throughout the simulated phylogeny pedigree data were recorded, which served to calculate the real (average) kinship coefficients between individuals (populations). In the analyses the average kinship estimates in each replicate were corrected through

$$f_{ij}' = \frac{f_{ij} - f_{\min}}{1 - f_{\min}}$$

where f_{ij}' is the population kinship corrected such that the population that existed just prior to first fission is assumed to be the founder population.

For each locus genetic similarities between all individuals in generation 50 were calculated and averaged to obtain per locus between population similarities. Subsequently, kinship estimates and contribution vectors were produced using the methods described above.

Additionally, per replicate a population kinship matrix **F** was constructed from pedigree data and contributions vectors **c** were calculated from this **F** matrix.

The four models described above were evaluated relative to the results from the pedigree-data, which were taken to be the 'true' results. The two models that required iterative solutions (WLM and WLMM) were run until the average absolute differences between the estimates of kinships in two subsequent iterations was $<10^{-8}$.

All models were applied to 4 cases ranging from 10 populations (N=10) typed for 10 loci (L=10) to 20 populations (N=20) typed for 20 loci (L=20). Where the number of loci was chosen to represent the upper and lower limit of numbers of loci generally used in genetic diversity studies.

For reference we added two cases, with N=10 and N=20 for 100 unlinked loci (L=100), where the number of alleles per locus was set to infinity, i.e. each individual in generation 0 received a unique set of alleles.

Application to field data

As an illustration of the standard use of the methods presented in this paper, we present here the results from a data set containing microsatellite data from Dutch populations of 10 cattle breeds. These 10 populations were genotyped for 11 microsatellites. Per locus similarity scores were calculated from the allele frequencies (Eding and Meuwissen, 2001a). The data is summarised in Table 4.1.

The Dutch Friesian is a black and white dual-purpose breed originating from the northern provinces of the Netherlands. Animals from this breed imported into the United States of

Table 4.1 Summary of the data on Dutch populations of cattle breeds.

Breeds	**N**	**Marker loci**	**# alleles**
Belgian Blue	210	BM1824	7
Dutch Red Pied	388	BM2113	12
Dutch Black Belted	90	ETH0010	9
Limousine	616	ETH0225	8
Holstein Friesian	2031	ETH0003	11
Galloway	402	INRA023	11
Dutch Friesian	417	SPS0115	7
Improved Red Pied	287	TGLA0122	23
Blonde d'Aquitaine	758	TGLA0126	8
Heck	169	TGLA0227	14
		TGLA0053	17

America contributed to the Holstein Friesian breed. The Dutch Black Belted is another Dutch dual purpose breed, that is capable of maintaining production when quality of feed is low (Felius, 1995). The Improved Red Pied breed is a beef breed developed from the dual-purpose Dutch Red Pied or Meusse-Rhine-Yssel cattle from the southern and eastern provinces of the Netherlands. The Heck auroch population in the data set is a (semi-)feral population kept in the province of Zeeland for landscape maintenance and nature development purposes. This population was introduced in Zeeland in 1983, with 25 founder animals (11 bulls and 14 cows), originating from two lines from former East Germany and Austria, respectively (de Bie and Bokdam, 1989). The Heck breed originally is a combination of four breeds, among them Spanish Fighting cattle and Corsican cattle (Felius, 1995). The animals of breeds not from Dutch origin (Belgian Blue, Limousine, Blonde d'Aquitain and Galloway) are all registered in the Dutch herd books of these breeds, although some gene flow from the country of origin is to be expected (notably for the Belgian Blue and the Limousine).

RESULTS

Simulated data

Tables 4.2 and 4.3 compare estimates from the WLS, ULM, WLM and WLMM model to those from the pedigree based kinship matrix. All methods overestimate the kinship coefficients ($\hat{\mathbf{F}} - \mathbf{F}$, i.e. estimates minus actual average kinships), where on average ULM has the highest bias, followed by the WLM in most cases. On average the bias was smallest for the WLMM, although in the cases where N=10, the WLS method performed equal or better then the WLMM. Disregarding the reference cases (L=100), under WLS, ULM and WLM the bias was smallest when N=10 and L=20 and largest when applied to N=20 and L=20. The bias is introduced by the restriction that all solutions are ≥0, which effectively sets the smallest kinship estimate to zero. Assuming a tree-like phylogeny and isolation of one main cluster from the other, the true kinship between a population from one cluster and a population taken from the other will be equal for all such breed combinations. The expectation of the estimate of these between breed kinships is zero, but the sampling error will introduce variance among these kinship estimates. Setting the smallest similarity from these breed combinations to zero will overestimate all other kinships, since they will all be positive. In essence the AIS probability is not estimated at the point of first population fission but somewhat prior to this point.

Correlations (r_F) between actual and estimated kinships were high for all models and increasing with number of loci considered. When N=10 the Weighted Least Similarity method outperformed the log-linear models in this regard, but for N=20 WLM showed a somewhat larger correlation on average for limited numbers of loci (L=10 and L=20). Generally however, differences in r_F between models were not significant ($p>0.05$).

For all models the correlation between estimated and actual contribution vectors (r_c) were moderate with large standard deviations, resulting in non-significant differences between models. Of the four models the WLM yielded on average the highest correlations, which was mainly due to the higher correlation when N=20.

The number of null-contributions in the corrected contribution vector, $n(c_i=0)$ (Tables 2 and 3), is an indication of how many breeds could not be ranked. The average number of null-contributions decreased from L=10 to L=100 for the models WLS, ULM and WLM, as was expected from the increased accuracy of the estimate of the **F** matrix. The number of null-contributions seems to be proportional to the number of population combinations. In the WLMM, the parameter λ was chosen such that at most only one contribution equals zero,

Table 4.2 Correlations between estimated and true of contribution vectors c and kinship matrices F with N=10 populations and L=10, 20 or 100 loci. $\hat{\mathbf{F}}-\mathbf{F}$ shows the bias of the kinship estimates, $\mathbf{r}_F$ and $\mathbf{r}_c$ are the correlations between estimated and true average kinship coefficients and between true and estimated contributions, respectively, $n(c_i = 0)$ shows the average number of null-contributions per replicate and std (c) shows the standard deviations of the contribution within replicates. $1-\hat{\mathbf{c}}'\mathbf{F}\hat{\mathbf{c}}$ shows the conserved genetic diversity when estimated contribution vectors are applied to the actual kinship matrix.

L=10	$\hat{\mathbf{F}}-\mathbf{F}$	$\mathbf{r}_F$	$\mathbf{r}_c$	$n(c_i = 0)$	std (**c**)	$1-\hat{\mathbf{c}}'\mathbf{F}\hat{\mathbf{c}}$
True	0.000	1.000	1.000	0.10	0.067	0.831
WLS	0.037	0.949	0.564	2.25	0.101	0.823
ULM	0.065	0.922	0.587	2.90	0.108	0.821
WLM	0.055	0.931	0.550	2.50	0.102	0.823
WLMM	0.035	0.938	0.572	$0.75^{+)}$	0.073	0.823
L=20						
True	0.000	1.000	1.000	0.10	0.067	0.831
WLS	0.015	0.970	0.645	1.60	0.089	0.826
ULM	0.058	0.951	0.637	2.10	0.098	0.825
WLM	0.035	0.960	0.634	1.60	0.088	0.827
WLMM	0.024	0.958	0.649	$1.00^{+)}$	0.072	0.826
L=100						
True	0.000	1.000	1.000	0.00	0.063	0.834
WLS	0.003	0.995	0.876	0.80	0.071	0.833
ULM	0.026	0.986	0.860	1.00	0.076	0.833
WLM	0.016	0.990	0.886	0.50	0.066	0.833
WLMM	0.014	0.989	0.889	$0.40^{+)}$	0.064	0.833

$^{+)}$ # $(c_i < 0.0001)$ repl.$^{-1}$

Table 4.3 Same as Table 4.2, except the number of populations N=20.

L=10	$\hat{\mathbf{F}}-\mathbf{F}$	$\mathbf{r}_\mathrm{F}$	$\mathbf{r}_\mathrm{c}$	$n(\mathrm{c}_\mathrm{i}=0)$	std ($\mathbf{c}_\mathbf{cor}$)	$1-\hat{\mathbf{c}}'\mathbf{F}\hat{\mathbf{c}}$
True	0.000	1.000	1.000	0.10	0.047	0.853
WLS	0.042	0.897	0.548	8.85	0.071	0.842
ULM	0.104	0.893	0.580	9.10	0.075	0.843
WLM	0.109	0.908	0.606	8.55	0.071	0.845
WLMM	0.014	0.878	0.583	$0.70^{+)}$	0.029	0.842
L=20						
True	0.000	1.000	1.000	0.10	0.047	0.853
WLS	0.033	0.949	0.656	6.85	0.065	0.846
ULM	0.081	0.936	0.662	7.55	0.070	0.846
WLM	0.055	0.949	0.687	6.35	0.063	0.848
WLMM	0.014	0.912	0.670	$1.00^{+)}$	0.031	0.845
L=100						
True	0.000	1.000	1.000	0.00	0.039	0.951
WLS	0.008	0.993	0.846	3.05	0.045	0.949
ULM	0.029	0.983	0.844	3.80	0.049	0.949
WLM	0.019	0.985	0.851	2.45	0.043	0.949
WLMM	0.006	0.969	0.870	$0.80^{+)}$	0.032	0.948

$^{+)}$ # ($\mathrm{c}_\mathrm{i} < 0.0001$) repl.$^{-1}$

hence WLMM number of null-contributions were either 1 or 0 (the latter is the case when λ=0 yields no negative contributions).

The average standard deviations of contributions per replicate was highest for the ULM and lowest for WLMM. The low standard deviations of contributions from the WLMM were expected due to the fact that bending regresses kinship estimates towards the mean, resulting in contributions, which have moved toward the mean as well. The standard deviation of the contributions under the WLMM still appear to be larger then the standard deviations of the actual contributions, when N=10. However, this is mainly caused by the fact that in most cases λ>0 resulted in one contribution being zero per replicate. Exclusion of this null-contribution from the calculation of the standard deviations per replicate resulted in standard deviations smaller then the standard deviation of the actual contributions. It may be noted that a small standard deviation of contributions is desirable, in that it implies a conservative estimation of the contributions. In particular few breeds will have null-contributions and will be lost for the core set.

The **c**-vectors obtained from the four models were tested for the amount of genetic diversity they actually conserved by calculating $1-\hat{\mathbf{c}}'\mathbf{F}\hat{\mathbf{c}}$, where **F** is the actual kinship matrix, obtained from pedigree data. Genetic diversity of the estimated core sets was consistently lower then actual core set diversity, but differences were small (for N=10 the largest difference was 0.010 for ULM). When N=20 the largest difference with actual core set diversity was 0.011 for WLS and WLMM). Differences in core set diversity among the four models were small also, the largest difference being 0.006 between WLM and ULM in the case where N=10 and L=10. As expected the difference among models and between models and actual core set diversity decreased as L increased. Overall core set diversity was highest, when WLM was applied.

Application to field data

In Table 4.4 the per breed optimal contributions are given for the four models described in this paper. There is general agreement among the four methods with regard to the top three contributing breeds, although contributions are less extreme under WLM and WLMM. Some substantial re-ranking occurs with regard to least contributing breeds. Notably the Galloway breed contributes both in WLS and ULM while it receives null-contributions in WLM and WLMM. In contrast, both the Dutch Friesian and the Blonde d'Aquitaine receive positive contributions under WLM and WLMM while receiving null-contributions under WLS and ULM. This can be explained by the iterative weighted adjustments of the MEK matrix under WLM and WLMM, reducing the effect of one locus (SPS0115) which showed a substantial amount of similarity between all breeds except for the Galloway. Excluding this locus from the analysis reduced the contribution of the Galloway sample to null under WLS and ULM as well as under WLM and WLMM (data not shown).

Table 4.4 Contribution vectors to a core set for 10 Dutch populations of cattle breeds, according to four different methods of analysis.

Breed	**WLS**	**ULM**	**WLM**	**WLMM**
Limousine	0.402	0.402	0.295	0.219
Holstein Friesian	0.290	0.304	0.268	0.215
Dutch Red Pied	0.181	0.215	0.194	0.152
Dutch Friesian	0	0	0.130	0.123
Blonde d'Aquitaine	0	0	0.035	0.099
Heck	0.052	0.066	0.080	0.097
Belgian Blue	0	0	0	0.047
Improved Red Pied	0	0	0	0.027
Dutch Black Belted	0	0	0	0.022
Galloway	0.076	0.013	0	0

Table 4.5 Loss in genetic diversity, when only a fixed set of breeds is kept (**Fixed**, consisting of BBL, LIM, HF,GAL and BA) or the fixed set plus one other population. In Italics the difference between the Safe +1 set and the Safe set is given.

Breed	WLS		ULM		WLM		WLMM	
	Div		Div		Div		Div	
Full set	0.9712		0.9593		0.9626		0.9654	
Safe set	0.9670		0.9537		0.9564		0.9586	
Safe +1								
Dutch Red Pied	0.9704	*0.0034*	0.9579	*0.0042*	0.9594	*0.0030*	0.9617	*0.0031*
Heck aurochs	0.9688	*0.0018*	0.9559	*0.0022*	0.9585	*0.0021*	0.9611	*0.0025*
Dutch Friesian	0.9670	*0.0000*	0.9540	*0.0003*	0.9584	*0.0020*	0.9609	*0.0023*
Improved Red Pied	0.9673	*0.0003*	0.9537	*0.0000*	0.9568	*0.0004*	0.9595	*0.0009*
Dutch Black Belted	0.9670	*0.0000*	0.9537	*0.0000*	0.9567	*0.0003*	0.9593	*0.0007*

One way of analysing diversity is by defining a set of populations that are not under threat of extinction (the Safe set) and compare the core set diversity of this safe set to the core set diversity of the Safe set plus one of the populations not in the safe set, for each 'non-safe' population in turn (Thaon d'Arnoldi *et al.*, 1998; Eding *et al.*, 2001). This gives the contribution of the threatened populations to genetic diversity on top of the diversity contained in the Safe set. The added advantage is that we can calculate contributions to the core set by these threatened breeds. The safe set was defined to consist of Belgian Blue, Limousine, Holstein Friesian, Galloway and Blonde d'Aquitaine cattle, since they are all population that are used in agriculture throughout Europe. Furthermore, neither of these populations is listed as threatened in the World Watch List for either the Netherlands or their country of origin (Scherf, 2000). Results of this analysis are given in Table 4.5.

Note that differences in diversity are very small. This can partly be explained by the readjustment of contributions after a population is removed from the full set. This readjustment will compensate to some extent for the loss in diversity caused by exclusion from the set. Inclusion of the Dutch Red Pied breed gave the largest increase in genetic diversity (0.003 under both WLM and WLMM) with an associated large contribution to the core set of Safe+1 (0.225 under WLM and 0.195 under WLMM). Note that both Galloway and Dutch black belted contribute to the genetic diversity of the safe set, while both had null-contributions or very low contributions to the Full set (Table 4.4).

There are some discrepancies between contributions to core set and actual contributions to core set diversity. For example, under WLM the Heck aurochs and the Dutch Friesians increase core set diversity by approximately the same amount (Table 4.5), while the contributions to the core set differ by a factor 2 (0.090 for Heck aurochs and 0.183 for Dutch Friesian, data not shown).

DISCUSSION

We developed a method that simultaneously estimates AIS probabilities and MEK using all pairwaise similarities simultaneously. These improved estimates of MEK resulted in a reduced number of null-contributions of populations in the core set. The Weighted Log-linear

Mixed Model (WLMM) was constructed to further reduce number of negative contributions. This resulted in more conservative estimates of the contribution vectors (Tables 2,3 and 5).

In terms of actual genetic diversity in the estimated core sets WLM gives the best overall results. However, WLM was not able to eliminate all null-contributions, while the true optimal contributions were all non-zero. Moreover, the number of null-contributions becomes progressively larger when the size of the study increases. As Tables 2 and 3 show, the number of null-contributions decreases when more markers are used in a study, indicating that populations receive null-contributions as a result of incomplete information (Eding *et al.*, 2001).

The WLMM method eliminates all but one of the null-contributions. WLMM calculates kinships as deviations from a 'mean kinship matrix'. This means contributions are calculated as deviations from equal contributions, since the 'mean kinship matrix' has an associated contributions vector whose elements are 1/N, where N is the number of populations. In cases where data is not sufficient, kinship estimates are regressed back towards their mean. Uninformative data result in a high weight for the mean kinship matrix. This leads to more equal contributions, indicated by the lower standard deviation of the contributions (Tables 2 and 3).

However, contributions that are regressed toward the mean (WLMM) also lead to slightly less genetic diversity as calculated by $1-\hat{\mathbf{c}}'\mathbf{F}\hat{\mathbf{c}}$, where **F** is the actual kinship matrix (Tables 2 and 3). The decrease in genetic diversity is a result of the structure of the mean kinship matrix, where all within and all between population kinships are assumed equal. A better-structured mean kinship matrix should result in improved contribution vectors. More research is needed to find such a better structure for the mean **F** matrix.

The results from the simulated data show the importance of the number of loci involved in the study. Previous studies have generally used data on between 10 and 20 loci, which is sufficient to accurately estimate genetic distances between populations (Cañón *et al.*, 2001; Laval *et al.*, 2001) or between and within population average kinships (Eding *et al.*, 2001; Lynch and Ritland, 1999). However, the correlations between actual and estimated contributions (Tables 2 and 3, r_c) show that these numbers of loci give only moderately accurate results when used to estimate optimal core set contributions.

The results from the field data show the importance of accounting for the sampling variance per locus when producing estimates of optimal core set contributions. The positive contributions of the Galloway breed under WLS and ULM can be attributed to one locus out of 11 (SP0115), that showed a high degree of similarity among all breeds except for the Galloway. Moreover, using ULM the expected similarity (based of estimates of f and s) deviated 1.4 standard deviations from the observed similarity for this locus. Under WLM this deviation was 2.9 standard deviations, indicating that SPS0115 has a large effect on the kinship estimates of the methods WLS and ULM. Under WLM and WLMM the influence of SPS0115 was reduced by its decreased weight, which resulted in the exclusion of the Galloway in favor of the Dutch Red Pied and the Dutch Friesian. As expected, using iterative weighted methods reduces the influence of such 'ill behaved' loci, reducing the risk of incorrect decision making with regard to conservation efforts.

Of the four methods described in this paper, WLMM is the most demanding in terms of computer resources. Moreover, the amount of time increases approximately quadratic with the number of populations in the study. In WLMM and in WLM solutions to the equations are found by iterating on the weights. WLMM performs this iteration while also searching for λ. These iterations increase the amount of calculations considerably. However, once $\hat{\mathbf{F}}$ has been obtained, calculating the core set or contributions to genetic diversity of individual populations or subsets of populations is straight forward and does not require large amounts of computer resources. Furthermore, this also holds when the methods are extended to calculate contributions to diversity for individual animals.

In conclusion, WLM estimates of MEK resulted in contributions to core sets that contained the highest diversity. WLMM gave slightly less diverse core sets, but many more breeds contributed to the core set, such that WLMM seems a more conservative method to construct core sets and to compare the diversity contained in alternative sets. The latter is because a null-contribution of a breed to a core set implies that this breed contributes no extra diversity to this set and will not be considered for genetic conservation.

ACKNOWLEDGEMENT

We would like to thank The Van Haeringen Laboratory in Wageningen, The Netherlands, who provided data on allele frequencies of the markers used in the application to field data.

APPENDIX 4.A

The diagonal elements of **W** are the variances of ln(1- $S_{ij,l}$), where $S_{ij,l}$ is the similarity for locus l in the pair of populations i and j. To obtain var[ln(1- $S_{ij,l}$)], ln(1- $S_{ij,l}$) can be approximated by the following Taylor series:

$$\ln\left(1-S_{ij,l}\right)\approx \ln\left(1-\hat{S}_{ij,l}\right)+\frac{\delta \ln\left(1-\hat{S}_{ij,l}\right)}{\delta\hat{S}_{ij,l}}\left(S_{ij,l}-\hat{S}_{ij,l}\right)$$

(A1a)

$$\approx \ln\left(1-\hat{S}_{ij,l}\right)-\frac{1}{1-\hat{S}_{ij,l}}\left(S_{ij,l}-\hat{S}_{ij,l}\right)$$

where $\hat{S}_{ij,l}=\hat{f}_{ij}+(1-\hat{f}_{ij})\hat{s}_l$. The variance of ln(1- $S_{ij,l}$) is obtained from:

$$\mathrm{var}\left[\ln\left(1-S_{ij,l}\right)\right]\approx \mathrm{var}\left[\ln\left(1-\hat{S}_{ij,l}\right)-\frac{1}{1-\hat{S}_{iij,l}}\left(S_{ij,l}-\hat{S}_{ij,l}\right)\right]$$

(A1b)

$$\approx\left(\frac{1}{1-\hat{S}_{ij,l}}\right)^2\mathrm{var}\left(S_{ij,l}\right)$$

From the binomial distribution: $\mathrm{var}\left(I_{xy}(k)\right)=\hat{S}_{ij,l}\left(1-\hat{S}_{ij,l}\right)$ and thus

$$\mathrm{var}\left(S_{ij,l}(k)\right)=\hat{S}_{ij,l}(k)\left(1-\hat{S}_{ij,l}(k)\right)/4$$

(A2a)
$$\mathrm{var}\left(S_{ij,l}\right)=\frac{\hat{S}_{ij,l}\left(1-\hat{S}_{ij,l}\right)}{4n_{ij}}=\left[\hat{f}_{ij}+(1-\hat{f}_{ij})\hat{s}_l\right]\left[(1-\hat{f}_{ij})(1-\hat{s}_l)\right]/4n_{ij}$$

where n_{ij} is the number of pairs of animals that contribute to $S_{ij,l.}$

The variance of ln(1- $S_{ij,l}$) reduces to:

$$\mathrm{var}\left[\ln\left(1-S_{ij,l}\right)\right]\approx\left(\frac{1}{1-\hat{S}_{ij,l}}\right)^2\mathrm{var}\left(S_{ij,l}\right)$$

(A2b)

$$\approx\frac{\left[\hat{f}_{ij}+(1-\hat{f}_{ij})\hat{s}_l\right]\left[(1-\hat{f}_{ij})(1-\hat{s}_l)\right]}{4n_{ij}\left[(1-\hat{f}_{ij})(1-\hat{s}_l)\right]^2}$$

$$\approx\frac{\hat{f}_{ij}+(1-\hat{f}_{ij})\hat{s}_l}{4n_{ij}(1-\hat{f}_{ij})(1-\hat{s}_l)}$$

Log-linear estimation

Chapter 5

GENETIC DIVERSITY MAINTAINED BY AFRICAN SUB-SAHARAN CATTLE BREEDS

Eding, Herwin[1,*] Olivier Hanotte[2], Theo H.E. Meuwissen[1] and J.E.O. Rege [3]

[1] Institute for Animal Science and Health, Box 65, 8200 AB Lelystad, The Netherlands

[2] International Livestock Research Institute (ILRI), P.O. Box 30709, Nairobi, Kenya

[3] ILRI, PO Box 5689, Addis Ababa, Ethiopia

Genetics Selection Evolution (Submitted)

[*] Institute for Animal Science and Health
Box 65, 8200 AB Lelystad, The Netherlands
Tel: +31-(0)320-238238; Fax +31-(0)320-238050
E-mail: j.h.eding@id.wag-ur.nl

ABSTRACT

Genetic diversity in a set of 59 African breeds of cattle was analysed using core set analysis. The breeds were from five regions and of three breed types (zebu, taurine and sanga). The results showed a patterns of substantial gene flow between breeds from the Eastern African regions, while taurine and Southern African sanga were closely related to each other, but relatively isolated from other breeds. Breeds that possess genetic diversity a-typical of the set being analysed, such as a breed influenced by zebu breeds in the taurine core set, received relatively high contributions at the cost of breeds more typical of the set.

Analysis of genetic diversity is the most logical and given the results the most beneficial approach. The best approach seems to be to assess the relative importance of individual breeds to genetic diversity in a species-wide core set, encompassing all breeds and populations. Such an approach requires the concerted input from all nations and regions involved.

INTRODUCTION

The African continent harbours a large proportion of the world's genetic diversity in livestock in general and cattle specifically (Scherf, 2000). Domesticated taurine populations of livestock are believed to have been first introduced into the African continent from the Near East (Epstein, 1971), although some evidence suggests an independent domestication of taurines occurred in North Africa, separately from the domestication of the ancestors to European and Asian taurine breeds (Bradley *et al.*, 1996). Further waves of introduction include taurine Shorthorns (*B. Taurus*) and humped zebu (*B. Indicus*) populations. Today some 186 million head of cattle in around 150 breeds (taurine, zebu and intermediate or sanga populations) can be found in sub-Saharan Africa (Hanotte *et al.*, 2000). Adapted to local conditions, these populations represent an important and diverse genetic resources.

Many of these populations are to a greater or lesser extent threatened in their existence (Scherf, 2000). One of the main causes is replacement of local breeds with exotic, high producing and high maintenance breeds like the Holstein Friesian (Oldenbroek, 1999). The threat to these populations can be reduced by conservation efforts, including setting up a gene bank to store germplasm of important and/or critically endangered populations.

Recently, a method was proposed to assess genetic diversity based on estimates of average kinship coefficients between and within populations (Eding *et al.*, 2001). Genetic diversity was defined as the maximum genetic variance of a population in Hardy-Weinberg-equilibrium that could be derived from the conserved set of breeds. The maximum genetic diversity can be obtained through the construction of a core set of breeds in which the average kinship is minimized. The relative importance of populations in terms of genetic diversity can be calculated as the contribution of each population to the this maximized genetic variance.

The main objective of this paper is to describe genetic diversity in African cattle and identify which African breeds or populations are important contributors to genetic diversity and could be considered for inclusion in a gene bank. A secondary objective of this paper is to study the effects of different strategies of conservation (sub-regional versus continental) on the priorities assigned to breeds of sub-Saharan African cattle breeds and the efficiency of the conservation efforts.

MATERIAL AND METHODS

Similarities based on genetic markers were calculated between- and within- populations using the similarity index proposed by Eding and Meuwissen (2001a) averaged over pairs of animals. The similarity index used was:

$$(1) \qquad S_{xyk} = \tfrac{1}{4}\left[I_{11} + I_{12} + I_{21} + I_{22}\right]$$

where I_{ij} is an indicator variable which is 1 when allele i on locus k in the individual x and allele j on the same locus in individual y are identical, otherwise it is 0. These similarities between pairs of individuals were then averaged to get the mean between (and within) population similarities S_{ijk}.

In this similarity index between two populations for a locus, the probability that the alleles are Identical By Decent (IBD) or kinship coefficient is confounded with the probability that the alleles are Alike In State (AIS). This can be represented as:

$$(2) \qquad (1 - S_{ijk}) = (1 - f_{ij})(1 - s_k)$$

Where S_{ijk} is the average pairwise similarity index between two populations i and j, f_{ij} is the average kinship between populations i and j which is equal to the probability of IBD (Malecot, 1948) and s_k is the probability of alleles AIS at locus k.

To estimate average between- and within- population kinships using all available information, a log-transformation was performed (Eding and Meuwissen, 2001b):

$$(3) \qquad \begin{aligned} \ln(1 - S_{ijk}) &= \ln[(1 - f_{ij})(1 - s_k)] + error_{ijk} \\ &= \ln(1 - f_{ij}) + \ln(1 - s_k) + error_{ijk} \quad \Leftrightarrow \\ y_{ijk} &= a_{ij} + b_k + error_{ijk} \end{aligned}$$

where $error_{ijk}$ is due to sampling. The values of a_{ij} and b_k can thus be estimated using log-linear models. Two models were used to obtain estimates for the between populations kinships, the Weighted Log linear Model (WLM) and the Weighted Log linear Mixed Model (WLMM; Eding and Meuwissen, 2001b). These are briefly described below.

Weighted Log linear Model

In matrix notation (3) can be written as:

$$(4) \qquad \mathbf{y} = \mathbf{X_a a} + \mathbf{X_b b} \ + \mathbf{e}$$

Where $\mathbf{X_a}$ and $\mathbf{X_b}$ are design matrices indicating relations and loci, respectively. Estimation errors of the solutions may be reduced by accounting for the error variances of the data vector **y** by weigthing each similarity score per locus with the expected variance on the similarity score. This results in more informative markers having a larger influence on the solutions of both *f* and *s*, and less informative markers having less influence. The WLM equations are:

$$(5) \qquad \begin{bmatrix} \mathbf{X_a'W^{-1}y} \\ \mathbf{X_b'W^{-1}y} \end{bmatrix} = \begin{bmatrix} \mathbf{X_a'W^{-1}X_a} & \mathbf{X_a'W^{-1}X_b} \\ \mathbf{X_b'W^{-1}X_a} & \mathbf{X_b'W^{-1}X_b} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}$$

where all observations are weighted by the matrix $\mathbf{W}^{-1}$. Matrix **W** is a diagonal matrix containing the expected variances of each observation (Eding and Meuwissen, 2001b):

$$(6) \qquad \operatorname{var}\left[\ln\left(1-S_{ijk}\right)\right] \approx \frac{\hat{f}_{ij} + (1-\hat{f}_{ij})s_k}{4n_{ij}(1-\hat{f}_{ij})(1-s_k)}$$

Where n_{ij} is the number of pairs of individuals in *i* and *j*. Since matrix **W** is diagonal, the weight of each observation is the reciprocal of (6). Because the weights are calculated from the estimates of *f* and *s*, this system of equations needs to be solved iteratively.

Weighted Loglinear Mixed Model

In order to reduce the sampling errors in the kinship estimates, which can lead to inconsistencies in the kinship structure, thereby causing populations to receive negative contributions to the core set, WLMM estimates are regressed towards their respective means, using Mixed Model methods. To this end, the means of the between- and within- population Marker Estimated Kinships (MEK) are introduced as an extra term in (3).

$$(7) \qquad \mathbf{y} = \left(\mathbf{Z\mu} + \mathbf{X_a a^*}\right) + \mathbf{X_b b} + \mathbf{e}$$

where μ is a 2 x1 vector of estimates of the between- and within- population MEK on the log scale; **Z** is a ½ N(N+1) x 2 design matrix for the means μ, N being the number of populations. The vector $\mathbf{a}^*$ contains deviations of the log-transformed kinship estimates from the mean such that $\mathbf{a} = \mathbf{Z}\mu + \mathbf{a}^*$.

The weighted log linear mixed model equations are:

$$(8) \qquad \begin{bmatrix} \mathbf{Z'W^{-1}y} \\ \mathbf{X_a'W^{-1}y} \\ \mathbf{X_b'W^{-1}y} \end{bmatrix} = \begin{bmatrix} \mathbf{Z'W^{-1}Z} & \mathbf{Z'W^{-1}X_a} & \mathbf{Z'W^{-1}X_b} \\ \mathbf{X_a'W^{-1}Z} & \mathbf{X_a'W^{-1}X_a} + \lambda\mathbf{I} & \mathbf{X_a'W^{-1}X_b} \\ \mathbf{X_b'W^{-1}Z} & \mathbf{X_b'W^{-1}X_a} & \mathbf{X_b'W^{-1}X_b} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{a^*} \\ \mathbf{b} \end{bmatrix}$$

where the term $\lambda\mathbf{I}$ is included to regress the deviations from the means, $\mathbf{a}^*$, back to zero, thereby regressing $\mathbf{a} = \mathbf{Z}\mu + \mathbf{a}^*$ back to its mean. The extent to which $\mathbf{a}^*$ is regressed to zero depends on λ. As with WLM, this system of equations is solved iteratively, in order to find the weights. The factor λ is chosen such that the smallest contribution to the core set equals zero, i.e. all contributions are larger or equal to zero.

After obtaining solutions to the equations, kinships were transformed back (see equation 3) through:

9) $$\mathbf{f} = 1 - e^{\mathbf{a}}$$

where $\mathbf{f}$ contains the (½)N(N+1) kinship estimates. These were subsequently used to construct the estimated kinship matrix $\hat{\mathbf{F}}$, which is used to calculate contributions to the core set.

Contributions to the core set

After obtaining kinship estimates, contributions of each population to a core set, which is constructed such that the kinship within the core set is minimal, were calculated according to Eding *et al.* (2001):

(10) $$\mathbf{c} = \frac{\mathbf{1}'\hat{\mathbf{F}}^{-1}}{\mathbf{1}'\hat{\mathbf{F}}^{-1}\mathbf{1}}$$

Where $\mathbf{c}$ is the vector, containing core set contributions of N populations and $\hat{\mathbf{F}}$ is an NxN matrix containing estimates of the average between population kinships. The total conserved genetic diversity, in terms of fraction of variation of the founder population, can be obtained from:

(11) $$\begin{aligned} Div &= 1 - \mathbf{c}'\hat{\mathbf{F}}\mathbf{c} \\ &= 1 - \frac{1}{\mathbf{1}'\hat{\mathbf{F}}^{-1}\mathbf{1}} \end{aligned}$$

Data

The data set that was used in this study was compiled at ILRI, Nairobi.The data set (Table 5.1) consisted of 59 cattle breeds phenotypically classified in three groups: taurine (*B. Taurus*), zebu (*B. Indicus*) and sanga (breeds which are considered to be intermediates between *B. Taurus* and *B. Indicus*). The data set also contained three zenga type breeds, which are considered to be intermediate between sanga and zebu breeds. On average, 35 individuals per breed were genotyped for 15 microsatellite loci. For more detailed information about breeds and loci see Hanotte *et al.*(2001).

Table 5.1 Overview of the breeds in this study and the loci used.

Zebu		***Sanga***		***Taurine***	
Arashie	*Abyssinia*	Abigar	*Abyssinia*	Sheko	*Abyssinia*
Arsi	*Abyssinia*	Afar	*Abyssinia*	Brune	*Northern Africa*
Bale	*Abyssinia*	Danakil	*Abyssinia*	Baladi	*Northern Africa*
Boraneth	*Abyssinia*	Raya-Azebu	*Abyssinia*	Baoule	*Western Africa*
Butana	*Abyssinia*	Ankole	*Lake Victoria*	Blonde	*Western Africa*
Nuba	*Abyssinia*	Kigezi	*Lake Victoria*	Kapsiki	*Western Africa*
Ogaden	*Abyssinia*	Watusi	*Lake Victoria*	Kuri	*Western Africa*
Highland Zebu	*Lake Victoria*	Africaner	*Southern Africa*	Namchi	*Western Africa*
Iringa Red	*Lake Victoria*	Barotse	*Southern Africa*	N'Dama	*Western Africa*
Kavirondo	*Lake Victoria*	Drankensberger	*Southern Africa*	Somba	*Western Africa*
Kenyan Boran	*Lake Victoria*	Kavango	*Southern Africa*	Friesian	*Europe*
Kilimanjaro	*Lake Victoria*	Koakoland	*Southern Africa*	Jersey	*Europe*
Madagascar Zebu	*Lake Victoria*	Landim	*Southern Africa*	Retinta	*Europe*
Orma Boran	*Lake Victoria*	Mashona	*Southern Africa*		
Zebu Malagasy	*Madagascar*	Nguni	*Southern Africa*		
Angoni	*Southern Africa*	Nkone	*Southern Africa*		
Gobra	*Western Africa*	Pedi	*Southern Africa*		
Maure	*Western Africa*	Tonga	*Southern Africa*		
Mbororo	*Western Africa*	Tuli	*Southern Africa*		
New Ongole	*Asia*	Sokoto Gudali	*Western Africa*		
Sahiwal	*Asia*			***Zenga***	
Nelore/Ongole	*Asia/Brazil*			Arado	*Abyssinia*
Gobra	*Western Africa*			Fogera	*Abyssinia*
Maure	*Western Africa*			Horro	*Abyssinia*

The 59 breeds in the data set were divided into regions (Hanotte *et al.*,2000) and breed type groups (Rege *et al.*, 1996). Analysis of the data was done on the entire dataset (including non-African breeds), African breeds only, by sub-region and by breed type. In the case of the taurine breed type, two separate analyses were done, one including, and one excluding, European taurine breeds.

RESULTS

Marker Estimated Kinships

Figures 5.1 and 5.2 form a graphical representation of the MEK for the entire dataset, including the Asian and European populations. Figure 5.1 is a Neighbor-Joining (NJ) tree calculated by transforming the MEK matrix into a kinship distance matrix (Eding *et al.*, 2001). Figure 5.2 which is a contour plot where darker shades indicate higher kinships between populations, was derived from the order of populations taken from the NJ-tree in Figure 5.1.

Overall, a clustering according to both breed type and region can be observed (Figure 5.2). Towards the middle of the tree a cluster of populations can be seen, which is comprised mainly of taurines (both European and West African) intermingled with South African sanga breeds, which agrees well with recent evidence (Hanotte *et al*, 2001) suggesting that the sanga are basically a sub-population of the original African Bos Taurus cattle. To the left of this cluster a number of West African zebus are clustered. These in turn, are closely linked to branches carrying Abyssinian sanga and zebus from the Lake Victoria region. Right of the middle, this latter cluster is continued with a cluster of zebu populations from both Lake Victoria and Abyssinia, ending in a cluster of Asian zebu plus Zebu Malagasy. The Madagascar Zebu is considered to be derived mainly from Asian zebu populations imported directly to the Island and may have had little, if any, contact with indigenous populations on the African mainland (Hanotte *et al*, 2001). This explains its clustering with Asian rather than African population.

From Figure 5.2 we can see that the clusters on the left and right of the combined taurine- southern African sanga cluster, are related more to each other than they are to the combined taurine- southern African sanga cluster, as indicated by the shades in Figure 5.2. However, a closer level of kinship can be observed between southern African sanga breeds and eastern African cattle breeds. MEKs within and between the taurine and southern African sanga breeds were generally higher than the MEKs within and between zebu and eastern African sanga breeds.

Figure 5.1 Kinship tree of the entire set of breeds, including European and Asian breeds. The tree is a Neighbor-Joining tree of the MEK-estimates using WLM.

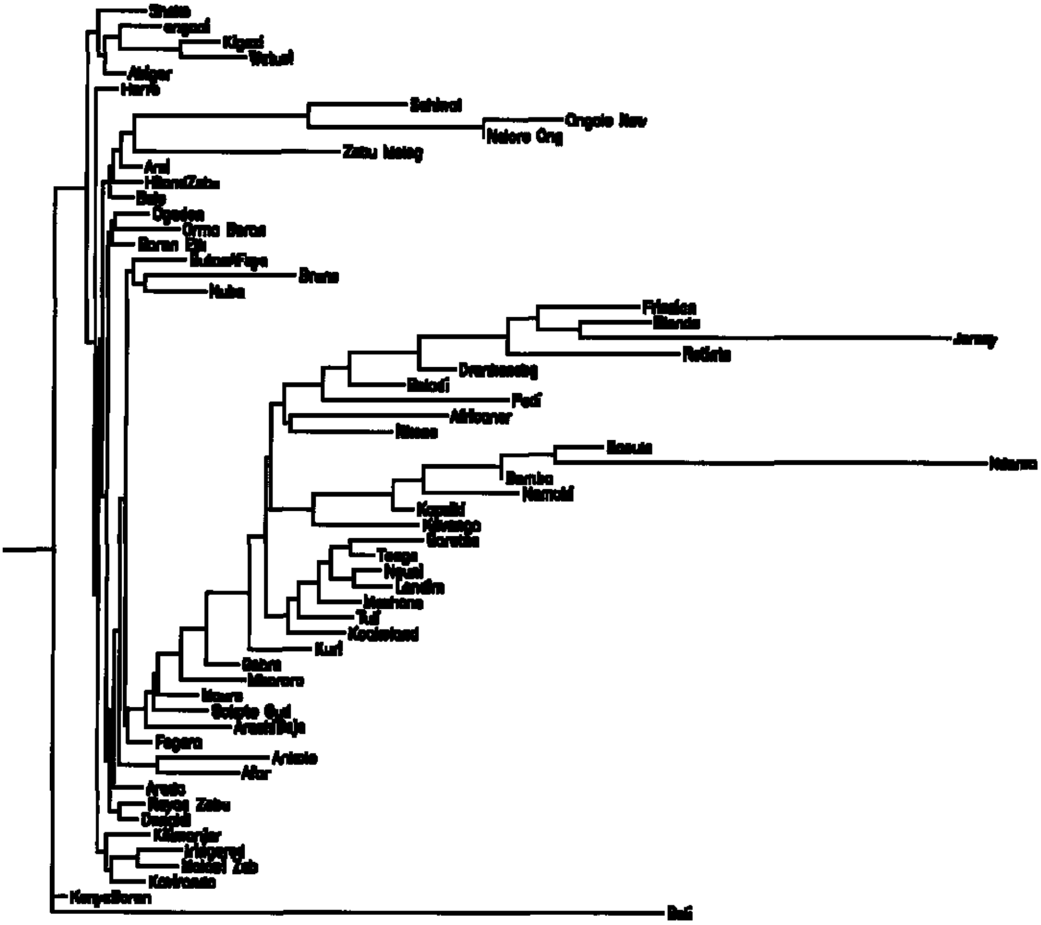
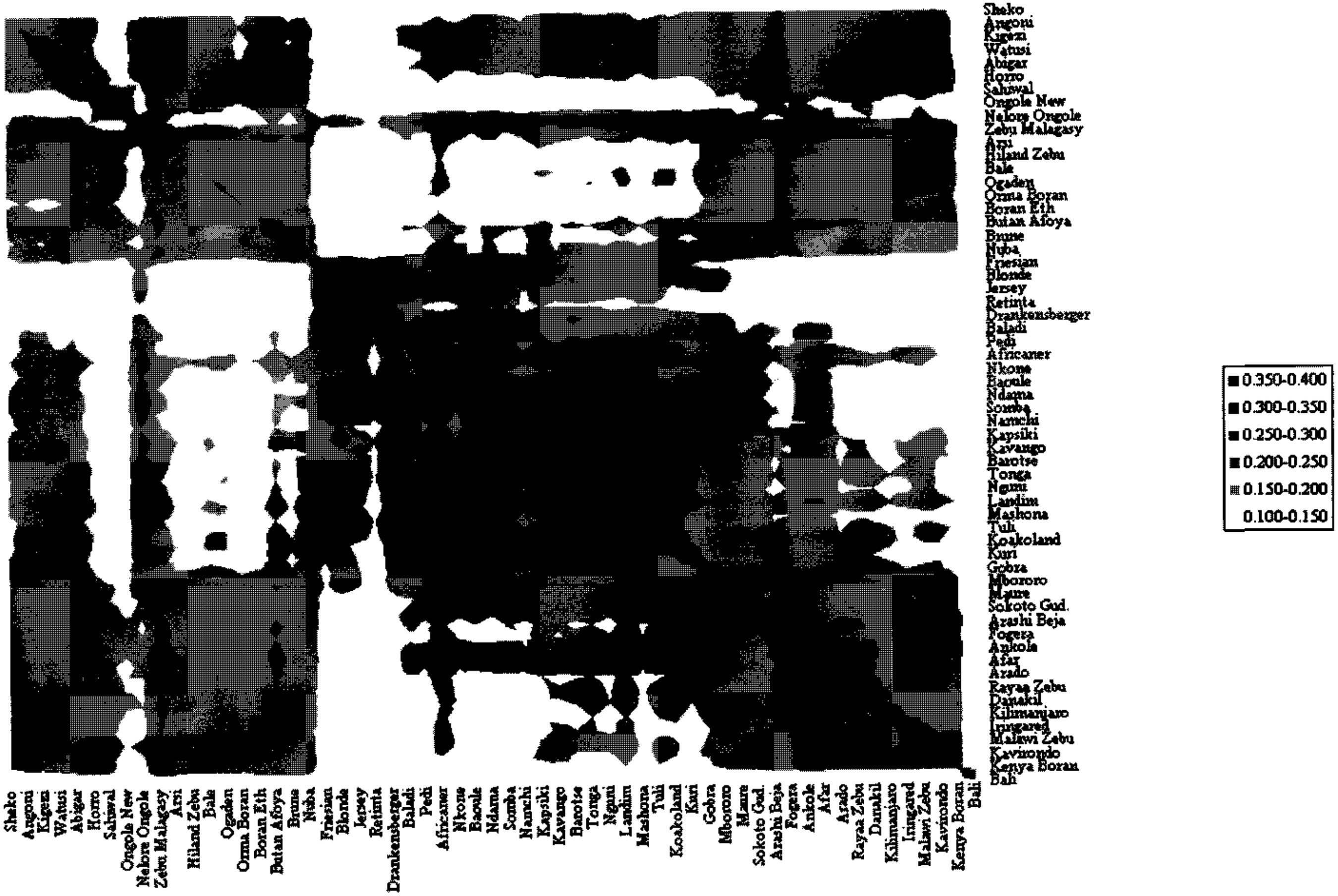


Figure 5.2 Contour plot of the matrix containing MEKs of the entire set of breeds (including European and Asian breeds) estimated using the Weighted Log linear Model (WLM). Darker shades indicating higher kinships.



Core set contributions

We first calculated core set contributions for all populations in the data set using both the WLM and the WLMM methods. Results are given in Table 5.2. Table 5.3 displays results for African populations only. Tables 5.4 and 5.5 display results when populations were analysed by breed type (zebu, sanga or taurine) and region, respectively. In all four tables the breeds are ranked according to their contribution, calculated with the WLMM method. For the WLMM method the values of parameter λ at which all contributions were zero or positive for the set of breeds analysed are also given. Note that for cases where Abyssinian populations, Abyssinian zebu breeds specifically, are included in the set of breeds, the value of λ is rather high: 481 for the entire data set (Table 5.2), 445 when only African breeds are analysed, (Table 5.3), 402 for zebu breeds (Table 5.4) and 359 for the Abyssinian region (Table 5.5). In the cases where λ was high when using WLMM, the number of null-contributions when using WLM was also high. In Table 5.2, where λ was highest (λ=481), 48 out of 59 breeds received null contributions when WLM was used.

When the entire set is considered, non-African breeds are given relatively high priority (Table 5.2). This can be explained by the fact that there are only a few non-African breeds and that these represent a portion of genetic diversity that is not present in Africa.

Although the ranking stays more or less intact in all analysed sets of breeds, there are some cases in which null-contributors in WLM receive a rather unexpected high contribution using WLMM (for instance the New Ongole and Sahiwal in Table 5.2, or the sanga breed Raya Azebu in Table 5.4). In both examples the breeds are closely related to breeds which received a considerable contribution using WLM (Nelore Ongole and Danakil, respectively), which suggests a redistribution of importance of breeds within a cluster. In this light, it is interesting to note that under WLM, the Asian cluster (New Ongole, Nelore Ongole and Sahiwal) receives a contribution of 0.096 for the Nelore Ongole only and using WLMM a contribution of 0.094, more evenly distributed over all three populations.

In some cases, breeds that received a relatively low contribution earlier when a larger set of breeds was analysed, are given priority in a limited set. For instance, African taurine breeds (notably the Baladi and the Blonde) and the South African sanga breed Drankensberger

Table 5.2 Contributions to a core set for all populations in the data set, including the Asian and European breeds as well as the Banteng population.

Breed	WLMM	WLM
Bali	0.103	0.312
Retinta	0.041	0.080
Nelore Ongole	0.036	0.096
New Ongole	0.034	0
Friesian	0.032	0.009
Kenyan Boran	0.028	0.121
Sheko	0.028	0.175
Orma Boran	0.026	0.011
Sahiwal	0.024	0
Jersey	0.024	0
Horro	0.023	0.026
Boran Ethiopia	0.023	0
Drankensberger	0.022	0
Highland Zebu	0.022	0
Baladi	0.021	0.132
Namchi	0.020	0.031
Iringared	0.019	0
Arsi	0.019	0
Blonde	0.018	0
Kilimanjaro	0.018	0
Ogaden	0.018	0
Brune	0.017	0
Madagascar Zebu	0.017	0
Nkone	0.016	0
Kigezi	0.016	0
Bale	0.016	0
Kavirondo	0.016	0
Kuri	0.016	0
Mashona	0.016	0
Pedi	0.015	0
Angoni	0.014	0
Danakil	0.014	0
Fogera	0.014	0
Arado	0.014	0
Koakoland	0.013	0.008
Nguni	0.013	0
Watusi	0.013	0
Abigar	0.013	0
Nuba	0.012	0
Raya-Azebu	0.012	0
Maure	0.011	0
Gobra	0.011	0
Tuli	0.011	0
Tonga	0.010	0
Sokoto Gudali	0.010	0
Baoule	0.009	0
Kavango	0.009	0
Butana	0.008	0
Afar	0.008	0
Arashie	0.008	0
Somba	0.008	0
Barotse	0.006	0
Kapsiki	0.005	0
Landim	0.005	0
Africaner	0.004	0
Mbororo	0.002	0
Zebu Malagasy	0.002	0
Ankole	0.000	0
N'Dama	0	0

Table 5.3 Contributions to a core set of all African populations in the data set. Populations are ranked in order of descending contributions under the WLMM method. The value of λ is given at which all contributions were equal or larger then zero.

Breed	**WLMM**	**WLM**	**Breed**	**WLMM**	**WLM**
	λ=445				
Kenyan Boran	0.042	0.369	Fogera	0.018	0
Blonde	0.040	0.060	Nuba	0.018	0
Sheko	0.038	0.237	Maure	0.018	0
Orma Boran	0.037	0.112	Watusi	0.017	0
Drankensberger	0.036	0.051	Gobra	0.017	0
Baladi	0.033	0.045	Raya-Azebu	0.017	0
Boran Ethiopia	0.031	0	Abigar	0.016	0
Horro	0.031	0	Koakoland	0.016	0
Highland Zebu	0.030	0	Kavango	0.015	0
Arsi	0.027	0	Nguni	0.015	0
Nkone	0.027	0.054	Baoule	0.015	0
Brune	0.026	0	Arashie	0.014	0
Iringared	0.026	0	Tuli	0.013	0
Namchi	0.026	0.072	Sokoto Gudali	0.013	0
Ogaden	0.025	0	Butana	0.013	0
Kilimanjaro	0.024	0	Tonga	0.013	0
Bale	0.024	0	Somba	0.012	0
Kuri	0.023	0	Kapsiki	0.007	0
Kavirondo	0.022	0	Barotse	0.007	0
Pedi	0.022	0	Afar	0.006	0
Madagascar Zebu	0.022	0	Afrikaner	0.004	0
Mashona	0.021	0	Mbororo	0.004	0
Angoni	0.020	0	Landim	0.001	0
Kigezi	0.020	0	Ankole	0.001	0
Danakil	0.020	0	Madagaskar Zebu	0.001	0
Arado	0.019	0	N'Dama	0	0

receive a relatively higher contribution when only African breeds are considered (Table 5.3) than when non-African breeds are also considered. Specifically, exclusion of non-African taurine breeds cause this re-ranking. The Drankensberger is known to have been influenced by Dutch cattle brought in by Dutch settlers in South Africa. Exclusion of European taurine has the effect of compensating the loss in diversity by giving higher priority to breeds closely related, in this case the Drankensberger. This effect can also be observed for the Blonde breed (Table 5.4; All Taurine versus African Taurine).

Regional versus breed type core sets

The ranking of a breed can differ considerably depending on what type of analysis is used. In the West African region the majority of breeds were taurine, with three zebu breeds (Maure, Gobra and Mbororo) and one sanga breed (Sokoto Gudali). When analysed by breed type, especially the Mbororo and the Sokoto Gudali received moderate to low contributions. However, their contribution to regional genetic diversity was considerable (Table 5.5). The ranking of the Gobra and Maure breed is reversed, indicating that the Gobra is genetically more influenced by West African taurine breeds than the Maure breed. Nevertheless, West African genetic diversity is largely dependent on the taurine populations, making a total proportionate contribution of 0.643, while zebus and sanga contributed 0.261 and 0.096, respectively.

From the Southern African region, all populations were sanga, except one: the Angoni breed. This translated into a relatively high contribution by Angoni in the southern African core set, even though the contribution of the breed to genetic diversity in African zebu populations was only modest (Table 5.4). The Nkone contributed more both in the southern African core set, and in the sanga core set, both in terms of ranking and in terms of actual contribution, indicating the possibility of some historic gene flow between the Nkone breed and other sanga breeds, not from the southern African region.

From the Lake Victoria region the core set comprised 10 populations, three of which were sanga and 7 were zebu populations. No substantial reranking occurred, compared to the ranking in the analysis by breed type (Table 5.4). The results seem to suggest that the zebu population is more important to genetic diversity in the region than the sanga breeds (total sanga contribution 0.192 using WLMM).

Table 5.4 Contribution to a core set of cattle populations per breed type. Populations are ranked in order of descending contributions under the WLMM method. The values of λ are given at which all contributions to the core sets were equal or larger the zero.

	WLMM	WLM
All Taurine	λ=126	
Sheko	0.158	0.498
Brune	0.135	0.088
Retinta	0.103	0.140
Kuri	0.099	0
Baladi	0.098	0.102
Friesian	0.085	0.117
Namchi	0.084	0.050
Jersey	0.073	0.005
Kapsiki	0.053	0
Somba	0.041	0
Blonde	0.041	0
Baoule	0.033	0
N'Dama	0	0

	WLMM	WLM
African Taurine	λ=109	
Sheko	0.192	0.454
Brune	0.177	0.136
Baladi	0.150	0.236
Kuri	0.118	0.030
Blonde	0.107	0.075
Namchi	0.093	0.070
Kapsiki	0.060	0
Somba	0.055	0
Baoule	0.049	0
N'Dama	0	0

	WLMM	WLM
Zebu	λ=402	
Kenyan Boran	0.092	0.496
Orma Boran	0.080	0.201
Gobra	0.068	0.132
Highland Zebu	0.060	0.100
Maure	0.059	0.026
Boran Ethiopia	0.058	0
Angoni	0.058	0.040
Arsi	0.057	0
Iringared	0.056	0
Nuba	0.054	0.006
Bale	0.054	0
Ogaden	0.053	0
Kavirondo	0.053	0
Kilimanjaro	0.047	0
Madagascar Zebu	0.047	0
Arashie	0.036	0
Mbororo	0.035	0
Butana	0.033	0
Zebu Malagasy	0	0

	WLMM	WLM
Sanga	λ=166	
Drankensberger	0.103	0.241
Danakil	0.094	0.333
Raya-Azebu	0.080	0
Nkone	0.077	0.109
Abigar	0.072	0.048
Kigezi	0.068	0.183
Watusi	0.067	0.036
Pedi	0.061	0.050
Sokoto Gudali	0.056	0
Mashona	0.053	0
Afar	0.051	0
Koakoland	0.038	0
Nguni	0.036	0
Tuli	0.035	0
Kavango	0.034	0
Ankole	0.032	0
Tonga	0.030	0
Barotse	0.008	0
Africaner	0.005	0
Landim	0	0

Table 5.5 Contribution to a core set of cattle populations per region. Populations are ranked in order of descending contributions under the WLMM method. The values of λ are given at which all contributions to the core sets were equal or larger the zero.

	WLMM	WLM
West Africa	λ=132	
Brune*	0.138	0.202
Baladi*	0.120	0.283
Maure	0.103	0.183
Sokoto Gudali	0.096	0.001
Blonde	0.095	0.049
Kuri	0.095	0.187
Gobra	0.094	0.026
Namchi	0.071	0.069
Mbororo	0.064	0
Kapsiki	0.044	0
Somba	0.043	0
Baoule	0.039	0
N'Dama	0	0

	WLMM	WLM
Lake Victoria	λ=156	
Kenyan Boran	0.187	0.529
Orma Boran	0.160	0.228
Kigezi	0.119	0.119
Highland Zebu	0.117	0.110
Kavirondo	0.097	0
Kilimanjaro	0.085	0
Iringared	0.082	0
Watusi	0.081	0.014
Madagascar Zebu	0.073	0
Ankole	0	0

	WLMM	WLM
South Africa	λ=98	
Angoni	0.209	0.466
Nkone	0.140	0.196
Drankensberger	0.138	0.186
Mashona	0.092	0.071
Pedi	0.079	0.029
Koakoland	0.078	0.052
Tuli	0.066	0
Kavango	0.065	0
Nguni	0.060	0
Tonga	0.049	0
Africaner	0.019	0
Landim	0.005	0
Barotse	0	0

	WLMM	WLM
Abyssinia	λ=359	
Sheko	0.120	0.412
Horro	0.093	0.292
Arsi	0.078	0.056
Nuba	0.077	0.123
Boran Ethiopia	0.076	0.118
Bale	0.073	0
Ogaden	0.071	0
Abigar	0.069	0
Fogera	0.062	0
Arado	0.055	0
Danakil	0.052	0
Raya-Azebu	0.050	0
Butana	0.048	0
Arashie	0.046	0
Afar	0.030	0
Zebu Malagasy	0	0

*) These breeds originate from Northern Africa (Table 5.1), but were included in the Western African set for the analysis.

The largest contributor to the Abyssinian core set was the Sheko breed, a taurine breed that is very much influenced by zebu and sanga genes (Hanotte, pers. comm.). This explains its large contribution in both the Abyssinian core set and in the taurine core set. That is, as the only representative of taurines in the Abyssinian core set it adds a considerable amount of 'zebu-diversity' to the taurine core set.

It may be noted that the contributions of the zebu-sanga intermediates (zenga) (Horro, Fogera and Arado) are rather high, seemingly at the cost of contributions by the sanga breeds, which, in essence, are also crosses of zebu and taurine populations. Overall, in the Abyssinian region, as in the Lake Victoria region, zebu populations represent the bulk of genetic diversity, totalling 47% of the core set (21%, 20% and 12% for zenga, sanga and the Sheko taurine respectively).

Genetic diversity

In Table 5.6 the results are given of the loss of diversity when only one specific group is conserved. The loss in diversity arising from focusing only on a single category is given relative to the entire set of breeds ('Entire') and relative to the African set of breeds ('Africa').

The loss of genetic diversity is greater when WLM is used than when WLMM is used. This is due to the (sometimes considerable) regression back to the within and between mean population kinships, which also makes the estimates of contribution of breeds to genetic diversity more conservative.

With regard to breed types, the greatest loss of genetic diversity occurs when only the taurine set of breeds is conserved (1.8%, Table 5.6). When only zebu or only sanga breeds are conserved the loss of genetic diversity is three times less (0.5% for zebu type breeds and 0.6% for sanga type breeds), indicating that both zebu and sanga breeds are more important to African genetic diversity.

The results of Table 5.6 indicate that both the Abyssinian region and the West African region are the most diverse. In the latter case this can be attributed to the presence of the sanga and zebu breeds, since the taurine breed set, all but one of which are West African breeds, contribute less to genetic diversity. Less important to genetic diversity was the Southern African region. The loss in genetic diversity, when only breeds from this region only are conserved was 1.3 %, almost two times higher than the loss when only breeds from Abyssinia are conserved).

Table 5.6 Loss of genetic diversity when only one set of breeds is conserved, relative to the entire set of breeds (including European and Asian breeds) or the African set (containing only the African breeds.

	WLMM			**WLM**	
	Entire	Africa		Entire	Africa
Taurine	0.018	0.012	Taurine	0.053	0.014
Zebu	0.009	0.005	Zebu	0.054	0.015
Sanga	0.010	0.006	Sanga	0.066	0.028
West Africa	0.014	0.009	West Africa	0.071	0.033
Abyssinia	0.011	0.007	Abyssinia	0.063	0.023
Lake Victoria	0.016	0.011	Lake Victoria	0.055	0.016
South Africa	0.018	0.013	South Africa	0.077	0.040
Europe	0.059		Europe	0.173	
Africa	0.003		Africa	0.040	

When all non-African breeds are excluded from the set, the loss in genetic diversity was 0.3% using WLMM. This suggests that despite the substantial contributions given to non-African populations when the entire set was analysed, the actual genetic diversity they add to the African set of breeds is only limited.

DISCUSSION

The results from this study indicate that there is a large amount of genetic diversity to be found in cattle on the African continent (Table 5.2, Table 5.6). This study also indicates that it is dependent on the context of the study (what breeds are included, analysis over regions, per region or per breed type) which breeds are the more important ones for conservation. Inclusion of foreign breeds (for reference or otherwise) in the analysis will lead to priority given to these foreign breeds, because they represent a facet of genetic diversity, which is not present among African breeds. Additionally, in the regional analyses as well as in the analyses by breed type, priority is given to those breeds that are a-typical of their category, for instance the Sheko breed, which is the only substantial contributor of zebu genes to the taurine core set and the only contributor of taurine genes to the Abyssinian core set.

This effect might cause problems in the decision making process, when the intention is to conserve genetic material native to a region or purely of one breed type. The methods employed in this paper aim to maximize the conserved genetic variation in a given core set and hence tend to focus on those breeds that carry diversity not typical of the set, i.e. having been crossed or otherwise influenced by other populations, especially when these latter populations are not included in the core set. A clear definition of the goals (maximising local diversity, maximising diversity within a breed type or maximising diversity on a continental level) to be achieved within the framework of conservation efforts is advisable, to ensure that decisions based on the results are consistent with the conservation objectives.

In a number of cases the parameter λ when using WLMM was very high (>350), as were the number of null-contributions when using WLM. In essence this means that the mean within- and between- breed kinships (μ in equations 7 and 8) receives a weight more than 350 times higher than the MEK matrix. Results from Tables 5.4 and 5.5 suggest that these large values of λ seem to be caused by the Abyssinian zebu populations.

A possible cause of the large values of λ might be an unclear definition of breeds. It is known that there are problems in properly defining separate populations of African livestock, although this is not a phenomenon exclusive to the African continent (Scherf, 2000). Often the boundaries between breeds are very diffuse and there is massive gene flow between breeds. This explains the rather high kinship between populations of zebu and sanga breeds, especially in the Abyssinian and Lake Victoria region, whereas the taurine populations (with the obvious exception of the Sheko) and the Southern African sanga populations are more isolated. It also explains the relatively low within-breed kinships, relative to between-breed kinships, of these populations, which can be seen by the lighter shading of the diagonal in Figure 5.2 or the shorter branches in Figure 5.1. Generally, for these populations the within population kinship estimates do not differ from the between population kinships all that much. Similarity of within-breed with between-breed kinships suggests that there may be as much gene flow between as there is within the breeds in question.

WLM provides weighted least square estimates of the (log transformed) Marker Estimated Kinships and are hence the most kinship estimates (Eding and Meuwissen, 2001b). It is for this reason that the MEKs from the WLM were used as the basis of Figures 5.1 and 5.2. Nevertheless, in cases where between population kinships differ little from between population kinships the error of the estimates becomes an important cause of irregularities in the MEK matrix such that a large λ is needed to remedy this situation.

It has been suggested that zebu and taurine breeds separated from a common ancestor at least 200,000 years ago (Loftus *et al.*, 1993). With regard to the comparison in terms of genetic diversity between taurines and zebu one would expect some error due to mutation. Both WLM and WLMM assume a pure drift model and mutation was not explicitly accounted for. However, in the case of African cattle populations we do not believe mutation is a major concern when studying genetic diversity, for the following reason: Both Figures 5.1 and 5.2 suggest a rather large gene flow between zebu and sanga breeds and between sanga and some taurine breeds (see also MacHugh *et al.*, 1997). Since this gene flow is of recent date and is still ongoing, the effects of gene flow on genetic diversity in African cattle are far greater than the effects of mutation.

Eding and Meuwissen (2001b) noted earlier, that WLMM provides more conservative contributions per breed to a core set. The results of this study are in agreement with this. For instance, a redistribution of contributions takes place, such that a diminished contribution for one breed leads to an increased contribution to a breed in the same cluster, thus ensuring that the genetic diversity contained in a cluster is conserved as optimally as possible.

The results in Tables 5.3, 5.4 and 5.5 give a clear indication which breeds are the major contributors to genetic diversity of the entire continent (Kenyan Boran, Blonde, Sheko, Orma Boran and Drankensberger), per region (Brune and Baladi for West Africa, Kenyan Boran and Orma Boran for Lake Victoria, Angoni and Nkone for Southern Africa and the Sheko and Horro for the Abyssininan region) or per breed type (Sheko and Brune taurines, Kenyan Boran and Orma Boran for zebus and the Drankensberger and Danakil for sanga breeds).

There is no clear indication on whether it is preferable to set up a core set per breed type or per region, although the results from Table 5.6 seem to indicate that a per breed type core set is more advantageous. It should be noted, however, that the per breed type core sets contain more populations than the per region core sets, which could explain the lower losses in diversity (on average) when all but one category are eliminated from the set.

The most efficient core set is the continental core set (Table 5.2). Most of the breeds that are considerable contributors to the African continental core set also contribute substantially to regional or breed type core sets. A continental core set requires concerted input from all regions. If this is not the case, regional optimisation may be better. However, in that case the regional gene bank will also attempt to conserve genes from neighbouring regions. Possibly the best way of assessing the contribution of breeds of a particular type or region to genetic diversity would be by estimating their contributions to a species-wide core set, encompassing all breeds and populations of a species. This would eliminate the effect where breeds with foreign influences receive disproportionate contributions. In the absence of such a species wide data set, this effect needs careful consideration when designing conservation schemes.

Chapter 6

GENERAL DISCUSSION

In this thesis genetic diversity contained in a set of breeds is defined as the maximum of genetic variation of a population in Hardy-Weinberg equilibrium derived from the breeds in a set. This definition leads to the methods outlined in Chapters 3 and 4 that maximise the genetic diversity. The contributions of individual breeds to this Hardy-Weinberg population with maximum genetic variation, which is called the core set, are optimised. Furthermore, the fraction of the genetic variation in the old base population that is maintained in the core set yields a quantitative measure of the conserved genetic diversity. This quantitative measure of genetic diversity can subsequently be used as a criterion to compare alternative conservation plans.

Core sets

When the concept of core sets was introduced in plant breeding (Frankel and Brown, 1984) it was assumed that the core set would be discrete in nature: either a strain or breed was included in the core set or it was not. Strains or breeds that were included were assumed to have equal contributions. In this thesis the concept of core sets is generalized, such that the contributions to a core set vary, giving larger contributions to more diverse populations. This allows the calculation of the maximum amount of genetic diversity contained in a set of breeds as well as the relative importance of breeds and strains to the conservation of this genetic diversity. It also gives the core set some capability to compensate for the loss or exclusion of one or more breeds, by adjusting the contributions of breeds still in the core set. Chapter 3 showed that optimum contributions of individual breeds to core sets are needed in order to let the aforementioned definition of genetic diversity satisfy Weitzman's criteria for a good measure of diversity.

Genetic diversity contributed by a single population

How much genetic diversity is lost when a breed is excluded from the core set is an additional measure of the importance of a breed. This was illustrated in Chapters 3 and 4 by the definition of a 'Safe set' containing breeds of the set that were considered safe from extinction. The genetic diversity content of a breed not in this Safe set was subsequently assessed by calculating the core set consisting of breeds in the Safe set plus the extra breed and comparing the genetic diversity of this 'Safe + 1' set to the genetic diversity of the Safe set. Thus, the genetic diversity content was calculated for each breed not in the Safe set.

Such comparisons of genetic diversities of different sets of breeds can be used to compare alternative conservation strategies. The definition of the Safe set to which other populations are added, depends on which breeds are certainly conserved. For instance, a Safe set could be defined in which a number of breeds are included *a priori* due to unique traits, whose conservation is regarded as essential. The genetic diversity of the Safe set will thus provide a base line relative to which populations not in the Safe set are assessed. Also, a conservation plan that is expected to save a population with probability 0.5 can be evaluated in this way: the expected diversity is 0.5*(diversity of the Safe set) + 0.5*(diversity in Safe +1).

It is important to realise that the estimated amount of genetic diversity of each breed not in the Safe set is dependent on the make up of the Safe set (Chapter 5). The marginal genetic diversity of a population not in the Safe set depends on what populations are included in the Safe set. Moreover, marginal diversities of the populations in the Safe set cannot be calculated without re-defining the Safe set, which also changes the baseline genetic diversity (i.e. the genetic diversity of the Safe set) from which these marginal diversities are calculated.

The results in Chapters 3 and 4 show that the absolute values of genetic diversity contributed by individual breeds are rather small, while the losses in numbers of founder genome equivalents, N_{fe}, are substantially larger. This discrepancy is noteworthy, because both genetic diversity and N_{fe} are derived from the average kinship within a set a breeds. Mathematically, this discrepancy is easy to understand. N_{ef} equals $1/\bar{f}$ while the fraction of genetic diversity left equals $\left(1-\bar{f}\right)$. It is easy to see a sharp decrease of N_{ef} results in a moderate decrease in genetic variation. In genetic terms, the number of founder genome equivalents is equal to the effective number of alleles when loci are assumed to have an infinite number of alleles in the founder population. Therefore, N_{fe} relates directly to the effective number of alleles on a locus still surviving in the present population. As was observed in Chapter 3, the results indicate that a substantial loss in N_{fe} does not seem to affect the amount of genetic variation still present in the overall population very much. When conservation of a sufficient number of alleles per locus is a consideration in a conservation program, it might be advisable to express losses from excluding breeds from the core set in terms of N_{fe} instead of genetic diversity. Doing this does not affect the ranking of breeds with respect to their contribution to diversity, but the relative contribution of a breed to N_{fe} is larger then its relative contribution to genetic variation.

Criteria for conservation

Ruane (1999) lists seven criteria, which might be used in the selection of breeds within a species for conservation programs. These are 1) the degree of endangerment, 2) adaptation to a specific environment, 3) traits of economic importance, 4) unique traits, 5) cultural or historical value, 6) genetic uniqueness. How the method presented in this thesis relates to these criteria will be discussed below.

Degree of endangerment

Obviously, a breed that is widely used does not need conservation. Not only because of safe numbers but also because a breed that is widely used usually has a relatively large (effective) population size. Conversely, breeds with small numbers of breeding animals typically have high rates of inbreeding, limiting the genetic variation within the population. This will lead to small contributions to the core set, unless the lack of within population diversity is compensated by a measure of genetic variation unique to the endangered breed. Such is the case of the Heck aurochs in the set of cattle populations in Chapter 4 (Table 4.4). Generally, however, endangered populations will be at a disadvantage, when their contribution to genetic diversity is measured, precisely because of their endangered status.

In Chapters 3 and 4, the populations were divided in a group of endangered breeds and a group of safe breeds (the Safe set in Chapters 3 and 4). Endangered breeds can then be evaluated relative to the genetic diversity they add to the genetic diversity contained in the set of breeds that is not endangered. The examples in chapters 3 and 4 illustrate that in the absence of other, related breeds at risk, the contribution of a single endangered breed can be quite substantial. The endangerment status of a population needs to be known *a priori*, although the within population kinship gives an indication of its genetic endangerment

Adaptation to specific environments

Adaptation to a specific environment will lead to a preference for keeping animals of a breed possessing this adaptation. If other breeds do not possess this adaptation this breed will be isolated from other breeds to some extent (Ruane, 1999), which will be reflected in a relatively low mean kinship of this breed to other, non-adapted breeds. The core set method approaches genetic diversity in a general way and hence maximises the expected genetic variance of the average genome. It is therefore possible that specific adaptations are not detected by the methods of Chapters 3 and 4, especially when the adaptation is due to only a few genes.

Economically important or unique traits

The concept of livestock core sets introduced in Chapter 3 maximises the genetic variation present in a set of breeds, without referring to specific traits. The variance present in a population for any polygenic trait is proportional to 1 – the average kinship within a population (Falconer and MacKay, 1996). The advantage of using kinships to measure and optimise genetic diversity lies in the fact that this single parameter describes the genetic variance of an 'average' quantitative trait.

There is a risk, however, that in the process of selecting breeds for conservation using the core set method, specific combinations of alleles are missed (Chapter 3). These specific combinations could represent unique traits, for instance the high fertility of Meishan pigs (Bidanel *et al.*, 1990), or the resistance to Trypanosomiasis ('sleeping disease') in Ndama and other Western African taurine breeds (Murray *et al.*, 1991). While the genetic variance in the core set is maximised, and therefore alleles that make up these unique traits are expected to be present in the core set if the trait is polygenic, it might require a rather substantial breeding effort to recover these specific combinations of alleles.

Cultural or historical value

Whether a breed is important culturally or historically depends on the period of time a breed has existed in a region, the importance given to products of specific breeds (Gandini, 1999), and the extent to which inhabitants of a region or members of a tribe identify themselves with a breed (Ruane, 1999). This aspect of diversity in livestock is difficult to quantify, even subjectively and as presented in Chapters 3 and 4 the core set method does not take historical or cultural value into consideration. In the section *Accounting for additional criteria* below we modify the core set method to give some extra value to specific breeds

Genetic uniqueness of breeds

Genetic uniqueness or taxonomic distinctiveness (Ruane, 1999) is considered an important criterion in conserving genetic variation in livestock. Genetically divergent populations are more likely to differ considerably in allele frequencies and haplotypes and therefore in (levels of expression of) traits. Considering taxonomic distinctiveness when prioritising breeds for conservation should ensure conservation of a range of haplotypes.

Genetic uniqueness is expressed in the form of genetic distances. Usually, when considering genetic distances between populations of livestock, the time scale is such that genetic drift is considered to be the main force (Eding and Laval, 1999) and hence, genetic distances based on differences in allele frequencies (i.e. based on genetic drift) are used to assess taxonomic distinctiveness. If genetic drift is the main force of change of allele frequencies, however, breeds with extreme allele frequencies have probably experienced more drift and thus more inbreeding and therefore display less genetic variation.

The goal of conserving genetically divergent populations is to conserve genetic variation (Ruane, 1999). The core set method was designed to capture as much as possible of the variation that was present in the ancient parental populations.

Genetic distances versus kinships

Genetic distances are closely linked to mean population kinships. The expression given in Equation (5) in Chapter 2, which expresses genetic similarities between random breeding populations in terms of allele frequencies can be found in expressions of (genetic drift based) genetic distances in one form or another (Eding and Laval, 1999). Expression (2) in Chapter 2 expresses these same genetic similarities in terms of kinships and probabilities of alleles Alike In State. Hence, it follows that drift based genetic distances can be expressed in terms of kinships and probabilities of alleles AIS. In most genetic distances the between population similarities are scaled with the within population kinships, confounding the two (Chapter 2). This can result in incorrect assessment of genetic diversity content within (a cluster of) breeds when methods are used that are based on genetic distances, including the Weitzman method (see example in Chapter 2).

Basing conservation decisions on genetic distances will lead to selection of breeds with more extreme genotypes (Chapter 2). Equating genetic distances to genetic diversity implicitly assumes that all within population kinships are equal. This assumption leads to the result that given the same between population kinships, those populations will be selected, which actually have higher within population kinships, i.e. that are more inbred. Hence, using genetic distances favours the conservation of inbred lines. It may be argued that this approach is desirable, since the genetic variation between populations is maximal. Moreover, the total genetic variance in a group of inbred lines is twice the genetic variance of the parental population (Falconer and MacKay, 1996). However, all this additional genetic variance is lost when the lines do not survive the high levels of inbreeding. Therefore, we defined genetic diversity as the variance maintained in a population in Hardy-Weinberg equilibrium that can be bred from the conserved set of breeds. The Hardy-Weinberg equilibrium requirement implies random mating across populations in the core set. This in turn implies that any between population variation due to inbreeding will be lost and thus does not contribute to the definition of genetic diversity proposed in Chapters 3 and 4.

Accounting for additional criteria

From the above one can conclude that the core set method is not constructed to account for most of the Ruane's criteria. The method is oriented on the objective conservation of genetic variation. As such the core set method will be more successful when the traits are genetically determined by many genes. But this also implies that it is possible that specific traits, whether they are adaptive, economically important or unique, may be missed in the core set. Moreover, the method does not account for cultural or historical value at all.

The criteria listed above could be accounted for explicitly by giving extra weight to breeds that have extra values due to Ruane's criteria. To this end the objective of the conservation program is re-defined to account for specific adaptations, traits and cultural/historical value. Let the relative value of breeds to these objectives be given by a vector **v**, such that the overall objective of conservation can be written as:

$$\mathit{Objective} = \mathbf{c}'\mathbf{v} + \lambda\mathbf{c}'\mathbf{F}\mathbf{c}$$

Instead of maximizing genetic diversity alone, the above objective is then maximized, e.g. by the algorithm of Meuwissen (1997). The variable λ contains the weight, which is given to genetic diversity relative to the values of the breeds in **v**. The weight λ can probably not be assessed in an objective manner, and choosing λ would therefore remain part of 'the art of genetic conservation'. This emphasizes the fact that ultimately decisions in conservation plans will depend also on subjective considerations. The contributions to genetic diversity as defined in this thesis are merely an extra argument for or against conservation of specific breeds.

Diversity on the individual level

The methods introduced in Chapters 3 and 4 are easily extended to the level of individuals. Having identified breeds for conservation, the method can be applied within a population to identify those animals that contribute to the diversity of the population. To this end either pedigree information or genetic marker data or both (Chapter 2) can be used. Individual contributions can be calculated either for a single population or over multiple populations.

In the case where pedigree information is used within a single population, the relationship matrix is calculated directly form pedigree data. Contributions to the core set can readily be calculated from this relationship matrix. As indicated in Chapter 2 this method provides the most accurate estimates of individual contributions, due to the kinship estimates being more accurate when they are taken from pedigree data (at least with the commonly used number of markers). If pedigree data is used in a multiple breed analysis, the pedigree data for a population will have to be augmented with data to account for between breed diversity. Also in Chapter 2 we show how this could be done using Wright's F-statistics, where pedigree based kinships are substituted for F_{IS} and F_{ST} is taken from the between population analysis.

In the case where only marker genetic information is available the log linear estimation method introduced in Chapter 4 will use all available information to estimate kinships between individuals. As we already discussed in Chapter 2, estimates of kinships between individuals from marker data lack accuracy when the number of markers is limited. Hence the use of the WLMM method is preferred, because it yields more conservative estimates of contributions.

Application of the core set method in conservation programs

When designing a conservation plan, the first step should be the definition of a group of breeds that are not endangered and therefore form future genetic resources without the need for conservation efforts. Then a second group of breeds should be defined containing those breeds that are regarded as essential for conservation, either due to specific traits or cultural/historical value. Together these to groups should act as a framework within which the core set method can be applied. Criteria for conservation other then genetic diversity as defined in this thesis can be accounted for, either by forming a Safe set of breeds *a priori* from the two pre-selected groups or by defining the objective as described in the section

'Accounting for additional criteria'. The core set method can then be used to choose those breeds not in either group that maximise the amount of genetic diversity that is conserved. Thus, the conservation program can be set up such, that traits deemed important are preserved, while maintaining a maximum of genetic variation in the overall conserved population.

There remains the question of what breeds to include in the analysis. As we have shown in Chapter 5, if a breed represents some genetic diversity also found in other breeds not included in the analysis, such a breed will receive a high contribution. Regional core sets will therefore have high contributions by breeds that have experienced gene flow from breeds outside the region. Ideally, one would have a species-wide set of within and between population kinship estimates to properly assess the importance of individual breeds to genetic diversity of the entire species.

Inclusion of foreign breeds in the analysis will diminish the priority of local breeds with foreign influences. If it can be assumed that these foreign breeds are either not endangered or subject to other conservation programs, the contributions of these breeds can be safely ignored in favour of regional breeds. However, this latter assumption stresses the point that for an efficient conservation effort co-operation between countries, regions and even continents is required. This co-operation should at least encompass data sharing between regions. However, the results of Chapter 5 indicate that a closer co-ordination of conservation efforts is advisable.

As we stated at the beginning of this chapter, the results from the core set method, when it is applied without accounting for any other considerations (endangerment, specific traits, cultural/historical value, etcetera) can be used as an upper limit representing the maximum of genetic variation that can be conserved. Hence, these results can serve as a starting point in testing conservation plans for efficiency in conserving genetic diversity. The risk of loss of specific traits, combinations of genes or even single alleles imply that using the results of the core set method as the only consideration for conservation is not advisable. On the other hand, focussing entirely on specific adaptations and traits is also not advisable. This becomes apparent from, for instance, QTL mapping studies in breed crosses, which revealed that the best breeds do *not* contain all the favourable alleles for a trait (Paterson, 1998; Fulton *et al.*, 2000).

CITED LITERATURE

Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize, *Theoretical and Applied Genetics* **85**, 1055-1062

Bidanel J.P., J.C. Caritez and C. Legault (1990) Ten years of experiments with Chinese pigs in France. 1. Breed evaluation. *Pig News Inform* **11**: 345-348

Bie S. de and J. Bokdam (1989) *Heckrunderen op de slikken van Flakkee II*, Vakgroep Natuurbeheer. Wageningen Agricultural University, Wageningen, The Netherlands.

Bradley D.G., David E. MacHugh, P. Cunningham and R.T, Loftus (1996) Mitochondrial diversity and origins of African and European cattle, *Proceedings of the National Acadamy of Science* **93**, 5131-5135

Bretting P.K. and M.P. Widrlechner (1995) Genetic markers and horticultural germplasm management, *HortScience* **30**, p1349-1356

Caballero A. and M.A. Toro (2000) Interrelations between effective population size and other pedigree tools for the management of conserved populations, *Genetical Research* **75**, 331-343

Cañón J., P. Alexandrino, I. Bessa, C. Carleos, Y. Carretero, S. Dunner, N. Ferran, D. Garcia, J. Jordana, D. Laloë, A. Pereira, A. Sanchez, K. Moazami-Goudarzi (2001) Genetic diversity measures of local European beef cattle breeds for conservation purposes, *Genetics Selection Evolution* **33**, 311-332

Crooijmans R.P.M.A., A.B. Groen, A.J.A van Kampen, S. van der Beek, J.J. van der Poel and M.A.M Groenen (1996) Microsatellite polymorphism in commercial broiler and layer lines estimated using pooled blood samples, *Poultry Science* **75**, 904-909

Crow J.F. and M. Kimura (1970) *An introduction to population genetics theory*, Harper&Row publishers.

Eding H, R. P.M.A. Crooijmans, M.A.M Groenen and T.H.E. Meuwissen (2001) Assessing the contribution of breeds to genetic diversity in conservation schemes, *Genetics Selection Evolution, submitted*

Eding H. and T.H.E. Meuwissen (2001a) 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 H. and T.H.E. Meuwissen (2001b) Estimation of marker based kinships to construct core sets for gene banks, *Genetics Selection Evolution, submitted*

Eding J.H. and G Laval (1999) Measuring the genetic uniqueness in livestock, in: *Genebanks and the conservation of farm animals genetic resources*, J.K. Oldenbroek (ed.) ID-DLO, The Netherlands.

Epstein H., (1971) *The origin of the domesticated animals of Africa*, Africana, New York

FAO (1998) *Primary Guidelines for Development of National Farm Animal Genetic Resources Management Plans*, FAO, Rome, Italy.

Falconer D.S. and T.F.C. Mackay (1996) *Introduction to quantitative genetics*, Longman House, Harlow.

Felius M. (1995) *Cattle breeds of the world*, Misset Uitgeverij bv, Doetinchem, The Netherlands

Felsenstein J. (1995) *PHYLIP, (Phylogeny Inference Package) Version 3.5c*, Universtity of Washington, http:\\evolution.genetics.washington.edu/phylip.html

Frankel O.H. and A.H.D. Brown (1984) Plant genetic resources today: a critical appraisal, *Crop genetic resources: conservation and evaluation*, George Allen & Unwin; London; United Kingdom

Frankham R (1994) Conservation of genetic diversity for animal improvement, *Proceedings. of the 5th WCGALP* **21**, 385-392

Fulton T.M., S. Grandillo, T. Beck-Bunn, E. Fridman, A. Frampton, J. Lopez, V. Petiard, J. Uhlig, D. Zamir and S.D. Tanksley (2000) Advanced backcross QTL analysis of a Lycopersicon esculentum X Lycopersicon parviflorum cross, *Theoretical and Applied Genetics* **7**, 1025-1042

Haig S.M., J.D. Ballou and S.R. Derrickson (1990) Management options for preserving genetic diversity: reintroduction of Guam rails to the wild, *Conservation Biology* **4**, 290-300

Hammond K. (1994) Conservation of Domestic Animal Diversity: global overview, *Proceedings of the 5th WCGALP*, **21**, 423-430

Hanotte O., C.L. Tawah, D.G. Bradley, M. Okomo, Y. Verjee, J. Ochieng and J.E.O. Rege (2000) Geographic distribution of frequency of taurine *Bos taurus* and indicine *Bos indicus Y* specific alleles amongst sub-Saharn African cattle breeds, *Molecular Ecology* **9**, 387-396

Hanotte O., D.G. Bradley, J. Ochieng, Y Verjee, E.W. Hill and J.E.O. Rege (2001) African pastoralism: Genetic imprints of origins and migration, *Science, submitted*

Hayes J.F. and W.G. Hill (1981) Modification of estimates of parameters in the construction of genetic selection indices ('bending'), *Biometrics* **37**:3, 483-493

Hedrick P.W. (1974) Genetic similarity and distance: comments and comparisons, *Evolution*, **29**: 2, 362-366.

Holland, J.H. (1975) *Adaptation in natural and artificial systems*, Ann Arbor, MI: The University of Michigan Press, Chigago

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

Jacquard A. (1983) Heritability: one word, three concepts, *Biometrics*, **39**, 465-477

Johnston L.A. and R.C. Lacy (1995) Genome resource banking for species conservation: Selection of sperm donors, *Cryobiology* **32**, 68-77

Lacy, R.C. (1989) Analysis of founder representation in pedigrees: founder equivalents and founder genome equivalence, *Zoo Biology* **8**, 111-124.

Lacy, R. (1995). Clarification of genetic terms and their use in the management of captive populations. *Zoo Biology* **14**, 565-578

Laval G., N. Iannuccelli, C. Legault, D. Milan, M.A.M.Groenen, E. Giuffra, L. Andersson, P. H. Nissen, C. B. Jârgensen, P. Beeckmann, H. Geldermann, J.-L. Foulley, C. Chevalet, L. Ollivier (2001) Genetic diversity of eleven European pig breeds, *Genetics Selection Evolution* **32**, 187-203

Li C.C., D.E. Weeks and A. Chakravarti, (1993) Similarity of DNA fingerprints due to chance and relatedness, *Human Heredity* **43**, 45-52

Loftus T.L., D.E. MacHugh, D.G. Bradley, P.M. Sharp and P. Cunningham (1993) Evidence for two independent domestications of cattle, *Proceedings of the National Acadamy of Science* **91**, 2757-2761

Lynch M. (1988) Estimation of relatedness by DNA fingerprinting, *Mol.Biol.Evol* **5**, 584-599

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

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

MacHugh, D.E., M.D. Shriver, R.T. Loftus, P. Cunningham and D.G. Bradley (1997) Microsatellite DNA variation and evolution, domestication and phylogeography of taurine and zebu cattle (*Bos Taurus* and *Bos indicus*), *Genetics* **146**, 1071-1086

Malécot G (1948) *Les mathématiques de l'hérédité*. Masson, Paris.

Meuwissen T.H.E. (1997) Maximizing the response of selection with a predefined rate of inbreeding, *Journal of Animal Science* **75**:4, 934-940

Moazami-Goudarzi K., D. Laloë, J.P. Furet and F. Grosclaude (1997) Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites, *Animal Genetics* **28**, 338-345

Murray M., M.J. Stear, J.C.M. Trail, G.D.M. d'Ieteren, K. Agyemang and R.H. Dwinger (1991) Trypanosomiasis in cattle: prospects for control. In: Breeding for disease resistance in farm animals. Edited by J.B. Owen and R.F.E Axford. CAB International, Wallingford, United kingdom

Nagylaki, T. (1998) Fixation indices in subdivided populations, *Genetics* **148**, 1325-1332

Nauta M.J. and F.J. Weissing (1996) Constraints on allele size at microsatellite loci: Implications for genetic differentiation, *Genetics* **143**, 1021-1032

Nei M. (1972) Genetic distance between populations, *American Naturalist* **106**, 283-292

Oldenbroek J.K. ed. (1999) *Genebanks and the conservation of farm animal genetic resources,* ID-DLO, Lelystad, The Netherlands

Paterson A.H. (1998) *Molecular dissection of complex traits*, CRC Press, Boca Raton, USA

Reynolds J. (1983) Estimation of the coancestry coefficient basis for a short-term genetic distance, *Genetics* **105**, 767-779.

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

Scherf B.D. (2000) *World Watch List for domestic animals*, FAO, Rome, Italy

Slatkin M. (1995) A measure of population subdivision based on microsatellite allele frequencies, *Genetics* **139**, 457-462

Rege J.E.O., C.V. Yapi-Gnaore and C.L. Tawah (1996) The indigenous domestic ruminant genetic resources of Africa, *Proceedings 2nd Africa Conference on Animal Agriculture*, Pretoria, South Africa, 57-75

Takezaki N. and M. Nei (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA, *Genetics* **144**, 389-399

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

Thompson E.A. (1975) The estimation of pairwise relationships, *Annals of Human Genetics* **39**, 173-188

Toro M., L. Silio, J. Rodriganez and C. Rodriguez (1998) The use of molecular markers in conservation programmes of live animals, *Genetics Selection Evolution* **30**, 585-600

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

Wit J. de, J.K. Oldenbroek, H. van Keulen, D. Zwart (1995) Criteria for sustainable livestock production: a proposal for implementation, *Agriculture, Ecosystems and Environment* **53**, 219-229

Wright S. (1968) *Evolution and the genetics of populations*, Vol. II, Univ. of Chicago Press, London, United Kingdom

Zheng Y.Q., D. Lindgren, O. Rosvall, J. Westin (1997) Combining genetic gain and diversity by considering average coancestry in clonal selection of Norway spruce, *Theoretical and Applied Genetics* **95**, 1312-1319

SUMMARY

In this thesis a method is developed, that can be used to evaluate populations with regard to their contribution to genetic diversity. Genetic diversity in this thesis is defined as the maximum of genetic variation represented by a set of populations. This definition of genetic diversity is neutral with regard to specific traits or adaptations of breeds to specific environments and is meant to reflect the overall genetic variation present in (a set of) breeds. The method is based on the concept of kinship or coancestry, since this parameter determines how much of the total genetic variation is present in a population. Because there are a number of possible situations in which kinships cannot be calculated from pedigree records (between breed kinships, poor administration of breeds), this thesis focuses on kinships estimated from genetic marker information.

Chapter 2 introduces a method to estimate average kinships between and within populations from genetic marker data. A similarity index based on Malécots coefficient of kinship or equivalently, the probability of alleles identical by descent (IBD) is introduced. The expectation of similarity between individuals x and y for a locus l is: $\mathrm{E}(S_{xy,l}) = f_{xy} + (1\text{-}f_{xy})s_l$, where f_{xy} is the kinship between individuals x and y and s_l is the probability of alleles alike in state (AIS), not identical by descent. Averaging the similarity index over random breeding populations yields $S_{ijk} = \Sigma(p_{ik}p_{jk})$, or the average genetic similarity between populations i and j on locus k is the product sum of allele frequencies in both populations.

To obtain kinship estimates between populations the per locus similarities are corrected for the probability of alleles AIS, s_l, which is determined by the distribution of allele frequencies in the founder population. The founder population is defined as the population that is the most recent ancestor of all populations for which kinships are estimated. Furthermore, since the average kinship within the founder population is assumed to be 0, probabilities of alleles AIS can be inferred from the smallest between population similarities. Average marker estimated kinships (MEK) between populations are subsequently estimated by averaging similarities corrected for alleles AIS over loci. Applied between populations these MEK are sufficiently accurate when a limited number of marker loci is used. Applied between individuals, however, the results show that between 30 to 50 sufficiently polymorphic marker genes are needed to distinguish between full sibs, half sibs (and equivalents) and less related animals.

Chapter 3 applies the concept of core sets to populations of livestock. Genetic diversity is defined as the maximum genetic variation present in a random breeding population derived from populations in the core set. To estimate the maximum variation, relative contributions of populations to the core set are optimized in order to minimize the average kinship within the core set. To account for the fact that a number of populations are not expected to become extinct in the short term, the contributions of endangered populations to genetic diversity are also analyzed relative to a set of safe populations.

The optimized contributions to a core set are insensitive to definition of the founder population. As long as the average kinships are known proportionally, the optimization will yield the same results, irrespective of the absolute values of the between population kinships. The core set method is applied to a dataset of poultry breeds and lines. Kinships are estimated using a weighted least similarity method, where AIS probabilities are set equal to the minimum average between population similarity. Sampling errors of allele frequencies lead to errors on the kinship estimates, which can lead to an inconsistent kinship structure. These inconsistencies can lead to contributions to the core set that are smaller then zero.

Chapter 4 introduces log-linear estimation procedures to estimate kinships and probabilities of alleles AIS simultaneously, using all information available. Using all information increases the accuracy of estimation, which reduces the number of negative contributions. The models are: Unweighted Log-linear Model (ULM), Weighted Log-linear Model (WLM), where marker data is weighted to account for the amount of information per locus and Weighted Log-linear Mixed Model (WLMM), where the solution is restricted such that a maximum of one zero-contribution remains. These models are tested using simulated data and compared to the results from the Weighted Least Similarity method introduced in Chapter 3. The WLM method provides the most accurate estimations of kinships. However, the estimated contributions of the WLMM are more conservative and more equally distributed over all populations in the analysis, due the restriction on negative contributions.

Chapter 5 applies the methods and procedures developed in the previous chapters to marker genetic data on a set of 59 mostly African cattle populations. These populations are taken from different regions and breed types (zebu, sanga and taurines). Assessing breeds using the core set method gives priority to breeds with genetic influences (i.e. gene flow), which are a-typical of the set of breeds that is analysed.

Chapter 6 discusses the relation of results from an analysis using the core set method and other considerations for conservation. Applied without regarding other considerations, the core set method can be used to calculate the maximum amount of genetic variation that can be conserved. However, strict adherence to the results of this analysis can lead to the loss populations possessing specific traits or cultural/historic value. Such considerations need to be accounted for explicitly. This is especially the case for breeds at risk of extinction, since they usually do not contribute very much to overall genetic diversity.

When a number of breeds are designated for inclusion in the core set, either because they are not endangered, or they possess traits that are deemed important for conservation, the core set method can be used to set up conservation efforts that will ensure conservation of the maximum amount of genetic diversity.

NEDERLANDE SAMENVATTING

De aanleiding voor dit proefschrift is het feit dat er wereldwijd een trend is om lokale rassen landbouwhuisdieren te vervangen door een klein aantal wereldwijd gebruikte, hoog productieve rassen, waarmee over het algemeen ook intensief gefokt wordt. Een van de gevolgen van deze trend is het afnemen van de (genetische) variatie tussen soorten landbouwhuisdieren, met het risico dat er eigenschappen verloren gaan die van belang zijn voor bijvoorbeeld de aanpassing aan specifieke lokale omstandigheden. Als er sprake is van intensieve fokkerij met een beperkt aantal rassen neemt ook de inteelt binnen een soort versneld toe, hetgeen verlies van fitness of vermogen tot reproductie en/of het verlies van vermogen tot aanpassen aan nieuwe omstandigheden tot gevolg heeft.

In dit proefschrift wordt een methode uitgewerkt, waarmee rassen (en individuele dieren) beoordeeld kunnen worden op hun bijdrage aan genetische diversiteit. Genetische diversiteit is hier gedefinieerd als de hoeveelheid genetische variatie aanwezig in een verzameling rassen of populaties van een soort landbouwhuisdieren. Als rassen veel genetische overeenkomsten vertonen, zal de genetische variatie laag zijn. Het maximum aan variatie is te verkrijgen door de samenstelling van de verzameling rassen zo te kiezen dat er zo min mogelijk overlap in genetische eigenschappen aanwezig is.

Een dergelijke definitie is onafhankelijk van gekozen kenmerken of van aanpassingen van een ras aan een omgeving en is bedoeld om de genetische variatie in (een verzameling) rassen in een algemene zin aan te geven. Bij deze benadering wordt geen rekening gehouden met specifieke eigenschappen van rassen of populaties, vandaar de toevoeging 'in algemene zin'. De methode is gebaseerd op bloedverwantschappen tussen dieren. Daar bloedverwanten genetisch gezien overeenkomsten vertonen, geeft de gemiddelde bloedverwantschap in een populatie aan hoeveel genetische variatie er bestaat in een populatie. Hoe hoger de verwantschappen, des te lager de genetische variatie. Bloedverwantschappen kunnen berekend worden uit stamboomgegevens. Stamboom gegevens gaan echter vaak niet meer dan 5 à 10 generaties terug. Dit is niet voldoende om de bloedverwantschappen tussen rassen uit te rekenen. Daarom ligt de nadruk van dit proefschrift op het schatten van bloedverwantschappen op basis van informatie over merkergenen.

Hoofdstuk 2 begint met de introductie van een methode om de gemiddelde bloedverwantschappen tussen en binnen populaties te schatten op basis van merkergengegevens. Er wordt een *merkerindex* tussen dieren geïntroduceerd die aangeeft op hoeveel merkergenen de dieren gelijke allelen bezitten. Dieren met veel gelijke merkerallelen zijn nauwer verwant en dieren met weinig gelijke merkerallelen zijn weinig verwant.

Merkerallelen kunnen echter ook door toeval gelijk zijn. Om de verwantschapgraad re bereken moeten we voor dot toeval corrigeren. In Hoofdstuk 2 is een eerste aanzet gegeven om voor dit toeval te corrigeren. Hiertoe werd de verwantschappen tussen de meest onverwante rassen (gebaseerd op de merkerindex) op nul gesteld, waardoor de toevalskans op gelijkheid van merkerallelen door verwantschap berekend kan worden.

Gemiddelde verwantschappen tussen populaties kunnen vervolgens geschat worden door over een aantal loci de merkerindex te corrigeren voor de kans dat twee allelen gelijk zijn door toeval en te middelen. Toegepast op populaties zijn deze geschatte verwantschappen voldoende accuraat, ook als slechts een beperkt aantal merkergenen zijn gebruikt. Toegepast op individuen echter, laten de resultaten zien dat er zo'n 30 à 50 merkergenen nodig zijn om een onderscheid te kunnen maken tussen volle broers en zussen, halfbroers en –zussen (en equivalente verwantschappen) en minder verwante dieren.

Hoofdstuk 3 past het concept kernverzameling toe op populaties landbouwhuisdieren. Een kernverzameling is een verzameling populaties die zó is samengesteld dat zij de maximale hoeveelheid genetische diversiteit vertoont met zo min mogelijk overlap in genetische eigenschappen. In dit hoofdstuk wordt genetische diversiteit gedefinieerd als de maximale hoeveelheid genetische variantie in een populatie die is samengesteld uit de populaties in de kernverzameling. Om die maximale hoeveelheid variantie te kunnen schatten worden de relatieve bijdragen van de populaties aan de kernverzameling geoptimaliseerd. In feite wordt per populatie berekend hoeveel dieren aan de samengestelde populatie moeten worden toegevoegd om de genetische variatie in de samengestelde populatie de maximaliseren.

Voor de berekening van genetische diversiteit en kernverzamelingen is het noodzakelijk dat alle beschikbare populaties of rassen in de analyse voorkomen. Een aantal van die rassen zal niet vallen in de categorie bedreigd en zullen in ieder geval op de korte termijn beschikbaar blijven als genetische hulpbron. Om rekening te houden met het feit dat een aantal populaties niet op de korte termijn bedreigd worden met uitsterven, wordt in dit hoofdstuk de waarde van een bedreigd ras voor de genetische diversiteit uitgedrukt als de toename in diversiteit wanneer het *bedreigde* ras aan de verzameling *onbedreigde* rassen wordt toegevoegd.

De geoptimaliseerde bijdragen worden berekend uit de geschatte verwantschappen. Dientengevolge zou te verwachten zijn dat de geoptimaliseerde bijdragen afhangen van de hoogte van die verwantschappen. En de hoogte van de verwantschappen hangen af van hoe ver men terugkijkt in de stamboom. Hoe verder men terugkijkt hoe hoger de verwantschappen. De geoptimaliseerde bijdragen zouden dus afhankelijk kunnen zijn van hoever men terugkijkt. Dit 'terugkijken in de stamboom' gebeurt hier met merkergenen, maar ook dan kan men niet verder terugkijken dan tot op het moment dat de allereerste populaties zich opsplitsten. Met de kernverzamelingsmethode wordt de optimalisatie echter zodanig uitgevoerd dat de geoptimaliseerde bijdragen aan een kernverzameling onafhankelijk zijn van hoe ver men terugkijkt. Zo lang de geschatte verwantschappen tussen populaties zich verhouden als in de echte verwantschappen, zal de optimalisatie van bijdragen dezelfde resultaten geven, onafhankelijk van de absolute waarden van de geschatte verwantschappen. In Hoofdstuk 3 wordt deze kernverzamelingsmethode toegepast op gegevens over merkergenen van pluimveerassen en -lijnen. Verwantschappen worden in dit hoofdstuk geschat met behulp van een methode, waarbij de verwantschap tussen de populaties die als eerste splitsen op nul wordt gezet. In feite wordt hierdoor de verwantschap uitgedrukt t.o.v. de minst verwante rassen in de dataset. Ook worden de geschatte verwantschappen per merkergen gewogen met hun 'informatie inhoud'. Het punt is namelijk dat bij een locus dat veel verschillende allelen bezit, de kans kleiner is dat twee allelen gelijk zijn door toeval. Een gen met meer allelen bevat dus meer en betere informatie over de verwantschap tussen dieren dan een gen met minder allelen.

Een probleem bij de kernverzamelingsmethode is het feit dat in sommige gevallen rassen of populaties onterecht een negatieve bijdrage aan genetische diversiteit toebedeeld krijgen. Toevalsafwijkingen door de keuze van de gegenotypeerde dieren leiden tot variatie op de schattingen van verwantschappen. Toevalsvariatie op de schattingen kan leiden tot fouten in de verwantschappenstructuur, wat op zijn beurt de oorzaak is van negatieve bijdragen aan genetische diversiteit van sommige populaties, terwijl dergelijke bijdragen altijd of nul of positief horen te zijn. Om dit probleem te ondervangen, zijn in Hoofdstuk 4 statistische verschillende modellen toegepast. Deze modellen gebruiken alle beschikbare informatie om de verwantschappen preciezer te schatten. In Hoofdstuk 4 worden een aantal log-lineaire procedures om verwantschappen te schatten getest. De log-lineaire modellen leveren de meest accurate schattingen van verwantschappen tussen populaties, maar geven nog steeds een aantal negatieve bijdragen op. Daarom is ook een methode ontwikkeld die voorzichtiger is met het uitdelen van bijdragen en in feite de bijdrages gelijker verdeeld over alle populaties in de analyse. Deze methode gaf geen negatieve bijdragen, maar de bijdrages waren wel iets minder nauwkeurig geschat.

In hoofdstuk 5 worden de methoden en procedures uit de vorige hoofdstukken toegepast op merkergen-gegevens van 59 vooral Afrikaanse rundveepopulaties. Deze populaties werden gekozen uit verschillende regio's en rastypen (zebu, sanga en *bos taurus* lijnen). De resultaten laten zien dat toepassing van de kernverzamelingsmethode prioriteit geeft aan rassen die genetische buitenbeentjes zijn. Dat zijn vooral die rassen die invloeden van buiten Afrika hebben ondergaan.

In hoofdstuk 6 wordt de relatie besproken tussen resultaten van een analyse met behulp van de kernverzamelingsmethode met andere overwegingen die bij conservering van genetisch diversiteit van belang zijn, zoals de mate waarin populaties bedreigd zijn, het belang van specifieke kenmerken en de cultureel-historische waarde van een ras. Als de kernverzamelingsmethode toegepast wordt zonder rekening te houden met dergelijke andere overwegingen berekent de methode de genetische variatie in algemene zin. Strikte toepassing van de resultaten van een dergelijke analyse draagt echter het risico met zich mee dat rassen die in het bezit zijn van specifieke kenmerken en/of van cultureel-historische waarde zijn verloren gaan. Met dergelijke overwegingen zal expliciet rekening gehouden moeten worden. Dat is zeker het geval waar het bedreigde rassen betreft, omdat zij vaak niet veel bijdragen aan genetische variatie in algemene zin als gevolg van de inteelt in die rassen.

Als een aantal rassen vooraf aangewezen worden voor opname in de kernverzameling, zij het omdat zij niet bedreigd zijn, of omdat zij kenmerken bezitten die als belangrijk voor conservatie worden beschouwd, dan kan de kernverzamelingsmethode gebruikt worden om een plan van conservatie van genetische diversiteit op te zetten, waarmee de conservatie van zoveel mogelijk genetische diversiteit verzekerd is.

Curriculum vitae

Jacob Hendrik Eding werd geboren op 18 maart 1969, te Zaandam. Na het doorlopen van de HAVO aan het Melanchton college in Rotterdam en het Ichtus college in Enschede, behaalde hij zijn VWO diploma in 1989. In datzelfde jaar ving hij zijn studie Zoötechniek aan de Landbouw Universiteit Wageningen aan met als specialisatie Veefokkerij. Na twee afstudeervakken Veefokkerij en een stage bij Agriculture Canada te Ottowa, Canada, behaalde hij zijn bul in november 1995.

Na een korte periode als fokkerijmedewerker bij het Dienstencentrum Schapenfokkerij te Lelystad en als medewerker bij het Crisis- en Coördinatiecentrum Varkensopkoop te Helmond ten tijde van de varkenspestepidemie in 1997, begon hij als AIO bij de leerstoelgroep Fokkerij en Genetica van Wageningen Universiteit. Het promotie-onderzoek, zoals beschreven in dit proefschrift, werd uitgevoerd bij de divisie Dier en Omgeving van ID-Lelystad, onder begeleiding van dr. ir. Theo Meuwissen.

Sinds oktober 2001 is hij werkzaam als tijdelijk medewerker bij de divisie Dier en Omgeving.