J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Epistatic kinship a new measure of genetic diversity for short-term phylogenetic structures – theoretical investigations

C. Flury, M. Tietze & H. Simianer

Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen, Gottingen, Germany

Correspondence
Christine Flury, Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Gottingen, Germany. Tel: +49 551 39 56 28; Fax: +49 551 39 55 87; E-mail: cflury@gwdg.de

Received: 17 June 2005;
accepted: 7 December 2005

Summary

The epistatic kinship describes the probability that chromosomal segments of length x in Morgan are identical by descent. It is an extension from the single locus consideration of the kinship coefficient to chromosomal segments. The parameter reflects the number of meioses separating individuals or populations. Hence it is suggested as a measure to quantify the genetic distance of subpopulations that have been separated only few generations ago. Algorithms for the epistatic kinship and the extension of the rules to set up the rectangular relationship matrix are presented. The properties of the epistatic kinship based on pedigree information were investigated theoretically. Pedigree data are often missing for small livestock populations. Therefore, an approach to estimate epistatic kinship based on molecular marker data are suggested. For the epistatic kinship based on marker information haplotypes are relevant. An easy and fast method that derives haplotypes and the respective frequencies without pedigree information was derived based on sampled full-sib pairs. Different parameters of the sampling scheme were tested in a simulation study. The power of the method decreases with increasing segment length and with increasing number of segments genotyped. Further, it is shown that the efficiency of the approach is influenced by the number of animals genotyped and the polymorphism of the markers. It is discussed that the suggested method has a considerable potential to allow a phylogenetic differentiation between close populations, where small sample size can be balanced by the number, the length, and the degree of polymorphism of the chromosome segments considered.

Introduction

Phenotypic selection since domestication has created a wide diversity of breeds of domestic animal that are adapted to different climatic conditions and purposes (Andersson 2001). Today more than 20% of the roughly 6400 documented breeds are at risk of extinction (Scherf 2000). Due to limited financial and human resources, not all breeds can be given the same priority for conservation (Oldenbroek 1999). One – but not the only – important criterion (Ruane 1999) is the uniqueness of breeds. Genetic distance studies are based on evolutionary models, which often do not hold for the development of livestock breeds. Most of the approaches were developed for the description of the evolutionary differentiation

between species, while for livestock the differentiation occurred within species (Ruane 1999; Simianer 2002a).

The formation of today's breeds goes back to the 19th or even the beginning of the 20th century (Sambraus 2001). Thus the assumption of an evolutionary time span does not hold for breed differentiation. Based on the reduced divergence time the role of mutation in creating differences between breeds is expected to be small (Takezaki & Nei 1996; Toro & Caballero 2004).

Toro & Caballero (2004) summarized further problems of conservation decisions based on phylogenetic diversity like the complete ignorance of genetic variance within population, the failure of principles of phylogeny reconstruction to account for population admixture, the problems arising from varying distances among the markers used and the impact of the demographic history of a population. Also, markers used for genetic distances are assumed to represent neutral loci.

Ignoring the genetic variance within population often leads to the conservation of the most inbred population (Eding 2002). To overcome this weakness of genetic distances the authors proposed mean coefficients of kinship between and within populations as a tool to assess genetic similarity in livestock populations. The coefficient of kinship K_{st} is defined as the probability that two randomly sampled alleles from the same locus in two individuals S and T are identical by descent (ibd) (Malécot 1948). Another concept for the estimation of genetic similarity between individuals is the coefficient of relationship R_{st} specified by Wright (1922). The link between the two parameters is $R_{st} = 2K_{st}$. Kinship coefficients can be calculated based on pedigree information (Cockerham 1967). As pedigree data are often not available for small livestock populations, some authors suggested the estimation of kinship coefficients based on marker information (Caballero & Toro 2000; Eding & Meuwissen 2001). Having non-unique founder alleles the correction for alleles identical by state, but not ibd is crucial. Lynch (1988) proposed a similarity index to overcome this problem for single loci. Eding & Meuwissen (2001) showed that marker-based estimates of kinship yielded higher correlations with pedigree-based kinships than genetic distance measures.

Coefficients of kinship refer to the ibd probability for a randomly chosen single locus or an average overall loci (Simianer 1994). This presumes independently segregating loci. For the genetic control of important traits the formation of gene complexes over multiple loci and epistatic interactions is important (Brockmann *et al.* 2000). Various studies investigate the properties of conserved haplotypes around a functional polymorphism. Haplotype sharing is important in the context of ibd mapping of quantitative trait loci (Meuwissen & Goddard 2000; Nezer *et al.* 2003). The length of conserved haplotypes depends on the timespan since separation or rather the number of recombination events. Visscher (2003) suggests that linkage disequilibrium (LD) created by crossbreeding may still persist in many of todays livestock populations, because crossbreeding was commonly practised from 50 to 100 generations ago. Coppieters *et al.* (1999) and Farnir *et al.* (2000) found strong evidence for long range LD for all autosomes of the Holstein Friesian population, with LD extending over regions >20 cM. Beside other factors they explain the disequilibrium particularly with drift, due to the small effective population size of the Holstein Friesian population.

In this study, we assume the existence of LD for small livestock populations and propose a diversity measure based on shared haplotypes within and between populations. Therefore, the coefficient of kinship will be extended from single loci to chromosomal segments of length x in Morgan. This leads to a new similarity index called epistatic kinship, which describes the probability of chromosomal segments being ibd. A similar measure was proposed by Hayes *et al.* (2003) as chromosome segment homozygosity for the estimation of past effective population size.

In the *Methods* section this parameter will be defined and algorithms to calculate epistatic kinship, epistatic relationship coefficient, epistatic inbreeding and the epistatic kinship matrix will be presented. An extension from the average homozygosity (Falconer & Mackay 1996) to average expected epistatic kinship is derived. The properties of the average epistatic kinship as a tool for the analysis of short-term phylogenetic structures are investigated for a known simulated pedigree structure in the first *Results and discussion* section. In the second *Results and discussion* section the epistatic kinship will be estimated based on marker information. Typing of animals result in genotypes, thus a method to derive haplotypes from genotyping information is needed. Different algorithms to infer haplotypes exist and are discussed by Niu (2004). For some algorithms pedigree information is a prerequisite, others who run without pedigree information are often complex and computing intensive (Windig & Meuwissen 2004). An easy and fast method to derive haplotypes without pedigree information or in simple standard pedigrees [e.g.

only full-sib pairs (FSP) are available] is suggested. The efficiency of the differentiation of close populations based on average epistatic kinship was compared for reconstructed versus true haplotypes.

Methods

Epistatic kinship, epistatic relationship and epistatic inbreeding

We define K_{st} as Malécot's (1948) kinship coefficient between individuals S and T, reflecting the probability that a randomly chosen allele at a given locus of individual S is ibd with a randomly chosen allele at the same locus in animal T. Consider now a randomly chosen chromosome segment of length x in Morgan. We chose at random one of the two homologous strands of this chromosome segment in individuals S and T respectively. We define K_{st}^x as the probability, that these two strands are ibd and call this parameter 'epistatic kinship'. This name is derived from the use of the same parameter to estimate epistatic effects in gene clusters, which is described in a companion paper (Flury *et al.* 2006).

The extension from single locus to chromosomal segments requires a correction for the probability that crossing over occurs. Under the assumption that crossingover events follow a Poisson distribution, the probability that an entire chromosome strand of length x is inherited without crossing over is e^{-x}. Consider an offspring T of animal S with the two strands t_1 and t_2 at the considered region. The probability that a randomly chosen strand of T, say t_i where i is either 1 or 2, is ibd with a randomly chosen strand s_j, $j = 1$ or 2, of animal S is $K_{st}^x = K_{st} \times e^{-x}$ thus $0.25e^{-x}$. Note that for $x = 0$ the value of $e^{-x} = 1$ and the probability equals the kinship coefficient $K_{st} = 0.25$, hence Malécot's kinship coefficient is a special case of the epistatic kinship coefficient for $x = 0$.

It is straightforward to extend the analogy of Malécot's kinship coefficient K_{st} and Wright's (1922) relationship coefficient $R_{st} = 2K_{st}$ to epistatic kinship and epistatic relationship, i.e. $R_{st}^x = 2K_{st}^x$.

There is also an analogy to the usual inbreeding coefficient F_j as defined by Wright (1922). For the extension to chromosome segments, we have to account for crossingover events in the formation of the parental gametes.

Epistatic inbreeding can be derived from the epistatic kinship of an individual by itself. Consider animal J with sire S and dam D and denote the two homologous strands of individual J at a given chromosome segment as s and d, reflecting the paternal and maternal origin. We sample at random two strands (with replacement) of individual J. The sampled pairs are, with equal probability 0.25, $\{s,s\}$, $\{s,d\}$, $\{d,s\}$ or $\{d,d\}$ respectively. In half of the cases, $\{s,s\}$ and $\{d,d\}$, the two sampled strands are clearly ibd because the same strands of animal J were sampled. For the sampled pairs $\{s,d\}$ and $\{d,s\}$, the chromosome segments are only entirely ibd if they were already ibd in the parents, of which the probability is K_{sd}^x, and if they were both inherited without crossing over. Hence, for a chromosome segment of length x,

$$K_j^x = 0.5 \times 1 + 0.5 \times K_{sd}^x \times (e^{-x})^2 = 0.5(1 + e^{-2x}K_{sd}^x).$$

Using this result

$$2 \times K_j^x = 1 + e^{-2x}K_{sd}^x = 1 + F_j^x,$$

which leads to the definition of the epistatic inbreeding coefficient

$$F_j^x = e^{-2x}K_{sd}^x = 0.5e^{-2x}R_{sd}^x.$$

The epistatic relationship matrix

The epistatic relationship matrix A^x for N individuals is a matrix of dimension $N \times N$ where element

$$A_{ij}^x = R_{ij}^x \text{ for } i \neq j \quad \text{and} \quad A_{ii}^x = 1 + F_i^x.$$

Note that for $x = 0$ the epistatic relationship matrix becomes the well-known numerator relationship matrix.

Analogously to the tabular method to set up the numerator relationship matrix (Emik & Terrill 1949), the following procedure is suggested.

The animals are numbered by age from 1 to N such that the oldest animal is number 1. A pedigree list is defined giving for each animal the sire and dam number. All animals appearing as sires and dams also have to have an animal number between 1 and N. Unknown parents are denoted by '0'.

Using this pedigree list, the following algorithm is performed:

(i) set $i = 1$ and $A_{11}^x = 1$;

(ii) set $i = i + 1$, read sire s and dam d of animal i from the pedigree list;

(iii) set $A_{ii}^x = 1 + 0.5e^{-2x}A_{sd}^x$ if s and d are $\neq 0$, otherwise set $A_{ii}^x = 1$;

(iv) let j go from 1 to $i - 1$, set $A_{ji}^x = 0.5e^{-x}(A_{js}^x + A_{jd}^x)$. If $s = 0$ $(d = 0)$ use $A_{js}^x = 0$ $(A_{jd}^x = 0)$. Finally set $A_{ij}^x = A_{ji}^x$.

(v) If $i < N$ continue with step 2.

After going through these steps for all animals, the epistatic relationship matrix is complete. The junction between the epistatic relationship matrix A^x and the epistatic kinship matrix K^x is $K^x = 0.5A^x$.

Expected epistatic kinship within and between populations

Assuming an ideal population of size N, the average homozygosity F_t in generation t can be computed by the recursive formula (Falconer & Mackay 1996):

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_{t-1}. \quad (1)$$

This equation is of two parts: the first expression $1/2N$ is the 'new' homozygosity which is generated in the meiotic sampling of the gametes leading to generation t, and $(1 - (1/2N))F_{t-1}$ is the 'old' homozygosity which was built up in generations 1 to $t - 1$.

If we use the same rationale to derive the expected epistatic kinship for a chromosome segment of length x, we have two processes, which overlay each other: in each generation, new epistatic kinship is generated by the sampling process, while at the same time old epistatic kinship is partly destroyed through crossing over.

In generation t, $2N$ chromosome segments are sampled from the pool of chromosome segments in generation $t - 1$. Each chromosome segment will show no crossing over with probability e^{-x}. Therefore, the probability that two randomly chosen chromosome segments in generation t are new epistatic homozygotes is $e^{-2x}/2N$. Old epistatic homozygotes may lose this property in any subsequent generation. The probability that an old epistatic homozygote existing in generation $t - 1$ stays homozygote in generation t is e^{-2x}. Combining these findings, the average expected epistatic kinship $\bar{K}_t^x$ in generation t can be calculated by the recursive formula

$$\begin{aligned}\bar{K}_t^x &= \frac{e^{-2x}}{2N} + e^{-2x}\left(1 - \frac{1}{2N}\right)\bar{K}_{t-1}^x \\ &= e^{-2x}\left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\bar{K}_{t-1}^x\right]. \quad (2)\end{aligned}$$

Note that the recursion (1) for single loci is a special case with $x = 0$. The resulting function of $f(t) = \bar{K}_t^x$is convex and asymptotically goes for $t \to \infty$ to

$$\bar{K}_{\max}^x = \frac{e^{-2x}}{e^{-2x} + 2N(1 - e^{-2x})}. \quad (3)$$

If a population is split in subpopulations in generation t' and these subpopulations are maintained without genetic exchange, no new epistatic kinship will be generated between these populations. The average epistatic kinship on the level of the time of fission will be maintained as the epistatic kinship between these populations if $x = 0$, but this old epistatic kinship will erode with the rate e^{-2x} in every generation through crossing over with $x > 0$. Thus, the between population expected average epistatic kinship in generation g after fission is

$$\bar{K}_{t'+g}^x = e^{-2xg}\bar{K}_{t'}^x. \quad (4)$$

Note that the rate of erosion of epistatic kinship between separated populations is independent of the population size.

Epistatic kinship based on pedigree information

It is suggested to use the epistatic kinship to differentiate phylogenetically close populations. The hypothesis is that this metric is more sensitive to small phylogenetic distances caused by short-time since separation than conventional distance metrics, which are based on mutation and/or genetic drift as the diversity generating process. It was assumed, that the full pedigree of two subpopulations back to a common base population was known. Samples were taken from the two subpopulations in the latest generation and it was tested, whether the average epistatic kinship between populations differed from the average epistatic kinship within populations.

The test was based on a random sample of M individuals in each of the two populations. For these individuals, L chromosome segments of length x were considered. For each pair of the $2M$ individuals the epistatic kinship was calculated using the tabular method described above.

For the statistical test, it was necessary to take the number of informative comparisons into account. An illustration and the corresponding approximations for the number of informative comparisons within populations N_w and between populations N_b are given in the Appendix.

Because in each comparison four different pairs of chromosome segments can be compared, the number of pairwise comparisons within (V_w) and between (V_b) populations are:

$$V_w = N_w \times 4L,$$

$$V_b = N_b \times 4L.$$

Note that the number of comparisons within and between populations is a linear function of the number of chromosome segments considered, L and a quadratic function of the number of animals sampled, M.

The average ibd probability within populations is denoted as p_w and the average ibd probability between populations is denoted as p_b.

To test the hypothesis

$$H_0 : p_w = p_b = p_0 \quad \text{versus} \quad H_a : p_w > p_b$$

the chi-squared test statistic was calculated using the basic formula

$$X^2 = \frac{(p_w V_w - p_0 V_w)^2}{p_0 V_w} + \frac{[(1-p_w)V_w - (1-p_0)V_w]^2}{(1-p_0)V_w} + \frac{(p_b V_b - p_0 V_b)^2}{p_0 V_b} + \frac{[(1-p_b)V_b - (1-p_0)V_b]^2}{(1-p_0)V_b}.$$

Using the average ibd probability p_0 under the null hypothesis

$$p_0 = \frac{p_w \times V_w + p_b \times V_b}{V_w + V_b}$$

the expected test statistic is

$$E(X^2) = \frac{(p_w - p_0)^2 \times V_w + (p_b - p_0)^2 \times V_b}{p_0} + \frac{(p_w - p_0)^2 \times V_w + (p_b - p_0)^2 \times V_b}{(1-p_0)}. \quad (5)$$

As this test statistic is not based on actual, but expected numbers of ibd segments under a specific realization of the alternative hypothesis, we denote $E(X^2)$ as the expected test statistic and assume that a higher value of this parameter corresponds with a higher power.

Epistatic kinship based on marker information

In applications to real-life data, the pedigree of animals from different populations back to common ancestors from one common base population rarely is available. Therefore, it is necessary to assess the ibd status of chromosome segments based on genotyping information from marker sets spanning a given chromosome segment length. Typing individuals for certain markers results in genotypes. For the estimation of the epistatic kinship within and between populations haplotypes are relevant. Haplotype reconstruction for individuals without known relationship is of limited efficiency. Therefore, it was assumed that genotyping was performed for FSP. Drawing FSP for the sample is possible without pedigree information for multiparous species like pigs before weaning.

For the proposed method the genotypes of each pair are compared and it is postulated that alleles which are common between full sibs potentially are ibd. In the comparison of genotypes three different cases can occur. In the first case there is no common allele found for at least one locus in the two genotypes of the pair. In this case inferring the haplotypes is not possible and the pair is not informative. The second case occurs when for the pair under consideration exactly one common haplotype is possible. In the third case different combinations of common haplotypes are possible, because of common alleles at least at one locus for equally heterozygous animals. If this is the case for m loci, 2^m different common haplotype combinations are possible. For the informative cases 2 and 3 the possible common haplotypes were derived. In case 3, the different possible common haplotype combinations were assigned with probability 2^{-m} respectively.

The statistical test conducted is based on the assumption that ibd haplotypes are more likely found within than between populations. Consider a situation where two samples of animals are taken. The null hypothesis is that the two samples originate from the same population, while the alternative hypothesis is that the two samples originate from different populations.

To verify this, a test statistic based on the accumulation of pairwise individual comparisons are suggested.

We compare two animals, I and J, at one chromosome segment, which, for simplicity of illustration, is assumed to be made up from two loci only. The observed genotypes are $G_i = \{1,2;1,2\}$ and $G_j = \{1,2;1,3\}$. Haplotype reconstruction results for both animals in $k = 2$ alternative haplotype combinations denoted as

$$G_i = \begin{Bmatrix} H_{ik1} \\ H_{ik2} \end{Bmatrix} \quad \text{and} \quad G_j = \begin{Bmatrix} H_{jk1} \\ H_{jk2} \end{Bmatrix}.$$

The possible haplotype combinations and their corresponding probabilities are:

$$G_i = \begin{Bmatrix} H_{i11} \\ H_{i12} \end{Bmatrix} = \begin{Bmatrix} 1-1 \\ 2-2 \end{Bmatrix}, \quad p_{i1} = 0.5$$

$$G_j = \begin{Bmatrix} H_{j11} \\ H_{j12} \end{Bmatrix} = \begin{Bmatrix} 1-1 \\ 2-3 \end{Bmatrix}, \quad p_{j1} = 0.5$$

$$G_i = \begin{Bmatrix} H_{i21} \\ H_{i22} \end{Bmatrix} = \begin{Bmatrix} 2-1 \\ 1-2 \end{Bmatrix}, \quad p_{i2} = 0.5$$

$$G_j = \begin{Bmatrix} H_{j21} \\ H_{j22} \end{Bmatrix} = \begin{Bmatrix} 2-1 \\ 1-3 \end{Bmatrix}, \quad p_{j2} = 0.5.$$

Next, each of the four possible haplotypes of animal I is compared with each of the four possible haplotypes of animal J. At this stage it is not relevant, whether the two individuals are from the same or from different samples. If two haplotypes are identical, the product of the corresponding haplotype probabilities is accumulated in the variable S_{ij}. In the present example, $H_{i11} = H_{j11}$ and $H_{i21} = H_{j21}$, so that

$$S_{ij} = p_{i1}p_{j1} + p_{i2}p_{j2} = 0.25 + 0.25 = 0.5.$$

For all within population comparisons, the average value of this variable is denoted as $\bar{S}_w$,while for all between population comparisons, the average value is denoted as $\bar{S}_b$.As under the alternative hypothesis we assume that common haplotypes are more likely within than between populations,

$$S = \bar{S}_w - \bar{S}_b \qquad (6)$$

is a suitable test statistic.

To verify the loss of information due to haplotype reconstruction, this test was applied in two forms:

(i) It was assumed that the true haplotypes were observed, i.e. not only the genotypes, but also the specific haplotype combination of an animal was observable. In this case, only one of the possible haplotype combinations received the probability 1 and all other possible haplotype combinations have the probability 0. Based on these probabilities, the test statistic S was calculated and is henceforth indicated as S_t (t standing for 'true').

(ii) To account for the uncertainty of haplotype reconstruction, the haplotype probabilities derived from full-sib genotypings as indicated above were used, the resulting test statistic is indicated as S_r, where r represents 'reconstructed'.

In both cases, the expected value under the null hypothesis (the two samples originate from the same population) is $E(S_t) = E(S_r) = 0$, while under the alternative hypothesis, we would expect that S_t and S_r take positive values. The distributions of the test statistics under the null hypothesis need to be determined empirically, either through simulation or through a permutation test approach (Doerge & Churchill 1996).

Simulation

An existing FORTRAN-Code was extended for the simulations in this study. A base population of 50 males and 50 females was generated. All animals were assumed to be unrelated and genotypes at the required number of loci were assigned at random, assuming the base population to be in Hardy–Weinberg and linkage equilibrium.

Under the null hypothesis, 15 generations of random mating and constant population size were simulated. For testing purposes the number of offsprings was doubled for the creation of the last generation.

Under the alternative hypothesis the population was randomly split after seven populations of random mating in two subpopulations of 50 males and 50 females each. For this purpose, the number of offspring was temporarily doubled in generation 7. From generation 8–16, random mating was conducted within these two subpopulations.

In the considered chromosome segments, crossingover events were assumed to follow a Poisson distribution without genetic interference, thus Haldane's mapping function (Haldane 1919) was applied. For the distribution of the family sizes Poisson distribution was assumed. Under both hypotheses the offsprings of the last generation were simulated as FSP, this full-sib structure was used for the reconstruction of haplotypes.

Under the null (alternative) hypothesis, a total of 1700 (2500) individuals were generated in one replicate. For these animals, the full pedigree and the simulated genotypes were stored.

For each assumed scenario, 1000 replicates were generated and analysed. To compute the empirical threshold value, the five and one percentile of the test statistic was calculated from the results of the simulation under the null hypothesis. The empirical power then was estimated by determining the proportion of replicates exceeding these empirical thresholds under the alternative hypothesis.

Scenarios studied

For the expected test statistic, $E(X^2)$ was calculated using equation (5), based on the average epistatic kinship within and between subpopulations. As this quantity is totally independent of the genotypes, it is only necessary to assume a chromosome segment length x, for which the values $x = 0$, 0.05, 0.10, 0.15, 0.20 were considered. Note that the results for

$x = 0$ reflect the outcome using the classical single-locus kinship as introduced by Malécot (1948).

For the marker-based estimation of epistatic kinship with the test statistics S_t and S_r a fixed set of six equidistant markers per chromosome segment were used, where for simplicity all markers had the same number of alleles, and each allele had the same probability to be drawn in the formation of the base population.

The following quantities were varied:

(i) the number of alleles per marker was set to $N_a = 2$, 4 and 6, where $N_a = 2$ reflects the situation with single nucleotide polymorphisms (SNPs) and $N_a = 6$ is a model for microsatellites;

(ii) the length of a chromosome segment was set to $x = 0.01$, 0.05, 0.10, 0.15, 0.20;

(iii) the number of chromosome segments was set to $N_{seg} = 1$, 3 and 6;

(iv) the number of FSP per sample was set to $N_{fsp} = 10$, 30 and 50.

Results and discussion

Epistatic kinship based on pedigree information

In Figure 1 the behaviour of the average epistatic kinship is depicted for all generations for the chromosome segment sizes $x = 0$ and $x = 0.2$ respectively. From generations 1 to 7 the epistatic kinship within the common base population is illustrated. After fission the epistatic kinship between the two subdivided populations is compared with the average epistatic kinship within populations 1 and 2. Figures 1a,b show that the empirical results from the simulation (dots) coincided perfectly with the theoretical expectations (lines) from equations (2–4).

With $x = 0$ (Figure 1a) only one locus is considered and the graph shows the average kinship within and between populations with common origin. The within population average kinship increases linearly with a rate of approximately $1/2N = 1/200 = 0.005$ per generation, leading to an average kinship of 0.073 in generation 16. The average kinship between the two subpopulations is fixed to the level achieved at the point of fission, i.e. 0.035 in generation 8, and remains constant henceforth.

With $x = 0.2$ (Figure 1b) the epistatic kinship within population loses the linear behaviour over generations. After generation 9 the increase of kinship within population resulting from co-ancestry is almost balanced by the loss of ibd status because of crossing over. In generation 16, the expected asymptotic value obtained from equation (2) was achieved to 99.6%.

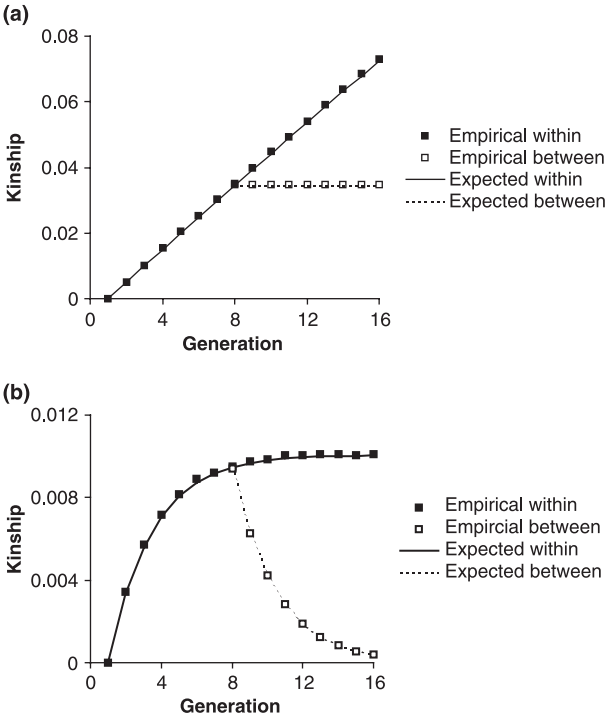

Figure 1 (a) Empirical and expected average epistatic kinship within and between population, $x = 0.00$ M. (b) Empirical and expected average epistatic kinship within and between population, $x = 0.20$ M.

$$\bar{K}^{0.2}_{max} = \frac{e^{-0.4}}{e^{-0.4} + 200(1 - e^{-0.4})} = 0.010064.$$

While after fission in generation 7 the degree of homozygosity between populations remains constant for $x = 0$, it quickly erodes with $x = 0.2$ with the rate $e^{-0.4} = 0.6703$ per generation, so that more than 97% of the expected epistatic kinship present at the time of fission are lost nine generations later.

In the first generations after fission, the difference between expected epistatic kinship within and between populations diverges faster for large chromosome segments compared to short chromosome segments (with the single locus case $x = 0$ as the extreme). However, the suggested test statistic is based on the comparison of expected numbers of ibd segments within and between populations. Here, not the ratio, but the absolute difference of observed ibd cases is relevant, hence it becomes essential that the absolute level of ibd probabilities is much higher for the single locus case (0.037 at generation 7) compared with the 20 cM case (0.009 at generation 7).

This difference in the level of the number of cases is reflected in the parameter $E(X^2)$ whose characteris-

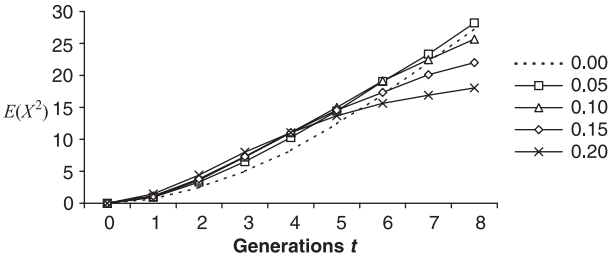

Figure 2 $E(x^2)$ for $x = 0.00$, 0.05, 0.10, 0.15 and 0.20 M for one to eight generations after fission.

Table 1 $E(X^2)$ for $x = 0.00$, 0.05, 0.10, 0.15 and 0.20 M for one to eight generations t after fission

	Generation t after fission							
x(M)	1	2	3	4	5	6	7	8
0.00	0.68	2.46	5.00	8.27	12.46	16.98	22.05	27.33
0.05	0.93	3.31	6.50	10.28	14.42	19.06	**23.31**	**28.22**
0.10	1.06	3.63	7.23	11.03	**15.00**	**19.13**	22.45	25.67
0.15	1.13	3.82	7.31	11.03	14.56	17.35	20.12	22.00
0.20	**1.47**	**4.42**	**8.00**	**11.05**	13.69	15.62	16.91	18.06

tics are depicted in Figure 2 for $M = 10$ individuals and $L = 5$ chromosome segments for the five different values for x. The curve for $x = 0$, i.e. considering one locus only, results in each generation with a lower $E(X^2)$ than the curve for some $x > 0$. Further, it can be seen that we have different most informative segment lengths for different generations since fission. This is also shown in Table 1, where the values of $E(X^2)$ are given for the chosen chromosome segment lengths. For each generation after fission, the highest value is printed in boldface. It is obvious, that in the first generations, the highest value is obtained for larger chromosome segments. With the number of generations increasing, the most informative chromosome segment length decreases. It can be concluded as a general rule that the closer two populations are expected to be (in terms of generations since fission), the longer the segment length should be chosen. For a large number of generations since divergence, very short segments or, in the extreme, single locus ibd status appears to be optimal.

Epistatic kinship based on marker information

The frequencies of the three cases 1, 2 and 3 for the haplotype reconstruction method are depicted in Figure 3 for $N_a = 2$, 4 and 6 alleles per locus and the segment length x from 0.01 to 0.20 M. Case 1

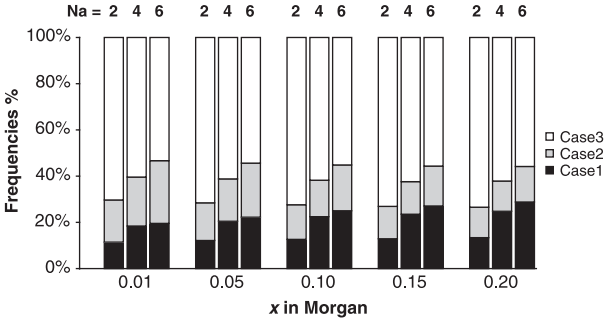

Figure 3 Frequencies for the three cases for $N_a = 2$, 4, 6 and segment length in Morgan.

describes the pairs without a common allele in the genotype of at least one locus, thus the cases where inferring the haplotypes is not possible and the genotyping information cannot be used. Case 2 describes the pairs where exactly one common haplotype is possible and case 3 where two or more common haplotypes are possible. The frequency of case 1 is increasing with increasing segment length and to that effect the frequency of case 2 is decreasing. The sum of cases 2 and 3 reflects the frequency of informative comparisons and it is decreasing from 80.5% (for $x = 0.01$) to 71.2% (for $x = 0.20$) with increasing segment length.

Case 1 is expected to have a high impact on the efficiency of the haplotype reconstruction method. Again the influence of the segment length becomes obvious. Due to higher probability of recombination events the number of non-informative FSP increases with increasing segment length. Further, non-informative comparisons increases with increasing number of alleles per locus. For a segment of 0.20 M and $N_a = 6$ the frequency of case 1 is almost 29%.

An overview of the power calculations for all different combinations of x, N_{seg}, N_a, N_{fsp} for true haplotypes are given in Table 2. One hundred and eighteen of 135 different combinations simulated result in a power >90% (shaded fields in Table 2). This underlines the high potential of the marker-based epistatic kinship for short-term phylogenetic studies.

Table 3 reports the results for the epistatic kinship based on reconstructed haplotpyes. For reconstructed haplotypes 75 of the 135 different combinations simulated yield in a power >90% (shaded fields in Table 3). The loss in power between the epistatic kinship with true haplotypes and reconstructed haplotypes is high (up to 57%) for the scenario where only 10 FSP are genotyped for one segment. Here the power based on reconstructed haplotypes is

Table 2 Power for the different scenarios based on true haplotypes

		$N_{fsp} = 10$			$N_{fsp} = 30$			$N_{fsp} = 50$		
N_{seg}	x	$N_a = 2$	$N_a = 4$	$N_a = 6$	$N_a = 2$	$N_a = 4$	$N_a = 6$	$N_a = 2$	$N_a = 4$	$N_a = 6$
1	0.01	0.657	0.735	0.696	0.990	0.998	0.999	0.999	0.999	1.000
	0.05	0.605	0.742	0.751	0.987	1.000	1.000	0.999	1.000	1.000
	0.10	0.603	0.732	0.680	0.989	1.000	1.000	1.000	1.000	1.000
	0.15	0.550	0.669	0.690	0.985	1.000	1.000	1.000	1.000	1.000
	0.20	0.512	0.702	0.655	0.990	1.000	0.999	0.994	1.000	1.000
3	0.01	0.916	0.994	0.996	1.000	1.000	1.000	1.000	1.000	1.000
	0.05	0.939	0.991	0.997	1.000	1.000	1.000	1.000	1.000	1.000
	0.10	0.935	0.984	0.985	1.000	1.000	1.000	1.000	1.000	1.000
	0.15	0.883	0.985	0.981	1.000	1.000	1.000	1.000	1.000	1.000
	0.20	0.835	0.966	0.958	1.000	1.000	1.000	1.000	1.000	1.000
6	0.01	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.05	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.10	0.996	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
	0.15	0.987	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000
	0.20	0.975	0.995	0.992	1.000	1.000	1.000	1.000	1.000	1.000

Shaded region represents power >90%.

<35% for all segment lengths and for all N_a, thus for this sample size the suggested method has its limitations.

The method for haplotype reconstruction used in this study does not account for LD in the populations. This leads to a certain loss of information by the estimation of the haplotype frequencies. Excoffier & Slatkin (1995) suggested an expected maximization (EM) algorithm which performed well in the presence of LD. A study comparing the efficiency of the epistatic kinship applying the haplotype reconstruction based on the EM algorithm is in preparation.

The haplotype reconstruction based on full-sib information lacks some generality. For multiparous species such as pig (as this method will be applied in a pig diversity study) it is possible to draw FSP without pedigree information. For other species (e.g. cattle) this might become a problem. For randomly sampled animals or other simple pedigree structures such as parent–offspring pairs, the planned implementation of the EM-algorithm is supposed to lead to a general solution.

Other than in using marker-based estimated kinships (Lynch 1988) we do not correct for the probability that an identical haplotype may be only identical by state, but not ibd. This possibility is neglected, because the probability of such a case is minor. With equal allele frequencies in the base population, the probability that two haplotypes made from N_{loc} loci with N_a alleles each is identical in the founder population is $N_a^{-N_{loc}}$. As in our study, the number of loci per haplotype was fixed to $N_{loc} = 6$, this probability varies between 1.56×10^{-2} for $N_a = 2$ and 2.14×10^{-5} for $N_a = 6$. Therefore, identity of haplotypes is expected to be almost exclusively due to ibd and correction is unnecessary.

Tables 2 and 3 highlight that the power of the marker-based epistatic kinship depends on the segment length x in Morgan. The power is decreasing with increasing x. The decrease in power with increasing x is smaller than expected, although Table 2 shows that a power >65% is feasible for a single segment of 0.20 M when genotyping highly polymorphic markers. Whereas for true haplotypes, the power reduction is mainly due to a reduced rate of identity by descent due to recombination in the generations between fission and the final generation, the loss of power between true and reconstructed haplotypes is because of failure or disturbance of haplotype reconstruction through crossingover events in the formation of the FSP.

The lower power for $N_a = 2$ with true haplotypes (Figure 2a) underlines the information loss with SNPs because of their biallelic nature (Vignal *et al.* 2002). The loss of power in this case is caused by the high proportion of ambiguous haplotypes. This becomes evident by the fact that at each locus 50% of the animals are expected to be homozygous for a biallelic SNP, while this rate is only 16.7% with a microsatellite with six loci. As homozygous loci add no information in discriminating between haplotypes, the informativeness of reconstructed haplotypes is minor for biallelic markers due to the low

Table 3 Power for the different scenarios based on reconstructed haplotypes

		$N_{fsp} = 10$			$N_{fsp} = 30$			$N_{fsp} = 50$		
N_{seg}	x	$N_a = 2$	$N_a = 4$	$N_a = 6$	$N_a = 2$	$N_a = 4$	$N_a = 6$	$N_a = 2$	$N_a = 4$	$N_a = 6$
1	0.01	0.264	0.228	0.240	0.779	0.826	0.831	0.943	0.957	0.951
	0.05	0.340	0.220	0.243	0.781	0.779	0.812	0.920	0.976	0.981
	0.10	0.313	0.165	0.212	0.808	0.795	0.750	0.933	0.980	0.931
	0.15	0.293	0.170	0.132	0.790	0.794	0.749	0.933	0.976	0.951
	0.20	0.288	0.166	0.138	0.785	0.782	0.667	0.923	0.975	0.955
3	0.01	0.550	0.550	0.465	0.988	0.991	0.997	1.000	1.000	1.000
	0.05	0.582	0.447	0.481	0.992	1.000	0.996	1.000	1.000	1.000
	0.10	0.574	0.451	0.391	0.995	0.999	0.994	1.000	1.000	1.000
	0.15	0.599	0.358	0.322	0.985	0.998	0.999	1.000	1.000	1.000
	0.20	0.534	0.433	0.282	0.990	0.993	0.994	0.998	1.000	1.000
6	0.01	0.837	0.731	0.752	1.000	1.000	1.000	1.000	1.000	1.000
	0.05	0.844	0.702	0.661	1.000	1.000	1.000	1.000	1.000	1.000
	0.10	0.811	0.713	0.607	1.000	1.000	1.000	1.000	1.000	1.000
	0.15	0.784	0.710	0.571	1.000	1.000	1.000	1.000	1.000	1.000
	0.20	0.728	0.707	0.580	1.000	1.000	1.000	1.000	1.000	1.000

Shaded region represents power >90%.

heterozygosity. This confirms the suggested analogy of one microsatellite being equivalent to two to three SNPs in linkage studies suggested by Evans & Cardon (2004).

For reconstructed haplotypes the increase from two to four alleles leads to a loss in power. This loss can be explained with the increase in non-informative comparisons between FSP when increasing the number of alleles per locus (Figure 3). Again, the need of a more powerful method for haplotype reconstruction is highlighted.

Classical distance measures reflect differences between populations, which are mainly due to genetic drift and mutation (Oldenbroek 1999). In our approach, mutation is totally disregarded. Yue *et al.* (2002) estimated the mutation rate of microsatellites in swine to be 7.5×10^{-5} per generation. Using this rate, the probability that a mutation occurs in a haplotype of six microsatellite loci over 10 generations is <0.5%.

Mutations may occur though, in the chromosome segments considered. A segment of 0.2 M contains an average of 2×10^7 bp. Nachmann & Crowell (2000) estimated the human mutation rate to be 2.5×10^{-8} per nucleotide and generation. Assuming this value to be valid for mammals in general, the probability is 8% that such a mutation occurs in a 20 cM interval in one generation, and the probability that at least one base change because of a mutation appears in 10 generations is 56.6% and thus non-negligible. However, this mutation will never be detected unless it affects a marker site, which was shown to be highly unlikely above or if it causes a major reorganization of the chromosome, e.g. through a translocation, deletion or inversion of a major chromosome segment, which is equally unlikely to appear *de novo* in viable offspring.

The second 'classical' driving force of population divergence is genetic drift which of course also operates on chromosome segments. However, in the assumed scenario of a limited number of generations since fission, drift is a much weaker process than crossing over, especially when longer chromosome segments are considered. As shown in equation (4), crossing over reduces the rate of epistatic kinship between populations in every generation with the rate e^{-2x}, independent of the population size. Disregarding crossing over, the drift variance of chromosome segment frequency is $\mathrm{var}(p_1) = (p_0(1 - p_0))/2N_e$ (Falconer & Mackay 1996), where p_0 is the initial frequency of the chromosome segment and p_1 is the frequency in the subsequent generation. To give an example: with $p_0 = 0.2$ and $N_e = 100$, the frequency of the chromosome segment in the next generation will lie with a 95% probability between $p_1 = 0.1446$ and $p_1 = 0.25540$ respectively. Drift is an undirected mechanism, which may both increase and decrease the chromosome segment frequency in a population. For a comparison of chromosome segment frequencies between lines, the probability that both frequencies change through drift by, say, more than 10% in the same direction (from $p_0 = 0.2$ to $p_1 = 0.22$ or $p_1 = 0.18$) is only 5.8%. Crossing over strictly reduces the probability of chromosome

homozygosity. In the example discussed, we expect a change of epistatic kinship between lines from $p_0 = 0.2$ to $p_1 = 0.18$ already with a chromosome segment length of $x = 0.053$. As this process, other than drift, is independent of effective population size, we expect the epistatic kinship based approaches to have higher sensitivity in cases where the effective size of the populations to compare is high.

As was argued in reference to Table 1, the suggested method even allows to 'adapt' the sensitivity of the method by choosing the optimal chromosome segment length depending on the (expected) number of generations since divergence, with long (20 cM and more) segments for less than four generations and short (5 cM and less) segments for more than seven generations.

The suggested approach is primarily targeted to the analysis of short-term phylogenies through subdivision of populations. Although this does not necessarily imply that the populations included are small, this will often be the case, leading to a relative small degree of polymorphism due to drift and eventually selection. Based on the results in Tables 2 and 3 we suggest to overcome this information loss nature by genotyping multiple segments. The number of segment genotyped N_{seg} has an immediate impact on the efficiency of the approach. Especially genotyping three segments instead of a single segment raises the power distinctively.

Another important factor is the sample size N_{fsp}, i.e. number of FSP drawn in each population. An increase in the tested animals from 10 to 30 FSP per population genotyped for one segment with microsatellites leads to doubled power for reconstructed haplotypes. Those findings with marker-based epistatic kinship support the linear influence of the number of segments typed and the squared influence of the sample size found when estimating the epistatic kinship with pedigree information.

At this point it is important to make some practical and economic considerations. Consider a case where two populations are compared based on $N_{seg} = 1$ segment of length $x = 0.05$ M with six microsatellite markers with $N_a = 6$ alleles based on $N_{fsp} = 10$ FSP. In this case, the power to statistically prove the difference between the two populations on the 5% error level is 0.751 based on true, but only 0.243 based on reconstructed haplotypes respectively (Tables 2 and 3). This result can either be improved by typing three instead of one segment or by considering 30 instead of 10 FSP. In both cases, the number of necessary genotypings is tripled. While in both cases the power based on true haplotypes increases to >0.99, the power based on reconstructed haplotypes is increased to only 0.481 with $N_{seg} = 3$ chromosome segments, while with $N_{fsp} = 30$ full sibs it is 0.812. Thus, the alternative to increase the number of FSP is much more efficient, which again reflects the quadratic effect of sample size.

However, increasing the sample size often has considerable extra cost, especially if samples have to be collected under field conditions. On the other hand, adding chromosome segments yields almost no extra cost, given the required markers are established in the laboratory (remember that the total number of genotypings is identical). The results in Table 2 show that with only 10 FSP per population and three to six chromosome segments carrying polymorphic markers, sufficient power to differentiate populations can be achieved. With further improvement of the haplotype reconstruction algorithm based on Excoffier & Slatkin's (1995) approach, it will be possible to get closer to this results when the analysis is based on reconstructed haplotypes. This demonstrates the potential of the suggested method to develop analytical tools of high sensitivity based on limited samples to be used in phylogenetic studies of domesticated, feral or wild populations.

References

Andersson L. (2001) Genetic dissection of phenotypic diversity in farm animals. *Nat. Genet.*, **2**, 130–138.

Brockmann G.A., Kratzsch J., Haley C.S., Renne U., Schwerin M., Karle S. (2000) Single QTL effects, epistasis, and pleiotropy account for two-thirds of the phenotypic F2 variance of growth and obesity in DU6i x DBA/2 mice. *Genome Res.*, **10**, 1941–1957.

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

Cockerham C.C. (1967) Group inbreeding and coancestry. *Genetics*, **56**, 89–104.

Coppieters W., Blott S., Farnir F., Grisart B., Riquet J., Georges M. (1999) From Phenotype to Genotype: Towards Positional Cloning of Quantitative Trait Loci in Livestock? In: J.C.M. Dekkers, S.J. Lament, M.F. Rothschild (eds) *From Jay L. Lush to Genomics: Visions for Animal Breeding and Genetics*. Iowa, Iowa State University.

Doerge R.W., Churchill G.A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics*, **142**, 285–294.

Eding J.H. (2002) *Conservation of Genetic Resources. Assessing Genetic Variation Using Marker Estimated Kinships*. Wageningen Agricultural University, Wageningen.

Eding H., Meuwissen T.H.E. (2001) Marker based estimates of between and within population kinships for the conservation of genetic diversity. *J. Anim. Breed. Genet.*, **118**, 141–159.

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

Evans D.M., Cardon L. (2004) Guidelines for genotyping in genome wide linkage studies: single nucleotide polymorphism maps versus microsatellite maps. *Am. J. Hum. Genet.*, **75**, 687–692.

Excoffier L., Slatkin M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.

Falconer D.S., Mackay T.F.C. (1996) Introduction to Quantitative Genetics. Longman Group Ltd, Essex.

Farnir F., Coppieters W., Arranz J.-J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., Georges M. (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.*, **10**, 220–227.

Flury C., Täubert H., Simianer H. (2006) Extension of the concept of kinship, relationship and inbreeding to account for linked epistatic complexes. *Livest. Prod. Sci.* (in press).

Haldane J.B.S. (1919) The combination of linkage values and the combination of distance between the loci of linkage factors. *J. Genet.*, **8**, 299–309.

Hayes B.J., Visscher P.M., McPartlan H., Goddard M.E. (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.*, **13**, 635–643.

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

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

Meuwissen T.H.E., Goddard M.E. (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, **155**, 421–430.

Nachmann M.W., Crowell S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.

Nezer C., Collette C., Moreau L., Brouwers B., Kim J.J., Giuffra E., Buys N., Andersson L., Georges M. (2003) Haplotype sharing refines location of an imprinted quantitative trait locus with major effect on muscle mass to a 250-kb chromosome segment containing the porcine IGF2 gene. *Genetics*, **165**, 277–285.

Niu T. (2004) Algorithms for inferring haplotypes. *Genet. Epidemiol.*, **27**, 334–347.

Oldenbroek J.K. (1999) *Genebanks and the Conservation of Farm Animals Genetic Resources*. DLO Institute for Animal Science and Health, Lelystad.

Ruane J. (1999) A critical review of the value of genetic distance studies in conservation of animal genetic resources. *J. Anim. Breed. Genet.*, **116**, 317–323.

Sambraus H.H. (2001) *Farbatlas der Nutztierrassen*. Verlag Eugen Ulmer, Stuttgart.

Scherf B.D. (Ed), (2000) *World Watch List for Domestic Animal Diversity*. FAO, Rome.

Simianer H. (1994) Derivation of single locus relationship coefficients conditional on marker information. *Theor. Appl. Genet.*, **88**, 548–556.

Simianer H. (2002a) Vorstudie zum Projekt 'Molekulargenetische Differenzierung verschiedener Rotviehpopulationen'. Molekulargenetische Differenzierung verschiedener Rotviehpopulationen. E. u. L. Bundesministerium für Verbraucherschutz. Landwirtschaftsverlag GmbH Münster-Hiltrup, Münster, vol. 493, pp. 7–32.

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

Toro M., Caballero A. (2004) *Characterisation and Conservation of Genetic Diversity Between Breeds*, 55th EAAP Annual Meeting, Bled, Slovenia.

Vignal A., Milan D., SanCristobal M., Eggen A. (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.*, **34**, 275–305.

Visscher P.M. (2003) *Principles of QTL Mapping*, Manual, PhD thesis. Salzburg University, Edinburgh.

Windig J.J., Meuwissen T.H.E. (2004) Rapid haplotype reconstruction in pedigrees with dense marker maps. *J. Anim. Breed. Genet.*, **121**, 26–39.

Wright S. (1922) Coefficients of inbreeding and relationship. *Am. Nat.*, **56**, 330–339.

Yue G.H., Beeckmann P., Geldermann H. (2002) Mutation rate at swine microsatellite loci. *Genetica*, **114**, 113–119.

Appendix

The necessity to account for the number of informative comparisons is illustrated with the following example: Consider individuals A and B in population 1 and C and D in population 2. We find that for one chromosome segment A is ibd with both C and D. B is also found to be ibd with C, then B has to be ibd with D as well. In this case, only three of the four comparisons between populations are in fact informative.

For the number of informative comparisons for a given chromosome segment we derived the following approximations

$$N_w = 2\left[(M-1) + \left(\frac{M^2}{2} - \frac{3M}{2} + 1\right)(1 - p_w^2)^{(M-2)}\right]$$

$$N_b = M + M(M-1)(1-p_b^3),$$

where N_w is the number of effective segment-specific pairwise comparisons within populations, N_b is the number of effective segment-specific pairwise comparisons between populations, p_w is the average ibd probability within populations, p_b is the average ibd probability between populations.

For p_w and p_b, the corresponding values calculated with recursion (2) and equation (4) can be used. Note that the proportion of informative segment-specific comparisons within and between populations is inversely proportional to the ibd probabilities p_w and p_b respectively. Note further that for $p_w = p_b = 0$, $N_w = M(M-1)$ and $N_b = M^2$, i.e. the effective number equals the true number of comparisons.