Genet. Sel. Evol. 39 (2007) 159–179

© INRA, EDP Sciences, 2007
DOI: 10.1051/gse:2006040

Original article

Haplotype kinship for three populations of the Goettingen minipig

Christine FLURY[a*], Steffen WEIGEND[b], Xiangdong DING[a], Helge TÄUBERT[a], Henner SIMIANER[a]

[a] Institute of Animal Breeding and Genetics, Georg-August-University of Göttingen, Albrecht-Thaer-Weg 3, 37075 Göttingen, Germany
[b] Institute for Animal Breeding, Federal Agricultural Research Centre (FAL), Hoeltystrasse 10, 31535 Neustadt-Mariensee, Germany

(Received 17 May 2006; accepted 9 November 2006)

Abstract – To overcome limitations of diversity measures applied to livestock breeds marker based estimations of kinship within and between populations were proposed. This concept was extended from the single locus consideration to chromosomal segments of a given length in Morgan. Algorithms for the derivation of haplotype kinship were suggested and the behaviour of marker based haplotype kinship was investigated theoretically. In the present study the results of the first practical application of this concept are presented. Full sib pairs of three subpopulations of the Goettingen minipig were genotyped for six chromosome segments. After haplotype reconstruction the haplotypes were compared and mean haplotype kinships were estimated within and between populations. Based on haplotype kinships a distance measure is proposed which is approximatively linear with the number of generations since fission. The haplotype kinship distances, the respective standard errors and the pedigree-based expected values are presented and are shown to reflect the true population history better than distances based on single-locus kinships. However the marker estimated haplotype kinship reveals variable among segments. This leads to high standard errors of the respective distances. Possible reasons for this phenomenon are discussed and a pedigree-based approach to correct for identical haplotypes which are not identical by descent is proposed.

genetic diversity / short term phylogeny / kinship / haplotype kinship / identity by descent

1. INTRODUCTION

Genetic diversity is required for populations to cope with environmental change and therefore the maintenance of genetic diversity is a primary objective in the management of threatened populations [15]. Numerous projects have been conducted in different livestock breeds with the goal to help decision

* Corresponding author: cflury@gwdg.de

Article published by EDP Sciences and available at http://www.edpsciences.org/gse or http://dx.doi.org/10.1051/gse:2006040

makers to identify genetically unique breeds to be included in conservation activities [28]. In subdivided populations like livestock species total genetic diversity consists of within and between subpopulation diversity. Within population diversity can be described with observed and expected heterozygosities, allelic diversity (*i.e.* the average number of alleles per locus) and the percentage of polymorphic loci [18, 32]. Between breed diversity is mostly assessed on the basis of genetic distances, for which allele frequencies are used as basic information.

In the last years genetic distances estimated from polymorphic microsatellite markers have been the most popular method for the assessment of the phylogenetic structure in animal genetic resources [1, 32]. However, genetic distances have statistical and biological properties which are often based on assumptions which do not hold for livestock populations. Without the consideration of those limitations, genetic distance values might become misleading and lose the explanatory power for genetic diversity in livestock breeds. The properties and limitations related to the subject of the study are presented in the next section, for more detailed discussion a reference to the literature is made [7, 19, 21].

Genetic distances have a base in population genetics, initially they have been developed with species in mind, thus for an evolutionary time span. For the creation of livestock breeds this assumption does not hold, as they have been domesticated and improved by man [32]. Most of today's breeds go back to the 19th or the beginning of the 20th century and crossbreeding was commonly practised 50 to 100 generations [29, 33] ago. Therefore the role of mutation in creating differences is assumed to be small and the often made assumption of no or negligible migration between populations is not applicable.

After the assessment of the uniqueness of different breeds with genetic distances a decision is required. Under limited financial resources for conservation activities the question is which breeds lead to the highest future genetic diversity. Weitzman [36] suggested a method that uses genetic and non genetic information to calculate the current diversity and the expected change in total diversity over a certain time horizon for a group of species [24]. The properties of this approach have been evaluated in detail [24,31]. The Weitzman approach was criticised by several authors [3, 6, 19] since it does not consider within population variability. Ignoring within population diversity is not only a drawback of the Weitzman method but of all diversity studies relying on genetic distances only. When neglecting the within breed diversity, the increase of genetic distances with increasing levels of inbreeding of populations might lead to the conservation of highly inbred populations [8]. To overcome this problem

Eding and Meuwissen [8] and Caballero and Toro [3] proposed to evaluate genetic variability within and between populations based on the kinship coefficient. Eding [6] evaluated marker estimated kinships between and within populations and proposed a distance which is equivalent to the Nei minimum distance [22]. The driving force for the kinship as a measure of genetic diversity is solely random drift. Thus, the short term evolution of livestock breeds is accounted for to some extent. However, drift is inversely proportional to the effective population size [12] so that the diversification of large populations will be slower than that of small populations. For decision making, a core set method based on average kinship coefficients was proposed [2, 9].

In this study the single locus concept of kinship is extended to chromosomal segments of a given length in Morgan units. A similar idea was applied for the estimation of past effective population size by Hayes *et al.* [17]. For the proposed measure based on segments identical by descent (ibd) called haplotype kinship (epistatic kinship in previous publications [13, 14]) a force additional to random drift becomes crucial – recombination. Thus it goes one step further, regarding "short" developing time of small populations. Algorithms were derived for the calculation of the haplotype kinship based on pedigree information [13]. Since pedigree information is often missing for small endangered livestock populations [28] the haplotype kinship was estimated based on marker information. Those investigations showed that the haplotype kinship is always more informative than the single locus approach in short term phylogenies and that with decreasing numbers of generations since fission, increasing segment lengths are more informative. This allows a further refinement of the method for the case when some population history is available and underlines the promising potential of the concept for the differentiation of short term phylogenies [14].

The goal of the present study was the practical evaluation of the haplotype kinship based on data from an existing population. The new measure was applied in a diversity study for three populations of the Goettingen minipig. The estimates for marker based haplotype kinship within and between the three subpopulations were derived. The expected values for the respective segment lengths were calculated based on pedigree information. Further haplotype kinship distances and the corresponding standard errors are presented.

2. MATERIALS AND METHODS

The Goettingen minipig was established in 1960 at the University of Goettingen for laboratory use. The goal was the development of a small pig

Table I. Average kinship coefficients within and between populations and the corresponding standard errors for the animals genotyped from populations *GE*, *DK1* and *DK2*.

	GE	**DK1**	**DK2**
GE	0.172 ± 0.029	0.148 ± 0.005	0.148 ± 0.003
DK1		0.176 ± 0.031	0.159 ± 0.005
DK2			0.178 ± 0.026

as a human model [16]. The founder population (*GE*) was separated in 1992 and an additional population was built up in Denmark (*DK1*). In 1998 the Danish population was split, resulting in the third population *DK2*. Today the three populations *GE*, *DK1* and *DK2* are kept closed under specific pathogen free conditions and without any exchanges between the populations. From the actual stock of the three populations *GE*, *DK1* and *DK2* tissue samples of randomly chosen full sib pairs were taken. An insight in the actual relationships within and between the three populations for the pedigree of the sampled animals is provided in Table I. The diagonal reflects the kinship coefficient within population and the corresponding standard error and the off-diagonals reflect the between population kinship and the corresponding standard error.

From the two porcine genetic maps USDA_MARC_v1 and USDA_MARC_v2 six segments on five different chromosomes were chosen [25, 26]. The segments were defined based on five or six microsatellites. The first criterion for the choice of the markers was the segment length in Morgan. The additionally constant order of the markers on the two maps, the heterozygosity and the annealing temperature were considered.

The PCR products were obtained in a total volume of 9 μL using *Qiagen HotStarTaq Master Mix* Kit (Qiagen GmbH, Hilden, Germany). Each PCR tube contained 20 ng of genomic DNA, 0.3 μM of each primer, 3 mM tetramethylammoniumchloride, and 4 μL of master mix containing 1 $\times$ reaction buffer, 200 μM of each dNTP and 0.4 units Taq polymerase. The amplification protocol of the *Hot Start PCR* was the following: 15' 95 °C; [1' 94 °C; 1' Z °C; 1' 72 °C] $\times$ 35; 10' 72 °C; 4 °C. The annealing temperature Z varied from 55 ° – 63 °C. Amplified DNA fragments were visualised by 8% polyacrylamide gel electrophoresis using a LI-COR automated DNA analyser (LI-COR GmbH, Bad Homburg, Germany). The allele scoring between gels was standardised using internal DNA standard alleles. Standard alleles were calibrated in size using a commercially available external size ladder (MWG Biotech AG, Ebersberg, Germany). For comparability with other studies, a set of standard alleles is available.

The DNA content was not sufficient for some samples. Furthermore some markers did not amplify during PCR. Marker *SW775* was not polymorphic in the three populations. Therefore it was discarded from further analysis. Finally 334 genotypes (106 from *GE*, 108 from *DK1* and 120 from *DK2*) for six segments and 33 microsatellites were available for the statistical analysis. The six segments varied in their length between 0.050 and 0.081 Morgan (based on USDA_MARC _v2), which lead to an average segment length of 0.067 Morgan. The average number of alleles per segment ranged from 3.20 (segment 2) up to 5.20 (segment 6). The microsatellites defining the six segments are listed in Table II.

2.1. Haplotype determination

For the estimation of the marker based haplotype kinship the sequence of the markers of each locus is relevant. Therefore an efficient method for haplotype reconstruction is needed. Excoffier and Slatkin [10] used the Expectation Maximisation (EM) algorithm [4] for the derivation of haplotypes with several loci and several alleles per locus. The EM-algorithm uses information on linkage disequilibrium and pedigree information is not requested. To fully account for the available full sib information, an extended version of Excoffier and Slatkin EM-algorithm was developed [5]. The EM-algorithm may lead to biased haplotype frequencies if markers are not in Hardy-Weinberg equilibrium (HWE) [10,30]. Therefore the test for HWE implemented in ARLEQUIN (version 3.0 [11]) was conducted for each marker in the three populations. Finally, haplotype reconstruction was conducted for all 33 markers.

2.2. Haplotype kinship and single locus kinship

For the marker estimated haplotype kinship (*MEHK)* between and within populations y and z, the haplotypes of each full sib pair were compared with the haplotypes of all other full sib pairs. In the case of common haplotypes the product of the haplotype probabilities was summed up.

In a full sib pair i, we have $j = 2$ individuals with $k = 2$ gametes each in the chromosome segment considered. Suppose in the population there are $l = 1, ..., L$ different haplotypes for this segment. We denote the probability that gamete k of animal j in full sib group i is identical to haplotype l as P_{ijkl}. Note that $\sum_{l=1}^{L} P_{ijkl} = 1$. To compare full sib group i with full sib group i', we

Table II. Observed heterozygosity, expected heterozygosity and the *p*-value from HWE-test for the 33 microsatellites and the three populations.

		Population GE			Population DK1			Population DK2		
seg	**Marker**	obs. het	exp. het	*p*-value	obs. het	exp. het	*p*-value	obs. het	exp. het	*p*-value
1	*SW970*	0.71	0.69	0.6867	0.71	0.67	0.0153	0.78	0.71	0.0217
	SW216	0.65	0.62	0.5038	0.55	0.58	0.3311	0.58	0.55	0.1526
	SW780	0.69	0.65	0.3229	0.73	0.68	0.0752	0.78	0.71	0.0001
	SW962	0.68	0.65	0.8807	0.62	0.60	0.7081	0.59	0.59	0.3431
	S0082	0.68	0.66	0.5369	0.70	0.65	0.2169	0.61	0.61	0.1302
	SW157	0.69	0.64	0.0245	0.57	0.60	0.2363	0.57	0.60	0.0468
2	*SW1536*	0.62	0.64	0.6770	0.76	0.74	0.5643	0.73	0.71	0.0094
	SW210	0.44	0.42	0.6079	0.66	0.60	0.3373	0.48	0.53	0.1759
	SWR1113	0.04	0.05	1.0000	0.12	0.12	1.0000	0.02	0.02	1.0000
	SW288	0.59	0.56	0.7975	0.55	0.60	0.1115	0.59	0.49	0.1220
	SW69	0.22	0.22	0.6713	0.26	0.25	0.6886	0.16	0.15	1.0000
3	*SW328*	0.55	0.70	0.0205	0.43	0.69	0.0000	0.58	0.74	0.0003
	SWR2063	0.44	0.65	0.0004	0.31	0.62	0.0000	0.42	0.60	0.0000
	SWR925	0.42	0.52	0.0415	0.51	0.52	0.2343	0.63	0.63	0.3285
	SW63	0.74	0.73	0.1041	0.74	0.77	0.2263	0.76	0.74	0.4642
	SW342	0.62	0.66	0.0433	0.57	0.62	0.5303	0.63	0.62	0.6285
	SWR84	0.54	0.54	0.1771	0.54	0.62	0.0603	0.71	0.68	0.2237
4	*SW304*	0.58	0.52	0.2499	0.50	0.46	0.7136	0.63	0.62	0.7820
	SW732	0.38	0.32	0.0204	0.23	0.22	0.3503	0.18	0.17	0.5957
	SWR2152	0.59	0.61	0.1759	0.56	0.65	0.0000	0.62	0.57	0.0027
	SWR1210	0.46	0.49	0.3957	0.56	0.51	0.4663	0.53	0.45	0.1809
	SW1122	0.28	0.27	0.4578	0.32	0.31	0.6024	0.05	0.06	1.0000
	SW175	0.61	0.59	0.1241	0.69	0.65	0.0022	0.47	0.44	0.8195
5	*SW1823*	0.62	0.68	0.3043	0.81	0.77	0.9216	0.74	0.73	0.1490
	SW316	0.58	0.56	0.6253	0.61	0.58	0.7599	0.45	0.39	0.3394
	SW446	0.36	0.36	0.7591	0.50	0.53	0.4530	0.33	0.30	0.6469
	SWR987	0.53	0.50	0.2636	0.54	0.57	0.5197	0.52	0.53	0.4209
	SW122	0.58	0.51	0.4174	0.47	0.46	0.8333	0.48	0.50	0.0752
6	*SW139*	0.61	0.60	0.9008	0.56	0.64	0.4054	0.65	0.62	0.8457
	SWR978	0.31	0.28	0.2984	0.18	0.18	1.0000	0.23	0.24	1.0000
	SW1315	0.75	0.73	0.7559	0.67	0.68	0.2843	0.66	0.75	0.1537
	S0094	0.76	0.69	0.3516	0.64	0.69	0.0197	0.59	0.69	0.0020
	SW1066	0.64	0.63	0.0057	0.63	0.69	0.0000	0.64	0.67	0.0000

$p < 0.01$

sum up all products of haplotype probabilities, *i.e.*

$$S_{ii'} = \sum_{l=1}^{L} \sum_{j=1}^{2} \sum_{j'=1}^{2} \sum_{k=1}^{2} \sum_{k'=1}^{2} P_{ijkl} P_{i'j'k'l}. \quad (1)$$

This statistic can vary between 0 (if all haplotypes with a probability >0 differ between the two full sib groups) and 16 (if all four individuals are homozygous for the same haplotype).

The *MEHK* are derived for each of the six segments separately and summed up. Finally the sum is averaged over the number of segments.

Pedigree information for the genotyped animals was available back to 1975. This led to a total pedigree consisting of 2081 animals. With the algorithm proposed for the derivation of the haplotype kinship based on pedigree [13] the expected values for segment length $x = 0.01$ up to 0.15 Morgan were derived in 1 cM steps. For the pedigree estimated haplotype kinship the abbreviation *PEHK* is used. The average segment length for the six segments based on the 33 markers is $x = 0.0665$, thus the corresponding *PEHK* were also derived for this average.

Marker estimated kinship (*MEK*) was derived for all 33 microsatellites. In analogy to *PEHK* the pedigree estimated kinship (*PEK*) was also derived for the single locus case.

For a better understanding of the differences between the single locus approach, *i.e.* the kinship coefficient and the haplotype kinship, regressions of the *MEK* values and the *MEHK* values on the corresponding expected values were calculated. In analogy to Eding and Meuwissen's [8] average similarity indices, pairwise comparisons between the genotypes at the 33 marker loci of the 334 genotyped animals were conducted. No correction for alleles being identical by state but not identical by descent was implemented, since the fraction is assumed to be the same in all three populations. The similarity indices found for each pair were compared with the pedigree based expected kinship coefficients for the same individuals resulting in 55 611 pairwise comparisons. Secondly pairwise comparisons were conducted for all 334 animals and the six segments and again the expected haplotype kinships for $x = 0.0665$ Morgan, *i.e.* equal the average segment length was derived for the 55 611 pairs.

In both approaches, the baseline similarity *i.e.* the probability of identity by state without identity by descent can be estimated by the intercept of the linear regression. The intercept of the regression of the *MEHK* on the *PEHK* of each segment separately is therefore proposed as a correction factor for the probability of identical haplotypes which are not identical by descent. After subtraction of the intercept from each element of the *MEHK*-matrix for the segment under consideration, the resulting values are considered as corrected marker estimated haplotype kinship, indicated by *MEHK_corr*.

The same correction factor, *i.e.* the intercept of the regression from the single locus similarity on the pedigree estimated kinship was applied for the derivation of the corrected marker estimated kinship (*MEK_corr*) in analogy to Eding and Meuwissen [8]. These authors assume that the markers used for the estimation of the kinship are not linked. To overcome this problem with our data,

the *MEK_corr* values were derived drawing one marker at random for each of the six segments over 10 000 replicates.

2.3. Genetic distances

Eding and Meuwissen [8] suggested the following kinship distance D_{ij} between two populations i and j based on kinship coefficients

$$D_{ij} = f_{ii} + f_{jj} - 2f_{ij} \tag{2}$$

where: f_{ii} = the average kinship coefficient within population i;

f_{jj} = the average kinship coefficient within population j;

f_{ij} = the average kinship coefficient between population i and j.

The average kinship coefficient between the two populations stays constant after population fission, thus the distance between the two populations is determined by the increase of within population kinship.

In a previous study [14] it was shown, that this stability of between population kinship is not given when considering segments. Haplotype kinship between populations is decreasing after population fission due to recombination, while the within population haplotype kinship reaches an equilibrium value in which drift-generated new haplotype homozygosity is equal to recombination-based erosion of old haplotype homozygosity. Therefore we suggest a different distance metric, which will be shown to be approximately linear with the number of generations since fission under certain conditions.

Consider a population which at the time of fission has the average haplotype kinship K^x_o. This population is split in two subpopulations i and j with effective population size N_i and N_j, respectively. If we assume that fission has taken place in generation t, then the average haplotype kinship both within subpopulations, denoted as $K^x_{i(t)}$ and $K^x_{j(t)}$, and between subpopulations, denoted as $K^x_{ij(t)}$, are equal to K^x_o.

Flury *et al.* [14] have shown that, for generation $t + 1$ the expected average haplotype kinship in a closed population i can be calculated as

$$K^x_{i(t+1)} = \mathrm{e}^{-2x}\left[\frac{1}{2N_i} + \left(1 - \frac{1}{2N_i}\right)K^x_{i(t)}\right] \tag{3}$$

and the expected average haplotype kinship between populations i and j is

$$K^x_{ij(t+1)} = \mathrm{e}^{-2x}K^x_{ij(t)} \tag{4}$$

for generation T after fission. The expected haplotype kinship between breeds then is

$$K^x_{ij(T)} = \mathrm{e}^{-2xT} K^x_o. \tag{5}$$

A distance measure should be based on the relation of between and within breed haplotype kinship, which is the case for

$$d^x_{ij} = \frac{K^x_i K^x_j}{\left(K^x_{ij}\right)^2}.$$

As was also shown by Flury *et al.* [14] the haplotype diversity in a closed population for $t \to \infty$ asymptotically approaches an equilibrium value

$$K^x_{i(\infty)} = \frac{\mathrm{e}^{-2x}}{\mathrm{e}^{-2x} + 2N_i(1 - \mathrm{e}^{-2x})}. \tag{6}$$

At this stage "new" homozygosity is generated in the same rate as "old" diversity is destroyed through recombination. For small x the equation simplifies to the approximate expectation $1/(4Nc + 1)$ by Hayes *et al.* [17]. Expression (6) is seen as a refinement of the chromosome segment homozygosity [17] which does not tend to overestimate haplotype diversity for large segments, as postulated by Wang [35].

It can be shown that this equilibrium value is approached rapidly if the chromosome segment is not too small. Therefore, close to the equilibrium $C = K^x_i K^x_j$ will remain approximately constant over generations and the change of the diversity is only depending on the kinship between populations, and d^x_{ij} is

$$d^x_{ij} \approx \frac{C}{(K^x_{ij})^2}.$$

Making use of equation (5), the diversity in generation T after fission is

$$d^x_{ij(T)} \approx \frac{C}{(\mathrm{e}^{-2xT} K^x_o)^2} = \frac{C}{\mathrm{e}^{-4xT}(K^x_o)^2}.$$

Taking the natural logarithm of this diversity, we get

$$\ln(d^x_{ij(T)}) \approx \ln(C) - \ln(\mathrm{e}^{-4xT}) - \ln(K^x_o)^2 = \ln(C) - 2\ln(K^x_o) + 4x \times T.$$

This shows that the natural logarithm of d^x_{ij} is an approximately linear function of the number of generations since fission, with slope $4x$. Therefore, we suggest the use of the diversity

$$D^x_{ij} = 2\ln(d^x_{ij}) = \ln(K^x_i) + \ln(K^x_j) - 2\ln(K^x_{ij}). \tag{7}$$

This measure is zero at the time of fission and increases approximately linearly with the slope $4x$ per generation.

To assess the expected distance $E(D_{ij}^x)$, based on the pedigree information *PEHK* values were used in equation (7). For marker based distance estimates, $\hat{D}_{ij}^x$, *MEHK* values were put in equation (7).

The variance for the *MEHK* distances was estimated as

$$Var(\hat{D}_{ij}^x) = Var(\ln(\hat{K}_i^x)) + Var(\ln(\hat{K}_j^x)) + 4 \times Var(\ln(\hat{K}_{ij}^x)) + 2 \times Cov(\ln(\hat{K}_i^x), \ln(\hat{K}_j^x)) - 4 \times Cov(\ln(\hat{K}_i^x), \ln(\hat{K}_{ij}^x)) - 4 \times Cov(\ln(\hat{K}_j^x), \ln(\hat{K}_{ij}^x)).$$

The required variances and covariances were calculated based on the obtained haplotype kinships within and between populations. The square root of the variance was taken as the standard error of the *MEHK* distances. Again, the distances and the respective standard errors were calculated for *MEHK* and *MEHK_corr* separately.

3. RESULTS AND DISCUSSION

Table II reports the results from HWE-testing for the 33 genotyped markers and the three populations. Markers with significant derivation from HWE (p-values < 0.01) are marked grey. HWE departures in all of the three populations were found for the microsatellites *SWR2063* and *SW1066*. *SW328* and *SWR2152* show a significant excess of homozygotes in populations *DK1* and *DK2*. Additionally, *SW175* is not in HWE in population *DK1* and *SW780*, *SW1536* and *S0094* are not in HWE in population *DK2*.

Excoffier and Slatkin [10] mentioned that the use of markers which are not in HWE might lead to biased haplotype frequencies when applying the EM-algorithm. In contrast to this, Tenesa *et al.* [30] observed that departures from HWE do not lead to a notable degree of bias in the estimates of haplotype frequencies using the EM-algorithm. Neglecting the eight markers which are not in HWE (Tab. II), 24% of the initial available marker information would be lost. The decreasing number of markers defining the six segments and the decrease in the average number of alleles per locus force the occurrence of identical haplotypes, which leads to a lower resolution of the suggested method. Therefore the use of all 33 markers is advised.

The relation between the similarity indices (*MEK*) of the 55 611 pairwise comparisons between the 334 animals and the respective pairwise kinship

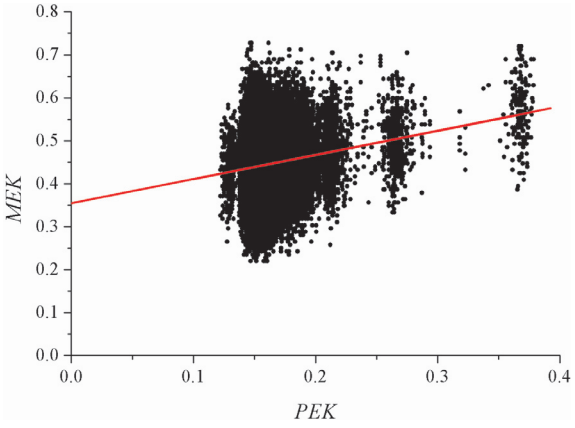

Figure 1. Relation between the average *MEK* for the 33 markers and the pedigree based kinship coefficient: 55 611 pairwise comparisons between the 334 individuals and the linear regression.

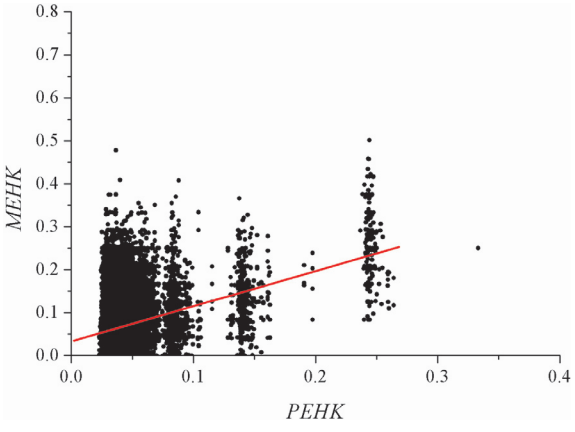

Figure 2. Relation between the *MEHK* for the 6 segments (*i.e.* 33 markers) and the *PEHK* for $x = 0.0665$ Morgan: 55 611 pairwise comparisons between the 334 individuals and the linear regression.

coefficients based on pedigree information are depicted in Figure 1. The estimated linear fit was

$$Y = 0.35461 + 0.56197X$$

with stability index $R^2 = 0.0291$.

Analogously, Figure 2 shows the relationship between the 55 611 pairwise comparisons of the *MEHK* and the *PEHK*. The estimated linear fit for this regression is $Y = 0.03319 + 0.81818X$ with stability index $R^2 = 0.0796$.

The stability index for the regression of *MEK* on pairwise kinship coefficients (R^2 = 0.0291) was much lower than the stability index for the pairwise comparisons of *MEHK* against *PEHK* (R^2 = 0.0796), although both regressions show a large amount of residual variation. The intercepts of the regressions can be used to correct for the probability of loci or haplotypes being identical by state but not identical by descent. This probability is much higher for single loci (0.355) compared to haplotypes (0.033). Together with the higher stability index, the correction applied to haplotype based kinship measures appears much more reliable than the correction for single locus based kinship measures.

The history of development of the three populations is rather short. After the bottleneck the extension of the population size was important since there was an increasing demand for miniature pigs on the market for laboratory animals. Under such circumstances the role of drift in creating differences is small and methods based on drift only – like marker estimated kinship – may fail to reflect the differences between populations properly.

In Figure 3 the average *MEHK* for the six different segments and their average (at segment length = 0.0665 Morgan) is presented. The *MEHK* within the three populations *GE*, *DK1* and *DK2* are depicted in Figure 3a) and the marker estimated kinship between the three populations in Figure 3b), respectively. The line reflects the averaged *PEHK*, *i.e.* the expected values based on pedigree information averaged over the three populations. Based on the close relatedness between the three populations, the expected values for the within and the between population *PEHK* were very similar, thus the averaged *PEHK* – value is given as a single curve in Figures 3a) and 3b), respectively.

The results for the *MEHK* are variable. With decreasing segment length the haplotype kinship is supposed to increase due to higher probability of identical haplotypes. This expectation is confirmed by the trend of increasing *MEHK* with decreasing segment length x in Figures 3a) and 3b). Despite this, an upward bias of the average marker based estimation in comparison with the pedigree based expectation was observed. The second and fourth segments heavily deviate from the expected values at the corresponding segment lengths (*i.e.* *PEHK* at 0.056 and 0.071) within and between populations, respectively.

The intercept of 0.03 in Figure 2 already suggests a certain overestimation applying marker based haplotype kinship due to segments being identical by state. For further quantification, regressions of the *MEHK* on the corresponding *PEHK* were derived for the six segments separately yielding length-specific correction factors for each segment. The corresponding intercepts, the slopes and the stability indices of the regressions and their average are given in

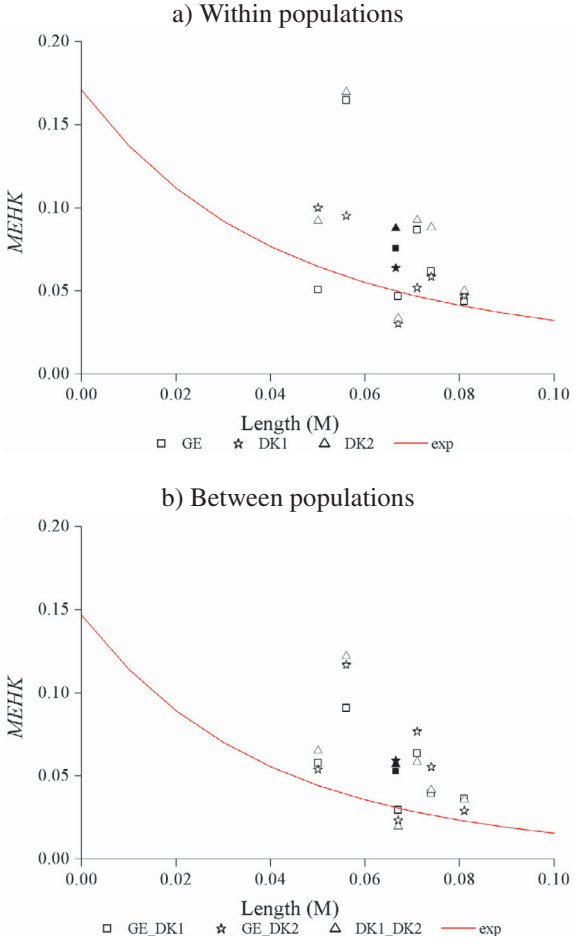

Figure 3. *MEHK* for the six segments and their average (■ ★ ▲) and the function of the expected values (——) within a) and between b) the three populations.

Table III together with the respective value for the regression from the average *MEK* on the *PEK*. Note that correction factors vary between 0.015 and 0.080 and the low stability indices for segments 2 and 4. The segment-specific correction factors were applied leading to corrected marker estimated haplotype kinship (*MEHK_corr)*.

The results of the *MEHK_corr* within and between populations for each segment and their average are depicted in Figures 4a) and b), respectively. The figures show that variability between segments is reduced without losing the expected trend of increasing *MEHK* with decreasing segment length. Also,

Table III. Intercept, slope and stability index for the linear fit of the regressions from *MEHK* on *PEHK* for the six segments separately, their average and the average of the regressions from *MEK* on *PEK* in the last line.

	Intercept	Slope	R^2
Segment 1	0.019	0.940	0.017
Segment 2	0.080	0.940	0.011
Segment 3	0.004	0.675	0.025
Segment 4	0.050	0.600	0.007
Segment 5	0.024	0.942	0.026
Segment 6	0.015	0.819	0.023
MEHK-all	0.033	0.818	0.080
MEK-all	0.355	0.562	0.029

Table IV. Haplotype kinship matrices based on pedigree information *(PEHK)* in a) and based on marker information for all 33 markers *(MEHK)* in b) with the corresponding standard errors and for the corrected (*MEHK_corr*) in c) respectively, for the average segment length $x = 0.0665$.

a) *PEHK*

	GE	DK1	DK2
GE	0.050	0.030	0.029
DK1		0.052	0.035
DK2			0.049

b) *MEHK*

	GE	DK1	DK2
GE	0.076 ± 0.019	0.053 ± 0.009	0.059 ± 0.014
DK1		0.064 ± 0.011	0.057 ± 0.015
DK2			0.088 ± 0.019

c) *MEHK_corr*

	GE	DK1	DK2
GE	0.044 ± 0.008	0.021 ± 0.004	0.027 ± 0.004
DK1		0.032 ± 0.011	0.025 ± 0.006
DK2			0.056 ± 0.009

the observed values are in much better agreement with the expected values, indicating that the suggested correction works well.

In Table IVa) the elements of the *PEHK*-matrix are listed for the average segment length at 0.0665 Morgan, where the diagonal reflects the within population kinship for each of the three populations and the off-diagonals the corresponding between population kinships. The respective elements of the uncorrected *MEHK*-matrix and their standard errors are given in Table IVb). Analogously, in Table IVc) the elements of the *MEHK_corr*-matrix and their standard errors are given. Comparing the standard errors of the elements of the uncorrected *MEHK* matrix in Table IVb) with the standard errors of the

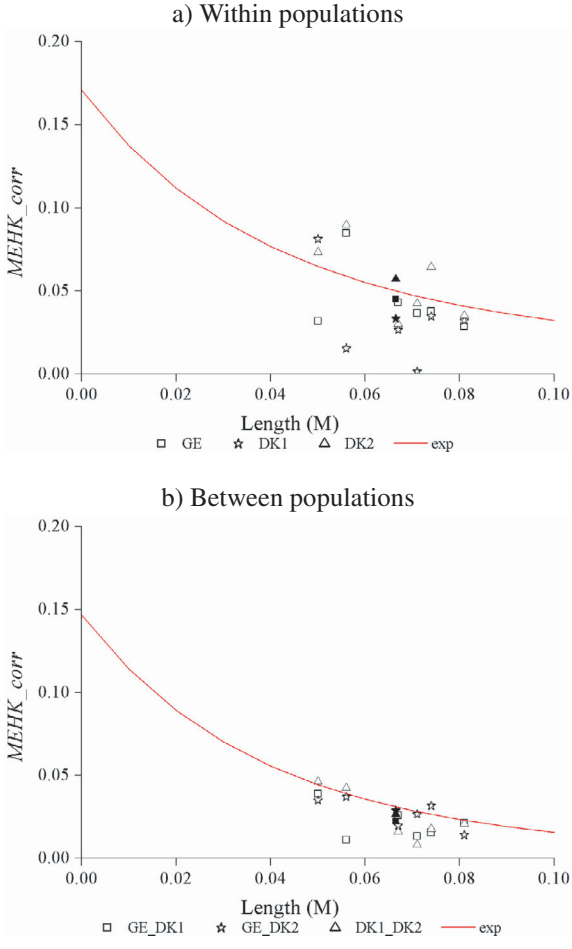

Figure 4. *MEHK_corr* for the six segments and their average (■★▲) and the function of the expected values (——) within a) and between b) the three populations.

MEHK_corr matrix in Table IVc) further indicates the higher accuracy of the corrected estimates.

The distance matrices corresponding to Table IV are given in Table V, again with the standard errors in Vb) and Vc), respectively.

To compare the efficiency of the single locus consideration [8] with the haplotype kinship, the corrected marker estimated kinship (using the intercept of the regression depicted in Figure 1 as correction factor) and the corresponding distances are given in Table VIa) and VIb).

Based on pedigree information, the two Danish populations *DK1* and *DK2* are less distinct than *DK1* with *GE,* and *DK2* with *GE,* respectively. The same

Table V. Distances for *PEHK* in a) *MEHK* and its standard errors in b) and the corrected *MEHK_corr* and its standard errors in c) for the average segment length $x = 0.0665$.

a) *PEHK*

	DK1	DK2
GE	0.997	1.051
DK1	0	0.717

b) *MEHK*

	DK1	DK2
GE	0.547 ± 0.201	0.640 ± 0.260
DK1	0	0.540 ± 0.278

c) *MEHK_corr*

	DK1	DK2
GE	1.157 ± 0.516	1.190 ± 0.231
DK1	0	1.029 ± 0.437

Table VI. The corrected marker estimated kinship matrix (*MEK_corr*) in a) and the corresponding distances derived according to the formula given by Eding and Meuwissen [8] in b) and the respective standard errors.

a) *MEK_corr*

	GE	DK1	DK2
GE	0.172 ± 0.085	0.132 ± 0.094	0.155 ± 0.096
DK1		0.141 ± 0.097	0.131 ± 0.103
DK2			0.184 ± 0.107

b) *MEK_corr - distances*

	GE	DK1	DK2
GE	0	0.050 ± 0.019	0.046 ± 0.022
DK1		0	0.063 ± 0.023
DK2			0

order was found for the *MEHK* and *MEHK_corr* distances, however these results are not confirmed considering *MEK* distances (Tab. VI). Especially, the two closest populations *DK1* and *DK2* were found to have the largest distance based on single locus kinship. This illustrates that the *MEK* fails to correctly reconstruct population divergence in short term phylogenies as in the example studied and underlines the promising potential of the haplotype kinship.

The overestimation of the marker based haplotype kinship (Tab. IVb)) leads to distances at a lower level (Tab. Vb)). Correction for identity by state without identity by descent removes this bias to a larger extent and leads to distance estimates with a slight upward bias compared to the pedigree based expectations, but well within the expected range (Tab. Vc)).

In theoretical investigations it was shown that the number of alleles per segment influences the power for the distinction between populations with the marker estimated kinship [14]. With decreasing number of alleles per locus the probability for identical haplotypes is increasing for the same average kinship between individuals. For the microsatellites defining the segments two and four, on average only 3.20 and 4.00 alleles were found in the three populations. Thus the low heterozygosity of the markers seems a possible explanation for

the high deviation from the pedigree based haplotype kinship and the *MEHK* within and between populations especially in the extreme case of segment 2. The adverse effect of the low heterozygosity is partly removed when the correction is applied, as demonstrated by the results obtained with the corrected *MEHK* in Figures 4a) and 4b).

Theoretical investigations yielded a high power for the distinction between populations with the *MEHK* under varying number of segments, number of full sib pairs genotyped and number of alleles per segment [14]. However, neutrality of the segments was assumed and therefore selection was not accounted for. In a QTL study of a Meishan × Goettingen minipig cross Wada *et al.* [34] found QTL for vertebra number and birth weight on chromosome 1, for teat number on chromosomes 1 and 7 and for backfat thickness on chromosome 7. For further investigation of the QTL on vertebra number F2 families of different Asian, European and miniature pig breeds were produced [20]. In this study the QTL on chromosome 1 was confirmed and an additional QTL for the same trait was found on chromosome 7 in six families but not in the Meishan × Goettingen minipig family. Rothschild and Plastow [27] reviewed the recent discoveries of gene mapping in commercial pigs and reported QTL for growth rate and immune response and the candidate gene of the ESR (Estrogen Receptor) on chromosome 1. The authors mention the associations between several traits and the pig major histocompatibility complex on chromosome 7.

Those findings suggest that the markers used for the definition of segments one and four (on chromosomes 1 and 7, respectively) might be influenced by selection. The main focus of selection in the three Goettingen minipig populations was set on decreasing body weight by keeping litter size at an acceptable level. The actual mean of piglets born alive is 5.68 ± 2.32 ($N = 140$) and 35.49 ± 9.05 ($N = 85$) for the 345- to 385-day weight in population *GE*. The deviations of number of piglets born alive and body weight in comparison with commercial pigs indicate the high selection pressure in the Goettingen minipig populations. This might also be an explanation for the fraction of markers deviating from HWE.

Therefore the knowledge of QTL and candidate genes should be taken into account in the choice of the segments, even though at the actual state of knowledge it might be a problem to define 6 segments with 5 to 6 microsatellites spanning a region of less than 0.10 Morgan which are selectively neutral. The aspect of selective neutrality for the choice of the segments is further ambivalent since selection can be an important force for the conservation of genomic regions, on which the haplotype kinship relies. The effect of selection on LD between linked loci was investigated by Nsengimana *et al.* [23] in five

populations of commercial pigs for regions of the two porcine chromosomes 4 and 7 where QTL affecting growth rate and fat deposition had been reported to be located. The effect of selection was not discarded by the authors, even though with a *p*-value of 0.06 no significance could be found.

4. CONCLUSIONS

The results of this study empirically confirm some of the theoretically derived properties of the suggested haplotype-based kinship and diversity measures [14]. It should be noted, though, that the reported results are merely an illustration of the methodology and, strictly speaking, can neither prove nor disprove the assumed properties. The study raises some aspects which need to be further studied and discussed.

(1) The hypothesis that the haplotype kinship is decreasing with increasing chromosome segment size is clearly confirmed (Figs 3a), 3b), 4a) and 4b)). If approximate information on the population history is available (*e.g.* number of generations since population fission), this allows the adaptation of the molecular tool, *i.e.* the length of chromosome segments genotyped, flexibly to the phylogenetic structure studied.

(2) The results also show that with chromosome segments the problem of being identical by state but not identical by descent is much less relevant compared to single locus approaches [8], but clearly is not negligible.

(3) The suggested correction based on the linear regression of the pedigree based haplotype kinship on the marker based haplotype kinship works well in the example studied, but depends on the availability of pedigree information. Correction factors which can be used in a situation where less information is available need to be developed.

(4) The kinship based distances, *i.e.* the consideration of single loci, fail to depict the known phylogenetic structure for the samples studied.

(5) The suggested diversity to our knowledge is the first such measure which was especially designed to study short term phylogenies, and which is not using genetic drift and mutation, but recombination as the major force creating population differences.

The suggested method will be especially useful, when SNP genotyping platforms will provide massive data on many chromosome segments spread across the entire genome. We expect that the method proposed here has a considerable potential to develop a better understanding of short-term phylogenetic structures in farm animal populations.

ACKNOWLEDGEMENTS

We gratefully acknowledge Prof. Gary Rohrer for the helpful comments on mapping positions and two anonymous reviewers for their constructive suggestions leading to an improved final version. The project was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

REFERENCES

[1] Baumung R., Simianer H., Hoffmann I., Genetic diversity studies in farm animals - a survey, J. Anim. Breed. Genet. 121 (2004) 361–373.

[2] Bennewitz J., Meuwissen T., A novel method for the estimation of the relative importance of breeds in order to conserve the total genetic variance, Genet. Sel. Evol. 37 (2005) 315–337.

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

[4] Dempster A.P., Laird N.M., Rubin D.B., Maximum likelihood from incomplete data via the EM-algorithm, J. Roy. Stat. Soc. 39 (1977) 1–38.

[5] Ding X., Zhang Q., Flury C., Simianer H., Haplotype reconstruction and estimation of haplotype frequencies from nuclear families with only one parent available, Hum. Hered. 62 (2006) 12–19.

[6] Eding J.H., Conservation of genetic resources. Assessing genetic variation using marker estimated kinships, Thesis Wageningen Agricultural University, Wageningen, 2002.

[7] Eding J.H., Laval G., Measuring the genetic uniqueness in livestock, in: Genebanks and the conservation of farm animals genetic resources, DLO Institute for Animal Science and Health, Lelystad, 1999, p. 33–58.

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

[9] Eding J.H., Meuwissen T., Linear methods to estimate kinships from genetic marker data for the construction of core sets in genetic conservation schemes, J. Anim. Breed. Genet. 120 (2003) 289–302.

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

[11] Excoffier L., Laval G., Schneider S., Arlequin ver. 3.0: An integrated software package for population genetics data analysis, Evolutionary Bioinformatics Online 1 (2005) 47–50.

[12] Falconer D.S., Mackay T.F.C., Introduction to quantitative genetics, 4 edn., Longman Group Ltd., Essex, 1996.

[13] Flury C., Taubert H., Simianer H., Extension of the concept of kinship, relationship, and inbreeding to account for linked epistatic complexes, Livest. Sci. 103 (2006) 131–140.

[14] Flury C., Tietze M., Simianer H., Epistatic kinship a new measure of genetic diversity for short-term phylogenetic structures – theoretical investigations, J. Anim. Breed. Genet. 123 (2006) 159–171.
[15] Frankham R., Ballou J., Briscoe D.A., Introduction to conservation genetics, University Press, Cambridge, 2002.
[16] Glodek P., Oldigs B., Das Göttinger Miniaturschwein Versuchstierkunde 7, Paul Parey, Berlin, 1981.
[17] Hayes B.J., Visscher P.M., McPartlan H., Goddard M.E., Novel multilocus measure of linkage disequilibrium to estimate past effective population size, Genome Res. 13 (2003) 635–643.
[18] Kantanen J., Olsaker I., Holm L.-E., Lien S., Vilkki J., Brusgaard K., Eythorsdottir E., Danell B., Adalsteinsson S., Genetic diversity and population structure of 20 north European cattle breeds, J. Hered. 6 (2000) 446–457.
[19] Laval G., SanCristobal M., Chevalet C., Measuring genetic distances between breeds: use of some distances in various short term evolution models, Genet. Sel. Evol. 34 (2002) 481–507.
[20] Mikawa S., Hayashi T., Nii M., Shimanuki S., Morozumi T., Awata T., Two quantitative trait loci on Sus scrofa chromosomes 1 and 7 affecting the number of vertebrae, J. Anim. Sci. 83 (2005) 2247–2254.
[21] Nagamine Y., Higuchi M., Genetic distance and classification of domestic animals using genetic markers, J. Anim. Breed. Genet. 118 (2001) 101–109.
[22] Nei M., Genetic distance between populations, Am. Nat. 106 (1972) 283–292.
[23] Nsengimana J., Baret P., Haley C.S., Visscher P.M., Linkage disequilibrium in the domesticated pig, Genetics 166 (2004) 1395–1404.
[24] Reist-Marti S.B., Simianer H., Gibson J., Hanotte O., Rege J.E.O., Weitzman's approach and conservation of breed diversity: an application to African cattle breeds, Conserv. Biol. 17 (2003) 1299–1311.
[25] Rohrer G.A., Alexander L.J., Keele J.W., Smith T.P., Beattie C.W., A microsatellite linkage map of the porcine genome, Genetics 136 (1994) 231–245.
[26] Rohrer G.A., Alexander L.J., Hu Z., Smith T.P., Keele J.W., Beattie C.W., A comprehensive map of the porcine genome, Genome Res. 6 (1996) 371–391.
[27] Rothschild M., Plastow G., Advances in pig genomics and industry applications, Ag. Biotech. Net. 1 (1999).
[28] Ruane J., A critical review of the value of genetic distance studies in conservation of animal genetic resources, J. Anim. Breed. Genet. 116 (1999) 317–323.
[29] Sambraus H.H., Farbatlas der Nutztierrassen, 6 edn., Verlag Eugen Ulmer, Stuttgart, 2001.
[30] Tenesa A., Knott S.A., Ward D., Smith D., Williams J.L., Visscher P.M., Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes, J. Anim. Sci. 81 (2003) 617–623.
[31] Thaon d'Arnoldi C., Foulley J.-L., Ollivier L., An overview of the Weitzman approach to diversity, Genet. Sel. Evol. 30 (1998) 146–161.
[32] Toro M., Caballero A., Characterisation and conservation of genetic diversity between breeds, in: Book of Abstracts of the 55th EAAP Annual Meeting, 5–9 September 2004, Vol. 10, Bled, Slovenia, p. 28.

[33] Visscher P.M., Principles of QTL mapping, manual PhD-course, held in Salzburg, Austria, Edinburgh, 2003.
[34] Wada Y., Akita T., Awata T., Furukawa T., Sugai N., Ishii K., Ito Y., Kobayashi E., Mikawa S., Yasue H., Inage Y., Kusumoto H., Matsumoto T., Miyake M., Murase A., Shimanuki S., Sugiyama T., Uchida Y., Yanai S., Quantitative trait loci (QTL) analysis in a Meishan × Goettingen cross population, Anim. Genet. 31 (2000) 376–384.
[35] Wang J., Estimation of effective population sizes from data on genetic markers, Philos. T. Roy. Soc. B. 360 (2005) 1395–1409.
[36] Weitzman M.L., On diversity, Q.J. Econ. 107 (1992) 363–405.

To access this journal online:
www.edpsciences.org