Pedigree analysis for conservation of genetic diversity and purging

R. WELLMANN[1]* AND I. PFEIFFER[2]

[1] *Department of Mathematics, University of Kassel, D-34109 Kassel, Germany*
[2] *UniKasselTransfer, University of Kassel, Gottschalkstrasse 22, D-34109 Kassel, Germany*

Summary

We present an approach to describe and evaluate changes in genetic diversity and to calculate bounds for improvement. This pedigree-based analysis was applied to the Kromfohrländer dog (FCI Gr9 Sec10). Pedigrees trace back to the foundation of the breed and were available for 5527 individuals. Based on this dataset the population structure and historical bottlenecks were studied. Distributions of allele frequencies were estimated by Monte Carlo simulation. To monitor changes in mating systems throughout the breeding history, the homozygosity of alleles was compared with their expectations in Hardy–Weinberg equilibrium. Different breeding lines were identified by hierarchical cluster analysis and were characterized by ancestor contributions. Our calculations showed that the founder event in 1945 was followed by two bottlenecks. One was caused by strong selection in a very small population, and the other was triggered by rigorous disease management. The necessary amount of purging that arised due to the bottlenecks was also discussed.

1. Introduction

Conservation of genetic diversity and purging of defect alleles are often primary goals in canine breed management as many of these breeds suffer from historical bottlenecks. The maintenance of genetic diversity enables future generations of breeders to improve traits, which must be neglected at the present time and reduces the necessary amount of purging.

The most widely accepted notion of genetic diversity is the gene diversity, defined by Nei (1973). We define the gene diversity GD(j) of an age cohort j as the probability that two alleles chosen at random from the age cohort are not identical by descent (IBD). This gene diversity corresponds to the gene diversity of Nei in a model where all alleles from the founder population are assumed to be different. Since this is not the case in real populations, other notions of genetic diversity are also considered. Because of overlapping generations, all our calculations are based not on generations, but on age cohorts. This approach is more sensitive to detection of effects of the historical changes in population management. The most effective ways to improve or to maintain genetic diversity is to set the

contributions of individuals to values that minimize the average coancestry of the progeny (Ballou & Lacy, 1995; Caballero & Toro, 2000) and/or to establish a sperm bank that keeps individuals with underrepresented genotypes available for stud as long as they are needed. Thus, even more important than the genetic diversity of the population is the potential genetic diversity of an age cohort. It is defined in section 2 as the maximum gene diversity that could have been achieved by optimal contributions of the breeding animals. This is not only most important in practice, but unlike other notions of diversity, it also has the desirable property of not being able to be increased by removing individuals from the population. In this paper, we study the historical development of genetic diversity in the Kromfohrländer breed since 1945 and the potential for improvement.

Less clear than the optimal choice of the contributions is the optimal choice of the mating system. When selection against recessive alleles is intended, knowledge of the allele frequencies would be of great help for deciding which mating system to follow, since the effectiveness of selection under a given mating system depends on the allele frequency. Most recessive alleles, first and foremost disease alleles, can be expected to descend from only one founder who carried

---

* Corresponding author. e-mail: wellmann@mathematik.uni-kassel.de

the allele only once. Such alleles are said to be rare within our paper. We estimate the distribution of the frequency of a rare neutral allele by Monte Carlo simulation. Neutrality can be assumed for example for alleles that cause heritable diseases that break out at an old age.

Various mating systems with directly opposed objectives have been proposed for the conservation of genetic diversity. The most common mating systems are line breeding and outbreeding. Line breeding means mating of related individuals, combined with intense selection for viability and fertility, but also the traits for which the common ancestors stood out from the crowd. Line breeding could cause temporarily a subdivision of the population. The available tools for analysing genetic diversity in subdivided populations were shown by Caballero & Toro (2002) and by Toro and Caballero (2005). The main advantage of line breeding is to expose recessive alleles to selection, as demonstrated by Robertson (1952). This comprises both desirable as well as undesirable alleles. It is well known that line breeding can remove undesirable recessive alleles with a large effect on viability within a few generations from the population if the inbreeding is moderate, so that line breeding is a kind of purging. Recently, various studies have dealt with different kinds of purging, see Leberg & Firmin (2008) and the references in it. The authors noted that responses vary much among replicates in laboratory investigations of purging. The outcome of purging is not predictable since the genetic basis of inbreeding depression is not known for most populations. Recessive deleterious alleles and overdominance could cause inbreeding depression. However, recessive alleles seem to be most important, as stated by Charlesworth & Charlesworth (1999).

Whereas purging by creating a bottleneck is clearly not advisable, purging via moderate inbreeding, combined with intense selection for viability and fertility, could be an appropriate breeding strategy, if the alleles under selection are recessive with low frequencies and care is taken to conserve the genetic diversity throughout the purging process. Subsequent crossing of lines should secure that the population reaches at least the viability and fertility from before the purging event, no matter whether purging was successful or not. This breeding strategy was applied successfully by Falconer (1971) and Eklund & Bradford (1976), who improved litter size in mice at the selection plateau. Kimura & Crow (1963) demonstrated that line breeding is more appropriate to conserve genetic diversity than outbreeding if the contributions of the individuals to the next generation are close to their optimal values. Caballero & Toro (2000) noted that this was repeatedly shown in the literature. It results from the fact that line breeding reduces the Mendelian drift. However, otherwise outbreeding is more appropriate to conserve genetic diversity.

Outbreeding occurs when mated pairs are less related than if they were chosen at random. Although outbreeding also is usually associated with selection for viability, it is not effective if the alleles are recessive with small frequencies. This is because many defect allele carriers are heterozygous and can thus not be identified and removed from the breeding stock. But outbreeding may be also effective for purging defect alleles, if the frequencies of the defect alleles are large. As pointed out by Ballou & Lacy (1995), another effect of outbreeding is that underrepresented genotypes become mixed with overrepresented genotypes and then it is no longer possible to increase the underrepresented genotypes. The main advantage of outbreeding is a reduction of the prevalence of heritable diseases in the next generation. Moreover, the frequent use of popular sires is less harmful to genetic diversity when they are outbred. In this paper, we discuss the amount of reduction that would have been achieved by outbreeding alone in the Kromfohrländer breed. We identify the prevalent mating systems and the different lines in the Kromfohrländer breed. We also discuss the reasons causing the bottlenecks and the necessary amount of purging to compensate for the negative effects of the bottlenecks.

2. Materials and methods

(i) Materials

The dataset was provided by the German Rassezuchtverein der Kromfohrländer eV and consisted of pedigrees and additional information on 5527 dogs. All European Kromfohrländer subpopulations were included in the database, but some dogs without offspring that were born before 1970 were missing. The Kromfohrländer, one of the most recent German dog breeds, originated in 1945 and has been recognized by FCI since 1955 (Group 9, section 10). It is used as a companion dog. The Kromfohrländer originally was a rough coated breed, but smooth coated dogs appeared early. In 1961, first dogs were exported to Finland. The Finnish population was isolated from the German population until 1988 due to strict quarantine legislation. In accordance with the German Rassezuchtverein der Kromfohrländer eV policies, the use of stud dogs was restricted to 6 litters since April 2003, and ancestors must occur only once in a two generation pedigree as well as only one sibling of the same litter and in the first and third generation no same ancestor.

(ii) Methods

The stochastic model depends only on the pedigrees and on the dates of birth. No disease records are

needed. The numbers of genes in the model depend on the desired evaluation. Founder alleles are pairwise not IBD, but they could be identical by state. Some evaluations assume that the states of the founder alleles are at random. Apart from that, only the alleles of the non-founders are at random, not the pedigrees. The passing of alleles through the pedigree is modelled in the usual way by assuming no selection and Mendelian transmission. All evaluations are based on age cohorts. An age cohort could be an arbitrary fixed subset of the population, e.g. the birth cohort B_t in year t, or the current population P_t at time $t \in \mathbb{R}$. Thereby, the current population is assumed to consist of all individuals up to an age of 9 years.

The **gene diversity** of an age cohort j satisfies the equation

$$GD(j) = 1 - E\left(\sum_{k=1}^{2\mathcal{F}} q_{kj}^2\right) = 1 - \bar{f}_j, \tag{1}$$

where $\bar{f}_j$ is the mean pairwise coancestry (kinship) in age cohort j, q_{kj} is the fraction of alleles at one locus in age cohort j that are IBD with the kth founder allele, $\mathcal{F}$ is the number of founders and E denotes the expectation (see Caballero & Toro, 2000). The gene diversity of an age cohort was obtained by computing the coancestry matrix with function `kinship()` from R-package `kinship`.

Although gene diversity is the most widely accepted notion of genetic diversity, there are two concerns about model reliability. First of all, this gene diversity corresponds to the gene diversity of Nei (1973) only if all alleles from the founder population are assumed to be different, and secondly, the sampling of the founders from an ancestral population is not modelled. The **allelic diversity** $E(n_t)$, which is defined as the expected total number n_t of alleles at one locus in population P_t that are not IBD, has the same disadvantages. We compare gene diversity with other notions of genetic diversity to meet concerns.

The **drift diversity** $DD_d(j)$ of age cohort j is defined as

$$DD_d(j) = 1 - 2^d E(|p_j - 0.5|^d), \tag{2}$$

where p_j is the random frequency of a neutral allele a in age cohort j and $d > 0$. Thereby, the founders are assumed to be chosen at random from a large ancestral population in which allele a has the initial frequency 0·5 and is in the Hardy–Weinberg equilibrium (HWE). Thus, drift diversity depends only on the expected departure of an allele frequency from the initial allele frequency. For $d = 2$, the diversity depends on the variance of allele frequencies, which is studied for example in Kimura & Crow (1963). Since

$$\frac{1}{2}DD_2(j) = 1 - E(p_j^2 + (1-p_j)^2), \tag{3}$$

the diversity $DD_2(j)$ is nothing but the gene diversity for the biallelic model. But here the allele frequencies

in the founder population are at random. If one is interested in absolute deviations rather than in squared deviations, also $d = 1$ can be used. Drift diversities were estimated according to formula (2) by Monte Carlo simulation. The genotypes of all descendants along the pedigree were simulated according to Mendelian rules, assuming no selection on this allele. The average absolute deviation of the allele frequency to its expectation 0·5 was calculated for all birth cohorts from 50 000 repetitions. The drift diversities were estimated from the result.

We define the **potential gene diversity** $PD(t)$ of birth cohort B_t as the maximum gene diversity that could have been achieved within the birth cohort by optimal contributions of the breeding animals. Because of eqn (1), these contributions minimize the average coancestry within the birth cohort. We have

$$\bar{f}_{B_t} = \frac{1}{(2N_t)^2}(c_t^T D_t c_t + 2N_t - c_t^T \text{Diag}(D_t)), \tag{4}$$

where D_t is the coancestry matrix of all r reproductive individuals and c_t is the vector of their absolute contributions, i.e. c_t is the vector with the numbers of offspring of each breeding animal. N_t is the number of individuals in the birth cohort and $\text{Diag}(D_t)$ is a vector that contains the diagonal elements of D_t. A proof can be found in the Appendix. With

$$\tilde{f}_t: \mathbb{R}_{\geqslant 0}^r \to [0, 1],$$
$$\tilde{f}_t(c) = \frac{1}{(2N_t)^2}(c^T D_t c + 2N_t - c^T \text{Diag}(D_t)),$$

we have

$$PD(t) = 1 - \tilde{f}_t(c_t^{\min}),$$

where $c_t^{\min}$ minimizes the function $\tilde{f}_t$ under appropriate side conditions. The side conditions should be that the sum over all contributions from a given sex is equal to N_t and that the contributions are natural numbers. But within calculations, the contributions could be arbitrary non-negative real numbers and it was assumed that all individuals from the current population P_t could be used for breeding. We used function `solve.QP()` from R-package `quadprog` to find the minimum of the objective function. As discussed in Toro & Caballero (2005), the removal of subpopulations from a population would increase the gene diversity, if allele frequencies become more equalized. By contrast, the potential gene diversity cannot be increased by removing individuals from the population because the removal of individuals would add additional constraints to the objective function.

Caballero & Toro (2000) derived a similar objective function for monoecious populations. If it is used for

a diecious population, then we get

$$\text{PD}(t) \leqslant 1 - \left(\frac{c_t^{\min}}{2N_t}\right)^{\mathrm{T}} D_t \left(\frac{c_t^{\min}}{2N_t}\right) \leqslant 1 - g_t(\tilde{c}_t^{\min}) = \text{PD}_\infty(t),$$

where $\tilde{c}_t^{\min}$ minimizes $g_t(c) = c^{\mathrm{T}} D_t c$ under the side conditions $c_k \geqslant 0$ and $\sum_{k=1}^r c_k = 1$. Thus, the objective function of Caballero & Toro (2000) gives the upper bound $\text{PD}_\infty(t)$ for the potential gene diversity of birth cohort B_t. Since the formula does not depend on the size of the birth cohort, the definition extends to a continuous timescale. It can be considered as the maximum gene diversity that could have been derived from the reproductive individuals at time t in an infinitely large birth cohort. Moreover, $\text{GD}(P_t) \leqslant \text{PD}_\infty(t)$. Therefore, $\text{PD}_\infty(t)$ is called the potential diversity of the population at time t.

An upper bound $\text{BD}(t)$ for the gene diversity that could be achieved in a birth cohort also has been introduced by Lacy (1995) as the diversity that would be achieved when all alleles still existing in the population are brought to equal frequencies. We have

$$\text{GD}(P_t) \leqslant 1 - E\left(\sum_{k=1}^{n_t} \left(\frac{1}{n_t}\right)^2\right) = 1 - E\left(\frac{1}{n_t}\right) = \text{BD}(t).$$

However, Lacy recognized that this upper bound (which he called the potential gene diversity) can never, even in theory, be achieved. In contrast, the potential gene diversity defined in this paper can be achieved, if the optimal contributions are natural numbers.

The **distribution of the allele frequency** of a rare neutral allele in the current birth cohort 2007 was estimated by Monte Carlo simulation according to MacCluer *et al.* (1986) from 100 000 repetitions. It is the conditional distribution, given that the allele descends from the founder of interest and that the allele is still segregating. The probability of fixation and the probability of elimination were also estimated.

The **mean frequency of homozygous carriers** of rare neutral alleles was studied by assuming that all founders had the same number of such alleles. Take $\mathcal{A}$ to be the set of all rare neutral alleles and let $\#\mathcal{A}$ be the number of these alleles. Take $q_{aa}(j)$ to be the frequency of homozygous carriers of allele $a \in \mathcal{A}$ and let $q_a(j)$ be the frequency of allele $a \in \mathcal{A}$ in age cohort j. Then the expectation of the mean frequency of homozygous carriers is given by

$$\text{Hom}(j) = E\left(\frac{1}{\#\mathcal{A}} \sum_{a \in \mathcal{A}} q_{aa}(j)\right),$$

whereas the expectation in the case of HWE is given by

$$\text{Hom}_{\text{HW}}(j) = E\left(\frac{1}{\#\mathcal{A}} \sum_{a \in \mathcal{A}} q_a(j)^2\right).$$

These parameters can be calculated from inbreeding coefficients and coancestries by using the formulae $\text{Hom}(j) = \frac{1}{2\mathcal{F}} \bar{F}_j$ and $\text{Hom}_{\text{HW}}(j) = \frac{1}{2\mathcal{F}} \bar{f}_j$, where $\bar{F}_j$ is the mean inbreeding coefficient in age cohort j. The first formula can be obtained with eqn (8) from the Appendix and the proof of the other formula is similar to the proof of eqn (1).

A hierarchical **cluster analysis** was carried out to identify the different lines of the population. Individuals whose inbreeding coefficients exceeded a predefined threshold value were clustered by function `agnes()` from R-package `cluster`. The distance matrix $1 - D$ was obtained from the kinship matrix D, which was calculated by function `kinship()` from R-package `kinship`.

The **most influential ancestors**, who are defined as the ancestors with the largest contribution to a given subpopulation, are also identified. In other papers (e.g. Cole *et al.*, 2004), the most influential ancestors are defined as the individuals with the largest kinship to the population. But this would be misleading since many of them did not produce offspring. The genetic contribution of a particular ancestor to a given subpopulation is the expectation of the fraction of genes which has been derived from this ancestor. Only direct descent is involved.

Bottlenecks increase the probability of individuals to be affected by heritable diseases and result in the **necessity of purging**. The question of interest is, how many undesirable alleles must be purged such that the probability of an individual to be affected by a deleterious allele falls below the probability from before the bottleneck. We assume that all deleterious alleles are recessive. At first, we look at some rare, neutral and independent recessive alleles from different genes, one from each founder. For any individual x within the stud book, the probability $P(x, 1)$ that the individual is not affected by one of these alleles can be approximated by the formula

$$P(F_x, 1) = \left(1 - \frac{F_x}{2\mathcal{F}}\right)^{\mathcal{F}}, \tag{5}$$

where F_x is the inbreeding coefficient of individual x. The approximation is derived in the Appendix. We assume that every founder i had $d \geqslant 1$ undesirable but neutral, rare recessive alleles $a_{i,1}, \ldots, a_{i,d}$, ordered for example by decreasing deleteriousness. For a given individual x the numbers $N_x(a_{i,j}) \in \{0, 1, 2\}$ of different alleles $a_{i,j}$ are assumed to be independent. The probability $P(x, d)$ that none of the $\mathcal{F} \cdot d$ alleles is homozygous is given by

$$P(x, d) = P(\forall j = 1, \ldots, d \quad \forall i = 1, \ldots, \mathcal{F}: N_x(a_{i,j}) < 2).$$

We call this the probability of the individual to be correct. Because of the independency we have $P(x, d) = P(x, 1)^d$. We calculate the correctness probability

$P(x, d)$ from the inbreeding coefficient by using the approximation

$$P(F_x, d) = \left(1 - \frac{F_x}{2\mathcal{F}}\right)^{\mathcal{F} \cdot d}.$$

Now we assume that breeders successfully purge the first $p \cdot 100\%$ of the undesirable alleles from each founder while ignoring the other ones, so that only the other ones remain neutral. Then, purging could compensate the negative effects of the bottlenecks, if the correctness probability after purging is larger than the correctness probability before the bottleneck, i.e. if

$$P(F_{\text{after}}, (1-p)d) \geqslant P(F_{\text{before}}, d).$$

Thereby, F_{before} denotes the inbreeding coefficient before the bottleneck, or alternatively the maximum inbreeding coefficient that is acceptable without purging and F_{after} denotes the expected inbreeding coefficient after purging. The inequality holds if and only if

$$1 - p \leqslant \frac{\ln(1 - \lambda F_{\text{before}})}{\ln(1 - \lambda F_{\text{after}})},$$

where $\lambda = \frac{1}{2\mathcal{F}}$. By using L'Hospital's rule we obtain as a simple approximation that the fraction p of purged alleles should satisfy

$$p \geqslant \frac{F_{\text{after}} - F_{\text{before}}}{F_{\text{after}}}. \tag{6}$$

3. Results

(i) *Genetic diversity*

Figure 1 shows a scatter plot of the gene diversity versus the drift diversities of the birth cohorts. The figure suggests that gene diversity is identical with drift diversity, if squared deviations are used. It is not difficult to show that both notions of genetic diversity are indeed identical, that is,

$$\text{GD}(j) = \text{DD}_2(j) \tag{7}$$

for any age cohort j. Thus, drift diversities generalize gene diversity. A proof can be found in the Appendix. There it is also shown that the drift diversity $\text{DD}_2(j)$ would coincide with gene diversity even if another initial frequency is used in definition (2). Figure 1 shows a strong monotone dependency also for $d = 1$. Therefore, further investigations may be based on the notion that is mathematically most tractable.

Figure 2 visualizes the development of the gene diversity $\text{GD}(P_t)$ of the population throughout the entire Kromfohrländer breeding history by using a continuous timescale. The diversity remained almost constant until recognition of the breed in 1955 because most dogs born before recognition had the same
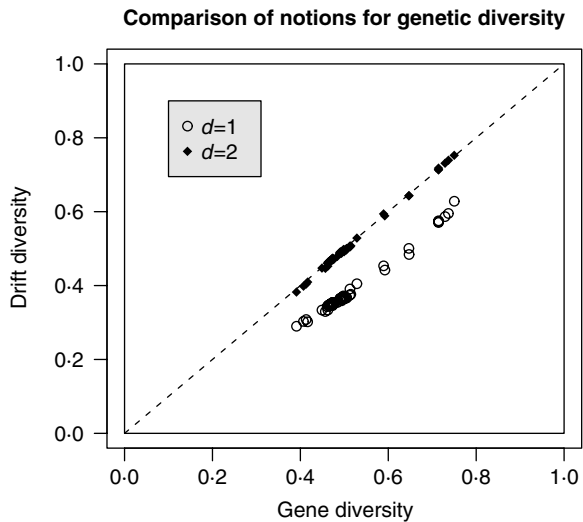

Comparison of notions for genetic diversity

Fig. 1. Comparison of genetic diversities.

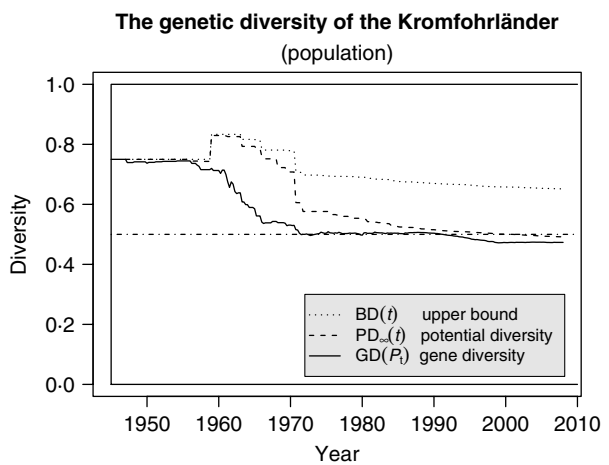**The genetic diversity of the Kromfohrländer**
(population)

Fig. 2. Gene diversity and potential gene diversity of the population.

parents. Recognition was followed by a sudden drop of gene diversity. Since 1970, the gene diversity was almost constant, except for a small but sustained decline after 1990. However, more important than the gene diversity of the population is the potential gene diversity $\text{PD}_\infty(t)$ of the population, given by the dashed line in Fig. 2. The strong rise of the potential gene diversity in 1959 was due to an additional founder. It can be seen that nowadays no substantial improvement of gene diversity can be achieved by optimal contributions of the breeding animals. The upper bound $\text{BD}(t)$ of Lacy (dotted line) heavily overestimates the scope for improvement since it does not account for the effects of mixing rare and common alleles and it accounts for genetic drift only via the number of different alleles.

Figure 3 shows the development of the gene diversity $\text{GD}(B_t)$ of the birth cohorts. It can be seen that the small but sustained decrease, that was detected in Figure 2, took place in 1990. The increase of potential
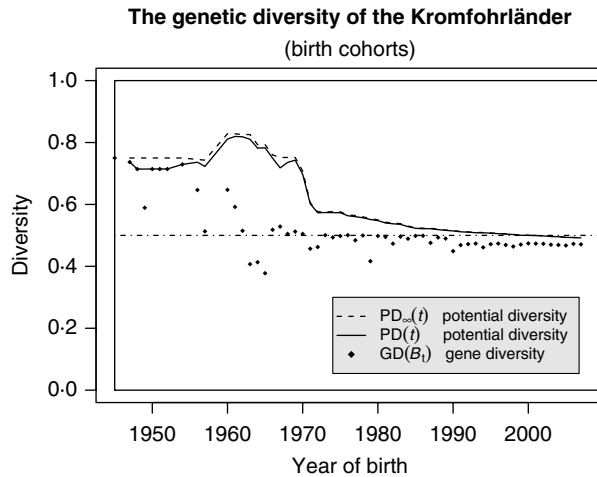
The genetic diversity of the Kromfohrländer

(birth cohorts)

Fig. 3. Gene diversity and potential gene diversity of the birth cohorts.

Inbreeding coefficients of the Kromfohrländers

Fig. 5. Development of inbreeding coefficients.

Density of the frequency of a segregating allele

that originates from a particular founder.

	Peter	Fiffi	Elfe
Prob. of elimination:	45·9 %	46·1 %	56·0 %
Prob. of fixation:	0·4 %	0·4 %	0·0 %
expected frequency:	0·38	0·38	0·21

Fig. 4. Distributions of allele frequencies.

The mean frequency of homozygous carriers

of a rare neutral allele within the population.

Fig. 6. Mean frequency of homozygous carriers of a rare neutral allele.
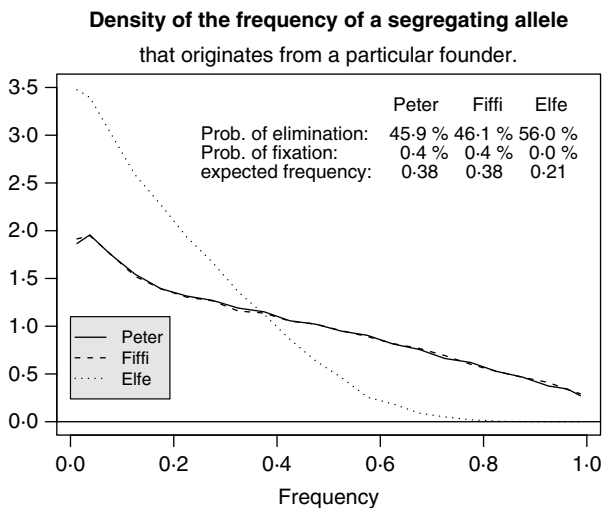
diversity $PD(t)$ in 1968 and 1969 resulted from an increased number of birth. The upper bound $PD_\infty(t)$ did not increase in 1968 and 1969 since it does not depend on the size of the birth cohorts.

(ii) *Distribution of allele frequencies*

The distribution of the frequency of a rare neutral allele in the current population depends on the founder from whom the allele originates. The founders of the Kromfohrländer are Peter, Fiffi and Elfe. Their relative contributions to the current population are 0·41, 0·41 and 0·18. The distribution of the frequency of a rare neutral allele that originates from a particular founder is shown in Fig. 4. An allele is eliminated with a probability of round about 50 %, no matter from whom the allele originates. The probability of fixation is 0·4 % and it is negligible if the allele originates from Elfe. But alleles that originate from Peter
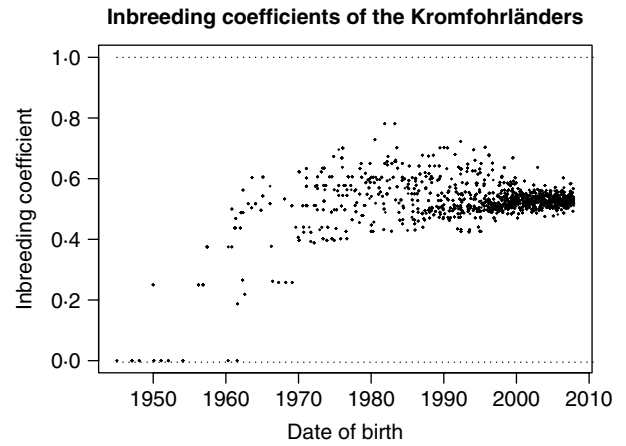
or Fiffi have large frequencies with high probability, if not undergone selection until now.

(iii) *Mating system*

Figure 5 shows a scatter plot of the inbreeding coefficients versus the dates of birth for all dogs in the database. The inbreeding coefficients are based on all generations back to the formation of the stud book, which accounts for the high values. Inbreeding coefficients increased until 1985. After that, the mean inbreeding coefficient decreased slightly and the variation of inbreeding coefficients decreased substantially. Nowadays, there exist no dogs with inbreeding coefficients larger than 0·6.

In order to identify the prevalent mating system at a given time, the expectation of the mean frequency of homozygous carriers of a rare neutral allele within the population is compared with its expectation in HWE. The expectations $Hom(P_t)$ of the mean frequency of homozygous carriers are given by the continuous line in Fig. 6, whereas the dashed line shows their
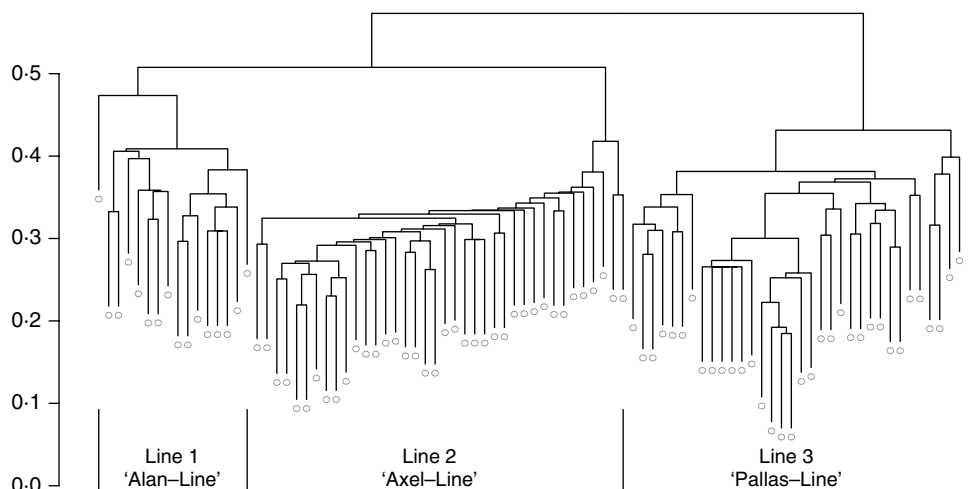
Fig. 7. The lines of the Kromfohrländer.

expectations $Hom_{HW}(P_t)$ in the case of HWE. Recall that the population consists of all individuals up to an age of 9 years, so that effects of changes in population management can be seen only with a delay. The loss of genetic diversity and the small number of founders account for the increase of homozygous carriers from 1955 to 1965. It is not due to line breeding since the expected frequencies were below their expectations in HWE, which suggests that outbreeding occurred. This outbreeding was due to the additional founder Elfe who had his first litter in 1960. Around 1985, line breeding or inbreeding was the dominating breeding system, since the expected frequencies were larger than their expectations in HWE. But thereafter, it shifted in the direction of outbreeding. Note that the mean fraction of homozygous carriers could be diminished only very little by the shift in the direction of outbreeding.

Dogs with inbreeding coefficients larger than 0·6 that were born between 1970 and 1990 were clustered by means of their pedigrees in order to identify the different lines of the Kromfohrländer breed. But only one dog from each litter was included. Closely related individuals belong to the same branch of the clustering tree shown in Fig. 7. It can be seen that there existed three different lines. The first group consists mainly of German smooth coated dogs, the second group consists mainly of German rough-coated dogs and the third group consists of Finnish dogs.

Table 1 shows the founders and all dogs whose contribution to one of the lines or to the current population is at least 35%. It can be seen that the first line is founded by Alan and Betta vom Weddern. The second group is linebred to Axel van de Poort van Drenthe (contribution 72% to Line 2), and the third line is founded by the Finnish dog Pallas av Ros-Loge (contribution 66% to Line 3). The relatively small contribution of Alan and Pallas to the current population indicate that the Alan-Line (Line 1) and the

Pallas-Line (Line 3) are more historical, whereas the current population is dominated by descendants of the Axel-Line (Line 2).

(iv) *Bottlenecks*

Two decreases of genetic diversity were detected in Fig. 3. A major decrease is observed between 1955 and 1965 and a minor decrease was in 1990. The Kromfohrländer population was very small for several years after recognition of the breed and several dogs were used extensively for breeding (see Fig. 8 and Table 1). This caused the dramatic loss of genetic diversity between 1955 and 1965, shown in Figs 2 and 3.

Axel's contributions to the birth cohorts increased substantially in 1990, which is considered to be the reason for the second decrease. Since a subdivision of the population was suggested by our cluster analysis, relative population sizes and contributions of Axel to birth cohorts were calculated separately for the Finnish subpopulation, for the non-Finnish rough-coated kennels, and for the non-Finnish smooth-coated kennels. A kennel is considered as a smooth-coated kennel, if more litter parents were smooth coated than rough coated. Figure 9 shows that the increased influence of Axel results from a breakdown of the Finnish population, an expansion of the German rough-coated subpopulation, and an export of German dogs to Finland. The reasons were determined by questionnaire from breeders.

The lines were established in the 1970s by mating very closely related individuals (see Table 1). In the 1980s, no German breeder exported dogs to Finland because of strict quarantine legislation. Thus, Finnish breeders could not breed to less related dogs. But problems (e.g. cataract) accumulated due to the fixation of deleterious alleles and breeders did not find enough offspring for breeding. As a consequence, no Finnish breeder had a litter in 1990. After Finland

Table 1. *Genetic contributions*

Name	Born	Sire	Dam	Line 1	Line 2	Line 3	Pop.
Peter	1945	Unknown	Unknown	0·40	0·39	0·48	0·41
Fiffi	1945	Unknown	Unknown	0·40	0·39	0·48	0·41
Elfe	1959	Unknown	Unknown	0·19	0·21	0·03	0·18
Zottel vom Wellersberg	1947	Peter	Fiffi	0·49	0·48	0·61	0·50
Bento vom Wellersberg	1947	Peter	Fiffi	0·32	0·31	0·36	0·31
Benno vom Wellersberg	1956	Bento vom Wellersberg	Zottel vom Wellersberg	0·55	0·54	0·57	0·54
Fax vom Wellersberg	1960	Benno vom Wellersberg	Elfe	0·38	0·43	0·06	0·36
Dina vom Lenneberg	1962	Alex von Mazeppa	Billy vom Lenneberg	0·38	0·43	0	0·35
Pallas av Ros-Loge	1966	Dirk vom Lenneberg	Astrid von Mazeppa	0	0·01	0·66	0·12
Quick vom Wellersberg	1969	Fax vom Wellersberg	Dina vom Lenneberg	0·20	0·43	0	0·30
Anka vom Hasselrain	1969	Peter vom Wellersberg	Bona vom Ehrling	0·47	0·23	0	0·17
Alan vom Weddern	1970	Quant vom Wellersberg	Anka vom Hasselrain	0·47	0	0	0·04
Alta vom Weddern	1970	Quant vom Wellersberg	Anka vom Hasselrain	0	0·38	0	0·17
Betta vom Weddern	1971	Quick vom Wellersberg	Anka vom Hasselrain	0·38	0·05	0	0·12
Axel van de Poort van D.	1972	Quick vom Wellersberg	Alta vom Weddern	0	0·72	0	0·31

Fig. 8. Development of population size.
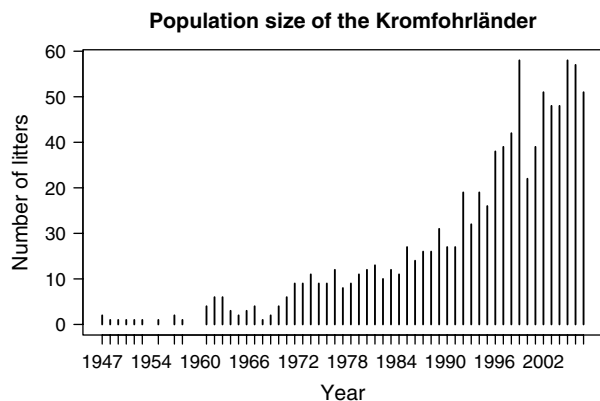

Fig. 9. Subpopulations.

relaxed the strict quarantine legislation in 1988, dogs from the Axel-Line were exported to Finland and had their first litters in 1991. They were mated to the dogs that remained and could re-establish the breed in Finland. Apart from that, the unequal ancestor contributions in the current birth cohorts indicate little gene flow among subpopulations.

(v) *Necessary amount of purging due to bottlenecks*

We consider three rare, neutral and independent alleles from different genes, one from each founder. Figure 10 shows the scatter plot of the probability of an individual to be not affected by one of these alleles versus the inbreeding coefficient. The probabilities are estimated from 20 000 repetitions by computer simulation for the Kromfohrländer. The function $P(F_x, 1)$ from eqn (5) approximates these probabilities very well, if all alleles are recessive. But if the three alleles are dominant, then the dependency is the opposite. Although there exist dominant alleles with incomplete
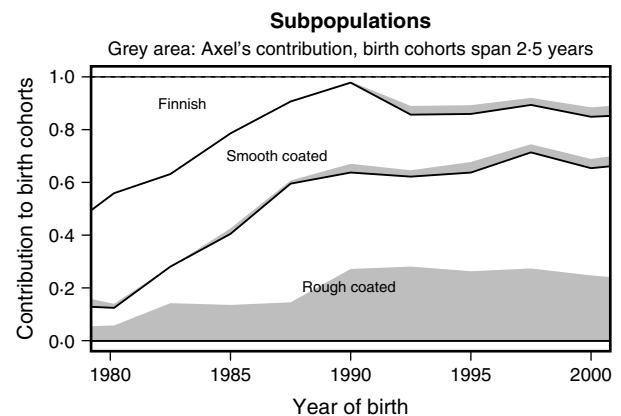
penetrance that cause heritable diseases, e.g. osteosarcoma in Scottish deerhounds (Phillips *et al.*, 2007), the majority of such diseases is caused by recessive alleles. If all deleterious alleles are recessive and if alleles with largest deleteriousness have priority, then the fraction that should be purged is given by eqn (6). Inbreeding coefficients of the Kromfohrländer increased from about $F_{before} := 0·25$ due to bottlenecks until they reached about $F_{after} := 0·5$, and thus, the correctness probability from before the bottleneck would be recovered by purging 50 % of the deleterious alleles. But note that a better recommendation could, in principle, be derived from disease records. In addition, less purging would be necessary if deleterious alleles with highest frequencies have priority or if some disease alleles are dominant.

4. Implications and discussion

The Kromfohrländer is based on a small number of founders and suffers from two bottlenecks. The first
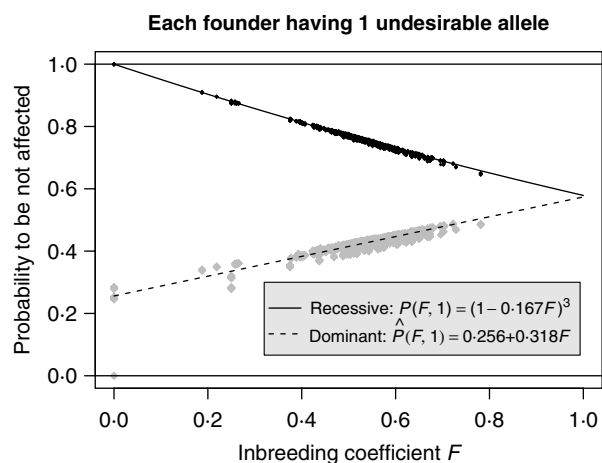
Each founder having 1 undesirable allele

Fig. 10. Probability to be not affected.

Legend (from figure):
Recessive: $P(F, 1) = (1 - 0.167F)^3$
Dominant: $\hat{P}(F, 1) = 0.256 + 0.318F$

bottleneck between 1955 and 1965 was generated by rigorous selection in a very small population. The second decrease of genetic diversity turned out to be caused by a combination of two factors: in the beginning, three lines were established by very close inbreeding, and thus the fixation of deleterious alleles could not be prevented in some lines. The breeders did not breed to less related dogs betimes. Secondly, selection within the whole population against partly fixed alleles was applied to this sublined population. This caused the breakdown of the Finnish subpopulation and triggered that gene flow was only from one line to the others but almost not backwards.

Except for the small decrease in 1990, the genetic diversity remained almost constant throughout the last 40 years. Selection against heritable diseases continued, so that the soundness of the breed can be expected to be better than before. Heritable diseases are still a problem, but the situation does not seem to be worse than in other dog breeds. It could not be deduced from the pedigrees, whether the bottlenecks still have a negative effect on the breed, because this is also dependent on the ability of the breeders to choose the right offspring for breeding and on the genetic basis of inbreeding depression. But we derived the magnitude of purging that is required to compensate the negative effects of the bottlenecks even if all undesirable alleles are recessive. Figure 10 showed that the probability of an individual to be affected by a heritable disease may increase or decrease with increasing inbreeding coefficient, depending on the mode of inheritance. Note that this and the theory of Awdeh & Alper (2005) and Awdeh *et al.* (2006) may explain why some purebred dog breeds have a higher median age at death than mixed breed dogs (see Proschowsky *et al.*, 2003).

Figure 3 showed for the Kromfohrländer that at the present time no substantial improvement of genetic diversity can be achieved by optimal contributions

of the breeding animals. Moreover, Fig. 6 showed that the shift from line breeding in the direction of outbreeding alone, would not have been effective in reducing the prevalence of heritable diseases with recessive inheritance. This is primarily due to the fact that all individuals are closely related to each other, so that real outbreeding was not possible. Thus, the usual approach to calculate optimal contributions for the breeding animals and to avoid inbreeding as far as possible would not be effective. Outcrossing with other dog breeds is also not an option. The small fixation probabilities of 0–0.4 % for a rare neutral allele showed that the elimination of undesirable alleles can likely be achieved by selection, i.e. without outcrossing. As suggested in the introduction, we propose to purge undesirable alleles. The subdivision of the population into at least two subpopulations (rough coated and smooth coated) should be sustained with the intention of purging until no further progress can be achieved. If the outcome is not satisfactory, then the subpopulations should subsequently be crossed for one generation. Figure 4 showed that alleles that originate from Peter or Fiffi have large frequencies with high probability, if they have not undergone selection until now. High frequencies enable effective selection against these alleles even without line breeding. Therefore, the avoidance of inbreeding within subpopulations should be followed as long as there exist heritable diseases with high prevalence and recessive inheritance. After that the current restrictions on line breeding should be relaxed in order to enable breeders to use more effective breeding schemes for purging the remaining undesirable alleles. But inbreeding must be moderate in order to not create new subpopulations in which deleterious alleles could become fixed. Line breeding should be combined with the calculation of optimal contributions for breeding animals or with establishing a sperm bank in order to not comprise the genetic diversity. However, optimal contributions that maximize gene diversity of the population are not yet well understood for populations with overlapping generations, but see the discussion in Caballero & Toro (2000).

The upper bound of Lacy turned out to overestimate the potential for improvement of gene diversity, but it could be useful to characterize the potential of a population to respond to selection. It is closely related to allelic diversity. An approach to conserve allelic diversity based on identity by state has been introduced by Vales-Alonso *et al.* (2003).

Drift diversities were considered as more realistic notions of genetic diversity than gene diversity since the model accounts for the sampling of the founders. But for $d = 2$, drift diversity and gene diversity turned out to be identical so that gene diversity can be interpreted as drift diversity in order to account for the

sampling. The definition of drift diversity could be further extended by assuming other joint distributions of the founder alleles $a_1, \ldots, a_{2\mathcal{F}}$. We think that this enables to model the structured ancestral populations discussed in Templeton (1980) as well as populations, where the pedigrees are not traced back to the foundation of the breed, but the development of population size is known.

In several populations, individuals with underrepresented genotypes are underrepresented for good reasons, so that breeding via optimal contributions may not be enforceable in due time, even if the potential for improvement of genetic diversity is large. We think that for such populations the maintenance of potential diversity is more important than the conservation of gene diversity since the potential diversity decreases even if the gene diversity is maintained.

Appendix

(i) *Proof of eqn (4)*

Take $D = (f^P_{ik})_{i,k=1, \ldots, r}$ to be the coancestry matrix of the reproductive individuals and for $i, k \in \{1, \ldots, N\}$ let f_{ik} be the coancestry of individuals number i and k from birth cohort j. Let $s_i, d_i \in \{1, \ldots, r\}$ be the sire and the dam of individual i and let $(e_1, \ldots, e_{2N}) = (s_1, \ldots, s_N, d_1, \ldots, d_N)$. Let $c = (c_1, \ldots, c_r)^T$, where $c_a = \#\{i : e_i = a\}$ is the contribution of the reproductive individual number a to the birth cohort. With eqns (1) and (2) from Caballero & Toro (2000) we obtain

$$\bar{f}_j = \frac{1}{N^2} \sum_{i \neq k} f_{ik} + \frac{1}{N^2} \sum_{i=1}^{N} f_{ii}$$

$$\stackrel{(1),(2)}{=} \frac{1}{(2N)^2} \left(\sum_{i,k=1}^{N} (f^P_{s_i s_k} + f^P_{s_i m_k} + f^P_{m_i s_k} + f^P_{m_i m_k}) + 2N \right.$$

$$\left. - \sum_{i=1}^{N} (f^P_{s_i s_i} + f^P_{m_i m_i}) \right)$$

$$= \frac{1}{(2N)^2} \left(\sum_{i,k=1}^{2N} f^P_{e_i e_k} + 2N - \sum_{i=1}^{2N} f^P_{e_i e_i} \right).$$

By reordering the summands it follows that

$$\bar{f}_j = \frac{1}{(2N)^2} \left(\sum_{(a,b) \in \{1, \ldots, r\}^2} \sum_{(i,k):(e_i, e_k)=(a,b)} f^P_{e_i e_k} + 2N \right.$$

$$\left. - \sum_{a=1}^{r} \sum_{i \in \{i : e_i = a\}} f^P_{e_i e_i} \right)$$

$$= \frac{1}{(2N)^2} \left(\sum_{(a,b) \in \{1, \ldots, r\}^2} c_a c_b f^P_{ab} + 2N - \sum_{a=1}^{r} c_a f^P_{aa} \right)$$

$$= \frac{1}{(2N)^2} \left(c^T D c + 2N - c^T \mathrm{Diag}(D) \right).$$

(ii) *Derivation of approximation (5)*

The inbreeding coefficient F_x of an individual x is the probability that both alleles at one locus are IBD. By considering one gene for which all the alleles $a \in \mathcal{A}$ in the founder population are different, the inbreeding coefficient is nothing but

$$F_x = \sum_{a \in \mathcal{A}} P(N_x(a) = 2),$$

where $N_x(a)$ is the number of a-alleles in individual x. We can write $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, where $\mathcal{A}_i$ is the set of alleles that descent from the ith chromosome of a founder. Because of symmetry, we can write

$$F_x = 2 \sum_{a \in \mathcal{A}_1} P(N_x(a) = 2). \tag{8}$$

Note that the probabilities would not change, if the alleles $a \in \mathcal{A}_1$ are located at pairwise different, independent genes.

Now let $\mathcal{B}$ be a set of rare, neutral and independent alleles from different genes, one from each founder. We have

$$P(x, 1) = P(\forall a \in \mathcal{B} : N_x(a) < 2)$$

$$= \prod_{a \in \mathcal{B}} (1 - P(N_x(a) = 2))$$

$$\sim \left(1 - \frac{\sum_{a \in \mathcal{B}} P(N_x(a) = 2)}{\mathcal{F}} \right)^{\mathcal{F}} \stackrel{(8)}{=} \left(1 - \frac{F_x}{2\mathcal{F}} \right)^{\mathcal{F}},$$

where the approximation '$\sim$' was obtained by the first-order Taylor -approximation of the exponential function.

(iii) *Proof of eqn (7)*

Let a be an allele with frequency $p \in (0, 1)$ in the ancestral population. Let a_k be the kth Founder allele at this gene and let q_{kj} be the fraction of alleles in age cohort j that descend from the kth Founder allele. We assume that $P(a_k = a) = p$ for all $k = 1, \ldots, 2\mathcal{F}$.

Let

$$\mathrm{DD}_{2,p}(j) = 1 - \frac{1}{p(1-p)} E\big((p_j - p)^2\big). \tag{9}$$

It is assumed that the founders are drawn randomly from a large population in HWE, that is, a_k and a_i should be independent for $i \neq k$. The assumptions of no selection and randomly drawn founders mean in particular that $q_{kj} q_{ij}$ should be independent of a_k and a_i as well as q_{kj} and a_k should be independent. Note that $\mathrm{DD}_2(j) = \mathrm{DD}_{2,0\cdot5}(j)$.

We can write $p_j = \sum_{k=1}^{2\mathcal{F}} q_{kj} 1_{a_k = a}$, where $1_{a_k=a}$ denotes the indicator function and $\sum_{k=1}^{2\mathcal{F}} q_{kj} = 1$. Since

$E(p_j) = p$, and

$$E(p_j^2) = \sum_{i,k} p^2 E(q_{kj}q_{ij}) - \sum_{k=1}^{2\mathcal{F}} p^2 E(q_{kj}^2) + \sum_{k=1}^{2\mathcal{F}} p E(q_{kj}^2)$$

$$= p^2 + p(1-p) \, E\left(\sum_{k=1}^{2\mathcal{F}} q_{kj}^2\right),$$

we have

$$\mathrm{DD}_{2,p}(j) = 1 - \frac{1}{p(1-p)} \left(E(p_j^2) - 2pE(p_j) + p^2 \right)$$

$$= 1 - E\left(\sum_{k=1}^{2\mathcal{F}} q_{kj}^2\right) = GD(j).$$

References

Awdeh, Z. L. & Alper, C. A. (2005). Mendelian inheritance of polygenic diseases: a hypothetical basis for increasing incidence. *Medical Hypotheses* **64**, 495–498.

Awdeh, Z. L., Yunis, E. J., Audeh, M. J., Fici, D., Pugliese, A., Larsen, C. E. & Alper, C. A. (2006). A genetic explanation for the rising incidence of type 1 diabetes, a polygenic disease. *Journal of Autoimmunity* **3**, 174–181.

Ballou, J. D. & Lacy, R. C. (1995). Identifying genetically important individuals for management of genetic diversity in captive populations. In *Population Management for Survival and Recovery* (ed. J. D. Ballou, M. Gilpin & T. J. Foose), pp. 76–111. New York: Columbia University Press.

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

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

Charlesworth, B. & Charlesworth, D. (1999). The genetic basis of inbreeding depression. *Genetical Research* **74**, 329–340.

Cole, J. B., Franke, D. E. & Leighton, E. A. (2004). Population structure of a colony of dog guides. *Journal of Animal Science* **82**, 2906–2912.

Eklund, J. & Bradford, G. E. (1976). Genetic analysis of a strain of mice plateaued for litter size. *Genetics* **85**, 529–542.

Falconer, D. S. (1971). Improvement of litter size in a strain of mice at a selection limit. *Genetical Research* **17**, 215–235.

Kimura, M. & Crow, J. F. (1963). On the maximum avoidance of inbreeding. *Genetical Research* **4**, 399–415.

Lacy, R. C. (1995). Clarification of genetic terms and their use in the management of captive populations. *Zoo Biology* **14**, 565–577.

Leberg, P. L. & Firmin, B. D. (2008). Role of inbreeding depression and purging in captive breeding and restoration programmes. *Molecular Ecology* **17**, 334–343.

MacCluer, J. W., VandeBerg, J. L., Read, B. & Ryder, O. A. (1986). Pedigree analysis by computer simulation. *Zoo Biology* **5**, 147–160.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the USA* **70**, 3321–3323.

Phillips, J. C., Stephenson, B., Hauck, M. & Dillberger, J. (2007). Heritability and segregation analysis of osteosarcoma in the Scottish deerhound. *Genomics* **90**, 354–363.

Proschowsky, H. F., Rugbjerg, H. & Ersbøll, A. K. (2003). Mortality of purebred and mixed-breed dogs in Denmark. *Preventive veterinary medicine* **58**, 63–74.

Robertson, A. (1952). The effect of inbreeding on the variation due to recessive genes. *Genetics* **37**, 189–207.

Templeton, A. R. (1980). The theory of speciation via the founder principle. *Genetics* **94**, 1011–1038.

Toro, M. A. & Caballero, A. (2005). Characterization and conservation of genetic diversity in subdivided populations. *Philosophical Transactions of the Royal Society B* **360**, 1367–1378.

Vales-Alonso, J., Fernandez, J., González-Castaño, F. J. & Caballero, A. (2003). A parallel optimization approach for controlling allele diversity in conservation schemes. *Mathematical Biosciences* **183**, 161–173.