Journal of Heredity 2012:103(2):197–205
doi:10.1093/jhered/esr135
Advance Access publication January 24, 2012

Extending Pedigree Analysis for Uncertain Parentage and Diverse Breeding Systems

Robert C. Lacy

From the Department of Conservation Science, Chicago Zoological Society, Brookfield, IL 60513.

Address correspondence to R. C. Lacy at the address above, or e-mail: rlacy@ix.netcom.com.

Abstract

Breeding programs aimed at conserving genetic diversity in populations of wildlife or rare domestic breeds rely on detailed pedigree analysis for selection of breeders that will minimize the loss of alleles, reduce the accumulation of inbreeding, and maintain gene diversity. Commonly, techniques use a matrix of kinship coefficients to derive measures of genetic variation, inbreeding, and the value of individuals as breeders. Although these techniques were first developed for use on known pedigrees of diploid individuals, the concepts and methods can be extended to apply to any entity that contains genes derived from definable sources (e.g., individual parents, social groups, colonies, gene banks) via a definable mechanism of heredity (e.g., sexual reproduction between separate sexes, hermaphroditic selfing, autozygous production of homozygous or haploid offspring, cloning). Individuals with partly unknown ancestry or multiple possible parents can also be incorporated into kinship calculations, based on probabilistic assignment of parental contributions. This paper presents the algorithms used in new PMx software to extend traditional pedigree analysis techniques used for complete pedigrees of sexually reproducing, diploid species to deal with missing information due to unknown or uncertain parentage, and other breeding systems such as clones, selfing hermaphrodites, and haploid offspring or autogamy.

Key words: *conservation genetics, inbreeding coefficient, kinship, population management, software*

Over the past few decades, methods of pedigree analysis have been developed and applied to the genetic management of captive populations of animals (Flesness 1977; Foose and Ballou 1988; Lacy 1994, 1995; Lacy et al. 1995; Caballero and Toro 2000; Ballou et al. 2010). Procedures currently in use for analyzing pedigrees and selecting breeding pairs can result in the optimal retention of gene diversity, minimal or near minimal loss of alleles, and provide the opportunity for optimal avoidance of inbreeding in future generations (Ballou and Lacy 1995; Fernández et al. 2001, 2003, 2004; Lacy 2009). All pedigree calculations must start with the "founders"—those individuals at the beginning of the pedigree with no known parents, ancestors, or relatives other than their own direct descendants in the population. For captive populations of wildlife, the founders are usually the wild-caught progenitors, although captive-born animals obtained from unrelated breeding programs may also be considered founders for a given captive population. The founders are assumed to be unrelated and equally valuable genetically (i.e., founders are equally likely to contain unique or rare alleles and alleles conferring high fitness). Pedigree analyses of the descendants of the founders are based on Mendelian inheritance, with each descendant receiving half of its genome from each of its 2 parents.

The techniques in use for pedigree analysis and management are highly efficient for meeting the genetic objectives of captive breeding programs for wildlife and are easily applied to completely known pedigrees of sexually reproducing diploid individuals. However, the tools have not been fully generalized for nondiploid or asexual forms of reproduction. Moreover, full application of the methods requires that the parentage of every nonfounder individual be known. Often some pedigree information is missing for a captive population because the parents of some individuals within the population were not recorded or multiple possible parents existed within breeding groups. Individuals with unknown or uncertain ancestry can be excluded from a breeding program but doing so will result in the loss of any genetic variation unique to those individuals and will risk increased inbreeding because of the reduced size of the breeding population. Conversely, if there are unknown relationships among founders of the pedigreed population, the incorrect assumption that all founders are unrelated and of equal genetic value can lead to inadvertent inbreeding and less than optimal retention of gene diversity in the breeding program. Willis (1993) presented procedures for deciding when unknown parents should be treated as unrelated founders rather than excluding animals with unknown ancestry from the breeding program. In addition, Rudnick and Lacy (2008) demonstrated that often retention of gene diversity will be nearly optimal even if there are some unknown kinships among founders. Although kinships among founders will alter the kinships among descendants, after several generations of captive breeding the kinships are dominated by connections through common ancestors in the more recent generations. The impact of any kinships among the original founders becomes diluted as the breeding program proceeds through generations.

Ballou and Lacy (1995) described modifications to pedigree analysis equations that can be used to omit from consideration those genes that descend from unknown or undocumented sources. The procedures for omitting the portions of genomes with unknown ancestry result in estimates of kinship, inbreeding, and genetic value (uniqueness within the managed population) that are based only on those portions of the genome which can be traced back to documented founders. Thus, animals with partly unknown ancestries are treated as though their genomes are less than diploid, having received incomplete sets of genes from one or both parents. Lacy (2009) and Oliehoek (2009) demonstrated that kinships based on partial pedigree information can be used to guide breeding programs that retain high levels of gene diversity, although long-term effectiveness of the breeding program declines to not much better than random breeding when more than about 10% of parents each generation are unknown.

Although the above referenced methods of pedigree analysis provide options for management of breeding programs when some parents are not known, comparable methods have not generally been available for populations in which individuals are not diploid with biparental inheritance, or when individuals within the managed population cannot be individually selected and paired for breeding (Leus et al. 2011). Thus, the pedigree analysis techniques currently used in wildlife conservation breeding programs are not directly applicable to analyzing and managing genetic variation at sex-linked genes or for genetic management of triploid species (e.g., some parthenogenetic *Ambystoma* salamanders), unisexual species (e.g., many *Cnemidophorus* whiptail lizards), or tetraploid species (e.g., salmonid fishes). Many species, such as some colonial birds, herd ungulates, bats, fishes, and invertebrates, are optimally or obligatorily maintained and bred in groups, with multiple females, multiple males, or both. In such cases, pairing may not be fully under the control of the population manager, even if parentage can be recorded. At times, parentage can only be specified as being from among a number of alternatives, with probabilities of alternative assignments being estimated or assumed to be equal. Wang (2004) provided methods for estimating genetic diversity and optimal genetic management for specific cases of sexually and asexually reproducing organisms managed within groups that descend from precisely defined source stocks, with regular censuses and discrete generations. However, there remained a need for methods and computer programs that provide highly flexible kinship analyses for pedigree analysis and management of populations with any of a wide variety of breeding systems, pedigree structures, and completeness of data. As will be described below, the equations of Ballou and Lacy (1995) for defining kinships and other genetic metrics for partly unknown ancestries can be extended to become applicable to cases in which genomes in fact do contain less (or more) than a full diploid complement of genes.

To implement some of the above referenced and new methods for pedigree analysis, the PMx software package was recently released (Ballou et al. 2011; Lacy et al. 2011; software and manual available at www.vortex9.org/PMx.html). PMx provides demographic and genetic analysis of pedigreed populations. The genetics module of PMx calculates measures of genetic diversity (proportional gene diversity relative to the source population, founder alleles retained, inbreeding, kinships) for the population, subpopulations, and individuals. It also provides tools for selection of the numbers and identities of breeders for the purpose of optimal retention of gene diversity within constraints of managed population size and demographic characteristics. This paper presents the algorithms used in PMx to extend traditional pedigree analysis techniques used for complete pedigrees of sexually reproducing, diploid species to deal with missing information due to unknown or uncertain parentage and other breeding systems such as clones, selfing hermaphrodites, and haploid offspring or autogamy. Full derivations of all equations are not given here, as they are straightforward applications of probability theory to the problems at hand regarding the probabilities that alleles sampled from individuals with specified pedigree relationships will be identical by descent.

The kinship calculations in PMx are extended to provide estimates of approximate kinship values and levels of population diversity for cases in which the pedigree entities are groups or colonies rather than individually identified and managed organisms. Full description of extensions of pedigree methods in PMx for pedigrees of groups (including kinships between groups and individuals) is beyond the scope of this paper and will be provided elsewhere.

Materials and Methods

Pedigree Analysis and Management of Diploid Organisms with Fully Specified Pedigrees

Analysis and management of pedigrees of captive breeding populations is based primarily on genetic measures derived from the matrix of all pairwise kinships (Ballou and Lacy 1995; Lacy 1995), and the foundational methods are summarized here for completeness, as background for describing the extensions to more complex situations. The kinship or coancestry coefficient, symbolized f_{ij}, between any 2 individuals i and j is the probability that an allele sampled at random from individual i is identical to an allele sampled at random from the same locus in j due to descent from an ancestor common to both individuals (Falconer and Mackay 1996).

For the purpose of determining measures of genetic diversity relative to the base population, the calculations can begin with an assumption that the founders are unrelated and noninbred, so that pairwise kinships among founders and founder inbreeding coefficients are set to 0, and each founder's kinship to itself is $f_{ii} = 0.5$. If kinships among founders have been estimated from data on DNA markers or otherwise, those empirically based kinships and founder inbreeding coefficients can be used as the starting values for pedigree analysis (Fernández et al. 2005). When founder kinships are specified to be other than 0, then the subsequent calculations provide genetic measures relative to an earlier or alternate baseline, such as that defined by an assumption of Hardy–Weinberg equilibrium among the allele frequencies at assayed loci.

With the pairwise kinship matrix for all founders in place, all kinships for the descendant population can be readily calculated by sequentially calculating the kinship of each descendant to all prior individuals in the pedigree by applying the following relationship:

$$f_{ij} = 0.5(f_{\mathrm{sj}} + f_{\mathrm{dj}}), \tag{1}$$

in which f_{ij} is the kinship between animal i and animal j, $f_{\mathrm{s}j}$ is the kinship between the sire of i and animal j, and $f_{\mathrm{d}j}$ is the kinship between the dam of i and animal j.

The kinship of animal i to itself is given by

$$f_{ii} = 0.5 + 0.5 f_{\mathrm{sd}}. \tag{2}$$

The inbreeding coefficient (F_i) of animal i is defined as the probability that the 2 homologous alleles at a random locus are identical due to descent from an ancestor of both the sire and dam, and it is equal to the kinship between the parents (f_{sd}):

$$F_i = f_{\mathrm{sd}}. \tag{3}$$

The value of the kinship matrix is that it directly provides the expected rate of loss of gene diversity (G, defined as the heterozygosity expected under Hardy–Weinberg equilibrium; Nei 1973), and it provides a measure of genetic value of each animal (the mean kinship of the animal to the population). The mean kinship of animal i is

$$\mathrm{MK}_i = \sum f_{ij}/N, \tag{4}$$

in which the summation is over all animals, j, including the kinship of animal i to itself and N is the number of living individuals in the population.

The mean of all kinships in the population is equal to the proportional loss of gene diversity, relative to the gene diversity (G_0) of the source population from which the founders were randomly sampled: mean MK = $(\Sigma \mathrm{MK}_i)/N$ $= 1 - (G/G_0)$, in which the summation is over all animals, i. Thus, the gene diversity is given by

$$G/G_0 = 1 - [(\sum \mathrm{MK}_i)/N]. \tag{5}$$

Consequently, breeding those animals with the lowest MK values will result in the optimal retention of gene diversity in the population. It also achieves optimal or nearly optimal retention of founder alleles, and it minimizes inbreeding in future generations (Ballou and Lacy 1995; Lacy 2009; Ivy and Lacy 2012).

For the purpose of documenting the success of a breeding program in retaining the gene diversity of the source population, a common convention is to include only the living, descendant population (not the founders) in the summations that define MK and G above (Lacy 1995).

Extension of Pedigree Analysis for Partly Known Ancestries

Ballou and Lacy (1995) presented the extensions of the above formulas for calculating kinships and inbreeding coefficients for partial genomes, in which some ancestors are unknown and it is desired to use only the known part of the genome for estimating genetic measures. The fraction of the genome of animal i that is known is given by

$$k_i = (k_{\mathrm{s}} + k_{\mathrm{d}})/2, \tag{6}$$

in which k_{s} and k_{d} are the fractions of the genomes of the sire and dam that are known. For each founder, k is assigned 1; although its parents are unknown, it is assumed to be unrelated to all other founders. For individuals born within the pedigreed population but for which neither parent was recorded, $k = 0$. The kinship, f_{ij}, between an animal i and animal j is given by

$$f_{ij} = c_{\mathrm{s}} f_{\mathrm{s}j} + c_{\mathrm{d}} f_{\mathrm{d}j}, \tag{7}$$

in which $c_{\mathrm{s}} = k_{\mathrm{s}}/(k_{\mathrm{s}} + k_{\mathrm{d}})$ and $c_{\mathrm{d}} = k_{\mathrm{d}}/(k_{\mathrm{s}} + k_{\mathrm{d}})$ are the proportions of the offspring's traceable genome that are contributed by the sire and dam, respectively. Note that c_{s}, c_{d}, and f_{ij} are undefined when both parents have unknown ancestry.

The kinship of animal i to itself is the probability that 2 alleles sampled from a genetic locus are identical by descent, either because the same physical allele derived from either the dam or the sire was sampled twice or because the maternal and paternal alleles were sampled and they were identical due to descent from a common ancestor of the 2 parents. It is given by

$$f_{ii} = c_{\mathrm{s}}^2 + c_{\mathrm{d}}^2 + (2 c_{\mathrm{s}} c_{\mathrm{d}}) f_{\mathrm{sd}}. \tag{8}$$

Equations 7 and 8 are extensions of Equations 1 and 2, allowing for the possibility that the sire and dam contribute unequally to the known portion of progeny genomes.

The mean kinship (MK) of animal i must now be weighted by the proportions of genomes that are known (k):

$$\mathrm{MK}_i = \sum (k_j f_{ij}) / \sum k_j, \tag{9}$$

in which the summations are over all living animals, j, including the focal animal, $j = i$.

The population mean kinship, which equals the proportional loss of gene diversity from the source population to the known (traceable) descendant population, is similarly weighted:

$$1 - G/G_0 = \text{mean MK} = \sum (k_i \text{MK}_i)/\sum k_i; \quad \text{so that}$$

$$\begin{aligned} G/G_0 &= 1 - \left(\sum (k_i \text{MK}_i)/\sum k_i\right) \\ &= 1 - \left[\left(\sum \left(k_i \sum (k_j f_{ij})\right)\right)/\left(\sum k_i\right)^2\right], \qquad (10) \end{aligned}$$

with summations over all *i* and *j* living animals in the population. Equations 9 and 10 are extensions of Equations 4 and 5, allowing for the possibility that the ancestries of some animals are only partly known, so that animals differ in the extent to which they contribute to the known gene pool.

It is sometimes the case that an unknown parent was not likely an individual who is in the known pedigree or related to any of the known founders. For example, an individual for which there is no information about its parents may have been born before there was any known breeding of the species in captivity or it may have been obtained from another population that was not believed to have any breeding individuals related to the pedigreed population. In those cases, individuals with unknown parents should usually be treated as unique founders, the same as are wild-caught individuals, and they would be assigned $k = 1$. In other cases, individuals with unknown parents may be thought to be possibly unrelated to all other founders, and it may be desirable to assign some genetic value to them rather than omitting them from the pedigree calculations. In such cases, k for these individuals can be set to some value between 0 and 1 to indicate the value to be assigned to genes descended from the individual.

Generalization to Probabilistic Pedigrees with Multiple Possible Sires or Dams

Uncertain (probabilistic) parents pose no special challenge for the calculations. They simply require that the proportional contribution of each possible parent (c_s and c_d values) be the probability that the individual was actually a parent multiplied by relative contribution made to the offspring's genome if it were a parent. With respect to the probabilities that are kinships, inbreeding coefficients, and other genetic measures, there is no fundamental difference between a known parent and a possible parent: In either case, for any given allele sampled from a locus, there is a definable probability that the allele was obtained from the parent or possible parent. For example, an allele sampled from a sexually produced diploid progeny has a 0.50 probability of having been obtained from a known sire; it has a 0.25 probability of having come from 1 of 2 equally likely possible sires; and it has a 0.20 probability of having come from a possible sire that has a 40% chance of having been the actual sire. These probabilities are all treated equivalently in the genetic calculations, as they are each the statistical expectation that a sampled allele was contributed by a given possible parent. As a result, calculations of kinships must be summed across all possible dams and all possible sires rather than just a single sire and dam pair.

The parent list does not need to be exhaustive. For example, it may be that there are 3 animals listed as possible sires, each with 20% probability, leaving a 40% chance that some other (unspecified) animal was the sire. This incomplete information about the set of possible parents results in unequal contributions of the set of possible dams versus the set of possible sires, as was already handled in the formulas above for cases of partially unknown ancestry.

The equations above can therefore be further generalized to account for our uncertainty among multiple possible sires and/or dams, as follows. Let p_d be the probability that individual d was the true dam. Let p_s be the probability that individual s was the true sire. Note that in the derivations below, p and k are always used as the product, pk, representing the probability that an allele came from a known part of a possible parent. Lack of a completely known ancestry ($k < 1$) is conceptually the same as a parental set that is incomplete ($\sum p < 1$): in either case a portion of the alleles descend from unknown ancestors in the population. For each possible sire, its expected contribution to the progeny is

$$c_\text{s} = p_\text{s}k_\text{s}/\left(\sum (p_\text{s}k_\text{s}) + \sum (p_\text{d}k_\text{d})\right). \qquad (11)$$

For each possible dam, $c_\text{d} = p_\text{d}k_\text{d}/(\sum(p_\text{s}k_\text{s}) + \sum(p_\text{d}k_\text{d}))$. The proportion known of an individual is

$$k_i = \left(\sum (p_\text{s}k_\text{s}) + \sum (p_\text{d}k_\text{d})\right)/2, \qquad (12)$$

which is a generalized form of Equation 6. Kinship of individual *i* to any other individual *j* is

$$f_{ij} = \sum (c_\text{s}f_{\text{s}j}) + \sum (c_\text{d}f_{\text{d}j}), \qquad (13)$$

the weighted mean of the kinships of *j* to the possible dams and possible sires of *i* and the generalized form of Equation 7. Kinship of individual *i* to itself is

$$f_{ii} = \sum (c_\text{s}^2) + \sum (c_\text{d}^2) + 2\sum c_\text{s}\left(\sum c_\text{d}f_\text{sd}\right), \qquad (14)$$

the generalized form of Equation 8, with the inner summation over all possible dams and the outer summation over all possible sires. Inbreeding coefficient of animal *i*:

$$F_i = \left(\sum \left(p_\text{s}k_\text{s}\sum p_\text{d}k_\text{d}f_\text{sd}\right)\right)/\left(\sum \left(p_\text{s}k_\text{s}\sum p_\text{d}k_\text{d}\right)\right), \qquad (15)$$

so that F_i is the weighted mean of all possible dam-by-sire combinations of kinships.

Oliehoek (2009) presented methods for the case in which the possible sires (or dams) are assumed to be equally likely to have been the true sire (or dam), and his equations are special cases of the more general methods presented here.

Application to Other Types of Mating Systems

Equations 11–15 above provide general forms for calculating kinships from complex pedigrees of biparental sexually reproducing individuals. With minor adjustments for special

Table 1 General equations for relative contributions of parents, proportion of genome known, kinships, and inbreeding coefficients for various systems of mating

	Biparental sexual	Uniparental sexual	Cloning	Haploidy
c_d	$p_\mathrm{d}k_\mathrm{d}/(\sum(p_\mathrm{s}k_\mathrm{s})+\sum(p_\mathrm{d}k_\mathrm{d}))$	$p_\mathrm{d}k_\mathrm{d}/\sum(p_\mathrm{d}k_\mathrm{d})$	$p_\mathrm{d}k_\mathrm{d}/\sum(p_\mathrm{d}k_\mathrm{d})$	$p_\mathrm{d}k_\mathrm{d}/\sum(p_\mathrm{d}k_\mathrm{d})$
k_i	$(\sum(p_\mathrm{s}k_\mathrm{s})+\sum(p_\mathrm{d}k_\mathrm{d}))/2$	$\sum(p_\mathrm{d}k_\mathrm{d})$	$\sum(p_\mathrm{d}k_\mathrm{d})$	$\sum(p_\mathrm{d}k_\mathrm{d})$
f_{ij}	$\sum(c_\mathrm{s}f_{\mathrm{s}j})+\sum(c_\mathrm{d}f_{\mathrm{d}j})$	$\sum(c_\mathrm{d}f_{\mathrm{d}j})$	$\sum(c_\mathrm{d}f_{\mathrm{d}j})$	$\sum(c_\mathrm{d}f_{\mathrm{d}j})$
f_{ii}	$\sum c_\mathrm{s}^2+\sum c_\mathrm{d}^2+2\sum c_\mathrm{s}\sum(c_\mathrm{d}f_\mathrm{sd})$	$0.5+0.5\sum(c_\mathrm{d}f_\mathrm{dd})$	$\sum(c_\mathrm{d}f_\mathrm{dd})$	1
F_i	$(\sum(p_\mathrm{s}k_\mathrm{s}\sum p_\mathrm{d}k_\mathrm{d}f_\mathrm{sd}))/(\sum(p_\mathrm{s}k_\mathrm{s}\sum(p_\mathrm{d}k_\mathrm{d})))$	$\sum(c_\mathrm{d}f_\mathrm{dd})$	$\sum(c_\mathrm{d}F_\mathrm{d})$	1

c_d = proportional contribution of possible dam, d, to the progeny genome. The comparable contribution for each possible sire, s, is c_s. k_i = proportion of a genome that can be traced back to known founders. f_{ij} = kinship of focal animal, i, to each prior animal, j, in the pedigree. f_ii = kinship of animal i to itself. F_i = inbreeding coefficient of i. Summations in the formulas are over all possible dams, d, or all possible sires, s.

forms of inheritance, these same methods—simplified if there is no distinction between "dams" and "sires"—can be applied to the determination of kinships for almost any other form of reproduction. A number of useful cases are described here, and the modified formulas for their kinship calculations are presented in Table 1.

Uniparental, Sexual Reproduction (Selfing, Mating of a Hermaphroditic Animal with Itself, or Parthenogenesis with Fusion of Independent Products of Meiosis)

Cases of uniparental reproduction must be considered carefully because there are several different genetic mechanisms that can be employed. These result in different patterns of inheritance and have very different consequences for inbreeding (Maynard Smith 1978). It may be difficult to determine which mechanism occurred unless analysis of DNA markers is conducted. In the case of selfing of a hermaphroditic individual, 2 genetically independent gametes unite to produce a diploid individual, resulting in homozygosity at half of the loci. This is a special case of sexual reproduction in which the set of possible dams is the same as the set of possible sires. Thus, the separate summations in the equations above for the set of sires and set of dams collapse into a single set of dams. Because whichever individual was the actual dam must be the same as the individual that was the actual sire, kinship to oneself (f_{ii}) is elevated and constrained relative to a case in which the dam and sire are independently selected from the set of possible parents.

Clones

Cloning (or apomictic parthenogenesis) creates a duplicate genome and therefore simply replicates all kinships, with the pairwise kinship of a clone to its parent being the same as the kinship to self, f_{ii}. When it is uncertain which of a set of possible parents produced the clone, then the kinships are weighted means of the values for the possible dams. Note that the case of identical siblings can be handled by first calculating the kinships for one offspring and then calculating the values for its identical siblings by treating them as clones of the first offspring.

Haploidy

The production of individuals from unfertilized gametes has been observed in some snakes and lizards (Dubach et al. 1997; Watts et al. 2006; Booth et al. 2011). Whether the resulting progeny were formed by fusion of identical gametes to create completely homozygous diploid individuals (automictic parthenogenesis with fusion of identical haploid products of meiosis) or production of haploid individuals (as in the males of haplodiploid systems)—the genetic consequences for kinships will be the same. Kinships to other individuals are the same as with the other uniparental forms of reproduction, but the kinship to oneself and inbreeding coefficients are both 1.

Results

To illustrate kinship calculations for cases of unknown and uncertain parentage and various forms of sexual and asexual reproduction, the pedigree of 12 individuals shown in Figure 1 was analyzed with PMx. The pedigree contains individuals produced from partly unknown parentage (#6), uncertain parentage (#10), cloning (#8), selfing (#11), haploidy (#9), and haplodiploidy (#12). Table 2 shows all pairwise kinships for this pedigree, calculated with assignment of unknown parents as founders (below the diagonal) or with omitting contributions from unknown parents (above the diagonal). The relative value of each individual to retention of gene diversity of the captive population is given by its mean kinship to the nonfounders (Equation 9), and these values are given in Table 2. The gene diversity of the

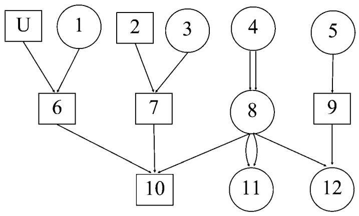

Figure 1. Sample pedigree for testing kinship calculations. "U" indicates an unknown sire. Individuals 1–5 are founders, assumed to be noninbred diploids that are unrelated to each other. Individual 4 was cloned to produce 8. Individual 9 is a haploid progeny of 5. The sire of Individual 10 was either 6 or 7, with an assumption of equal probability. Individual 11 was produced by a hermaphroditic selfing of 8.

Table 2 Kinships for the pedigree in Figure 1

ID	1	2	3	4	5	6	7	8	9	10	11	12	MK
1	0.500	0	0	0	0	0.500	0	0	0	0.071	0	0	0.049
2	0	0.500	0	0	0	0	0.250	0	0	0.071	0	0	0.049
3	0	0	0.500	0	0	0	0.250	0	0	0.071	0	0	0.049
4	0	0	0	0.500	0	0	0	0.500	0	0.286	0.500	0.250	0.235
5	0	0	0	0	0.500	0	0	0	0.500	0	0	0.250	0.118
6	0.250	0	0	0	0	0.5, 1.0	0	0	0	0.143	0	0	0.098
7	0	0.250	0.250	0	0	0	0.500	0	0	0.143	0	0	0.098
8	0	0	0	0.500	0	0	0	0.500	0	0.286	0.500	0.250	0.235
9	0	0	0	0	0.500	0	0	0	1.0	0	0	0.500	0.235
10	0.062	0.062	0.062	0.250	0	0.125	0.125	0.250	0	0.5, 0.510	0.286	0.143	0.216
11	0	0	0	0.500	0	0	0	0.500	0	0.250	0.750	0.250	0.274
12	0	0	0	0.250	0.250	0	0	0.250	0.500	0.125	0.250	0.500	0.255
MK	0.045	0.045	0.045	0.214	0.107	0.089	0.089	0.214	0.214	0.196	0.250	0.232	0.184, 0.209

Below the diagonal are kinships if unknown parents are assumed to be founders; above the diagonal are kinships if unknown ancestries are omitted. Kinships to self of 2 animals with partly unknown ancestries are different if unknown ancestry is omitted (second value) or not (first value). Marginal values are mean kinships to the nonfounders (6–12) and the mean of all 49 such kinships. MK values in the last row are the mean kinships to the 7 nonfounders when unknown parents are assumed to be founders (values on or below the diagonal). MK values in the last column are the mean kinships to the nonfounders when unknown ancestries are omitted (values on or above the diagonal). The MK values in the last column are weighted by the proportion of the genome of each individual that is known ($k_6 = 0.500$ and $k_{10} = 0.875$, whereas $k = 1$ for all other individuals).

nonfounders (#6–12) is calculated as one minus the mean of all their pairwise kinships, weighted by the portion of each genome that derives from known founders (Equation 10). This population gene diversity, as a proportion of the gene diversity of the source population from which the founders were assumed to have been randomly sampled, is $G = 0.816$ if the unknown parent is treated as a founder and $G = 0.791$ if unknown ancestry is excluded.

To confirm the accuracy of the methods, a "gene drop" simulation (MacCluer et al. 1986) of the transmission of founder alleles through the pedigree (provided within PMx) was repeated for 1 000 000 iterations. The gene diversity calculated from resultant allele frequencies in the 7 nonfounder individuals ($G = 1 - \sum p_i^2$, for founder allele frequencies p_i) was confirmed to be 0.816 when alleles from the unknown parent was included and was 0.789 when alleles from the unknown parent were excluded. The small imprecision of the estimated gene diversity (0.789 vs. 0.791) when unknown parents are excluded was expected. A small bias when using the equations in Table 1 occurs because the iterative kinship calculations and resultant gene diversity estimates assume homogeneity of genetic processes across loci. However, for individuals that have partly missing ancestry, some loci may be diploid and others haploid (rather than, e.g., an individual with 1 unknown grandparent somehow being 1.5-ploid at each locus), and there may be different distributions of ancestral alleles among the loci having different ploidy. For example, an individual, *i*, with $k_i = 0.5$ may have had 1 unknown parent (e.g., sire $k_\mathrm{s} = 0$) and 1 fully known parent (e.g., dam $k_\mathrm{d} = 1$) or may have had 2 parents each with $k_\mathrm{s} = k_\mathrm{d} = 0.5$. For offspring of *i*, these 2 possibilities have different consequences for kinships and inbreeding. In the first case, individual *i* is considered haploid at all loci, and 2 offspring that receive from *i* a known allele at a locus will necessarily have received the same allele from the grandmother d. In the second case, individual *i* will be diploid at 25% of its loci, haploid at 50% of its loci, and null-ploid at 25% of its loci, and 2 offspring that receive a known allele at a given locus from *i* will have a 50% chance of receiving alleles derived independently from the paternal grandparents, d and s. This dependency on grandparent *k* is not accounted for in Equations 7 and 8. In small pedigrees in which unknown parents are at most a few generations deep, it would be possible to calculate unbiased kinships from probabilities of shared alleles based on the specific pedigree structure. However, for large pedigrees with many generations, calculating the effects of nonhomogeneity of loci (dependent on *k* in earlier generations) becomes prohibitively complex. The one-generation method of Equations 7 and 8, as given by Ballou and Lacy (1995), can be used to obtain approximate results that will likely be sufficiently accurate except when many individuals have only partly known ancestries.

Discussion

Genetic management of individuals has previously utilized various approaches to deal with uncertainty in parentage: unknown or uncertain ancestors could be assumed to be unrelated to all other animals in the population (Willis 1993), animals coming from unknown sources could be assigned some average estimated kinship (and therefore value) (Willis 2001), portions of genome descended from unknown or uncertain ancestors could be omitted from analysis (Ballou and Lacy 1995), kinship calculations can be averaged across those determined for the possible parents (Oliehoek 2009), or the most likely parent could be assumed to be the true parent. This last option is probably utilized much more often than is recognized, as breeders may record their best guess as to the parentage of an animal, without any indication that parentage is uncertain. When unknown parts of the genome are omitted from the calculations, kinships, and inbreeding that result from shared common ancestors that themselves descend entirely from unknown sources will not be included in the calculations. For example, if a male with unknown parents sires 2 offspring (which must, therefore, be related at least at the level of half siblings), the kinship of those 2 offspring through that sire will not be used in the calculations. Although this may seem to bias kinship calculations downwards, it may be that the offspring are actually related much more closely than half siblings (e.g., the sire may be a sibling of the dam). Errors caused by making unverifiable assumptions about missing parents can be avoided by omitting any kinship occurring through the sire that has $k = 0$ and basing instead the calculation of the kinships entirely on the known portion of the pedigree (Ballou and Lacy 1995).

The assumption of homogeneity of inheritance patterns among loci will lead to a small bias toward underestimating kinships among some descendants when there are missing parents in the pedigrees. However, the bias is usually much less than the substantial underestimation of kinships that can occur if unknown parents are included in calculations and assumed to be unrelated (i.e., new founders). The problem of bias in kinships due to nonhomogeneity of ploidy and allele distributions across loci does not arise when there are multiple possible parents but all possible parents have been specified (including cases of breeding systems that produce actual nondiploid progeny) because the probability distribution of ancestral alleles is then the same at all loci in descendants.

The general methods for analysis and management of pedigrees therefore provide tools for incorporating uncertainty regarding parentage into pedigree analysis. Kinship, inbreeding, and gene diversity are all measures of probabilities of shared alleles (sampled from between or within genomes). Uncertainty in parentage alters these probabilities in definable ways. When a parent is completely unknown, a probability of shared alleles is based on only the known portion of the genome, as the unknown ancestry provides no information regarding identity by descent of sampled alleles. When an individual can be specified to be a possible parent, with a defined probability, then this probability information can be incorporated into kinship and inbreeding calculations. If an individual has a certain probability of being a parent, then the probability of an allele sampled from an offspring being descended from that possible parent is simply the probability of parentage multiplied by the probability of the sampled allele having been derived from that individual's genome if it was the parent. Although it may seem problematic to assume a nonzero kinship to possible parents, which may have contributed no alleles to the progeny, it should be remembered that the sharing of any given allele even with a known parent or other relative is always a probabilistic occurrence.

Although most breeding programs for wildlife are focused on sexually reproducing, diploid terrestrial vertebrates, increasingly conservation breeding programs are being established for fishes and invertebrates (Pearce-Kelly et al. 2007; Penning et al. 2009; Leus et al. 2011), and often these species have different or more flexible breeding systems, including self-compatible hermaphroditism, parthenogenesis of various forms, and cloning. With minor adjustments (often simplifications; see Table 1), the general algorithms for calculating kinships from uncertain and multiple parents can be used also for these other breeding systems.

The extensions of kinship calculations to individuals with partially unknown ancestry functionally treats individuals as receiving differential genetic contributions from the parents and as containing other than diploid complements of genes. Rather than several possible sires or dams, with varying proportions of their ancestries known, there may be multiple actual sires and dams that vary in the amounts that they contributed to the progeny genome. For example, a triploid toad (*Bufo baturae*) receives diploid ova from the dam, with one chromosome set recombining and the other clonal, and haploid sperm from the sire (Stöck et al. 2011). Equations 11–15 are applicable to such cases in which the individuals are not diploid and receive variable proportions of their genes from any number of parents employing any kind of breeding system. The values of p in the equations then become the proportional contributions to the progeny rather than the probability of being a parent. The contributions p_s and p_d can be considered the ploidy (which may be fractional or may be greater than 1) of the gametes or propagules that each sire and dam contributed.

Similar methods can also be used to estimate kinship relationships among groups in which the individuals are not separately identified. Kinships among groups of individuals are not conceptually different than kinships among individuals that received varying contributions from any number of possible parents. Therefore, Equations 11–15 can be applied with the parents and their offspring being groups (parental source and derived offspring groups), although if more than 2 groups are combined to create a new group, then the equations need to be extended to account for more parents than just biparental sires and dams. The kinship between groups i and j are still defined as the probability that 2 alleles sampled from the 2 gene pools are identical by descent, and the inbreeding of a group genome or gene pool can still be defined as the probability that 2 sampled alleles (that are not resampling the same physical allele) are identical by descent from a common ancestor. It should be noted that this definition of the inbreeding level of the group is not the same as the mean inbreeding coefficient of the individuals that comprise a group because that mean individual inbreeding ignores the probabilities of identity by descent between individuals within a group. The group inbreeding coefficient as defined above treats the group as a singular, homogeneous entity of any possible ploidy rather than as a collection of individuals.

Although the kinships among groups (and kinships between groups and individuals) can be calculated with the methods presented above, calculating the inbreeding coefficients of groups is more complicated because, unlike individuals, groups can change their genetic composition over time. Group size and composition change when individual members of the group die or are removed and when new members are added via reproduction within the group. Neither of these processes change the kinships to other groups or individuals because the probability that an allele will be sampled from the group genome is unchanged as the group passes through internal generations or has members removed at random (i.e., there is no change in the expectations of the allele frequencies).

PMx provides these kinship calculations among groups with defined ancestries (and between groups and individuals within the same pedigree). However, PMx makes the simplifying assumption that the genetic composition of each group is stable, and it makes no adjustment for the loss of genetic diversity (increased inbreeding) that will occur as a group undergoes internal generational turnover. As a first estimate, this is probably adequate for management for those cases in which groups are relatively stable over time, either because of low turnover or because record-keeping assigns new group identifiers frequently. If a new group is defined to be created by sampling from the parent group each time that all member individuals are replaced by their progeny, then genetic change occurs only during group formation and the calculations in PMx will be appropriate. The CERCI software (Burlingham-Johnson et al. 1994) produced by the Zoological Society of London provides precise calculations of genetic change within groups for a specific population management system in which generations are discrete (Wang 2004).

Traditional pedigree analysis and management is dependent on accurate and complete recording of the parents of each individual. The extension of pedigree analysis to accommodate uncertain parentage and alternative breeding systems in some ways relaxes the minimal data needed, but in other ways requires further information. Generalized pedigree analysis provides the opportunity to use additional information that cannot be utilized in simple pedigrees of diploid individuals. Missing information about some parents does not interfere with analyses, although the reduced knowledge will diminish the precision of genetic management. Knowledge about possible parents can be fully utilized, if the probabilistic contributions can be specified. To permit use of such data in pedigree analyses, breeding programs will need to record whatever is known about possible parents (including probabilities of parentage or other information from which probabilities might be inferred). The most recent version of the SPARKS studbook software (ISIS 2011) has the capability to record these data and to export them to PMx for analysis.

The methods presented here were developed so as to allow use of whatever information is available on the ancestry of individuals in a pedigreed population. However, it should be emphasized that incomplete or inaccurate information does reduce the ability to retain genetic diversity or to achieve other goals of the breeding program (Lacy 2009; Oliehoek 2009). To allow optimal pedigree analysis and management, complete pedigree information should be obtained whenever that is feasible.

Conclusions

When Nei (1973) proposed the term “gene diversity,” he noted that the concept was applicable to any organism, whether diploid, haploid, or polyploid. (This is one reason why the concept of gene diversity is far more general than the term “expected heterozygosity,” as heterozygosity is a concept that is defined in terms of diploid organisms.) The techniques for pedigree analysis that focus on gene diversity are equally applicable to any organism. The methods described here show that the concepts of kinship, inbreeding, and gene diversity can be generalized to any entity containing genes and applied to the genetic management of a population of such genomes. Thus, genetic management based on kinship approaches can be applied not only to diploid individuals with fully known ancestries (at least back to the defined founders) but also to animals with partially unknown ancestries, to animals with multiple possible parents, to haploid or polyploid individuals, to gene banks, to social groups, to breeding colonies, or to any managed population. Propagation can involve equal sexual exchange, unequal sexual exchange, fission, cloning, or any other form of inheritance for which the probabilities of allele transmission can be defined. Traditional methods of pedigree analysis are special cases of the general methods presented here.

Use of these techniques allows identification of the optimal individuals for propagation—those with the lowest mean kinships (Ballou and Lacy 1995; Ivy and Lacy 2012), as estimated from whatever knowledge is available about ancestries. The selection of optimal genomes for reproduction results in maximal retention of gene diversity, and therefore also of effective population size (Caballero and Toro 2000) and founder genome equivalents (Lacy 1989, 1995), both of which can be defined in terms of gene diversity. Tracking complete individual-based pedigrees is not possible for some species, so the less precise management afforded by general techniques is the best that can be achieved. For many species, however, individual pedigrees could be kept, but only at considerable cost of time and other resources, such as when molecular genetic analyses are used to determine parentage (Ivy et al. 2009; Ivy and Lacy 2010). It will be important to determine how much more rapidly gene diversity is lost and how other genetic goals are compromised when genetic management of breeding programs is based on kinships estimated from partial rather than complete pedigree information.

Funding

Institute of Museum and Library Services (CP Grant #1C-03-05-0027-05); Chicago Zoological Society.

Acknowledgments

Jamie Ivy, Jinliang Wang, and Jonathan Ballou provided helpful comments on the manuscript. They and many other colleagues also provided valuable insights toward the development of methods for the management of captive populations, and it would be difficult to ascribe the methods described above to any one or a few people.

References

Ballou JD, Lacy RC. 1995. Identifying genetically important individuals for management of genetic diversity in pedigreed populations. In: Ballou JD, Gilpin M, Foose TJ, editors. Population management for survival & recovery. Analytical methods and strategies in small population conservation. New York: Columbia University Press. p. 76–111.

Ballou JD, Lacy RC, Pollak JP. 2011. PMx: software for demographic and genetic analysis and management of pedigreed populations. Version 1.0. Brookfield (IL): Chicago Zoological Society.

Ballou JD, Lees C, Faust LJ, Long S, Lynch C, Bingaman Lackey L, Foose TJ. 2010. Demographic and genetic management of captive populations. In: Kleiman DG, Thompson KV, Baer CK, editors. Wild mammals in captivity: principles and techniques for zoo management. 2nd ed. Chicago (IL): University of Chicago Press. p. 219–252.

Booth W, Johnson DH, Moore S, Schal C, Vargo EL. 2011. Evidence for viable, non-clonal but fatherless Boa constrictors. Biol Lett. 7:253–256.

Burlingham-Johnson A, Clarke D, Pearce-Kelly P. 1994. CERCI: a computer system for the demographic and genetic analysis of captive invertebrates, fish and other populations of colony animals. Int Zoo Yearb. 33:278–283.

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

Dubach J, Sajewicz A, Pawley R. 1997. Parthenogenesis in the Arafuran file snake (*Acrochordus arafurae*). Herpetol Nat Hist. 5:11–18.

Falconer DS, Mackay TFC. 1996. Introduction to quantitative genetics. 4th ed. Harlow (UK): Longman.

Fernández J, Toro MA, Caballero A. 2001. Practical implementations of optimal management strategies in conservation programmes: a mate selection method. Anim Biodivers Conserv. 24:17–24.

Fernández J, Toro MA, Caballero A. 2003. Fixed contributions designs versus minimization of global coancestry to control inbreeding in small populations. Genetics. 165:885–894.

Fernández J, Toro MA, Caballero A. 2004. Managing individuals' contributions to maximize the allele diversity maintained in small, conserved populations. Conserv Biol. 18:1–10.

Fernández J, Villanueva B, Pong-Wong R, Toro MA. 2005. Efficiency of the use of pedigree and molecular marker information in conservation programs. Genetics. 170:1313–1321.

Flesness NR. 1977. Gene pool conservation and computer analysis. Int Zoo Yearb. 17:77–81.

Foose TJ, Ballou JD. 1988. Population management: theory and practice. Int Zoo Yearb. 27:26–41.

ISIS. 2011. SPARKS (Single Population Analysis and Records Keeping System). Version 1.6. Eagan (MN): International Species Information System.

Ivy JA, Lacy RC. 2010. Using molecular methods to improve the genetic management of captive breeding programs for threatened species. In: DeWoody JA, Bickham JW, Michler CH, Nicols KM, Rhodes OE, Woeste KE, editors. Molecular approaches in natural resource conservation and management. Cambridge (UK): Cambridge University Press. p. 267–295.

Ivy JA, Lacy RC. 2012. A comparison of strategies for selecting breeding pairs to preserve maximal genetic diversity in breeding programs. J Hered.

Ivy JA, Miller A, Lacy RC, DeWoody JA. 2009. Methods and prospects for using molecular data in captive breeding programs: an empirical example using parma wallabies (*Macropus parma*). J Hered. 100:441–454.

Lacy RC. 1989. Analysis of founder representation in pedigrees: founder equivalents and founder genome equivalents. Zoo Biol. 8:111–124.

Lacy RC. 1994. Managing genetic diversity in captive populations of animals. In: Bowles ML, Whelan CJ, editors. Restoration and recovery of endangered plants and animals. Cambridge (UK): Cambridge University Press. p. 63–89.

Lacy RC. 1995. Clarification of genetic terms and their use in the management of captive populations. Zoo Biol. 14:565–577.

Lacy RC. 2009. Stopping evolution: genetic management of captive populations. In: Amato G, DeSalle R, Ryder OA, Rosenbaum, HC, editors. Conservation genetics in the age of genomics. New York: Columbia University Press. p. 58–81.

Lacy RC, Ballou JD, Pollak JP. 2011. PMx: software package for demographic and genetic analysis and management of pedigreed populations. Methods Ecol Evol. Advance Access published September 5, 2011, doi:10.1111/j.2041-210X.2011.00148.x

Lacy RC, Ballou JD, Princée F, Starfield A, Thompson E. 1995. Pedigree analysis. In: Ballou JD, Gilpin M, Foose TJ, editors. Population management for survival & recovery. Analytical methods and strategies in small population conservation. New York: Columbia University Press. p. 57–75.

Leus K, Traylor-Holzer K, Lacy RC. 2011. Genetic and demographic population management in zoos and aquariums: recent developments, future challenges and opportunities for scientific research. Int Zoo Yearb. 45:213–225.

MacCluer JW, VandeBerg JL, Read B, Ryder OA. 1986. Pedigree analysis by computer simulation. Zoo Biol. 5:147–160.

Maynard Smith J. 1978. The evolution of sex. Cambridge (UK): Cambridge University Press.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A. 70:3321–3323.

Oliehoek P. 2009. Genetic conservation of endangered animal populations [PhD thesis]. [Wageningen (Netherlands)]: Wageningen University.

Pearce-Kelly P, Morgan R, Honan P, Barrett P, Perrotti L, Magdich M, Daniel BA, Sullivan E, Veltman K, Clarke D, et al. 2007. The conservation value of insect breeding programmes: rationale, evaluation tools and example programme case studies. In: Stewart AJA, New TR, Lewis OT, editors. Insect conservation biology. Oxfordshire (UK): CABI. p. 57–73.

Penning M, McGregor Reid G, Koldewey H, Dick G, Andrews B, Arai K, Garratt P, Gendron S, Lange J, Tanner K, et al., editors. 2009. Turning the tide: a global aquarium strategy for conservation and sustainability. Bern (Switzerland): World Association of Zoos and Aquariums.

Rudnick JA, Lacy RC. 2008. The impact of assumptions about founder relationships on the effectiveness of captive breeding strategies. Conserv Genet. 9:1439–1450.

Stöck M, Ustinova J, Betto-Colliard C, Schartl M, Moritz C, Perrin N. 2011. Simultaneous Mendelian and clonal genome transmission in a sexually reproducing, all-triploid vertebrate. Proc R Soc B Biol Sci. doi: 10.1098/rspb.2011.1738

Wang J. 2004. Monitoring and managing genetic variation in group breeding populations without individual pedigrees. Conserv Genet. 5:813–825.

Watts PC, Buley KR, Sanderson S, Boardman W, Ciofi C, Gibson R. 2006. Parthenogenesis in Komodo dragons. Nature. 444:1021–1022.

Willis K. 1993. Use of animals with unknown ancestries in scientifically managed breeding programs. Zoo Biol. 12:161–172.

Willis K. 2001. Unpedigreed populations and worst-case scenarios. Zoo Biol. 20:305–314.

Received August 2, 2011; Revised October 29, 2011; Accepted November 3, 2011

Corresponding Editor: Robin Waples