

The genetic impact of demographic decline and reintroduction in the wild boar (*Sus scrofa*): A microsatellite analysis

C. VERNESI*†, B. CRESTANELLO*‡, E. PECCHIOLI*†‡, D. TARTARI*, D. CARAMELLI†§, H. HAUFFE‡ and G. BERTORELLE*

*Department of Biology, Università di Ferrara, Via Borsari 46, 44100 Ferrara, Italy; †Biosfera – Conservation Biology, Research, and Education, Florence, Italy; ‡Centre of Alpine Ecology, Viote del Monte Bondone (TN), Italy; §Department of Animal Biology and Genetics, University of Florence, Italy

Abstract

The reintroduction of wild boar from central Europe after World War II has contributed substantially to the range expansion of this species in Italy, where indiscriminate hunting in earlier times resulted in extreme demographic reduction. However, the genetic impact of such processes is not well-understood. In this study, 105 individuals from Italian and Hungarian wild boar populations were characterized for nine autosomal microsatellite loci. The Hungarian samples, and two central Italian samples from protected areas (parks) where reintroduction is not documented, were assumed to be representative of the genetic composition of the source and the target populations in the reintroduction process, respectively. Animals hunted in the wild in the Florence area of Tuscany (Italy) were then studied to identify the effects of reintroduction. The results we obtained can be summarized as follows: (i) none of the populations analysed shows genetic evidence of demographic decline; (ii) the three parental populations from Italy and Hungary are genetically distinct; however, the low level of divergence appears in conflict with the naming of the Italian and the European subspecies (*Sus scrofa majori* and *Sus scrofa scrofa*, respectively); in addition, the Italian groups appear to be as divergent from each other as they are from the Hungarian population; (iii) most of the individuals hunted near Florence are genetically intermediate between the parental groups, suggesting that hybridization has occurred in this area, the average introgression of Hungarian genotypes is 13%, but $\approx 45\%$ of the genetic pool of these individuals can not be directly attributed to any of the parental populations we analysed; (iv) analysis of microsatellite loci, though in a limited number, is an important tool for estimating the genetic effect of reintroduction in the wild boar, and therefore for the development of conservation and management strategies for this species.

Keywords: genetic variation, hybridization, microsatellites, reintroduction, wild boar

Received 5 July 2002; revision received 20 November 2002; accepted 20 November 2002

Introduction

Until recently, reductions in the population size and geographical distribution of game species in Europe were simply counterbalanced by uncontrolled reintroductions by local authorities. Several species of mammals and birds were restocked using animals bred in captivity or collected

from distant regions (Caughley & Gunn 1996; Rhymer & Simberloff 1996), with no particular attention paid to the possible consequences for the autochthonous groups (hybridization or even extinction) and the ecological structure in general. The reintroduction of wild boar into Italy is a typical case in point.

Widespread in most of the Italian peninsula in historical times, wild boar populations underwent extreme reduction during the 18th and 19th centuries, due mainly to human persecution (Apollonio *et al.* 1988; Massei & Genov

Correspondence: Giorgio Bertorelle. Tel. +39 532 291743; Fax: +39 532 249761; E-mail: ggb@unife.it

2000). At the beginning of the 20th century, the geographical distribution of this species was restricted to Sardinia and some central–southern regions (Ghigi 1911). In the 1920s, wild boar from France began to naturally re-colonize northwestern Italy, and three subspecies were recognized at that time (De Beaux & Festa 1927): *Sus scrofa meridionalis* in Sardinia (the Sardinian wild boar), *S. s. majori* in central–southern Italy (the Maremma wild boar) and *S. s. scrofa* in most of Europe including northern Italy (the European wild boar). More recently, the Maremma and European wild boar have been considered ecotypes rather than subspecies, because of their low level of mitochondrial DNA (mtDNA) divergence (Randi 1995).

After World War II, a considerable expansion occurred in this species. Changes in agricultural practices, placement of artificial feeding sites, a decrease in predator numbers and global warming appear to have favoured the increase in wild boar populations (Massei & Toso 1993; Massei & Genov 2000), but extensive reintroduction of animals has probably played a critical role. Primarily for hunting, large numbers of wild boar from central Europe (Hungary, Poland and the former Czechoslovakia), and individuals hybridized in captivity with domestic pigs, were released to restock local populations (Apollonio *et al.* 1988). These translocations and the related demographic expansion have produced two main problems that urgently need to be addressed.

The first is related to the cohabitation and possible hybridization between individuals with different geographical origins, or between wild and domesticated forms. In general, the genetic integrity, and therefore, the very existence of the native groups is threatened by hybridization (Rhymer & Simberloff 1996). In this case, an additional risk is due to the fact that the endemic Maremma wild boar is smaller and less prolific than the European wild boar reintroduced into its range (Massei & Toso 1993). Phenotypic hybrids between native and introduced forms have been reported in many regions (Massei & Toso 1993; Pedrotti *et al.* 2001), and only two populations located in protected areas, the Maremma Regional Park in Tuscany and the Castelporziano Reserve near Rome, are regarded as being 'pure' Maremma wild boar (Apollonio *et al.* 1988).

The second problem associated with the reintroduction of wild boar in Italy is that the density of this species has reached such high levels that frequent agricultural damage is reported (e.g. Macchi *et al.* 1995) and disturbance to native fauna (especially deer and, through egg predation, birds) and flora is inevitable (Henry & Conley 1972; Singer *et al.* 1984).

Therefore, specific conservation and management programmes are needed to reduce hybridization and possible extinction of local populations and to minimize damage to crops and natural ecosystems. As commonly recognized for different species (e.g. Polziehn *et al.* 2000; Maudet *et al.*

2002), genetic information regarding the populations and the individuals involved in the reintroduction is of primary importance for developing and implementing such programmes.

Here, we present results of a study on the genetic variation and divergence of different groups of individuals from populations not affected by reintroduction plans. We also consider the process of hybridization by studying the genetic composition of individuals hunted in the wild. This analysis, based on nine highly variable nuclear microsatellite markers, is the first attempt to quantify the genetic impact of the demographic decline and translocation events in this species.

Materials and Methods

Samples

Hair, skeletal muscle or auricle samples were collected from 105 hunted or trapped wild boar in Italy and Hungary. Italian samples representative of the native population of the Maremma wild boar were collected in two protected areas where reintroduction can be reasonably excluded (Apollonio *et al.* 1988), the Maremma Regional Park (MRP, $n = 11$) and the Castelporziano Presidential Reserve (CPR, $n = 26$). Hungarian samples of the European wild boar were chosen because they are closely related to individuals reintroduced to central Italy in the past. These samples were collected in three different Hungarian regions (Budapest, Gemenc and Kab-Hegy), but we pooled them in a single group (HUN, $n = 29$) as they did not show any significant genetic structure (data not shown). Finally, a sample of 39 individuals hunted in the wild, mostly in the Florence area of Tuscany (FLO, $n = 39$) were analysed as representative of the present-day wild boar population of central Italy. Sample locations are reported in Fig. 1.

DNA extraction, amplification and typing

DNA was extracted from samples using a proteinase K digestion procedure followed by standard phenol/chloroform extraction (Sambrook *et al.* 1989) and concentration using Microcon-30 microconcentrators (Amicon). Nine polymorphic dinucleotide microsatellites were selected from a panel designed for the domestic pig (Alexander *et al.* 1996) to allow multiplex loading, and the annealing temperatures of the primers were adapted for this species (see Table 1). Amplification reactions were performed in a total volume of 20 μ L, with 0.5 μ M of each primer (labelled with HET, TET and FAM fluorescent dyes), 0.2 mM dNTP, 1 \times Taq buffer (Tris–HCl, pH 8.8; 166 mM (NH₄)₂SO₄; Tween 20® 0.1%), 1 U of Taq DNA polymerase (Polymed) 1.5 mM MgCl₂ and 2 μ L template DNA. The thermocycling regime consisted of 94 °C for 5 min, followed by 30–35 cycles of 94 °C for 30 s, T_a for 30 s, and 72 °C for 30 s, with

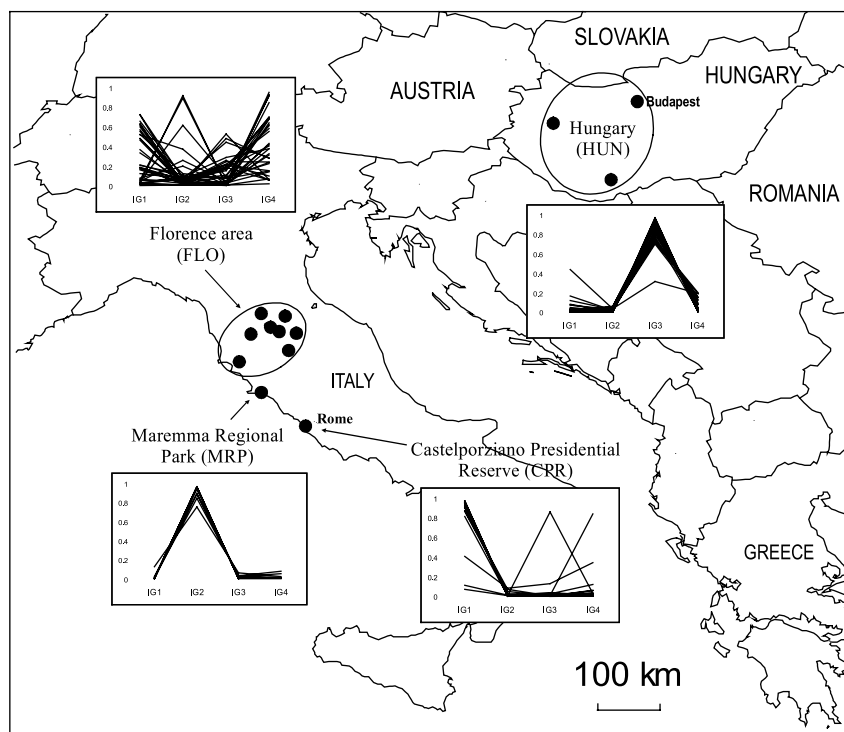


Fig. 1 Location and abbreviation (in parenthesis) of the four groups of wild boar analysed. The three Hungarian localities are not genetically differentiated and are pooled in the HUN sample. Possibly introduced or hybrid individuals hunted in the Florence area of Tuscany (Italy) are grouped in the FLO sample. Insets: the estimated genetic contribution of the inferred groups (from IG1 to IG4) into the individual genotypes; each line represents a single individual, and individuals sampled in the same original population are grouped in the same plot.

Table 1 Summary statistics for the loci and populations analysed. T_a = annealing temperatures in °C; k_T = total number of alleles for each locus; n = sample size (number of individuals); k = number of alleles in each population for each locus (in parenthesis the allelic richness); V = variance of the number of repeats; H_O = observed heterozygosity. Asterisks indicate significant departure from Hardy–Weinberg equilibrium. The last line reports the total number of alleles and the average heterozygosities and variances across loci

| Locus | T_a | Size range | k_T | CPR | | | | MRP | | | | HUN | | | | FLO | | | | |
|--------|-------|------------|-------|-----|-----------|------|-------|------|---------|------|-------|------|----------|------|-------|------|-----------|------|-------|------|
| | | | | n | k | V | H_O | n | k | V | H_O | n | k | V | H_O | n | k | V | H_O | |
| SW55 | 58 | 122–152 | 11 | 26 | 8 (6.1) | 10.9 | 0.65 | 11 | 3 (3.0) | 0.1 | 0.18 | 28 | 9 (7.4) | 23.3 | 0.68 | 38 | 10 (7.6) | 28.2 | 0.84 | |
| SW461 | 62 | 118–150 | 12 | 26 | 6 (5.0) | 1.2 | 0.77 | 11 | 8 (8.0) | 6.8 | 0.82 | 28 | 10 (7.9) | 8.3 | 0.89 | 39 | 11 (7.2) | 6.5 | 0.61* | |
| SW841 | 62 | 158–184 | 13 | 26 | 6 (4.9) | 1.5 | 0.62 | 11 | 6 (6.0) | 12.4 | 0.36* | 29 | 8 (6.4) | 7.4 | 0.76 | 39 | 13 (7.3) | 9.2 | 0.61* | |
| SW1465 | 57 | 65–101 | 10 | 26 | 4 (3.2) | 0.5 | 0.46 | 11 | 3 (3.0) | 0.5 | 0.54* | 28 | 9 (6.3) | 11.8 | 0.82 | 39 | 7 (5.3) | 1.7 | 0.51* | |
| SW1492 | 52 | 116–138 | 9 | 26 | 6 (4.6) | 1.4 | 0.69 | 11 | 6 (6.0) | 2.5 | 0.54* | 28 | 6 (4.3) | 0.7 | 0.50* | 39 | 8 (5.8) | 3.3 | 0.74* | |
| SW1514 | 55 | 138–162 | 11 | 26 | 7 (6.5) | 3.8 | 0.82 | 11 | 5 (5.0) | 1.5 | 0.64 | 29 | 8 (6.3) | 3.5 | 0.65 | 38 | 10 (8.2) | 7.3 | 0.79 | |
| SW2021 | 65 | 102–132 | 15 | 26 | 6 (4.8) | 3.2 | 0.54* | 11 | 6 (5.9) | 4.4 | 0.91 | 29 | 11 (7.7) | 11.8 | 0.90 | 39 | 12 (8.5) | 7.1 | 0.77* | |
| SW2496 | 52 | 184–228 | 19 | 26 | 14 (10.3) | 21.6 | 0.88 | 11 | 8 (8.0) | 12.3 | 0.64 | 29 | 12 (8.9) | 11.1 | 0.86 | 39 | 17 (10.1) | 17.4 | 0.85* | |
| SW2532 | 58 | 174–192 | 9 | 26 | 7 (5.0) | 8.5 | 0.54* | 11 | 1 (1.0) | 0.0 | 0.00 | 29 | 6 (5.4) | 6.9 | 0.72 | 39 | 6 (5.1) | 12.5 | 0.72 | |
| | | | | 109 | 64 | | 5.9 | 0.66 | 46 | | 4.5 | 0.52 | 79 | | 9.4 | 0.75 | 94 | | 10.3 | 0.72 |

a final extension of 72 °C for 7 min. Electrophoresis and analysis was performed using an ABI 377 automatic DNA sequencer and the GENESCAN Version 2.0.2. software (Perkin–Elmer). Sizing of polymerase chain reaction (PCR) products was determined using an internal size standard.

Data analysis

Simple statistics were computed to estimate genetic variability within each of the four populations considered

(MRP, CPR, HUN and FLO). The computation of the allelic richness, which corrects the observed number of alleles for the different sample sizes, was based on the rarefaction method, as described in Comps *et al.* (2001). Deviation from Hardy–Weinberg equilibrium (HWE) was tested separately for each locus and each population using an analogue of the Fisher's exact test which was implemented using a Markov chain (Guo & Thompson 1992).

Two different methods were used to detect population bottlenecks. One (Cornuet & Luikart 1996) is based on the

fact that the number of alleles is reduced faster than the expected heterozygosity (H_E) after a bottleneck (Maruyama & Fuerst 1985). In this situation, H_E (i.e. the gene diversity *sensu* Nei 1987) should be higher than the equilibrium heterozygosity (H_{EQ}) predicted in a stable population from the observed number of alleles. As suggested by the authors, for the samples where fewer than 20 loci are typed, the results obtained separately for each locus were combined using the Wilcoxon test (Cornuet & Luikard 1996; Piry *et al.* 1999). The second method (Garza & Williamson 2001) is based on the ratio M of the number of alleles to the range in allele size. The average of this value across loci can be used to detect bottlenecks when microsatellite data are available, because a reduction of the population size is expected to reduce the number of alleles faster than the range in allele size. Both methods test the statistical significance of the results only after specifying the parameters of the mutation model. Again, following the authors' suggestions (Piry *et al.* 1999; Garza & Williamson 2001), we assumed 90 or 95% of single-step mutations (mutations that increase or decrease the size of the allele by one repeat) and 10 or 5% of multiple-step mutations (mutations that increase or decrease the size of the allele by more than one repeat) with an exponential distribution of step numbers having an average of 3.5 steps. Unfortunately, testing the statistical significance of the average M also implies specification of the unknown parameter $\theta = 4N\mu$ of the population simulated to compute the critical value of M under the null hypothesis of no bottleneck (N and μ are the effective population size and the mutation rate, respectively). We tested two values of θ reasonable for microsatellites markers, i.e. $\theta = 5$ and $\theta = 10$ (Bertorelle & Excoffier 1998; Garza & Williamson 2001).

Genetic divergence between populations was analysed using the classical F_{ST} approach (Wright 1965; Weir & Cockerham 1984) and its analogue for microsatellites R_{ST} (Slatkin 1995). Whereas F_{ST} takes into account only the differences in allele frequencies observed in different populations, R_{ST} also includes the molecular information associated with different allele sizes.

Factorial correspondence analysis, FCA, was then applied to graphically represent the genetic distances between individual multilocus genotypes. This multivariate analysis is based on a χ^2 value computed by comparing the observed distribution of alleles in the genotypes with that expected if alleles were randomly assigned to individuals. New variables (factors) which explain decreasing values of the computed χ^2 are then elaborated and assigned to genotypes, and usually only the first two variables are plotted. This and other multivariate techniques such as principal component analysis (PCA) or multidimensional scaling (MDS) are commonly used for the same purposes (see, e.g. Kadwell *et al.* 2001; Randi *et al.* 2001) with similar results.

The hybridization process was quantitatively analysed following two different approaches. Assuming, *a priori*, that the FLO population is a hybrid among CPR, MRP and HUN, we first estimated the relative proportions of the parental groups in the hybrid. This estimation, based on the method developed by Bertorelle & Excoffier (1998) and extended by Dupanloup & Bertorelle (2001), can also incorporate the information related to different allele sizes assuming the single step mutation process. Second, we applied the Bayesian clustering method proposed by Pritchard *et al.* (2000) to identify the population structure most compatible with the data, and to estimate the relative contribution of each *inferred* population into the genotypes of each individual and into the original groups. This method is implemented by a Markov chain Monte Carlo algorithm that allows the estimation of the posterior distribution of the allele frequencies in the inferred populations. It does not assume any prior knowledge of the population structure and, under the assumption that individuals' genotypes come from unknown groups (the inferred populations) at HWE with all loci in linkage equilibrium, it reconstructs the number and the genetic composition of these groups. The genetic contribution of each inferred group to the predefined populations (CPR, MRP, HUN and FLO in our case) and also to the single individuals can be estimated. The Pritchard *et al.* (2000) approach is a powerful tool for analysing admixture when populations not included in the samples might have contributed to the genetic composition of the admixed individuals. The results presented here are based on a Markov chain with 10^6 iterations, following a burn-in period of 40 000 iterations.

Finally, we introduce an index of true hybridization (the true hybridization index; THI). When the genetic composition of a population can be assigned in different proportions to different parental (or inferred) groups, two possible situations can be envisaged: most individuals are hybrids or most individuals are similar to either of the parental groups. The first situation is a true hybridization, and the individuals in the hybrid swarm are hybrids by varying number of generations of backcrossing with parental types or mating among hybrids (Allendorf *et al.* 2001). Conversely, the second situation is typical of cryptic population structure, in which hybridization is very recent, it does not produce fertile individuals, or it does not occur. THI (see Appendix) ranges from 0 to 1, with larger values indicating a true hybridization process in the population.

The computer software we used to analyse the genetic data were MICROSAT Version 1.5d (E. Minch, Stanford University, 1999), ARLEQUIN 2001 (Schneider *et al.* 2000), GENETIX Version 4.02 (K. Belkhir, University of Montpellier, 2001), BOTTLENECK Version 1.2.02 (Cornuet & Luikard 1996), CRITICAL_M (C. Garza, <http://www.pfeg.noaa.gov/tib/carlos.htm>), ADMIX Version 2.0 (I. Dupanloup, [© 2003 Blackwell Publishing Ltd, *Molecular Ecology*, 12, 585–595](http://</p>
</div>
<div data-bbox=)

Table 2 The results of two tests for the identification of population bottlenecks

| Locus | CPR | | $M\ddagger$ | MRP | | M | HUN | | M | FLO | | M |
|--------|----------------------------------|---------|-------------|---------------------------|---------|------|---------------------------|--------|------|---------------------------|---------|------|
| | $H_E - H_{EQ}\dagger$ 10% m-s | 5% m-s | | $H_E - H_{EQ}$ 10% m-s | 5% m-s | | $H_E - H_{EQ}$ 10% m-s | 5% m-s | | $H_E - H_{EQ}$ 10% m-s | 5% m-s | |
| SW55 | -0.062 | -0.066 | 0.50 | -0.267* | -0.272* | 1.00 | -0.024 | -0.034 | 0.60 | -0.029 | -0.037 | 0.63 |
| SW461 | -0.013 | -0.020 | 1.00 | -0.013 | -0.015 | 0.89 | 0.002 | -0.003 | 0.63 | -0.023 | -0.024 | 0.69 |
| SW841 | -0.061 | -0.073 | 1.00 | 0.034 | 0.028 | 0.67 | 0.0 | -0.007 | 0.89 | -0.094* | -0.100* | 1.00 |
| SW1465 | -0.107 | -0.112 | 1.00 | 0.085 | 0.074 | 1.00 | -0.050 | -0.056 | 0.47 | -0.013 | -0.024 | 0.88 |
| SW1492 | -0.060 | -0.071 | 1.00 | -0.026 | -0.028 | 1.00 | -0.110 | -0.122 | 1.00 | -0.001 | -0.009 | 0.67 |
| SW1514 | 0.065 | 0.053 | 1.00 | -0.050 | -0.059 | 0.71 | -0.028 | -0.039 | 0.73 | 0.030 | 0.020 | 0.77 |
| SW2021 | -0.130 | -0.139* | 0.86 | -0.063 | -0.067 | 0.86 | -0.026 | -0.031 | 0.73 | -0.013 | -0.018 | 0.92 |
| SW2496 | -0.019 | -0.023 | 0.82 | -0.025 | -0.029 | 0.67 | -0.005 | -0.008 | 0.57 | -0.015 | -0.018 | 1.00 |
| SW2532 | -0.082 | -0.090 | 0.70 | — | — | — | 0.053 | 0.042 | 0.67 | 0.051 | 0.039 | 0.60 |
| | -(+)§ | -(+) | 0.88 | -(-) | -(+) | 0.85 | -(+) | -(+) | 0.70 | -(-) | -(+) | 0.79 |

$\dagger H_E$ = expected heterozygosity (= gene diversity, Nei 1987); H_{EQ} = heterozygosity expected in a stationary population with the observed number of alleles and assuming the mutation model described in the text with either 10 or 5% of multiple-step (m-s) mutations. Significant difference ($P < 0.05$) between H_E and H_{EQ} are indicated by an asterisk. Bottlenecked populations are expected to show a heterozygosity excess, i.e. ($H_E - H_{EQ}$) > 0 (Cornuet & Luikart 1996).

\ddagger The index M computed for each locus and averaged across all loci (last line). Small M -values indicate that some intermediate allelic states are unoccupied, possibly as a consequence of population bottlenecks (Garza & Williamson 2001); average M is significantly different from the value expected in a stable population ($P < 0.05$) only for the HUN sample and only when 5% of multiple-step mutations were assumed. §+ and - indicate that the one tail probability of a heterozygosity excess tested across loci with the Wilcoxon technique (Piry *et al.* 1999) is significant ($P < 0.05$) or not, respectively; the results obtained for the test for a heterozygosity deficiency are in parenthesis.

www.unife.it/genetica/Isabelle/admix2_0.html), and STRUCTURE (J. Pritchard, <http://pritch.bsd.uchicago.edu/>).

Results

A total of 109 different alleles were identified for the 9 microsatellites analysed. With the exception of SW2532 in the Maremma Regional Park (MRP), all loci are polymorphic, and in most cases highly so, in each of the four samples (see Table 1). Variation, measured using the heterozygosity, the allelic richness and the variance of the repeat size, appears slightly higher in the HUN and FLO samples compared with the samples from the restricted areas of CPR and MRP. Deviation from Hardy-Weinberg expectation was systematically detected only in the FLO sample, with 7 of 9 loci showing a significant deficit of heterozygotes. The HUN sample did not show any evidence of heterozygote deficit, again ruling out internal population structure for this population obtained by pooling the samples from three localities.

Two different tests failed to identify the genetic signature of a demographic decline in all the four groups of individuals (Table 2). The relative excess of H_E predicted in bottlenecked populations (Cornuet & Luikard 1996) was not observed either in the single locus analysis or in the Wilcoxon test that combines the results for the different loci. By contrast, a general deficiency of heterozygosity was noted for most loci in all populations. Similar results are provided by the index M . With the exception of only the

Table 3 Pairwise genetic distances between the four groups considered. F_{ST} and R_{ST} values are reported below and above the diagonal, respectively. All values differ significantly from 0 ($P < 0.01$)

| | CPR | MRP | HUN | FLO |
|-----|------|------|------|------|
| CPR | — | 0.42 | 0.07 | 0.09 |
| MRP | 0.17 | — | 0.27 | 0.20 |
| HUN | 0.11 | 0.16 | — | 0.02 |
| FLO | 0.05 | 0.09 | 0.04 | — |

HUN sample when 5% of multiple-step mutations are assumed, the average M -values fell always in the 95% acceptance region of M obtained simulating a stable populations with $\theta = 5$ or $\theta = 10$, and 5 or 10% of multiple step mutations.

The overall genetic divergence among the four populations was significant. About 10% of the genetic variance can be attributed to population differences, with F_{ST} and R_{ST} values of 0.08 ($P < 0.05$) and 0.12 ($P < 0.05$), respectively. Pairwise comparisons (Table 3) suggest that the genetic divergence between the three populations that were probably unaffected by reintroduction events (MRP, CPR and HUN) is similar, even if two of them are very close geographically. The high R_{ST} value obtained in the comparison between the MRP and CPR samples is due to the high single-locus R_{ST} values observed at three markers (SW841, SW1514 and SW2532), and it is probably related to

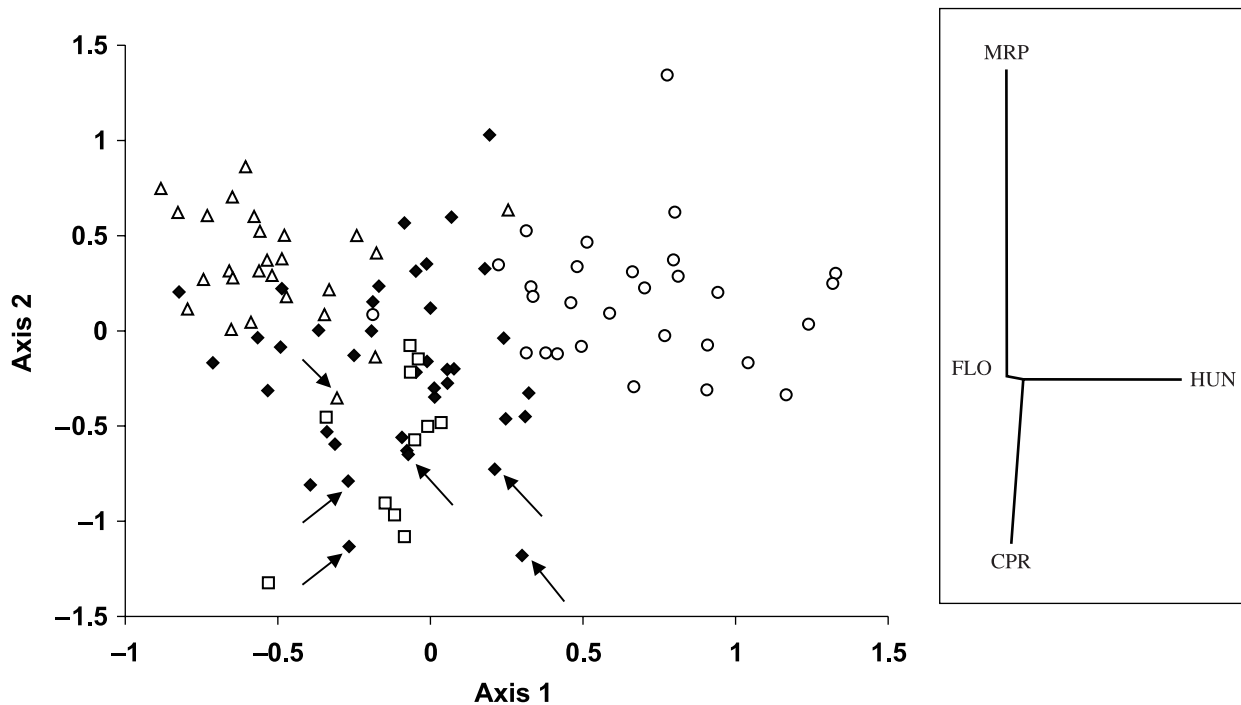


Fig. 2 Factorial correspondence analysis of the 105 individual genotypes. Δ , CPR; \circ , HUN; \square , MRP; \blacklozenge , FLO. Individuals with a genetic contribution from the inferred group IG4 > 0.8 (see Fig. 1) are indicated by arrows. Inset: Neighbour-joining tree based on pairwise F_{ST} distances.

sampling effects that strongly affect this molecular measure of differentiation. The FLO sample, as expected in hybrid populations, shows lower genetic distances when compared with the other samples.

A clear indication that the FLO population is genetically intermediate, and that several FLO individuals are genetically hybrids, is provided by the factorial correspondence analysis (Fig. 2). Individuals from MRP, CPR and HUN cluster in three different clouds, with FLO individuals falling within one of these clouds (suggesting a major contribution of one parental group) or between them (suggesting a composite genetic pool). Figure 2 also shows a neighbour-joining tree (Saitou & Nei 1987) based on pairwise F_{ST} distances: the position of FLO and its short branch are again consistent with the hybrid composition of this group.

If we assume that the FLO group is a genetic admixture of three parental populations, and that our samples MRP, CPR and HUN are representative of these populations, we can estimate the relative contribution of each parental group to the FLO sample. The results we obtain when the molecular information is ignored (mY_f in Table 4) suggest a similar fraction of the Hungarian and the Italian genetic pool, and a prevalence of the CPR population in the Italian contribution. However, when allele size is considered in the estimates (mY_m in Table 4), the contributions of CPR and HUN are smaller than 0 and larger than 1, respectively. These inconsistent values of the molecular estimator mY_m

Table 4 Estimated contributions of three predefined parental groups into the Florence area (FLO) sample. mY_f considers only the differences of allele frequencies, whereas mY_m makes use of the molecular information (number of repeats). Bootstrapped standard errors in parenthesis

| | CPR | MRP | HUN |
|--------|--------------|-------------|-------------|
| mY_f | 0.30 (0.07) | 0.18 (0.05) | 0.52 (0.07) |
| mY_m | -0.61 (0.83) | 0.02 (0.25) | 1.59 (1.05) |

are not uncommon when the simple model of admixture is violated (Bertorelle & Excoffier 1998), and should be interpreted with caution.

Quantitative analysis of the hybridization process provides different results when the data are analysed following the Pritchard *et al.* (2000) approach. The 105 individuals are compatible with $K = 4$ groups at HWE and linkage equilibrium (the inferred groups here called IG1, IG2, IG3 and IG4). The likelihood of $K = 4$ is in fact 1.32×10^9 times larger than the likelihood of $K = 5$, and much larger than the likelihood of any other value of K between 1 and 6. In other words, if we assume a uniform prior on K between 1 and 6, the posterior probability of $K = 4$ is virtually 1 (and 0 for any other value of K). But what do these four inferred groups correspond to? The genetic contributions (y -axes) of each inferred group (IG along the x -axes) into the individual genotypes is depicted in the insets of Fig. 1. Each

Table 5 Estimated proportion of each of the inferred groups in each of the predefined populations

| Pre-defined groups | Inferred groups | | | |
|--------------------|-----------------|------|------|------|
| | IG1 | IG2 | IG3 | IG4 |
| CPR | 0.86 | 0.02 | 0.05 | 0.07 |
| MRP | 0.03 | 0.93 | 0.02 | 0.02 |
| HUN | 0.05 | 0.03 | 0.86 | 0.06 |
| FLO | 0.29 | 0.13 | 0.13 | 0.45 |

line corresponds to an individual, and individuals from the same original locations are in the same graph. Clearly, almost all individuals from CPR, MRP and HUN have a major contribution from inferred groups IG1, IG2 and IG3, respectively. In other words, these three predefined groups are genetically homogenous clusters, approximately corresponding to the three inferred clusters identified using the Pritchard *et al.* (2000) method. Most of the individuals from the FLO sample, instead, appear genetically hybrid between two, three or four inferred groups. A simple equivalence between FLO and IG4 should therefore be excluded, as well as a simple model of hybridization among CPR, MRP and HUN that produced the FLO genotypes.

The average contributions of IGs into the original populations (Table 5) confirm that three predefined groups (CPR, MRP and HUN) correspond mainly to three distinct clusters, whereas the FLO genetic pool can be attributed to the four inferred groups in the following proportions: 0.29 (IG1 \approx CPR); 0.13 (IG2 \approx MRP); 0.13 (IG3 \approx HUN); 0.45 (IG4). Interestingly, the proportions estimated using the classical mY_f method (see Table 4) for the Italian populations (0.30 for CPR and 0.18 for MRP) are very similar to the estimated IG1 \approx CPR and IG2 \approx MRP contributions, whereas the Hungarian coefficient estimated by that method corresponds to the sum of the IG3 \approx HUN and IG4 proportions. The hybrid nature of most individuals in the FLO sample appears in Fig. 2, but is also supported by the high value of THI that we introduced in this study. In fact, THI reaches 0.69 in this sample; that is, more than two-thirds of the maximum value that would be obtained if all individuals were hybrids with relative contributions equal to that estimated for the whole sample. This value of THI is almost three times larger than the value (0.26) we would observe, for example, in the hypothetical admixed sample obtained by pooling artificially the individuals from CPR and MRP.

Discussion

Hybridization between genetically differentiated populations may have positive consequences for population viability when the potentially increased genetic variability

of the hybrid group translates into heterosis effects or reduced inbreeding depression (Frankham 1995; Coulson *et al.* 1998; Haig 1998). More often, however, hybridization is considered the cause of extinction of many populations in many animal and plant taxa, either because local adaptation is disrupted in hybrids (outbreeding depression) or simply because the genetic integrity of one group is lost (Rhymer & Simberloff 1996; Allendorf *et al.* 2001). The risk of extinction is probably higher when local populations are introgressed by gene pools from domesticated (usually inbred) animals (Lynch & O'Healy 2001). Therefore, the study of the genetic variation in source and target populations involved in large-scale reintroduction plans (which were common in the last century for many mammal species), appears crucial to the understanding of the genetic impact of such human-induced migrations, and to develop effective management programmes (Allendorf *et al.* 2001; Maudet *et al.* 2002).

The reintroduction of the wild boar in central Italy began in the 1950s, after approximately two centuries of demographic decline (Apollonio *et al.* 1988; Massei & Genov 2000). Individuals from central Europe, usually classified as *Sus scrofa scrofa*, were used to restock the local populations, usually classified as belonging to the *S. s. majori* subspecies. Releases of animals crossed in captivity with domestic pigs were also reported. Our study of nine microsatellite loci has clarified several aspects of such processes, suggesting that: (i) strong bottleneck effects can be excluded; (ii) the parental populations, two in Italy and one in Hungary, are genetically differentiated; (iii) the Italian individuals from areas affected by reintroduction appear to be genetic hybrids; (iv) hypervariable loci are useful in the wild boar to study the hybridization process and possibly to identify populations and individuals that deserve conservation efforts.

Genetic variation appears relatively high in all four populations of wild boar analysed, suggesting that the demographic decline of this species in the last centuries was not as extreme as for other species (e.g. Taylor *et al.* 1994; Houlden *et al.* 1996; O'Ryan *et al.* 1998). Even in the less variable Maremma Regional Park (MRP) sample, where only 11 individuals were typed, 5.1 alleles and 66% of heterozygotes are found at each locus on average. Higher values of genetic variability, possibly due to the different geographical origin of the individuals, are observed in the Hungarian (HUN) and the Florence area (FLO) samples. The three Hungarian locations we considered, however, are not genetically differentiated (and were pooled for that reason), and therefore the higher variability might be the consequence of a higher population size. Most individuals in the FLO sample are hybrids (see below), thus a higher variability in this group is expected.

Absolute levels of genetic variation should be treated with caution, as the typed loci were selected from a panel

developed for the domestic pig and excluded monomorphic or almost monomorphic markers. However, the results provided by two tests developed specifically to identify bottlenecks confirm the hypothesis that the wild boar populations we analysed have not experienced strong reductions in effective size.

The excess of heterozygosity (in the sense of gene diversity, i.e. H_E) compared with the value predicted at equilibrium from the observed number of alleles (a signature of demographic decline, Maruyama & Fuerst 1985), was not observed. By contrast, we identified some evidence of heterozygosity deficiency, possibly correlated with a demographic expansion (Maruyama & Fuerst 1984). The effects of genetic drift are in fact reduced in expanding populations (Slatkin 1994), and the relative proportion of rare alleles not removed by random drift but contributing only slightly to the heterozygosity is therefore higher than expected in stable populations. Cornuet & Luikart (1996) did not analyse in detail the effects of a population expansion on the relationship between the computed and the equilibrium heterozygosities. However, some of their simulations with mutation rates appropriate for microsatellite data show that even recent expansions (0.01 units of N generations ago, where N is the effective population size after the expansion) could actually produce a heterozygosity deficiency. These results, combined with the fact that wild boar have a very short generation time (females are fertile at about one year of age, Mauget & Pepin 1991), suggest that the heterozygosity deficiency we observed might be related to the demographic expansion which occurred in the last 50 years.

The absence of gaps in the allele size distributions, tested by the index M (Garza & Williamson 2001), also excludes a strong reduction in population size. The average values of M across loci varied between 0.70 and 0.88, whereas all data sets corresponding to population or species where bottlenecks were historically documented provided average M -values < 0.7 (Garza & Williamson 2001). These values proved to be statistically compatible with a simulated stable population (the behaviour of M in an expanding population is not known), with the only exception of the Hungarian population when a low value (5%) of multiple-step mutations was assumed. Pooling individuals from three geographical areas, although statistically not differentiated, might have contributed to create some gaps in the allele size distribution that decreased the value of M for the probably stable Hungarian population.

The Italian populations from the Maremma Regional Park and the Castelporziano Presidential Reserve, where major reintroduction can be excluded, are as genetically differentiated from each other as from the population in Hungary (pairwise F_{ST} values $\approx 15\%$). The simplest explanation for this result is a strong drift effect in restricted areas. However, the reasonably high levels of genetic

variation, the absence of genetic evidence for bottlenecks, and the historical records available for Castelporziano (Emiliani 1989), exclude this hypothesis. Therefore, it seems that the Italian samples used in this study as parental groups could be representative of two distinct and native central Italian populations. The similar degree of genetic divergence among these two Italian groups and the Hungarian populations, which is statistically significant but similar or lower than the divergence observed between populations of many other Ungulates (e.g. Arctander *et al.* 1996; Goodman *et al.* 2001; Vernesi *et al.* 2002), is in conflict with the morphological separation of the European (*S. s. scrofa*) and the Italian subspecies (*S. s. majori*) of wild boar.

The admixed composition of the group of individuals hunted in the Florence area where reintroduction is documented (the FLO sample), is supported by several lines of evidence. Its high genetic variability, the heterozygote deficit compared with Hardy–Weinberg expectations, the low genetic distances from all the other populations and the position in the tree topology (Fig. 2) are typical characteristics of a hybrid population (Chakraborty 1986; Cavalli-Sforza *et al.* 1994). However, only through more sophisticated analyses were we able to distinguish between population admixture without hybridization (a population structure in which individuals have the genetic composition of either of the parental groups) from a real hybridization process (in which most individuals have a genetic contribution from each parental population). The factorial correspondence analysis shows that most of the 39 individuals in the Florence area sample are genetic intermediates, with different relative contributions, between the three well-distinguishable parental groups from Italy and Hungary. The new index THI summarizes the high level of true hybridization in the Florence sample, although only the application of THI to other data will confirm its utility. More details of the genetic impact of reintroduction in the wild boar individuals hunted in the Florence area were then identified by the Bayesian clustering method proposed by Pritchard *et al.* (2000). Beside confirming the homogeneous composition of the parental populations and the hybrid characteristics of the FLO individuals, this analysis suggests a limited impact of reintroduction from central Europe but also that the wild boar hybrids have possibly a fourth component in their genetic pool not related to any of the sampled groups. About 13% of the genetic composition of the hybrids can be attributed, on average, to the Hungarian parental population, whereas 42% appears related to the Italian native groups (0.29% to CPR and 0.13% to MRP). The remaining 45% of the genetic pool of the hybrid individuals was attributed by the Pritchard *et al.* (2000) method to an additional *inferred* (i.e. not sampled) population.

The presence of a fourth unknown genetic component in the FLO hybrid population should not come as a surprise.

Classical methods of estimating admixture coefficients simply assume that all the parental groups are sampled, but this is probably not true in most cases. Identifying the origin of this fourth parental group, however, is not simple. Historical records and genetic data suggest at least five possibilities: a French component related to the colonization of northwestern Italy started in the 1920s, a central European component related to reintroduction events with individuals genetically differentiated from the Hungarian population in our sample, an additional native Italian component, a domestic pig component and, finally, a component erroneously identified as a parental group (see Pritchard *et al.* 2000) but actually related to the incipient process of divergence of the hybrid population. Unfortunately, the genetic data in our study do not provide clear indications in favour or against one of these hypotheses. The comparison between the results provided by the classical frequency-based admixture analysis (mY_f in Table 4) and by the Bayesian approach (Table 5) might suggest that the unknown genetic component in the hybrid populations has a central Europe origin. In fact, the estimated contribution mY_f of the Hungarian population in the hybrid (52%) corresponds approximately to the sum of the coefficients for the inferred groups IG3 (\approx HUN) and IG4 (the unknown parental group), whereas the two methods produce similar results for the contributions of the Italian parental groups. This result would appear logical if the unknown parental group was genetically related to the Hungarian population. However, an additional component from a central European population at least partially related to the Hungarian group seems unlikely when we note that the individuals in our sample more clearly related to the unknown inferred group (contribution of IG4 $>$ 0.8) appear in the multivariate analysis (see arrows in Fig. 2) closer to the Italian than to the Hungarian individuals. We are currently typing domestic pigs to verify the possible contribution of Tuscany breeds into wild boar hybrids, but more samples from central European regions are definitely needed.

Finally, it is important to note that analysis of microsatellite variation, even if the number of loci is relatively small, appears a very powerful tool to identify the genetic composition not only of populations, but also of single wild boar individuals. Selected hunting plans in high-density areas where this species causes serious damage to agricultural activities and disturbance to other species, as well as reintroduction and conservation projects aimed at reducing the chance of extinction of native stocks, could almost certainly be improved by carrying out preliminary genetic analyses.

Acknowledgements

We thank Mike Bruford, Guido Barbujani and Isabelle Dupanloup for helpful discussions, Mr Tonini ('Corpo Forestale dello Stato',

Maremma Regional Park) for technical assistance in sample collection and Laszlo Sugar (University of Veszprém, Hungary) for providing the Hungarian samples. The Istituto Nazionale della Fauna Selvatica kindly allowed us to collect samples from the Castelporziano Presidential Reserve. This work has been supported by a grant from the Fondazione Ente Cassa di Risparmio di Firenze.

References

- Alexander LJ, Rohrer GA, Beattie CW (1996) Cloning and characterization of 414 polymorphic porcine microsatellites. *Animal Genetics*, **27**, 137–148.
- Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrid: setting conservation guidelines. *Trends in Ecology and Evolution*, **16**, 613–622.
- Apollonio M, Randi E, Toso S (1988) The systematics of the wild boar (*Sus scrofa* L.) in Italy. *Bollettino Di Zoologia*, **3**, 213–221.
- Arctander P, Kat PW, Aman RA, Siegismund HR (1996) Extreme genetic differences among populations of *Gazella granti*, Grant's gazelle in Kenya. *Heredity*, **76**, 465–475.
- Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. *Molecular Biology and Evolution*, **15**, 1298–1311.
- Caughley G, Gunn A (1996) *Conservation Biology in Theory and Practice*. Blackwell Science, Cambridge, MA.
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.
- Chakraborty R (1986) Gene admixture in human populations: models and predictions. *Yearbook of Physical Anthropology*, **29**, 1–43.
- Comps B, Gomory D, Letouzey J, Thiebaut B, Petit RJ (2001) Diverging trends between heterozygosity and allelic richness during postglacial colonization in the European beech. *Genetics*, **157**, 389–397.
- Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.
- Coulson TN, Pemberton JM, Albon SD *et al.* (1998) Microsatellites reveal heterosis in red deer. *Proceedings of the Royal Society of London. Series B: Biology Sciences*, **265**, 489–495.
- De Beaux O, Festa E (1927) La ricomparsa del cinghiale nell'Italia settentrionale-occidentale. *Memorie della Società Italiana di Scienze Naturali e Museo Civico di Storia Naturale di Milano*, **III**, 263–322.
- Dupanloup de Ceuninck I, Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. *Molecular Biology and Evolution*, **18**, 672–675.
- Emiliani G (1989) *La tenuta di Castelporziano. Cenni storici*. Segretariato Generale della Presidenza della Repubblica.
- Frankham R (1995) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.
- Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, **10**, 305–318.
- Ghigi A (1911) Ricerche faunistiche e sistematiche sui mammiferi d'Italia che formano oggetto di caccia. *Natura*, **2**, 289–337.
- Goodman SJ, Tamate HB, Wilson R *et al.* (2001) Bottlenecks, drift and differentiation: the population structure and demographic history of sika deer (*Cervus nippon*) in the Japanese archipelago. *Molecular Ecology*, **10**, 1357–1370.

- Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.
- Haig SM (1998) Molecular contribution to conservation. *Ecology*, **79**, 413–425.
- Henry VG, Conley J (1972) Fall foods of European wild hogs in the southern Appalachians. *Journal of Wildlife Management*, **36**, 854–860.
- Houlden BA, England PR, Taylor AC, Greville WD, Sherwin WB (1996) Low genetic variability of the koala *Phascolarctos cinereus* in south-eastern Australia following a severe population bottleneck. *Molecular Ecology*, **5**, 269–281.
- Kadwell M, Fernandez M, Stanley HF *et al.* (2001) Genetic analysis reveals the wild ancestors of the llama and the alpaca. *Proceedings of the Royal Society of London, Series B: Biology Sciences*, **268**, 2575–2584.
- Lynch M, O'Hely M (2001) Captive breeding and the genetic fitness of natural populations. *Conservation Genetics*, **2**, 363–378.
- Macchi E, Sicureo B, Perronem A, Gallo Orsi U, Durio P (1995) The wild boar's impact on agriculture in Piedimont (Italy): a study on administrative reports. *Ibex Journal of Mountain Ecology*, **3**, 236–237.
- Maruyama T, Fuerst PA (1984) Population bottlenecks and non-equilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics*, **108**, 745–763.
- Maruyama T, Fuerst PA (1985) Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics*, **111**, 675–689.
- Massei G, Genov P (2000) *Il Cinghiale*. Calderini Edagricole, Bologna.
- Massei G, Toso S (1993) Biologia e gestione del cinghiale. *INFS, Documenti Tecnici*, **5**.
- Maudet C, Miller C, Bassano B *et al.* (2002) Microsatellite DNA and recent statistical methods in wildlife conservation management: applications in Alpine ibex [*Capra ibex (ibex)*]. *Molecular Ecology*, **11**, 421–436.
- Mauget R, Pepin D (1991) Energy intake, growth and timing of puberty in the European wild boar (*Sus scrofa*). In: *Global Trends in Wildlife Management* (eds Bobek B, Perzanowski K, Regelin W), pp. 205–209. Swiat Press, Krakow/Warsaw.
- Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- O'Ryan C, Harley EH, Bruford MW, Beaumont MA, Wayne RK, Cherry MI (1998) Microsatellite analysis of genetic diversity, within fragmented South African buffalo populations. *Animal Conservation*, **1**, 85–94.
- Pedrotti L, Dupré E, Preatoni D, Toso S (2001) Banca dati ungulati: status, distribuzione, consistenza, gestione, prelievo venatorio e potenzialità delle popolazioni di Ungulati in Italia. *Biologia E Conservazione Della Fauna*, **109**, 1–132.
- Piry S, Luikart G, Cornuet JM (1999) BOTTLENECK: a computer program for detecting recent reduction in the effective population size using allele frequency data. *Journal of Heredity*, **90**, 502–503.
- Polzheim RO, Hamr J, Mallory FF, Strobeck C (2000) Microsatellite analysis of North American wapiti (*Cervus elaphus*) populations. *Molecular Ecology*, **9**, 1561–1576.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
- Randi E (1995) Conservation genetics of the genus *Sus*. *Ibex Journal of Mountain Ecology*, **3**, 6–12.
- Randi E, Pierpaoli M, Beaumont M, Ragni B, Sforzi A (2001) Genetic identification of wild and domestic cats (*Felis silvestris*) and their hybrids using Bayesian clustering methods. *Molecular Biology and Evolution*, **18**, 1679–1693.
- Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.
- Saitou N, Nei M (1987) The neighbor-joining method. A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.
- Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Schneider S, Roessli D, Excoffier L (2000) *ARLEQUIN, Version 2.0: a Software for Population Genetic Data Analysis*. University of Geneva, Switzerland.
- Singer FJ, Swank WT, Clebsh EEC (1984) Effects of wild pig rooting in a deciduous forest. *Journal of Wildlife Management*, **48**, 464–473.
- Slatkin M (1994) Linkage disequilibrium in growing and stable populations. *Genetics*, **137**, 331–336.
- Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–462.
- Taylor AC, Sherwin WB, Wayne RK (1994) Genetic variation of microsatellite loci in a bottlenecked species: the northern hairy-nosed wombat *Lasiornhinus krefftii*. *Molecular Ecology*, **3**, 277–290.
- Vernesi C, Pecchioli E, Caramelli D, Tiedemann R, Randi E, Bertorelle G (2002) The genetic structure of natural and reintroduced roe deer (*Capreolus capreolus*) populations in the Alps and Central Italy, with reference to the mitochondrial DNA phylogeography of Europe. *Molecular Ecology*, **11**, 1285–1297.
- Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
- Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution*, **19**, 395–420.

This project is part of an ongoing collaboration among the Department of Biology of the University of Ferrara, the Centre of Alpine Ecology in Trento, and the Biosfera Association in Florence on the conservation genetics and management of several ungulates. Davide Tartari is a graduate student, Elena Pecchioli and Barbara Crestanello are PhD students, and Cristiano Vernesi is a postdoctoral fellow. Heidi Hauffe is the coordinator of the genetics laboratory of the Centre of Alpine Ecology. Giorgio Bertorelle is interested in the analysis of population genetics data for conservation purposes.

Appendix

Assume that the contribution of each of d parental (or inferred) groups can be estimated in each of n individuals sampled in an admixed population, and call this x_{ij} with i ranging from 1 to n and j ranging from 1 to d . Clearly,

$$\sum_{j=1}^d x_{ij} = 1 \quad \forall_i$$

$$\bar{x}_{i\cdot} = \sum_{j=1}^d \frac{x_{ij}}{d} = \frac{1}{d} \quad \forall_i$$

$$\sum_{i=1}^n \sum_{j=1}^d x_{ij} = \sum_{i,j}^{n,d} x_{ij} = n,$$

and the average contribution $\bar{x}_{\cdot j}$ of the parental population j into the admixed population is given by

$$\bar{x}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n x_{ij} = \frac{x_{\cdot j}}{n}, \text{ and the sum of all } \bar{x}_{\cdot j} \text{ is } \sum_{j=1}^d \bar{x}_{\cdot j} = \frac{1}{n} \sum_{j=1}^d$$

$$x_{\cdot j} = \frac{1}{n} \sum_{i,j}^{n,d} x_{ij} = 1.$$

Now, the same set of average contribution $\bar{x}_{\cdot j}$ can be observed in a mixed sample of nonhybrid individuals similar to either of the parental groups or in a sample of truly hybrid individuals. In the first case, however, the contribution of each parental group into each individual will be close either to 1 or 0 (with different individuals having the major contribution from different parental groups), whereas in the second case these individual contributions will be close to the average values $\bar{x}_{\cdot j}$. A simple quantity able to discriminate these two situations is therefore the average V of the individuals variances V_i , where

$$V_i = \frac{\sum_{j=1}^d (x_{ij} - \bar{x}_{i\cdot})^2}{d} = \frac{\sum_{j=1}^d (x_{ij} - 1/d)^2}{d} \quad \text{and}$$

$$\bar{V} = \frac{\sum_{i=1}^n V_i}{n} = \frac{\sum_{i,j}^{n,d} (x_{ij} - 1/d)^2}{nd} = \frac{1}{nd} \left(\sum_{i,j}^{n,d} x_{ij}^2 - n/d \right).$$

Clearly, V is expected to be high in samples of individuals with different origin, but low in samples of hybrid individuals.

It can be shown that V is also the variance of all x_{ij} contributions, since $\bar{x}_{i\cdot}$ is the same for all i individuals.

The maximum and minimum value that V_i can assume for a specific set of $\bar{x}_{\cdot j}$ are simple to compute. In fact, V_i cannot be larger than

$$V_{i(\max)} = \frac{(1 - 1/d)^2 + \sum_{j=1}^{d-1} (0 - 1/d)^2}{d} = \frac{d - 1}{d^2}$$

which is the value that occurs when the individual i has a genetic composition that can be completely attributed to either of the parental groups. This is also the maximum value of V , $\bar{V}_{(\max)}$, obtained when all individuals i have the same $V_i = V_{i(\max)}$. Alternatively, when the individual i has a set of contributions equal to the average contributions $\bar{x}_{\cdot j}$, V_i will be minimum. In this case:

$$V_{i(\min)} = \frac{\sum_{j=1}^{d-1} (\bar{x}_{\cdot j} - 1/d)^2}{d} = \frac{\sum_{j=1}^{d-1} (x_{\cdot j}/n - 1/d)^2}{d} = \frac{1}{nd} \left(\frac{\sum_j x_{\cdot j}^2}{n} - \frac{n}{d} \right),$$

and, as before, $\bar{V}_{(\min)} = V_{i(\min)}$ when all individuals have the same $V_i = V_{i(\min)}$. The variance $V_{(\min)}$ is 0 only when the d average contributions $\bar{x}_{\cdot j}$ are equal (and therefore equal to $1/d$).

The *THI*, ranging from 0 to 1 with larger values expected in populations where hybridization is common since several generations, can be therefore defined as the complement at 1 of the standardized variance V , i.e.

$$THI = 1 - \frac{\bar{V} - \bar{V}_{\min}}{\bar{V}_{\max} - \bar{V}_{\min}} = \frac{\bar{V}_{\max} - \bar{V}}{\bar{V}_{\max} - \bar{V}_{\min}}.$$

We note that the proposed standardization produces an index $THI = 1$ when all individuals have a set of contributions equal to the average contributions $\bar{x}_{\cdot j}$ even when these contributions $\bar{x}_{\cdot j}$ are different and unequal. A simplified computational expression of *THI* is as follows:

$$THI = \frac{n \left(n - \sum_{i,j}^{n,d} x_{ij}^2 \right)}{n^2 - \sum_j x_{\cdot j}^2}.$$