

# Is a Multivariate Consensus Representation of Genetic Relationships Among Populations Always Meaningful?

K. Moazami-Goudarzi\* and D. Laloë<sup>†,1</sup>

\*Laboratoire de Génétique Biochimique et de Cytogénétique, INRA, 78352 Jouy-en-Josas, France and <sup>†</sup>Station de Génétique Quantitative et Appliquée, INRA, 78352 Jouy-en-Josas, France

Manuscript received October 1, 2001  
Accepted for publication June 12, 2002

## ABSTRACT

To determine the relationships among closely related populations or species, two methods are commonly used in the literature: phylogenetic reconstruction or multivariate analysis. The aim of this article is to assess the reliability of multivariate analysis. We describe a method that is based on principal component analysis and Mantel correlations, using a two-step process: The first step consists of a single-marker analysis and the second step tests if each marker reveals the same typology concerning population differentiation. We conclude that if single markers are not congruent, the compromise structure is not meaningful. Our model is not based on any particular mutation process and it can be applied to most of the commonly used genetic markers. This method is also useful to determine the contribution of each marker to the typology of populations. We test whether our method is efficient with two real data sets based on microsatellite markers. Our analysis suggests that for closely related populations, it is not always possible to accept the hypothesis that an increase in the number of markers will increase the reliability of the typology analysis.

**A**NALYSIS of genetic relationships is useful for phylogenetic or biodiversity studies. For this purpose, it is customary to use genetic markers such as protein and blood group polymorphisms or DNA markers. Generally, the approach for the genetic analysis of these data consists of calculating genetic distances and constructing trees. For example, ESTOUP *et al.* (1995) highlighted the use of microsatellite loci for a precise dissection of the genetic structure of honey bee colonies. BARKER *et al.* (1997) analyzed the phylogenetic relationships of Asian water buffalos by comparing results from protein loci and microsatellite loci.

On the basis of theoretical studies, TAKEZAKI and NEI (1996) have shown that one of the important factors for analyzing the correct position of populations in a genetic study is the number of loci used. Within this framework, the main task is to obtain a mean or consensus phylogeny. The reliability of the results is addressed through studies of confidence limits on phylogenies using bootstrap as in FELSENSTEIN (1985) and EFRON *et al.* (1996) or Markov chain Monte Carlo methods as in LI *et al.* (2000).

Alternatively, representations of the genetic relationships among populations may be obtained by using multivariate procedures. These techniques condense the information from several alleles and loci into a few synthetic variables. The connection between tree proce-

dures and multivariate procedures is close (CAVALLI-SFORZA *et al.* 1994). The first splits in a tree generally correspond to the separation of populations generated by the first dimensions of the multivariate procedure. The use of these procedures in studies on genetic variation among populations has been pioneered by Cavalli-Sforza and his colleagues over the last 25 years to reconstruct the history of human populations (CHEN *et al.* 1985; CAVALLI-SFORZA and PIAZZA 1993; BOWCOCK *et al.* 1994; CAVALLI-SFORZA *et al.* 1994; CAVALLI-SFORZA 1997; UNDERHILL *et al.* 2000). Blood group, protein, and DNA markers were typed for large population samples around the world. These studies provided a broad picture of human populations from a genetic perspective, with additional descriptions based on archeological data and linguistics.

Multivariate procedures are particularly attractive when admixtures are known to have occurred among the populations under study, because construction of trees using admixed populations contradicts the principles of phylogeny reconstruction (FELSENSTEIN 1982). This situation is often found for genetic studies concerning breeds or populations from the same species, as in human population genetic studies (SAHA and TAY 1992; AYALA *et al.* 1994; SAHA *et al.* 1995; BOSCH *et al.* 1997; JORDE *et al.* 1997) or in studies concerning domestic breeds (GROSCLAUDE *et al.* 1990; MACHUGH *et al.* 1997; CYMBRON *et al.* 1999; YANG *et al.* 1999; CAÑON *et al.* 2000; HANSLIK *et al.* 2000; WIMMERS *et al.* 2000). For instance, BLOTT *et al.* (1998) used blood group and protein polymorphisms to examine the genetic relationships among 37 European cattle breeds. They have shown that rela-

<sup>1</sup>Corresponding author: Station de Génétique Quantitative et Appliquée, INRA, Domaine de vilvert, 78352 Jouy-en-Josas, France.  
E-mail: ugendla@dga2.jouy.inra.fr

tionships among breeds reflect their geographical origin and common ancestry rather than the agricultural use for which the breeds have been selected. MACHUGH *et al.* (1997) used different markers such as microsatellites, mtDNA, and a Y chromosomal probe to determine the evolutionary relationships among 20 different cattle breeds from Africa, Europe, and Asia. They confirmed the large divergence between *Bos taurus* and *B. indicus* lineages. In addition, they revealed a pattern of zebu genetic introgression in African taurine populations.

However, the reliability of multivariate analysis is never discussed. In this article, we investigate this problem by addressing the following specific questions: Is the consensus representation meaningful? Which markers influence the typology of populations? Finally, we study their interest and efficiency with bovine data.

## MATERIALS AND METHODS

**Fisher's exact test:** Equality of allelic frequencies per breed was tested for each marker by a Fisher's exact test of independence. *P* values were estimated by a Monte Carlo procedure (AGRESTI 1992). Tests were performed with the procedure FREQ of SAS 8.1.2 (SAS INSTITUTE 2000).

**Principal component analysis:** Among the many multidimensional analysis methods, principal component analysis (PCA) offers a simple and powerful mode of analysis of a set of population-by-gene frequency data. A detailed presentation of the general method can be found, for instance, in MARDIA *et al.* (1979).

Let us consider *g* populations and an *n*-allelic marker.  $p_{ik}$  denotes the allelic frequency of the *k*th allele in the *i*th population, and  $p_k$  the mean allelic frequency of the *k*th allele in all the populations. As advocated by CAVALLI-SFORZA *et al.* (1994) the use of an *a priori* standardization of the allelic frequencies  $p_{ik}$  by their estimated standard deviation  $\sigma_k = [p_k(1 - p_k)]^{1/2}$ , can be expected to improve the recovery of information. This standardization emphasizes contribution of rare alleles. The general term of the array *X* of standardized allelic frequencies is then equal to  $x_{ik} = (p_{ik} - p_k)/\sigma_k$ . This standardization is used throughout this article.

Each population is represented by a point in an *m*-dimensional space, the coordinates of which are the *m* standardized allelic frequencies. The core of the PCA is to find the most scattered directions, or principal components, of this cloud of points. Principal components (PCs) are eigenvectors of the variance-covariance matrix of standardized allelic frequencies. The sum of the eigenvalues is the trace of this matrix and it is sometimes called "total variance." PCs are ranked according to the fraction of total variance that each of them can independently explain: For instance, the first PC is by definition more informative than the second PC and all the PCs are independent from each other.

This method can be applied to the allelic frequencies of one marker (single-marker analysis) or of several markers (overall analysis) to get the induced typology of populations.

**Relationships among single-marker analyses:** Phylogenetic studies typically involve several markers and lead to several typologies, trees, or multidimensional plots. The first problem is to evaluate the relationship among these typologies, which is commonly called "interstructure." Use of PCA leads to an implicit euclidean distance between populations *i* and *j*:

$$d(i, j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}.$$

Note that it is equal to the square root of the distance of Barker (BARKER *et al.* 1993). Congruence or correlations among the different distance matrices resulting from the *m* different single-marker PCA can be calculated by their MANTEL (1967) correlations. An equivalent process in terms of probability is to use a standardized version that is based on the classical Pearson correlation coefficient (MANLY 1997, p. 174), *i.e.*, the correlation computed from two sets of  $g(g - 1)/2$  distances. Computation of the pairwise correlations between distances leads to a correlation matrix of order *m*. For a powerful description of the correlation structure, a PCA on the matrix of Mantel correlations is useful. Markers can then be represented in a plot, or correlation circle, the axes of which are the two first principal components. The coordinate of a marker along an axis is the correlation (or loading factor) of the marker with this axis (MORRISON 1976).

Relative positions of markers in this circle indicate the magnitude of association between these markers. For instance, if all markers are positively correlated, their correlations with the first axis are positive, and they will be positioned to the right half of the circle. If diagonalizing this matrix, Perron-Frobenius theorem ensures that all the coefficients of the first principal component are positive (*e.g.*, MARDIA *et al.* 1979). Conversely, dispersed positions of the markers inside the circle are a visual indicator of the incongruence of markers.

**Test of structure congruence:** If single-marker structures are not congruent, no compromise structure will exist, Mantel correlations between distances will be indiscriminately positive or negative, and the sum *S* of all the  $m(m - 1)$  Mantel correlations among the distances will not be significantly different from 0. To test the significance of *S*, a randomization test is conducted as follows: (1) calculation of the observed statistic  $S_{obs}$ , (2) random permutation of entire rows of each table *X* of standardized allelic frequencies, and (3) recalculation of  $S_{rnd}$  values. The probability that no compromise exists is calculated as  $(\text{number of } S_{rnd} \geq S_{obs} + 1) / (\text{number of randomizations} + 1)$ . The 1 in the numerator and the denominator represents the observed value for the statistic being evaluated, which is considered as a possible value of the randomization distribution.

Another method (MOAZAMI-GOUDARZI *et al.* 1997), based on a nonparametric analysis of the variance, considers the standardized distance  $d_{ijk}$  between populations *i* and *j* computed from the allelic frequencies of marker *k* as a result of the following one-way linear model:

$$d_{ijk} = d_{ij} + e_{ijk}.$$

If no congruent typology exists among the populations, they are roughly equidistant from each other, and there is no significant difference between the mean distances  $d_{ij}$ .

The null hypothesis  $H_0$  is then: All distances are equal. Conversely, if some typology exists, the alternative hypothesis  $H_1$  will be that at least one distance is different from the others. This test can be performed by numerous nonparametric methods, for instance, the Kruskal-Wallis test, based on Wilcoxon rank scores (CONOVER 1980), in the procedure NPARIWAY of SAS 8.1.2. (SAS INSTITUTE 2000).

If significant relationships are found for all the markers, a global analysis is meaningful and can be done by a PCA on the entire data. An interesting feature of PCA in this multimarker context is the possibility of evaluating the relative contribution of each marker in the structure of the principal components, to see if one of the principal components, and therefore part of the typology among populations, is due to the action of a few markers only or to the whole set of markers.

All computations relative to PCA were performed with the SAS 8 package (SAS INSTITUTE 2000).

**Application to data:** Now we test the efficiency of our

method on real data. We present results obtained by analyzing two data sets. Data set 1 was obtained from eight French and two European cattle breeds and data set 2 was obtained from 20 distinct populations from Africa (10) and Europe (10). *B. taurus*, *B. indicus*, and one crossbred population are included in data set 2.

Data set 1, taken from MOAZAMI-GOUDARZI *et al.* (1997), contains data for 17 microsatellite loci (INRA K, INRA 005, INRA 011, INRA 013, INRA 016, INRA 023, INRA 025, INRA 032, INRA 035, INRA 037, INRA 040, INRA 063, INRA 064, INRA 072, ETH 131, ETH 152, and ETH 225) genotyped in 10 cattle breeds (Breton Black Pied, Charolais, Holstein, Jersey, Limousin, Maine-Anjou, Montbeliard, Normand, Parthenais, and Vosgien). Samples were collected throughout France.

In data set 2, nine microsatellite loci (INRA K, INRA 005, INRA 016, INRA 032, INRA 035, INRA 063, INRA 072, ETH 152, and ETH 225) were studied in 20 different cattle populations including the 10 populations of data set 1 and 10 African populations [Somba, Lagunaire, N'dama, Baoulé, Kuri, Sudanese Fulani zebu, Red Bororo zebu, Shuwa zebu, Madagascar zebu, and Borgou (Shorthorn  $\times$  zebu crossbreds)]. West African populations were sampled in three neighboring countries: Somba cattle samples were collected in the Atacora highlands (Northwestern Benin/Northeastern Togo), which is the birthplace of this breed. Lagunaire cattle samples were collected in Benin, N'dama, Baoulé, and Borgou; Sudanese Fulani zebu samples in Burkina-Faso and Kuri; Red Bororo zebu and Shuwa zebu samples in Chad; and Madagascar zebu samples in Madagascar.

For Somba, Lagunaire, Borgou (Shorthorn  $\times$  zebu crossbreds), and Sudanese Fulani zebu samples, we used protocols described in MOAZAMI-GOUDARZI *et al.* (1997). Data concerning N'dama, Baoulé, Kuri, Red Bororo zebu, Shuwa zebu, and Madagascar zebu populations were taken from SOUVENIR ZAFINDRAJONA *et al.* (1999). Data concerning the 10 European breeds were taken from MOAZAMI-GOUDARZI *et al.* (1997).

## RESULTS AND DISCUSSION

**Fisher's exact test:** The frequency distributions of each microsatellite breed combination were significantly different as demonstrated by Fisher's exact tests ( $P < 10^{-4}$ ).

**Mantel correlations:** Mantel correlations have been computed for the two data sets (Tables 1 and 2). For data set 1, 73 out of the 136 Mantel correlations, *i.e.*, more than one-half of the coefficients, are negative. These results are confirmed by the correlation circle (Figure 1A). All the markers are scattered in the circle and only a few markers are in the same area (for example, INRA 011, INRA 013, INRA 035, and INRA 063). In addition, high  $P$  values equal to 0.48 for the permutation test on the sum of Mantel correlations and 0.60 for the Kruskal-Wallis test were obtained. In this case, tests for the existence of compromise structure are not significant. This lack of structure can be explained by the small level of differentiation among populations or by different typologies exhibited by different markers. As demonstrated by Fisher's exact tests, significant differences among breeds are exhibited by each microsatellite. For this data set, the second explanation is more likely. Further analyses are meaningless.

However, in data set 2, only 5 out of the 36 Mantel correlations, *i.e.*,  $\sim 1/7$ , are negative. This low proportion is a first indication of a congruent typology. These results are confirmed by the correlation circle (Figure 1B.). All the markers, except INRA 035, are clustered in the same area. Tests for the existence of a common typology are very significant.  $P$  values are equal to 0.0001 for both the Mantel correlation test and the Kruskal-Wallis test. Thus, the search for a compromise typology is meaningful and therefore we performed an overall PCA.

**Principal component analysis for data set 2:** The first three PCs account for 59%, *i.e.*, most of the variance (Figure 2A). The scatterplot is in Figure 3. The first PC accounts for 30% of the total variance and it clearly distinguishes three clusters: (1) the Kouri, Borgou, and zebu cluster; (2) the African taurine cluster; and (3) the French taurine cluster. The second PC summarizes 16% of the total variance and it clearly separates the African taurine breeds from the Madagascar zebu breed. The third PC describes 13% of the total variance and it strongly isolates the Madagascar zebu.

From a geographical point of view (Europa/Africa) and from a species point of view (*B. taurus*/*B. indicus*), the breeds included in this analysis are very well characterized. The first cluster is heterogeneous, since it includes the Borgou, a crossbred population between zebu and taurine breeds, and the Kuri, considered as a taurine population since it is humpless and has the small metacentric Y chromosome of *B. taurus*. This intermediate position of the Kuri breed has also been found by MAHÉ *et al.* (1999) and SOUVENIR ZAFINDRAJONA *et al.* (1999). Divergent explanations have been proposed. The genetic differences between Kuri and other taurine populations of western Africa could have a relatively remote historical origin. For example, the ancestors of the N'dama probably spread through northern and northwestern Africa while those of Kuri spread through the Sahara during the "green period." The gene frequencies of the original domesticated population of southwest Asia were probably closer to those of zebu than to those of modern taurines of western Africa. Thus, it is conceivable that Kuri has remained genetically closer to zebu. MACHUGH *et al.* (1997) have concluded that an east-to-west introgression gradient of microsatellite alleles exists in Indian *B. indicus* into African populations, including subpopulations of the taurine type N'dama. Another nonexclusive explanation could be that these results reflect recent crossbreeding with Shuwa and Red Bororo zebras located around Lake Chad.

**Influence of individual markers:** Contributions of markers to the first three PCs are in Figure 4. Five markers contribute roughly equally to the first PC: ETH 152 (18%), ETH 225 (15%), INRA 063 (15%), INRA k (15%), and INRA 032 (14%); *i.e.*, a total of 77%. Single PCA performed with these markers shows a good separa-



**TABLE 2**  
**Mantel correlations ( $\times 1000$ ) for data set 2**

Microsatellites	e.152	e.225	i.16	i.32	i.35	i.63	i.72	i.5	i.k
e.152	1000	747	306	468	124	247	404	285	607
e.225		1000	468	569	152	452	481	344	422
i.16			1000	266	-86	244	446	83	245
i.32				1000	-2	453	510	335	301
i.35					1000	106	-108	-230	-31
i.63						1000	328	228	213
i.72							1000	350	360
i.5								1000	237
i.k									1000

i code corresponds to INRA microsatellites and e code corresponds to ETH microsatellites.

tion, particularly between the zebu cluster and the other breeds.

Other markers, especially INRA 035, which contributes only for 3.6%, do not exhibit any clustering between taurine and zebu breeds. INRA 035 separates Charolais and Parthenais breeds from the others. No explanation was found for this specific structure. No null alleles were observed for any of the microsatellites used in this study. The allele of INRA 035 that we have sequenced was composed of a perfect, uninterrupted, and homogeneous TG. It is also known that gene selection processes may lead to the fixation of alleles and the polymorphism in the flanking region may be wiped out (SCHLÖTTERER and WIEHE 1999). Until now, no gene has been mapped in the region 16q11 where INRA 035 is localized (Bov-Map database <http://locus.jouy.inra.fr>). In addition, no homology between the whole sequence of this microsatellite and other mammalian sequences has been found (Blast search).

Contributions to other PCs are more disparate. In particular, the third PC, which strongly separates Madagascar zebu from other breeds, is supported mainly by two markers only: ETH 152 (27%) and INRA 032 (32%). Thus, the excentric position of the Madagascar zebu is not really reliable and should be confirmed by further analysis.

**Robustness of the compromise structure:** When a compromise typology exists, the majority of markers will contribute to the construction of the first PCs. These PCs will explain a great percentage of the variance, while the other PCs, which express specific actions of markers, are of less importance. In this case, the omission of noncongruent markers will not change the compromise typology since they do not participate in its construction. This analysis is robust against the presence/absence of an incongruent marker. For instance, in data set 2, ignoring INRA 035 in the analysis will not change the general compromise typology.

When each marker leads to an independent typology, each PC of the global PCA is associated with one marker and therefore each PC will explain roughly the same

variance proportion. In this case, the omission of markers will still lead to a nonmeaningful joint analysis. This situation corresponds to our data set 1. The variance proportions explained by each PC (Figure 2B) show a slow decrease of values that contrasts with the corresponding number for data set 2. Results concerning contributions of markers to the construction of PCs (Figure 5) are also different from the analysis of data set 2. PCs, even the first one, are not built by a majority of markers, but by only two of them.

**Comparison with the neighbor-joining tree:** It is interesting to compare our approach with the neighbor-joining tree of data set 1 (MOAZAMI-GOUDARZI *et al.* 1997). While all microsatellite loci showed significant differences among breeds, a lack of strong differentiation was observed by bootstrap resampling of loci. The most robust feature of the typology concerned the Holstein/Maine-Anjou grouping with a bootstrap value of 74%.

Note that this problem is not specific to French cattle breeds but common for geographically close populations. Similar results have been obtained from many different sources of data relevant to genetic diversity and evolutionary history of humans (CAVALLI-SFORZA and PIAZZA 1993; BOWCOCK *et al.* 1994; UNDERHILL *et al.* 2000) or domestic breeds (cattle, MACHUGH *et al.* 1997; goats, SAITBEKOVA *et al.* 1999; chicken, WIMMERS *et al.* 2000; dogs, KOSKINEN and BREDBACKA 2000).

This lack of structure is commonly explained by the fact that too few markers were used for the analysis. However, this explanation implicitly assumes that each marker reveals the same typology among populations. In this context, differences that may be observed among typologies will be considered as typical residual errors or white noise, the influence of which will decrease with the number of markers involved. Consequently, when the number of loci increases, estimation error variance of the distances among populations will decrease (FOULLEY and HILL 1999), and bootstrap values of dendrograms will increase (TAKEZAKI and NEI 1996).

Alternatively, this lack of structure could be explained by discrepancies among the typologies exhibited by each

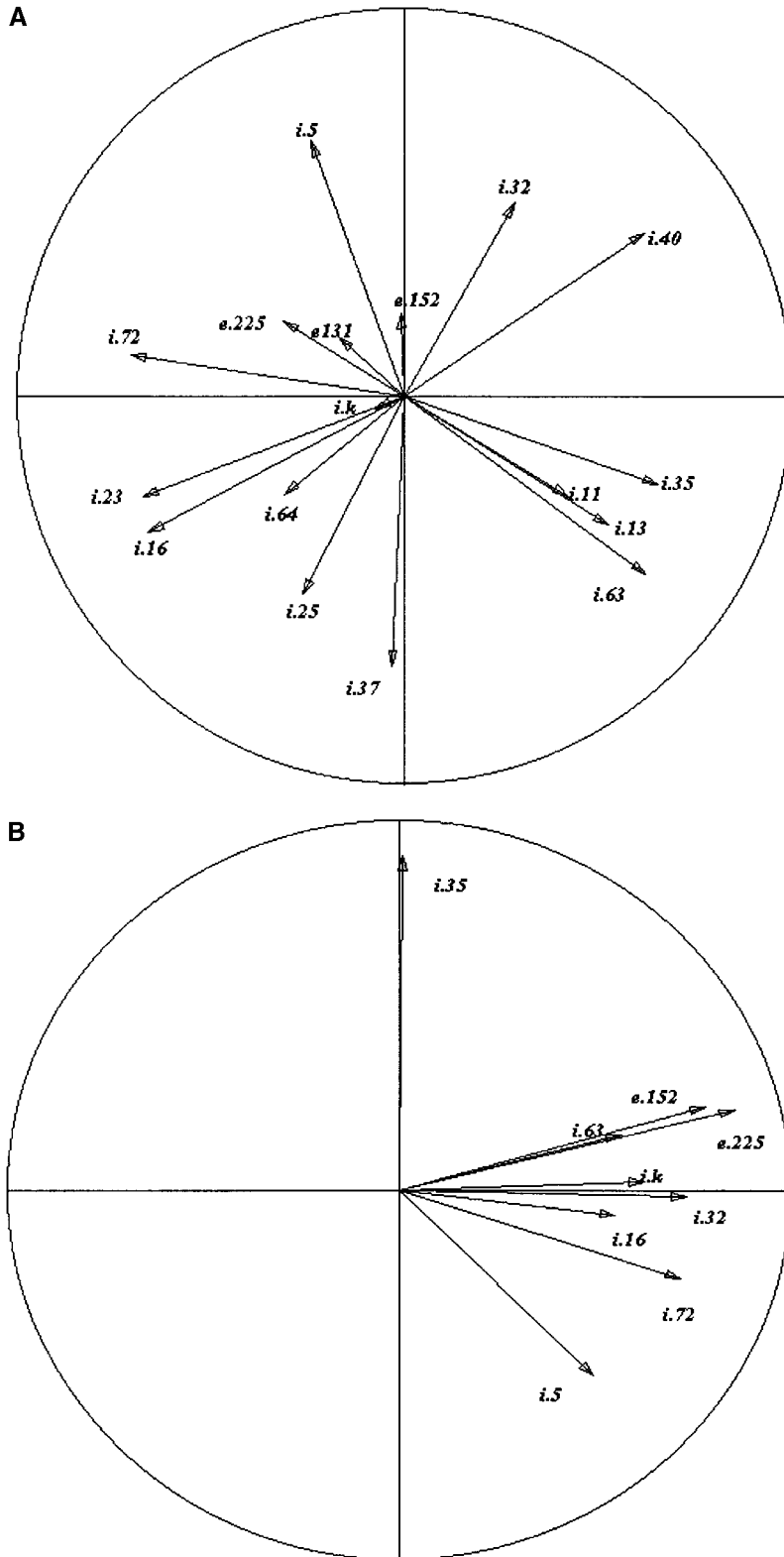


FIGURE 1.—(A) Correlation circle for data set 1 (8 French and 2 European cattle breeds genotyped for 17 microsatellites). (B) Correlation circle for data set 2 (20 cattle breeds: 5 African, 4 African zebu, 1 Borgou, 8 French, and 2 European genotyped for 9 microsatellites). At each arrow end the i code corresponds to INRA microsatellites and the e code corresponds to ETH microsatellites.

marker. This is highlighted by the fact that one-half of the Mantel correlations among markers are negative in data set 1. The study of LAVAL *et al.* (2000) illustrated this phenomenon. They analyzed the genetic diversity of 11 European pig breeds on the basis of 18 microsatel-

lites. They reported that these breeds exhibit a very strong differentiation and it was difficult to infer any reliable phylogeny among those populations. When the analysis was restricted to the 9 breeds for which genotypes were available at 25 loci, even less reliable results

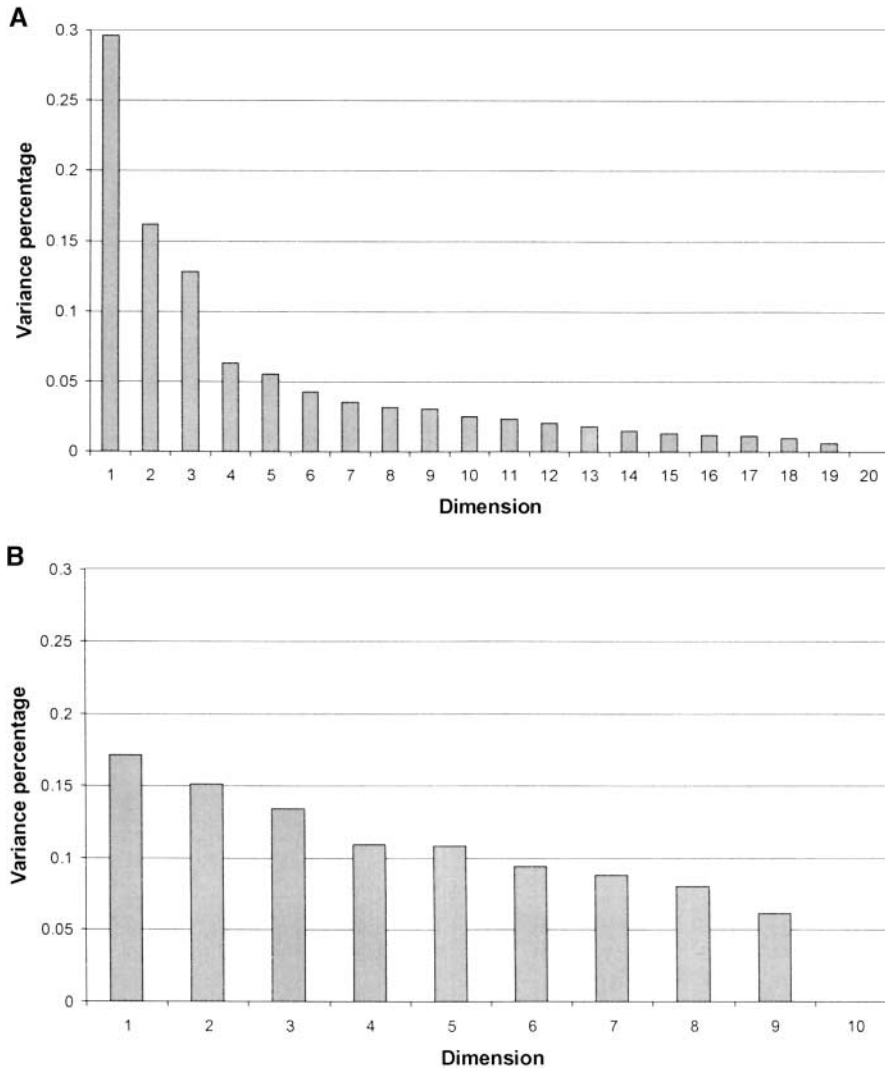


FIGURE 2.—Variance percentage explained by each dimension for data set 2 (A) and for data set 1 (B).

were obtained. Increasing the number of loci decreased the reliability of the phylogeny. LAVAL *et al.* (2000) suggested that populations had differentiated according to a radiative scheme of divergence. According to this scheme, expected genetic distances between populations would be equal, and any casual differences among distances might be due to random genetic drift. In this context, incongruent single-marker structures could be obtained. This hypothesis has to be confirmed, for example, by simulation studies.

**Assignment:** Several studies have demonstrated the high potential of microsatellites for discrimination among individuals (CORNUET *et al.* 1999; PRITCHARD *et al.* 2000; BJORNSTAD and ROED 2001; CAÑON *et al.* 2001). The lack of congruence among structures is not a disadvantage for assignment studies, where independence of discriminating factors is a great asset. For instance, in data set 1, INRA 035 discriminates the Parthenais breed from the others (allele 116 is present in this breed only, with a frequency of 21%; the frequency of allele 114 is 13% while it is <6% for other breeds), and INRA 013 isolates the Maine Anjou breed (the frequency of allele 188 is

37% in this breed while it is <10% in other breeds). These two markers are not congruent, but their combination makes it possible to discriminate both Parthenais and Maine-Anjou breeds. For this data set, the proportion of animals correctly assigned to their population of origin was 90% (MOAZAMI-GOUDARZI *et al.* 1998).

The contrasts between the good efficiency of assignment techniques and the weak structuralization among breeds are also found in other studies. GOLDSTEIN *et al.* (1999) analyzed microsatellite variations in populations of island foxes (*Urocyon littoralis*) on California's Channel Islands with 19 microsatellites. Out of the five nodes of the UPGMA consensus tree, only one could be considered as significant, with a bootstrap value of 0.96, while the other bootstrap values were 0.5, 0.49, and 0.74. However, assignment of individuals to their geographic origin was already perfect with 181 individuals out of 183 correctly assigned. CAÑON *et al.* (2000) studied the genetic structure of seven Spanish Celtic horse breeds by genotyping 13 microsatellites. Bootstrap values of the five nodes were 0.49, 0.51, 0.59, 0.64, and 0.79, while 77–97% of correct assignments were obtained. VILÀ *et*

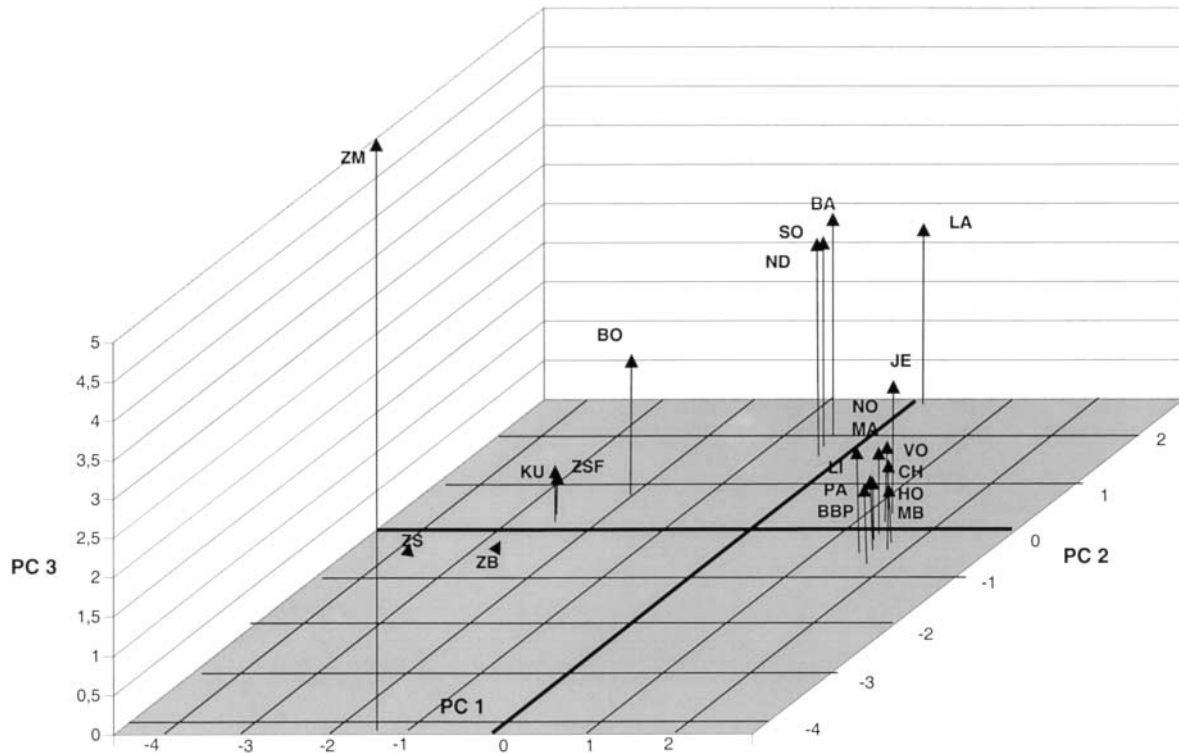


FIGURE 3.—Scatterplot diagram showing the first three PCs from allele frequencies for data set 2. The letter codes correspond to the populations sampled as follows: SO, Somba; LA, Lagunaire; BO, Borgou; ZSF, Sudanese Fulani zebu; ND, N'dama; BA, Baoulé; KU, Kuri; ZB, Red Bororo zebu; ZS, Shuwa zebu; ZM, Madagascar zebu; BBP, Breton Black Pied; CH, Charolais; LI, Limousin; MA, Maine-Anjou; MB, Montbeliard; NO, Normand; PA, Parthenais; VO, Vosgien; HO, Holstein; and JE, Jersey.

*al.* (2001) analyzed 15 microsatellites of 10 horse breeds and generally their bootstrap values were not significant while 95% of the individuals were assigned correctly.

### CONCLUSION

In this article, we have described a two-step process: The first step consists of performing single-marker analyses and studying relationships between them (interstructure) and the second step is building a compromise plot (intrastructure). We have focused on particular techniques, such as PCA, but other multivariate methods exist, such as correspondence analysis (HILL 1974), or multidimensional scaling (CARROLL 1981). The study of intrastructure and testing of the existence of a compromise typology were based on distances and Mantel correlations and can be applied outside a strictly multidimensional context.

We have illustrated our method with real data provided by microsatellites. These markers are substantially more complicated than assumed. Knowledge of the mutation processes at microsatellite loci is currently insufficient. Several factors have been found to be relevant to the evolution of these repeated sequences, such as asymmetry in the distribution of mutations, dependence of the mutation rate on the number of repeats, and purity of alleles or constraints on allele size (ESTOUP and

CORNUET 1999). The method described here is independent from the mutation model of the marker used. The advantage of this method is that it can be applied to various types of markers (*e.g.*, phenotypical markers, proteins, blood groups, microsatellites, amplified fragment length polymorphisms, single nucleotide polymorphisms . . .). Because of the heterogeneity of the genome, it is more realistic to use different genetic systems. This will decrease the potential of misinterpretation when investigating phylogeographic processes. It could be particularly useful in the context of protection of the biodiversity of domestic species, where the usefulness of molecular data alone in the identification of priorities for conservation is under debate (RUANE 1999). As advocated by CREPALDI *et al.* (2001), the combined use of random neutral markers and expressed genes will allow a global assessment of the genome that might complement information of quantitative and unique traits and serve as a rational basis for an objective choice of the genetic resources to be conserved. In a context where genomic heterogeneity is no longer a negative but an interesting parameter, applying methods that permit different compromise typologies to be revealed would be very appealing.

We acknowledge the assistance of the respective breeders associations in the collection of French cattle samples. We acknowledge the following persons for their help in planning and conducting the

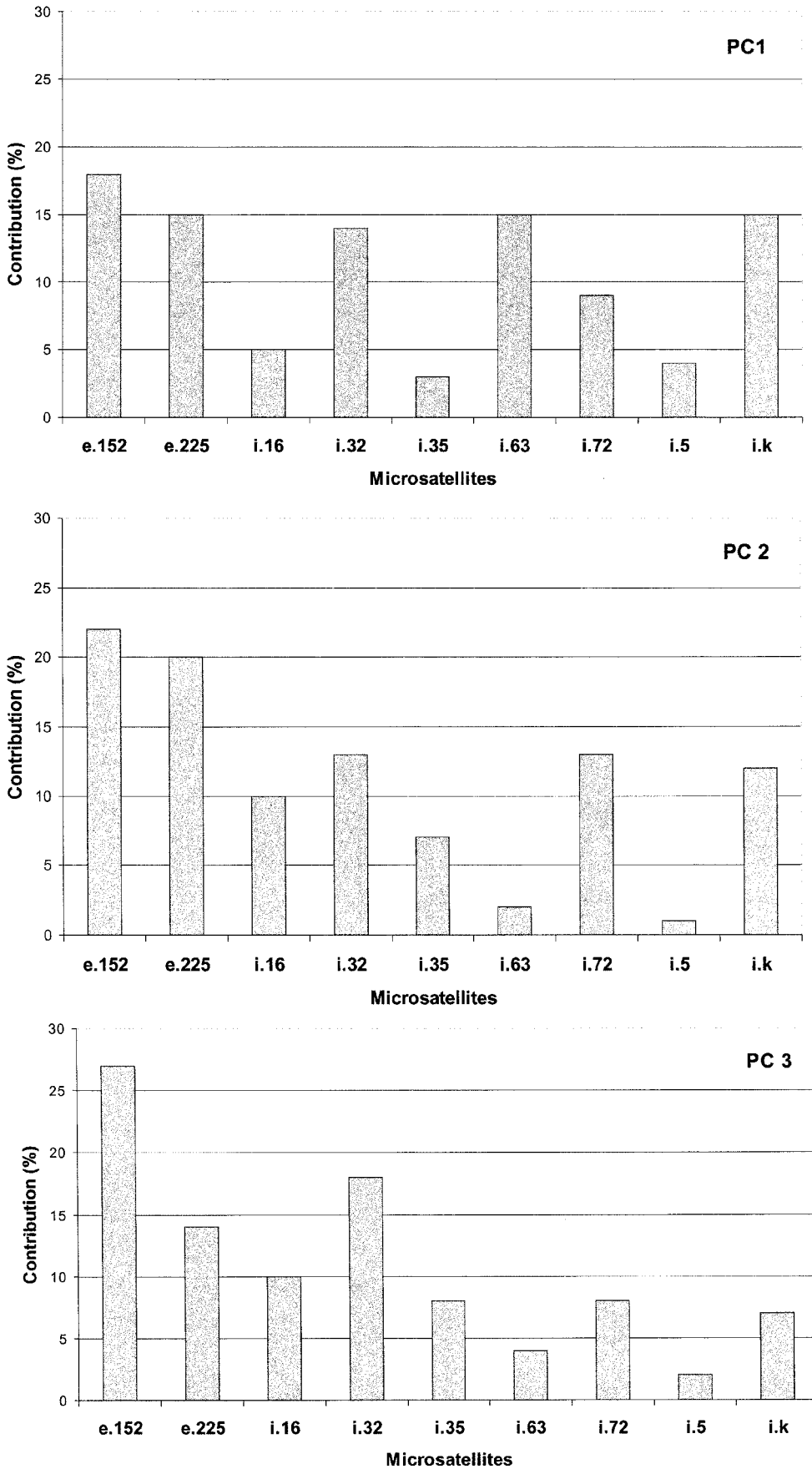


FIGURE 4.—Diagrams of the contribution of markers for the first three PCs for data set 2.

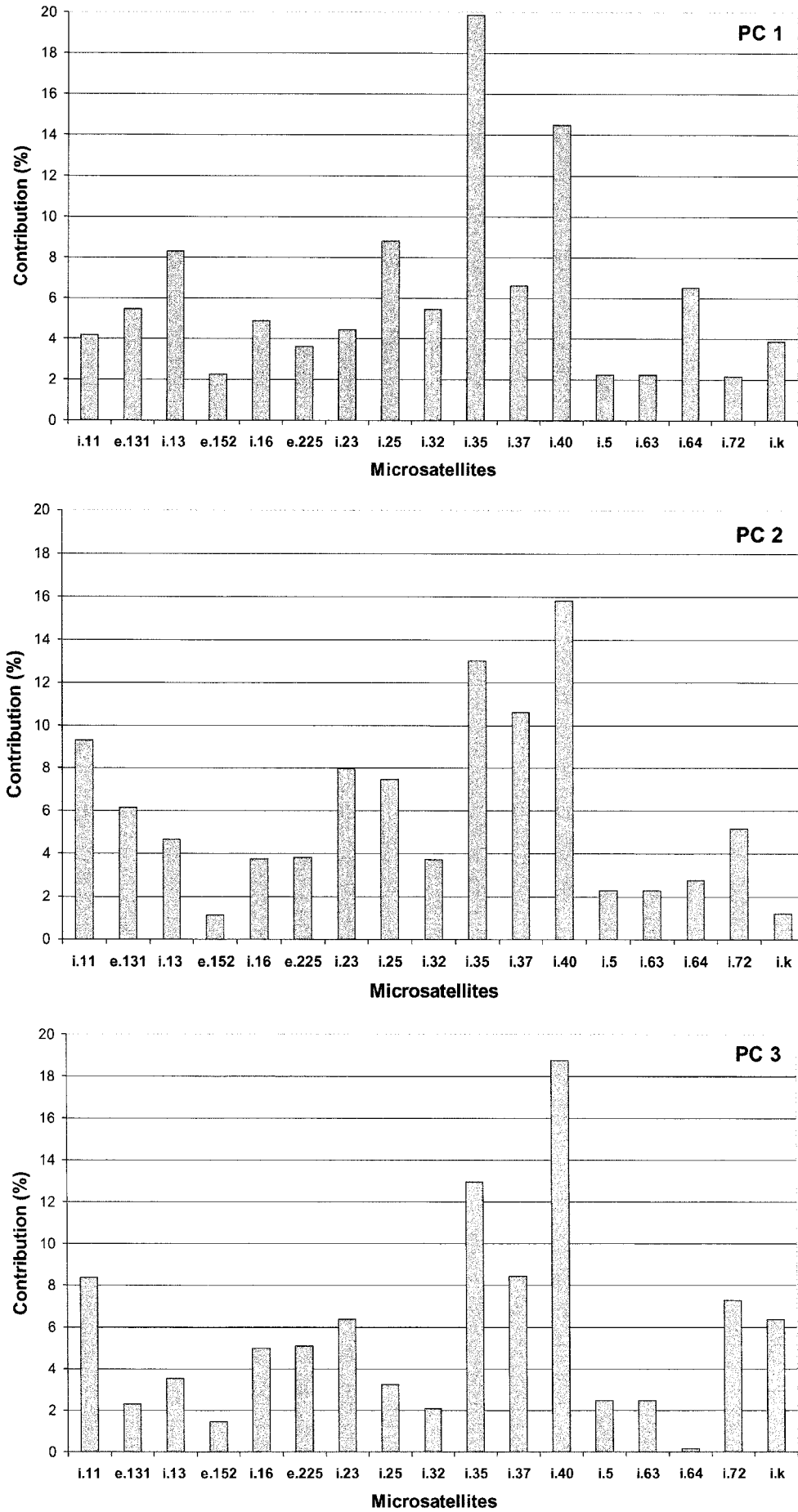


FIGURE 5.—Diagrams of the contribution of markers for the first three PCs for data set 1.

sampling missions for African samples: V. Codja (Bénin), N. T. Kouagou (Togo), I. Sidibé (Burkina-Faso), and P. Souvenir Zafindrajaona (Chad and Madagascar).

## LITERATURE CITED

- AGRESTI, A., 1992 A survey of exact inference for contingency tables (with discussion). *Stat. Sci.* **7**: 131–177.
- AYALA, F. J., A. ESCALANTE, C. O'HUIGIN and J. KLEIN, 1994 Molecular genetics of speciation and human origins. *Proc. Natl. Acad. Sci. USA* **91**: 6787–6794.
- BARKER, J. S., D. G. BRADLEY, R. FRIES, W. G. HILL, M. NEI *et al.*, 1993 An integrated global programme to establish the genetic relationships among the breeds of each domestic animal species. FAO Report, Rome.
- BARKER, J. S., S. S. MOORE, D. J. S. HETZEL, D. EVANS, S. G. TAN *et al.*, 1997 Genetic diversity of Asian water buffalo (*Bubalus bubalis*): microsatellite variation and a comparison with protein-coding loci. *Anim. Genet.* **28**: 103–115.
- BJORNSTAD, G., and K. H. ROED, 2001 Breed demarcation and potential for breed allocation of horses assessed by microsatellite markers. *Anim. Genet.* **32**: 59–65.
- BLOTT, S. C., J. L. WILLIAMS and C. S. HALEY, 1998 Genetic relationships among European cattle breeds. *Anim. Genet.* **29**: 273–282.
- BOSCH, E., F. CALAFELL, A. PEREZ-LEZAUN, D. COMAS, E. MATEU *et al.*, 1997 Population history of north Africa: evidence from classical genetic markers. *Hum. Biol.* **69**(3): 295–311.
- BOWCOCK, A. M., A. RUIZ-LINARES, J. TOMFORHDE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455–457.
- CAÑON, J., M. L. CHECA, C. CARLEOS, J. L. VEGA-PLA, M. VALLEJO *et al.*, 2000 The genetic structure of Spanish Celtic horse breeds inferred from microsatellite data. *Anim. Genet.* **31**: 39–48.
- CAÑON, J., P. ALEXANDRINO, I. BESSA, C. CARLEOS, Y. CARRETERO *et al.*, 2001 Genetic diversity measures of local European beef cattle breeds for conservation purposes. *Genet. Sel. Evol.* **33**: 311–332.
- CARROLL, J. D., 1981 INDSCAL, pp. 371–389 in *Introduction to Multidimensional Scaling*, edited by S. S. SCHIFFMAN, M. L. REYNOLDS and F. W. YOUNG. Academic Press, New York.
- CAVALLI-SFORZA, L. L., 1997 Genes, peoples, and languages. *Proc. Natl. Acad. Sci. USA* **94**: 7719–7724.
- CAVALLI-SFORZA, L. L., and A. PIAZZA, 1993 Human genomic diversity in Europe: a summary of recent research and prospects for the future. *Eur. J. Hum. Genet.* **1**: 3–18.
- CAVALLI-SFORZA, L. L., P. MENOZZI and A. PIAZZA, 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.
- CHEN, K. H., H. CANN, T. C. CHEN, B. VAN WEST and L. CAVALLI-SFORZA, 1985 Genetic markers of an aboriginal Taiwanese population. *Am. J. Phys. Anthropol.* **66** (3): 327–337.
- CONOVER, W. J., 1980 *Practical Nonparametric Statistics*, Ed. 2. John Wiley & Sons, New York.
- CORNUET, J. M., S. PIRY, G. LUIKART, A. ESTOUP and M. SOLIGNAC, 1999 New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**: 1989–2000.
- CREPALDI, P., R. NEGRINI, E. MILANESI, C. GORNI, M. CICOGLIA *et al.*, 2001 Diversity in five goat populations of the Lombardy Alps: comparison of estimates obtained from morphometric traits and molecular markers. *J. Anim. Breed. Genet.* **118**: 173–180.
- CYMBRON, T., R. T. LOFTUS, M. I. MALBEIRO and D. G. BRADLEY, 1999 Mitochondrial sequence variation suggests an African influence in Portuguese cattle. *Proc. R. Soc. Lond. Ser. B* **266**: 597–603.
- EFRON, B., E. HALLORAN and S. HOLMES, 1996 Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **93**: 7085–7090.
- ESTOUP, A., and J. M. CORNUET, 1999 Microsatellite evolution: inference from population data, pp. 49–65 in *Microsatellites, Evolution and Applications*, edited by D. B. GOLDSTEIN and C. SCHLÖTTERER. Oxford University Press, Oxford.
- ESTOUP, A., L. GARNERY, M. SOLIGNAC and J. M. CORNUET, 1995 Microsatellite variation in honey bee (*Apis Mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**: 679–695.
- FELSENSTEIN, J., 1982 How can we infer geography and history from gene frequencies? *J. Theor. Biol.* **96**: 9–20.
- FELSENSTEIN, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **29**: 783–791.
- FOULLEY, J. L., and W. G. HILL, 1999 On the precision of estimation of genetic distance. *Genet. Sel. Evol.* **31**: 457–464.
- GOLDSTEIN, D. B., G. W. ROEMER, D. A. SMITH, D. E. REICH, A. BERGMAN *et al.*, 1999 The use of microsatellite variation to infer population structure and demographic history in a natural model system. *Genetics* **151**: 797–801.
- GROSCLAUDE, F., R. Y. AUPETIT, J. LEFEBVRE and J. C. MÉRIAUX, 1990 Essai d'analyse des relations génétiques entre les races bovines françaises à l'aide du polymorphisme biochimique. *Genet. Sel. Evol.* **22**: 317–338.
- HANSLIK, S., B. HARR, G. BREM and C. SCHLOTERRER, 2000 Microsatellite analysis reveals substantial genetic differentiation between contemporary new world and old world Holstein Friesian populations. *Anim. Genet.* **31**: 31–38.
- HILL, M. O., 1974 Correspondence analysis: a neglected multivariate technique. *J. R. Stat. Soc. Ser. C* **23**: 340–354.
- JORDE, L., A. R. ROGERS, M. BAMSHAD, W. S. WATKINS, P. KRAKOWIAK *et al.*, 1997 Microsatellite diversity and the demographic history of modern humans. *Proc. Natl. Acad. Sci. USA* **94**: 3100–3103.
- KOSKINEN, M. T., and P. BREDBACKA, 2000 Assessment of the population structure of five Finnish dog breeds with microsatellites. *Anim. Genet.* **31**: 310–317.
- LAVAL, G., N. IANUCCELLI, C. LEGAULT, D. MILAN, M. A. M. GROENEN *et al.*, 2000 Genetic diversity of eleven European pig breeds. *Genet. Sel. Evol.* **32**: 187–203.
- LI, S., K. P. PEARL and H. DOSS, 2000 Phylogenetic tree construction using Markov chain Monte Carlo. *J. Am. Stat. Assoc.* **95**: 493–508.
- MACHUGH, D. E., M. D. SHRIVER, R. T. LOFTUS, P. CUNNINGHAM and D. G. BRADLEY, 1997 Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* **146**: 1071–1086.
- MAHÉ, M. F., G. MIRANDA, R. QUEVAL, A. BADO, P. SOUVENIR ZAFINDRAJAONA *et al.*, 1999 Genetic polymorphism of milk proteins in African *Bos taurus* and *Bos indicus* populations. Characterization of variants  $\alpha_1$ -Cn H and  $\kappa$ -Cn J. *Genet. Sel. Evol.* **31**: 239–253.
- MANLY, B. F. J., 1997 *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, London.
- MANTEL, N., 1967 The detection of disease clustering and generalized regression approach. *Cancer Res.* **27**: 209–220.
- MARDIA, K. V., J. T. KENT and J. M. BIBBY, 1979 *Multivariate Analysis*, pp. 213–246. Academic Press, New York.
- MOAZAMI-GOUDARZI, K., D. LALOË, J. P. FURET and F. GROSCLAUDE, 1997 Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites. *Anim. Genet.* **28**: 338–345.
- MOAZAMI-GOUDARZI, K., S. PIRY, D. LALOË, M. SOLIGNAC and J. M. CORNUET, 1998 Breed assignment in cattle using microsatellites. Twenty-sixth International Conference on Animal Genetics, Auckland, New Zealand.
- MORRISON, D. F., 1976 *Multivariate Statistical Methods*, Ed. 2. Wiley, New York.
- PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- RUANE, J., 1999 A critical review of the value of genetic distance studies in conservation of animal genetic resources. *J. Anim. Breed. Genet.* **116**: 317–323.
- SAHA, N., and J. S. TAY, 1992 Origin of the Koreans: a population genetic study. *Am. J. Phys. Anthropol.* **88**: 27–36.
- SAHA, N., J. W. MAK, J. S. TAY, Y. LIU, J. A. TAN *et al.*, 1995 Population genetic study among the Orange Asli (Semai Senoi) of Malaysia: Malayan aborigines. *Hum. Biol.* **67** (1): 37–57.
- SAITBEKOVA, N., C. GAILLARD, G. OBEXER-RUFF and G. DOLF, 1999 Genetic diversity in Swiss goat breeds based on microsatellite analysis. *Anim. Genet.* **30**: 36–41.
- SAS INSTITUTE, 2000 *SAS/STAT, User's Guide, Version 8*. SAS Institute, Cary, NC.
- SCHLÖTTERER, C., and T. WIEHE, 1999 Microsatellite, a neutral marker to infer selective sweeps, pp. 238–248 in *Microsatellites, Evolution and Applications*, edited by D. B. GOLDSTEIN and C. SCHLÖTTERER. Oxford University Press, Oxford.

- SOUVENIR ZAFINDRAJONA, P., V. ZEUB, K. MOAZAMI-GOUDARZI, D. LALOË, D. BOURZAT *et al.*, 1999 Study on the phylogenetic status of Lake Chad Kuri cattle using molecular markers. *Revue Elev. Méd. Vét. Pays Trop.* **52** (2): 155–162.
- TAKEZAKI, N., and M. NEI, 1996 Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144**: 389–399.
- UNDERHILL, P. A., P. SHEN, A. A. LIN, L. JIN, G. PASSARINO *et al.*, 2000 Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26**: 358–361.
- VILÀ, C., J. A. LEONARD, A. GÖTHERSTRÖM, S. MARKLUND, K. SANDBERG *et al.*, 2001 Widespread origins of domestic horse lineages. *Science* **291**: 474–477.
- WIMMERS, K., S. PONSUKSILI, T. HARDGE, A. VALLE-ZARATE, P. K. MARTUR *et al.*, 2000 Genetic distinctness of African, Asian and South American local chickens. *Anim. Genet.* **31**: 159–165.
- YANG, L., S. H. ZHAO, K. LI, Z. Z. PENG and G. W. MONTGOMERY, 1999 Determination of genetic relationships among five indigenous Chinese goat breeds with six microsatellite markers. *Anim. Genet.* **30**: 452–455.

Communicating editor: C. HALEY