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

K. Moazami-Goudarzi* and D. Laloe†

*Laboratoire de Génétique Biochimique et de Cytogénétique, INRA, 78352 Jouy-en-Josas, France and †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.

Analysis 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 procedures 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; Bösch 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; Cañón et al. 2000; Hanslík 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-

---

1Corresponding author: Station de Génétique Quantitative et Appliquée, INRA, Domaine de Vilvert, 78352 Jouy-en-Josas, France. E-mail: ugendla@dg2.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 Mar氏 et al. (1979).

Let us consider g populations and an n-allelic marker. \( p_k \) 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 aflari standardization of the allelic frequencies \( p_k \) 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 \( x_k = (p_k - \bar{p})/\sigma_k \). This standardization is used throughout this article.

Each population is represented by a point in an \( n \)-dimensional space, the coordinates of which are the \( n \) 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 topologies, trees, or multidimensional plots. The first problem is to evaluate the relationship among these topologies, which is commonly called “interstructure.” Use of PCA leads to an implicit euclidean distance between populations \( i \) and \( j \):

\[
d(i, j) = \sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2}.
\]

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 \( g \) \(-1/2 \) 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., Mandia 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 \( m(m-1)/2 \) 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 \( S_{\text{obs}} \), (2) random permutation of entire rows of each table \( X \) of standardized allelic frequencies, and (3) recalculation of \( S_{\text{rand}} \). The probability that no compromise exists is calculated as \((\text{number of } S_{\text{rand}} \geq S_{\text{obs}} + 1)/\text{(number of randomizations + 1)}\). The \( 1 \) in the numerator 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 \( d_{ij} \) between populations \( i \) and \( j \) computed from the allelic frequencies of marker \( k \) as a result of the following one-way linear model:

\[
d_{ij} = d_k + e_{ij}.
\]

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 \( d_k \).

The null hypothesis \( H0 \) is then: All distances are equal. Conversely, if some typology exists, the alternative hypothesis \( H1 \) 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 NPAR1WAY 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, Soudanese Fulani zebu, Red Bororo zebu, Shuwa zebu, Madagascar zebu, and Borgou (Shorthorn × 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; Soudanese 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 × zebu crossbreds), and Soudanese 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 Zafindrajaona 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 ($P < 10^{-5}$).

**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 Kruskall-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 typings 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 Kruskall-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 Zafindrajaona 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 southwestern 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 1
Mantel correlations (×1000) for data set 1

<table>
<thead>
<tr>
<th></th>
<th>i.11</th>
<th>e.131</th>
<th>i.13</th>
<th>e.152</th>
<th>i.16</th>
<th>e.225</th>
<th>i.23</th>
<th>i.25</th>
<th>i.32</th>
<th>i.35</th>
<th>i.37</th>
<th>i.40</th>
<th>i.5</th>
<th>i.63</th>
<th>i.64</th>
<th>i.72</th>
<th>i.k</th>
</tr>
</thead>
<tbody>
<tr>
<td>i.11</td>
<td>1000</td>
<td>205</td>
<td>525</td>
<td>-64</td>
<td>-75</td>
<td>187</td>
<td>-145</td>
<td>-22</td>
<td>-287</td>
<td>372</td>
<td>-93</td>
<td>364</td>
<td>-182</td>
<td>314</td>
<td>130</td>
<td>-175</td>
<td>-168</td>
</tr>
<tr>
<td>e.131</td>
<td>1000</td>
<td>239</td>
<td>147</td>
<td>249</td>
<td>308</td>
<td>58</td>
<td>24</td>
<td>231</td>
<td>-163</td>
<td>34</td>
<td>-26</td>
<td>200</td>
<td>-130</td>
<td>90</td>
<td>139</td>
<td>-143</td>
<td></td>
</tr>
<tr>
<td>i.16</td>
<td>1000</td>
<td>186</td>
<td>-113</td>
<td>-80</td>
<td>-421</td>
<td>163</td>
<td>18</td>
<td>135</td>
<td>94</td>
<td>121</td>
<td>-314</td>
<td>353</td>
<td>12</td>
<td>-487</td>
<td>35</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td>e.152</td>
<td>1000</td>
<td>-206</td>
<td>-18</td>
<td>140</td>
<td>250</td>
<td>127</td>
<td>-395</td>
<td>-212</td>
<td>74</td>
<td>-99</td>
<td>-263</td>
<td>-271</td>
<td>-275</td>
<td>171</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.25</td>
<td>1000</td>
<td>276</td>
<td>452</td>
<td>240</td>
<td>-346</td>
<td>-334</td>
<td>210</td>
<td>-435</td>
<td>102</td>
<td>-76</td>
<td>347</td>
<td>423</td>
<td>-176</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.23</td>
<td>1000</td>
<td>247</td>
<td>-54</td>
<td>-179</td>
<td>-89</td>
<td>-154</td>
<td>-36</td>
<td>414</td>
<td>-146</td>
<td>6</td>
<td>172</td>
<td>-207</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.32</td>
<td>1000</td>
<td>387</td>
<td>-256</td>
<td>-241</td>
<td>224</td>
<td>-329</td>
<td>-89</td>
<td>-286</td>
<td>90</td>
<td>402</td>
<td>-75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.35</td>
<td>1000</td>
<td>-107</td>
<td>-259</td>
<td>367</td>
<td>-216</td>
<td>-338</td>
<td>-87</td>
<td>37</td>
<td>-22</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.37</td>
<td>1000</td>
<td>172</td>
<td>347</td>
<td>-202</td>
<td>612</td>
<td>-135</td>
<td>-289</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.40</td>
<td>1000</td>
<td>-321</td>
<td>-331</td>
<td>351</td>
<td>48</td>
<td>-84</td>
<td>-116</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.63</td>
<td>1000</td>
<td>94</td>
<td>199</td>
<td>-141</td>
<td>-295</td>
<td>295</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.64</td>
<td>1000</td>
<td>-85</td>
<td>-438</td>
<td>-213</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.72</td>
<td>1000</td>
<td>349</td>
<td>-102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.k</td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

i code corresponds to INRA microsatellites and e code corresponds to ETH microsatellites.
TABLE 2
Mantel correlations (×1000) for data set 2

<table>
<thead>
<tr>
<th>Microsatellites</th>
<th>e.152</th>
<th>e.225</th>
<th>i.16</th>
<th>i.32</th>
<th>i.35</th>
<th>i.63</th>
<th>i.72</th>
<th>i.5</th>
<th>i.k</th>
</tr>
</thead>
<tbody>
<tr>
<td>e.152</td>
<td>1000</td>
<td>747</td>
<td>306</td>
<td>468</td>
<td>124</td>
<td>247</td>
<td>404</td>
<td>285</td>
<td>607</td>
</tr>
<tr>
<td>e.225</td>
<td>1000</td>
<td>468</td>
<td>569</td>
<td>152</td>
<td>452</td>
<td>481</td>
<td>344</td>
<td>422</td>
<td></td>
</tr>
<tr>
<td>i.16</td>
<td>1000</td>
<td>266</td>
<td>−86</td>
<td>244</td>
<td>446</td>
<td>83</td>
<td>245</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.32</td>
<td>1000</td>
<td>−2</td>
<td>453</td>
<td>510</td>
<td>335</td>
<td>301</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.35</td>
<td>1000</td>
<td>106</td>
<td>−108</td>
<td>−230</td>
<td>−31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.63</td>
<td>1000</td>
<td>328</td>
<td>228</td>
<td>213</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.72</td>
<td>1000</td>
<td>350</td>
<td>360</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i.k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Congruence of Population Structure

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 (BovMap 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-GONDONZI 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, Sattbekova 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
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 microsatellites. 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
were obtained. Increasing the number of loci decreased the reliability of the phylogeny. Lavall 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; Cañón 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. Cañón 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. Vila et al.
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. Codija (Bénin), N. T. Kouagou (Togo), I. Sidibé (Burkina-Faso), and P. Souvenir Zafindrajaona (Chad and Madagasgar).

LITERATURE CITED


MacHugh, D. E., M. D. Shriver, R. T. Loftus, P. Cunningham and D. G. Bradly, 1997 Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (Bos taurus and Bos indicus). Genetics 146: 1071–1086.


Takezaki, N., and M. Nei, 1996 Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144: 389–399.


Communicating editor: C. Haley