Empirical Evaluation of Genetic Clustering Methods Using Multilocus Genotypes From 20 Chicken Breeds

Noah A. Rosenberg,* Terry Burke,[†,1] Kari Elo,[‡] Marcus W. Feldman,[*,1] Paul J. Freidlin,[§,1] Martien A. M. Groenen,[,1] Jossi Hillel,[§,1] Asko Mäki-Tanila,[‡,1] Michèle Tixier-Boichard,[††,1] Alain Vignal,[‡‡,1] Klaus Wimmers[§§,1] and Steffen Weigend[***]**

**Department of Biological Sciences, Stanford University, Stanford, California 94305, [†]Department of Animal and Plant Sciences, Sheffield University, S10 2TN, United Kingdom, [§]Department of Genetics, The Hebrew University of Jerusalem, Faculty of Agriculture, Rehovot 76100, Israel, **Institute of Animal Sciences, Wageningen Agricultural University, 6700 AH Wageningen, The Netherlands, [‡]Agricultural Research Centre, Institute of Animal Production, FIN-31600 Jokioinen, Finland, [††]Institut National de la Recherche Agronomique, Centre de Recherches de Jouy-en-Josas, 78 352 Jouy-en-Josas Cedex, France, [‡‡]Institut National de la Recherche Agronomique, Centre INRA de Toulouse, 31326 Castanet Tolosan, France, [§§]Institute of Animal Breeding Science, Rheinische Friedrich-Wilhelms-Universitat, D-53012 Bonn, Germany and ***Institute for Animal Science and Animal Behaviour, Mariensee, 31535 Neustadt, Germany*

Manuscript received March 26, 2001
Accepted for publication August 1, 2001

ABSTRACT

We tested the utility of genetic cluster analysis in ascertaining population structure of a large data set for which population structure was previously known. Each of 600 individuals representing 20 distinct chicken breeds was genotyped for 27 microsatellite loci, and individual multilocus genotypes were used to infer genetic clusters. Individuals from each breed were inferred to belong mostly to the same cluster. The clustering success rate, measuring the fraction of individuals that were properly inferred to belong to their correct breeds, was consistently ~98%. When markers of highest expected heterozygosity were used, genotypes that included at least 8–10 highly variable markers from among the 27 markers genotyped also achieved >95% clustering success. When 12–15 highly variable markers and only 15–20 of the 30 individuals per breed were used, clustering success was at least 90%. We suggest that in species for which population structure is of interest, databases of multilocus genotypes at highly variable markers should be compiled. These genotypes could then be used as training samples for genetic cluster analysis and to facilitate assignments of individuals of unknown origin to populations. The clustering algorithm has potential applications in defining the within-species genetic units that are useful in problems of conservation.

CHARACTERIZATIONS of the population structure of species are useful in a variety of contexts. Genetic ascertainment of within-species population structure has been widely applied for classifying subspecies, for defining intraspecific conservation units, for understanding events in the history of a species, for identifying ongoing speciation events, and for testing hypotheses about evolutionary processes. In other situations, the presence of population structure poses a practical nuisance. For example, allele frequencies in reference groups are central to calculations in forensic studies, and it is difficult to identify appropriate reference groups in structured populations (National Research Council 1996). In case-control studies that test for statistical associations between a genotype at a particular locus and a phenotype, not taking into account population structure can lead to the false detection of associations (*e.g.*, Devlin and Roeder 1999).

Corresponding author: Noah A. Rosenberg, Program in Molecular and Computational Biology, University of Southern California, 1042 W. 36th Pl., DRB 155, Los Angeles, CA 90089-1113.
E-mail: noahr@usc.edu

[1]These authors are listed alphabetically.

Population structure assessment has often relied upon *a priori* groupings of individuals on the basis of phenotypes or sampling locations. A classification chosen by an investigator, however, might not accurately describe the genetic structure of the populations. Genetically similar groups of individuals might be labeled differently due to distinct geography, different phenotypes, or, in the case of human groups, cultural differences; however, a high level of geographic, phenotypic, or cultural diversity among a collection of populations need not imply that the groups are genetically divergent. Conversely, geographic overlap or phenotypic similarity may mask underlying genetic variation. Thus, a purely genetic analysis using no external information provides the most direct method of determining population structure. Only if a correspondence between genetic and geographic or phenotypic classifications is established can these characteristics also serve as appropriate classification tools.

The *structure* algorithm (Pritchard *et al.* 2000) constructs genetic clusters from a collection of individual multilocus genotypes, estimating for each individual the fractions of its genome that belong to each cluster. In contrast to methods that use genetic distances, sampling

locations, hypothesized genetic origins of individuals, phenotypic information, and the number of genetic clusters do not need to be specified before the algorithm is applied. With extensive simulations, Pritchard *et al.* (2000) demonstrated that the *structure* genetic cluster analysis method can accurately infer individual ancestries. For two data sets, one in which two genetic clusters were inferred and another for which three were inferred, they found that the inferred and expected population structures were roughly coincident.

In this article, we consider the utility of genetic cluster analysis on a large data set for which population structure is known, with the aim of making recommendations about its future uses. We employ a collection of 27-locus genotypes from 600 individuals representing 20 chicken breeds. This data set is substantially larger than previous data sets on which *structure* has been applied (Pritchard *et al.* 2000; Beaumont *et al.* 2001; Rosenberg *et al.* 2001) and it includes individuals from a larger number of genetic populations. Importantly, isolation of the breeds in different locations allows us to be sure that, in most cases, these breeds have been genetically separated from each other for at least 20–50 generations, so that we can test if cluster analysis successfully uncovers this genetic structure.

We first characterize the genetic differences among the populations. We then demonstrate that genetic cluster analysis has great ability to correctly ascertain the population structure for these data, and we compare the cluster analysis to a cladogram derived from the neighbor-joining algorithm. To assess the success of clustering as a function of the number of markers, we consider subsets of the loci chosen by different criteria of variability. We also consider the success of clustering as a function of the number of individuals used per population. Finally, we discuss recommendations on the use of genetic cluster analysis for ascertaining population structure, for applications in the assignment of individuals of unknown origin to populations, and for identifying genetically distinctive populations.

MATERIALS AND METHODS

Breeds: We genotyped 30 individuals from each of 20 breeds. These breeds form a subset of the populations studied in a survey of European chicken genetic diversity (Hillel *et al.* 1999; Tixier-Boichard *et al.* 1999; Weigend 1999). The breeds, which are designated by the same code numbers as in other studies (Hillel *et al.* 1999), represent five general classes, as described by Tixier-Boichard *et al.* (1999): feral (*Gallus gallus gallus* [102]); traditional unselected breeds of the Middle East (*Bedouin* [5]) and Northern Europe (*Icelandic landrace* [16]); traditional breeds selected for morphological traits and deriving from Central Europe (*Green-legged partridge* [27], *Orlov* [28], *Transylvanian naked neck* [26]), from the Mediterranean region (*Fayoumi* [4], *Old Scandinavian reference population* [18], *Padovana* [21]), from Northern Europe (*Jaerhoens* [19]), and from Western Europe (*Marans* [13]); lines selected for quantitative traits or economic indices, including commercial broilers (*Broiler dam line D* [50], *Broiler sire line B* [42]), experimental lines (*Godollo Nhx* [33], *High-Ab line* [51], *Sarcoma-susceptible line* [3402]), and commercial layers (*Brown-egg layer line C* [44], *Brown-egg layer line D* [45], *White-egg layer line A* [37]); and a highly inbred strain of white leghorn origin maintained at low population size (*C line* [32]).

Markers: Genotypes were used for 27 microsatellite markers spread across the chicken genome (listed in Table 1). Except for ADL278, LEI94, LEI166, LEI194, LEI228, and LEI234, it has previously been reported that these markers show high levels of polymorphism within and between breeds (Hillel *et al.* 1999). In general, pairs of markers among these are unlinked (see Groenen *et al.* 2000 for a map).

Genotyping: Genotyping was performed in the laboratories of T. Burke, M. A. M. Groenen, J. Hillel, and S. Weigend, with similar procedures used in all labs. The example procedure that follows is from the laboratory of S. Weigend. PCR products were obtained in a 25-μl volume using Ready-To-Go PCR Beads (no. 27-9555-01; Amersham Pharmacia Biotech Europe, Freiburg, Germany) and a thermal cycler (Mastercycler; Eppendorf, Hamburg, Germany). Two pairs of microsatellite primers were run in one tube. Each PCR tube contained 20 ng of genomic DNA, 10 pmol of each forward primer labeled with either IRD700 or IRD800 (MWG-Biotech, Ebersberg, Germany), 10 pmol of each unlabeled reverse primer, and 1 mm tetramethylammoniumchloride. The amplification involved initial denaturation at 95° (1 min), 35 cycles of denaturation at 95° (1 min), primer annealing at temperatures varying between 58° (1 min) and extension at 72° (1 min), followed by final extension at 72° (10 min). Specific DNA fragments produced by amplification were visualized as bands by 8% PAGE, which was performed with a LI-COR automated DNA analyzer (LI-COR Biotechnology Division, Lincoln, NE 68504). Electrophoregram processing and allele-size scoring were performed with the RFLPscan package (Scanalytics, Division of CSP, Billerica, MA).

Missing data: The proportion of missing data was 0.8%, and 12 of 27 loci had missing genotypes. For no locus were >3.5% of the possible genotypes missing. Missing genotypes were distributed across 88 individuals from 18 breeds. For no breed were >4.1% of its genotypes missing. Out of 600 individuals, 13 individuals originating from 6 breeds did not have available genotypes at >1 locus. These 13 individuals included 1 individual that was lacking genotypes at 9 loci and 3 individuals that were missing genotypes at 10 loci.

Statistical analysis: *Genetic differentiation:* For each pair of breeds, allele frequencies were tabulated at each locus, sequentially pooling the rarest alleles into one allelic class, until the average frequency for the two breeds exceeded 0.1 for each class. A chi-square association test statistic was computed for each locus, with the number of degrees of freedom equaling one fewer than the number of allelic classes. We counted how many loci produced test statistics below the 0.001 level.

Genetic distance between breeds was calculated using the negative logarithm of the proportion of shared alleles (PSA) in the two breeds (Bowcock *et al.* 1994), as implemented in *microsat* (Minch *et al.* 1998). For each locus, this measure sums the lower of the corresponding allele frequencies in the two breeds across all alleles. The sums are then averaged across loci, yielding an overall proportion of shared alleles. Note that this generalized PSA distance is based on allele frequencies rather than individual genotypes and, thus, it assumes independence between the two alleles of an individual at a given locus.

Clustering of breeds: Population structure was studied using two methods. First, we obtained an unrooted neighbor-joining cladogram (Saitou and Nei 1987) based on the PSA genetic distance matrix between populations, using the *neighbor* pro-

TABLE 1

Values of diversity statistics for each marker and rankings of markers according to the highest number of alleles, highest expected heterozygosity, highest values of F_{st}, and a random ordering

Marker	Rank based on expected heterozygosity	Rank based on total no. of alleles	Rank based on F_{st}	Rank based on random ordering	Expected heterozygosity	Total no. of alleles (no. of private alleles)	F_{st}
LEI228	1	1	21	21	0.924	41 (17)	0.281
LEI234	2	3	9	26	0.892	23 (9)	0.334
LEI194	3	7	14	4	0.885	15 (5)	0.314
LEI192	4	2	22	15	0.866	37 (15)	0.255
MCW34	5	5	26	6	0.859	18 (4)	0.228
LEI94	6	4	11	10	0.828	23 (9)	0.330
MCW206	7	6	7	27	0.779	16 (4)	0.341
ADL268	8	18	15	11	0.767	7 (1)	0.309
MCW183	9	11	6	18	0.739	11 (4)	0.346
MCW295	10	13	18	13	0.718	9 (1)	0.295
ADL278	11	19	3	3	0.677	7 (2)	0.376
MCW67	12	21	2	24	0.674	6 (2)	0.417
MCW37	13	23	12	17	0.673	4 (0)	0.320
MCW69	14	14	20	7	0.672	9 (1)	0.282
LEI166	15	22	8	5	0.668	5 (1)	0.337
MCW81	16	15	1	8	0.668	9 (3)	0.501
ADL112	17	17	24	20	0.628	8 (0)	0.247
MCW216	18	20	5	25	0.622	6 (1)	0.347
MCW78	19	16	25	1	0.613	8 (2)	0.228
MCW222	20	24	13	9	0.590	4 (0)	0.315
MCW14	21	10	17	19	0.576	12 (3)	0.297
MCW284	22	25	23	2	0.576	4 (0)	0.254
MCW111	23	12	16	14	0.551	11 (5)	0.303
MCW330	24	9	10	23	0.499	14 (5)	0.331
MCW98	25	26	19	12	0.476	3 (1)	0.287
MCW103	26	27	4	16	0.438	2 (0)	0.373
MCW248	27	8	27	22	0.421	14 (6)	0.189

Ties for the same number of alleles were broken by ranking the average number of alleles per population from largest to smallest.

gram (Felsenstein 1993) to construct the cladogram. We performed 1000 bootstraps across the set of loci to obtain a consensus cladogram.

The second approach utilized the program *structure*, which identifies clusters of related individuals from multilocus genotypes (Pritchard *et al.* 2000). First, we performed many runs of various lengths with different proposals for the number of genetic clusters (K) represented by the individuals genotyped, testing all values of K from 1 to 23. Clustering solutions of highest likelihood were obtained when the vast majority of genomic assignment was distributed over exactly 17, 18, or 19 clusters. We did not observe clustering solutions in which >19 clusters were assigned nontrivial fractions of the data. To choose the best value of K, we ran *structure* 20 times for 50,000 steps, after a burn-in period of 5000 steps, using each of $K = 17$, $K = 18$, and $K = 19$. Using the Wilcoxon two-sample test, both $K = 18$ and $K = 19$ produced higher likelihood solutions than $K = 17$ (two-sided $P = 0.03$ for $K = 18$ *vs.* $K = 17$; two-sided $P = 0.04$ for $K = 19$ *vs.* $K = 17$). For $K = 18$ and $K = 19$, solutions had similar likelihoods (two-sided $P = 0.86$). However, since runs with $K = 19$ occasionally produced solutions of *particularly* high likelihood that distributed individuals over all 19 clusters, 18 was insufficient for maximal clustering, and we used $K = 19$ for all subsequent analyses. Runs used in the determination of K were not considered in further analysis.

Evaluation of cluster analysis: Each individual was assigned to a specific breed using *structure* (Pritchard *et al.* 2000), following the five-step algorithm in Figure 1. In step 1, we chose the value of K, as described above. The aim of the remaining steps was to assign individuals to breeds and to evaluate the fraction of individuals correctly assigned. In step 2, we clustered individuals and associated each individual with the cluster that corresponded to the greatest fraction of its genome. In step 3, we associated breed labels with each of the inferred genetic clusters. In cases for which a cluster was labeled with multiple breeds, this step required additional subclustering runs of *structure*. These runs used only those individuals that were assigned to that cluster in step 2, and they lasted 20,000 iterations with a burn-in period of 5000. For subclustering runs, K equaled the number of breeds associated with the cluster.

Once the individuals were clustered to the greatest extent possible at the conclusion of step 3, we followed step 4 to assign each individual to a single breed. The "clustering success rate" (step 5) was then defined as the proportion of individuals correctly assigned to their breeds of origin.

Note that we assumed that individuals were maximally clustered after step 3. This assumption avoided additional subclustering runs: In principle, a cluster C that was associated only with breed B in step 3 might have been decomposable into subclusters. However, each of the resulting subclusters would

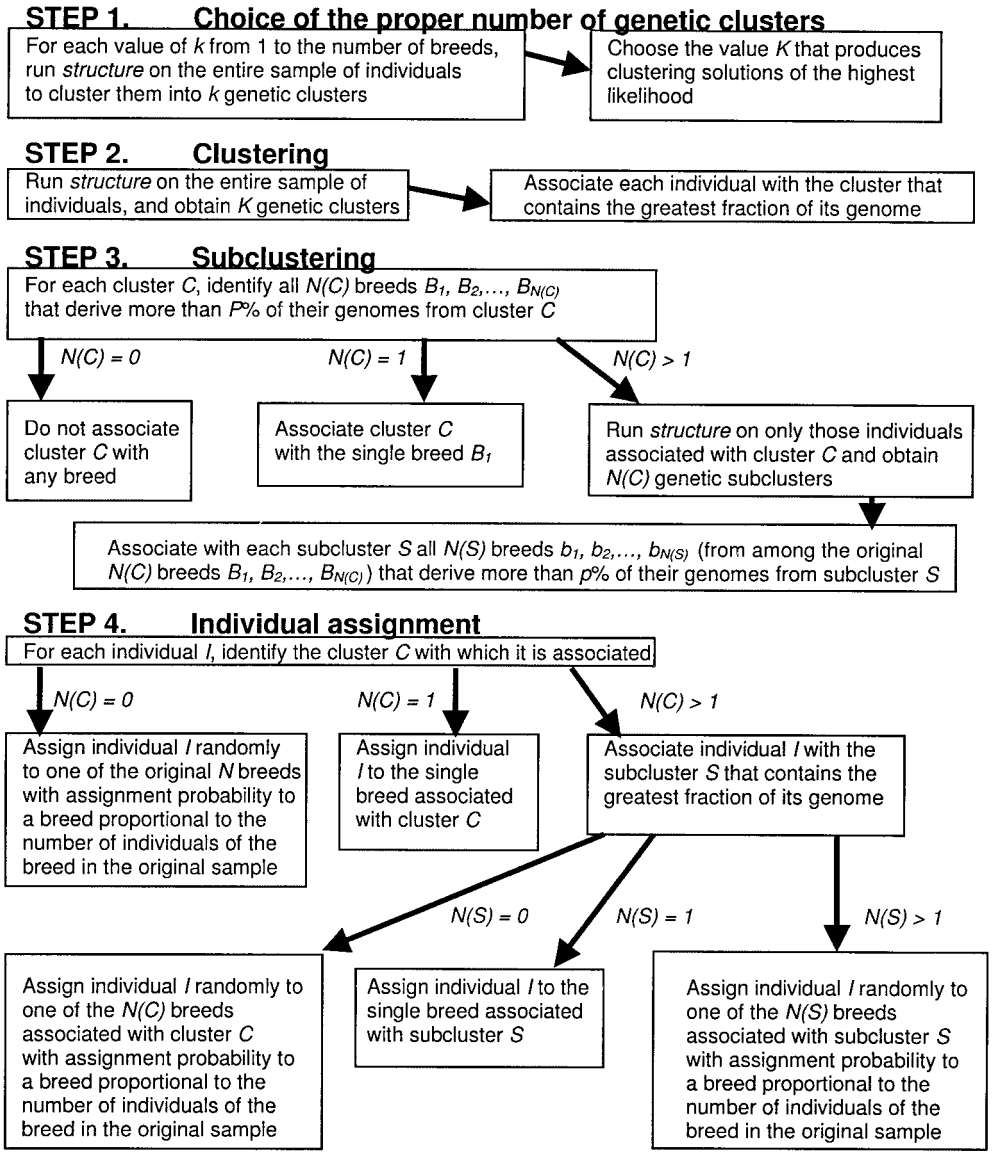

FIGURE 1.—Procedure by which cluster analysis was evaluated. For our data, all breeds had the same number of individuals in the sample. We used $K = 19$, $P = 25$, $p = 60$.

then be associated with either no breed or with the single breed *B*. Thus, this subclustering would not greatly affect the eventual assignment of individuals of cluster *C* to breeds. We also did not decompose any subclusters obtained in step 3 into "sub-subclusters." While it is conceivable that subclusters could be further divided, a single round of subclustering provided a convenient stopping point for the evaluation, allowing us to devise the precise procedure in Figure 1. Since only a small number of individuals would have been affected by sub-subclustering, the impact of this assumption on the clustering success rate was likely not very large. In the application of *structure* to data of unknown population structure, however, subclustering should be performed hierarchically, so that each cluster, subcluster, or lower-level grouping cannot be further decomposed.

Pairwise cluster analysis: We assessed populations two at a time with neighbor-joining tree diagrams of the individuals in two populations (MOUNTAIN and CAVALLI-SFORZA 1997). Pairwise distance between individuals was computed using one minus the proportion of alleles shared by the two individuals. Trees were obtained from distance matrices using the *neighbor* program (FELSENSTEIN 1993). If a tree could be partitioned into two connected pieces, each of which contained individuals from a single breed, the tree was considered "consistent with breed affiliation" (MOUNTAIN and CAVALLI-SFORZA 1997). If a tree was consistent with breed affiliation and if the partition of the tree was made by cutting the longest internal edge, the tree was deemed "strongly consistent with breed affiliation." If the partition was not necessarily made by cutting this edge, the tree was deemed "weakly consistent with breed affiliation."

For each pair of breeds, we also ran the cluster analysis using 20,000 iterations and a burn-in period of 5000, with $K = 2$. The clustering success rate was measured using the algorithm in Figure 1, though the criterion for subclustering was not met for any pair of populations.

Clustering success as a function of the number of markers: To determine properties of markers that make them effective in cluster analysis, we performed cluster analysis using subsets

of the original 27 markers according to several variability criteria. For each criterion, and for each value of M ($M = 1, 2, 3, \ldots 27$), we selected the M markers that exhibited the highest values of that criterion, and we performed cluster analysis using that subset of loci. In cases where two or more criteria produced the same subset, we only performed one analysis for that subset. The criteria included the following: (1) Expected heterozygosity—treating the whole sample as one group, for each locus we computed one minus the sum of the squares of the sample allele frequencies; (2) total number of alleles in the sample—if two or more loci had the same number of alleles, we broke ties by ranking markers in order of the mean number of alleles per breed; (3) F_{st}—we estimated F_{st} according to Weir (1996, Equation 5.3).

We also considered marker subsets taken in *reverse* order by expected heterozygosity, and we used a random ordering of the loci: For each value of M, we selected the M markers that were associated with the M highest random numbers. Rankings of markers are shown in Table 1. With the exception of orderings induced by the number of alleles and expected heterozygosity (Kendall coefficient = 0.464, $P = 0.0007$), we did not detect evidence for rank correlation among pairs of orderings (of course, the Kendall coefficient was -1 for the orderings by highest and lowest expected heterozygosity, and it equaled -0.464 for the orderings by highest number of alleles and lowest expected heterozygosity).

Clustering success as a function of the number of individuals: To see how cluster analysis performed with fewer individuals, for each value of N ($N = 5, 10, 15, 20, 25$), we repeated the analysis (with all markers and with marker subsets) using N randomly chosen individuals from each breed.

RESULTS

Genetic differentiation: For each pair of breeds, the null hypothesis that the two populations had equal allele frequencies was rejected at the 0.001 significance level for at least 6 loci (not shown). Even between the most closely related pairs of breeds, extremely significant differences were found. The only breed pairs for which 15 or fewer loci had significantly different allele frequencies at the 0.001 level were (44, 45), (5, 16), (16, 18), (18, 26), and (37, 3402). For several pairs, the null hypothesis of equal allele frequencies was rejected for at least 26 of 27 loci. These pairs included (4, 28), (4, 51), (4, 50), (26, 32), (28, 32), (32, 33), (32, 50), (32, 102), and (37, 102). Genetic distances were generally large as well (not shown), with only 10 pairwise comparisons <0.5 and with the average pairwise distance equaling 0.782. The lowest genetic distances were found for the following pairs: (44, 45), (5, 16), (16, 18), (18, 37), and (45, 51). The 25 largest genetic distances involved breeds 4, 19, 32, and 102.

Clustering of breeds: Due to the complexity of the relationships among the individuals in the data and the existence of numerous likely clustering solutions, different runs of *structure* identified different potential clusterings of the individuals (Table 2). Some features of the clustering were consistent across runs. Most strikingly, breeds 4, 19, 27, 32, and 102 always fell into their own clusters, while breeds 44 and 45 always shared the same cluster. Breeds 5, 13, 21, 26, 28, 51, and 3402 usually occupied their own clusters, and breeds 18 and 37 were often found together in a single cluster.

In 9 of the 100 runs performed, 19 clusters were assigned nontrivial fractions of the data. The remaining runs included 43, 44, and 4 runs for which 18, 17, and 16 clusters were occupied, respectively. In the 8 solutions of highest likelihood, breeds 44 and 45 shared a single cluster and each of the other 18 breeds occupied an exclusive cluster. The most frequent groupings (Table 2), including (5, 16), (16, 18), (18, 37), (33, 44, 45), (37, 3402), (42, 50), (44, 45), and (44, 45, 51), appeared in high-likelihood solutions, while rare groupings were obtained in low-likelihood solutions. None of the rare groupings (13, 26), (16, 33), (21, 26), (26, 42), (28, 42), or (33, 51) occurred in any of the 40 solutions of highest likelihood. The 10 lowest-likelihood runs contained the single instances that produced (21, 26) and (28, 42), as well as three of four instances in which the grouping (13, 26) was obtained.

Breeds that grouped into clusters generally fell close to each other in their placement on the neighbor-joining cladogram (Figure 2), although frequently clustered groups did not always form clades. Of the eight most commonly clustered sets, three did not form clades, namely (33, 44, 45), (16, 18), and (18, 37). Bootstrap confidence values for groupings in the cladogram were generally low.

Pairwise clustering: Although runs using all 20 breeds clustered pairs or triples of populations because 20 breeds were placed into 19 clusters, cluster analysis using only the individuals from 2 breeds separated them into 2 clusters. Of 190 pairs, 175 could be perfectly separated (Table 3). For the remaining 15 pairs, at most 5 individuals of 60 were placed incorrectly. However, for only 5 of these 15 pairs were individuals assigned to the wrong breed with high confidence ($>75\%$). The clustering success rate was also high for the two triads of populations that grouped together: For both (33, 44, 45) and (44, 45, 51), only individual 45_1 was misplaced (with breed 44).

For 188 of 190 breed pairs, the neighbor-joining tree was weakly consistent with breed affiliation (Figure 3). Of these 188 trees, 170 were strongly consistent with breed affiliation. The 2 pairs for which trees were not consistent (Figure 3), (5, 16) and (44, 45), were among the pairs for which clustering was imperfect. The 18 pairs for which trees were weakly consistent but not strongly consistent with breed affiliation were (5, 18), (5, 33), (5, 50), (5, 102), (13, 26), (13, 102), (16, 18), (16, 50), (16, 102), (18, 37), (27, 102), (28, 102), (33, 45), (33, 102), (42, 50), (42, 102), (50, 102), and (51, 102).

Evaluation of clustering: Using the complete set of 27 markers, cluster analysis obtained correct groupings of individuals with high accuracy (Figure 4). When only the most polymorphic markers were selected according to the greatest number of alleles or the highest expected

TABLE 2

Clustering behavior of 20 chicken breeds, based on 100 runs of *structure* using 19 proposed clusters

Breed code	Breed name	Breed category	No. of runs in which breed forms an exclusive cluster	Clustering companions of breed when not found in its own cluster, with frequencies
4	Fayoumi	Traditional selected, Mediterranean	100	
5	Bedouin	Traditional unselected, Middle East	93	16 only (7)
13	Marans	Traditional selected, Western Europe	96	26 only (4)
16	Icelandic landrace	Traditional unselected, Northern Europe	72	5 only (7)
				18 only (17)
				33 only (4)[b]
18	Old Scandinavian reference population	Traditional selected, Mediterranean	30	16 only (17)
				37 only (53)
19	Jaerhoens	Traditional selected, Northern Europe	100	
21	Padovana	Traditional selected, Mediterranean	99[a]	26 only (1)
26	Transylvanian naked neck	Traditional selected, Central Europe	93	13 only (4)
				21 only (1)
				42 only (2)
27	Green-legged partridge	Traditional selected, Central Europe	100	
28	Orlov	Traditional selected, Central Europe	99	42 only (1)
32	C line	Inbred	100	
33	Godollo Nhx	Commercial selected, experimental	87	16 only (4)[b]
				44 and 45 (10)[b]
				51 only (1)
37	White-egg layer line A	Commercial selected, layer	37	18 only (53)
				3402 only (10)
42	Broiler sire line B	Commercial selected, broiler	67	26 only (2)
				28 only (1)
				50 only (30)
44	Brown-egg layer line C	Commercial selected, layer	0	33 and 45 (10)[b]
				45 only (84)
				45 and 51 (6)
45	Brown-egg layer line D	Commercial selected, layer	0	33 and 44 (10)[b]
				44 only (84)
				44 and 51 (6)
50	Broiler dam line D	Commercial selected, broiler	70	42 only (30)
51	High-Ab line	Commercial selected, experimental	93	33 only (1)
				44 and 45 (6)
102	Gallus gallus gallus	Feral	100	
3402	Sarcoma-susceptible line	Commercial selected, experimental	90	37 only (10)

[a] In one run, breed 21 was distributed over two separate exclusive clusters.

[b] In two runs, breed 33 was distributed over two separate clusters, one of which was shared with 16 and the other of which was shared with both 44 and 45.

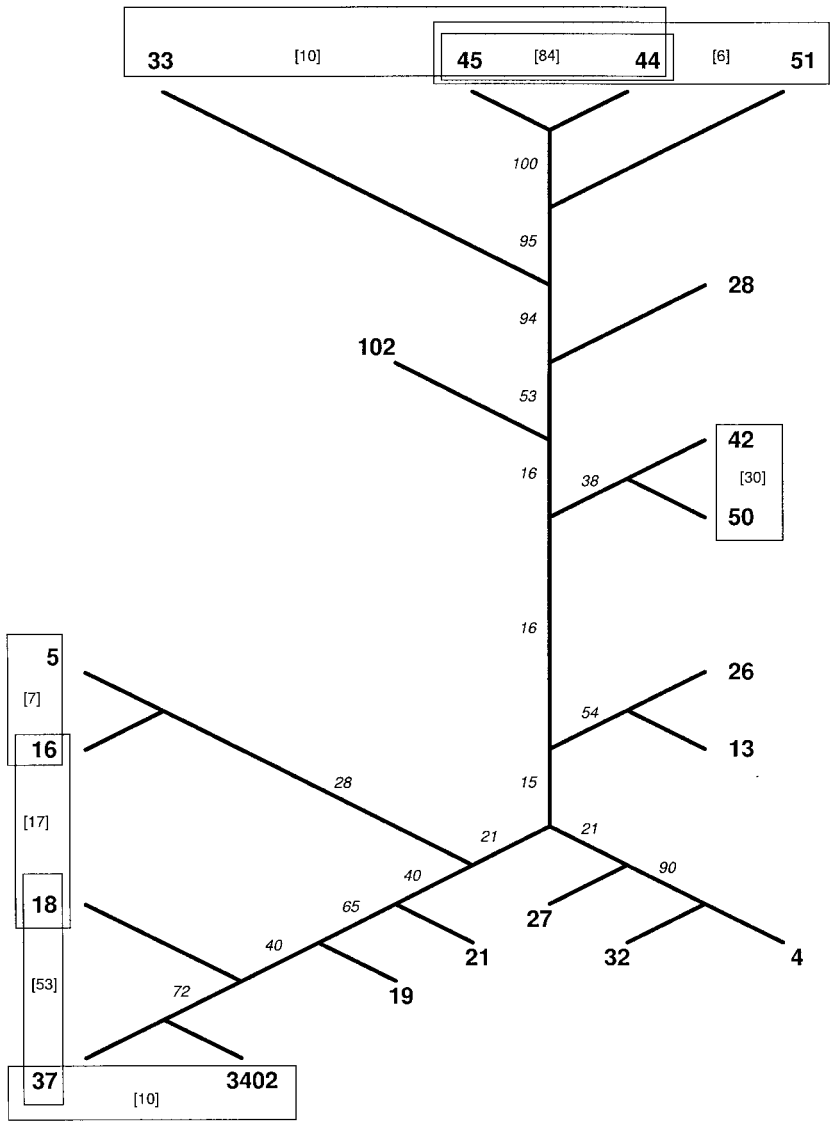

FIGURE 2.—Number out of 100 clustering solutions in which multiple breeds occupied the same clusters, superimposed on a neighbor-joining cladogram derived from the generalized proportion-of-shared-alleles distance measure. Breed names at branch termini are labeled in boldface type; bootstrap confidence values of breed groupings are italicized next to edges of the cladogram to which they correspond; and numbers of clustering solutions are surrounded by brackets inside rectangles enclosing the clusters to which they correspond. Bootstrap confidence values are taken over 1000 trees. Rectangles for clustering solutions are shown only for groupings that appeared in at least 5 of 100 runs. Branch lengths were chosen so that the figure could be conveniently represented.

heterozygosity, only ~8–10 markers were needed to attain 95% accurate clusterings. Once 11–12 markers were chosen, expected heterozygosity, number of alleles, and F_{st} performed similarly, achieving 95–98% in almost every run. Although the random ordering achieved 90% clustering accuracy with 10–12 markers, it required 17–20 markers to achieve 95%. The reverse ordering by expected heterozygosity required 14–15 loci to achieve 90% and 17–20 loci to reach 95%. When only a few markers were used, the discrepancy between the two most effective criteria and the other criteria was extremely high. The marker sets chosen by reverse order of expected heterozygosity performed particularly poorly compared to the other methods. However, as the number of markers increased, all criteria produced nondisjoint sets of markers, and when nearly all markers were used, the accuracy of clustering was ~98% for each criterion.

For further analysis and discussion, we used expected heterozygosity. This statistic and the number of alleles are highly correlated (in fact, the most variable marker and the set of seven most variable markers coincided according to these two criteria) and they produce similarly accurate clusterings. However, expected heterozygosity is more generally useful—for example, if single nucleotide polymorphisms are used, it provides a natural method to rank loci that all have two alleles.

Most breeds were clustered perfectly (Table 4), and although clustering solutions differed across runs, the same individuals tended to be misclassified across runs. For some breeds, including 4, 19, 27, and 32, all individuals were perfectly clustered using a small number of highly heterozygous loci. Others, including 5, 13, 16, 18, 44, 45, and 102, required many loci to obtain correct clustering. For breeds 5, 16, 45, and 102, large numbers of markers did not improve classification of a few specific individuals.

When subsets of the individuals were chosen, clustering accuracy declined. When 5 individuals were chosen from each breed, 27 markers were insufficient to obtain

TABLE 3

Evaluation of cluster analysis in separating pairs of breeds

Pair of breeds	Clustering success rate	Individuals incorrectly placed in pairwise analysis
5 and 16	0.917	*16_21*, 16_22, 16_23, 16_28, 16_56
5 and 50	0.983	5_9
5 and 102	0.933	102_2, *102_19*, *102_20*, *102_21*
13 and 102	0.950	*102_19*, *102_20*, *102_21*
16 and 18	0.983	16_28
16 and 102	0.983	102_21
26 and 102	0.967	102_19, 102_21
27 and 102	0.950	102_19, 102_20, 102_21
28 and 102	0.967	102_20, 102_21
33 and 102	0.950	102_19, 102_20, 102_21
44 and 45	0.983	*45_1*
44 and 102	0.950	102_19, 102_20, 102_21
45 and 102	0.950	102_19, 102_20, 102_21
50 and 102	0.950	*102_19*, *102_20*, *102_21*
51 and 102	0.983	102_21

Individuals in italics had at least 75% of their genomes assigned to the incorrect breed. Fifteen pairs for which clustering was imperfect are shown; 175 pairs for which the clustering success rate was 100% are not shown. Individual identifiers include the breed code, an underscore, and the specific individual within the breed (for example, 102_20 represents individual 20 in breed 102).

90% accuracy (Figure 5). When 10 individuals were chosen, 21 markers were sufficient to achieve 90%. When 15 or more individuals were selected from each breed, 90% accuracy was attained using the 12 most variable loci.

DISCUSSION

Correspondence of inferred and known population structure: When the full data set was used, inferred genetic clusters of individuals corresponded extremely well to predefined breed categorizations. Since similar likelihoods of many proposed clusterings make it difficult to label a "best" clustering of the data, we suggest that, for large data sets, cluster analysis should be performed multiple times before inferences are drawn. All solutions had in common that each cluster contained all or nearly all individuals from one or a few breeds. Upon further analysis, all clusters that contained more than one breed could be subdivided into a collection of subclusters, each of which matched a single breed.

While *structure* easily separated individuals into clusters that corresponded almost exactly to phenotypic labels, the bootstrap neighbor-joining cladogram was less capable of grouping subsets of the data with great regularity. Several possibilities can explain this discrepancy. First, while *structure* constructs genetic clusters from individual genotypes without reference to breed affiliation, cladograms assume that genetic clusters correspond to breed designations. This correspondence essentially holds, although 10–15 individuals frequently appeared more similar to breeds from which they did not originate. The inclusion of these individuals in breed groupings used for the neighbor-joining tree potentially decreases genetic distances of certain pairs and hence affects the cladogram. However, upon removal of all individuals that were sometimes placed incorrectly in pairwise clustering (Table 3), the cladogram was essentially unchanged (not shown); thus, these individuals cannot explain its poor reliability.

An alternate explanation for the performance of the neighbor-joining tree is the fact that in domesticated species such as chickens, population histories may not follow a bifurcating tree model, so that tree diagrams present a misleading or inaccurate representation of population relationships. The considerable frequency of gene exchange among historical chicken populations could potentially explain the low bootstrap values on internal edges of the tree and on edges that group feral and traditional breeds.

Finally, it is likely that *structure* simply uses individual genotypic data more efficiently than cladograms based on genetic distance matrices (Pritchard *et al.* 2000; Rosenberg *et al.* 2001). While genetic distance matrices compress all information about two populations into a single number, *structure* does not summarize the data in a unidimensional manner.

It has been argued that 30 markers are insufficient for distinguishing related populations using phylogenetic analysis (Moazami-Goudarzi *et al.* 1997). With high-resolution clustering analysis and with careful choices of markers, however, far fewer loci sufficed to separate all breeds. The chicken generation interval is short, ~1 year for these breeds, so that considerable genetic variation has built up within and among chicken breeds (Dunnington *et al.* 1994; Crooijmans *et al.* 1996; Ponsuksili *et al.* 1996, 1998, 1999; Mafeni *et al.* 1997; Takahashi *et al.* 1998; Vanhala *et al.* 1998; Hillel *et al.* 1999; Wimmers *et al.* 1999, 2000; Zhou and Lamont 1999; Kaiser *et al.* 2000). Thus, it is possible that chicken breeds could be easily separated partly due to high levels of intraspecific variation. However, the success of *structure* when compared to the cladogram in separating breeds makes it likely that the method was largely responsible.

In pairwise analysis of populations, clustering and neighbor-joining trees performed similarly. We note that neighbor-joining trees of individuals from two breeds are more useful if the strong criterion of separation is used. If genetic origins of individuals in two populations are known, the weak criterion of consistency for separating populations is applicable—two populations are separated if there exists a decomposition of the tree into two components, each corresponding to a population. However, if genetic origins are unknown beforehand, an objective method must be used to sepa-

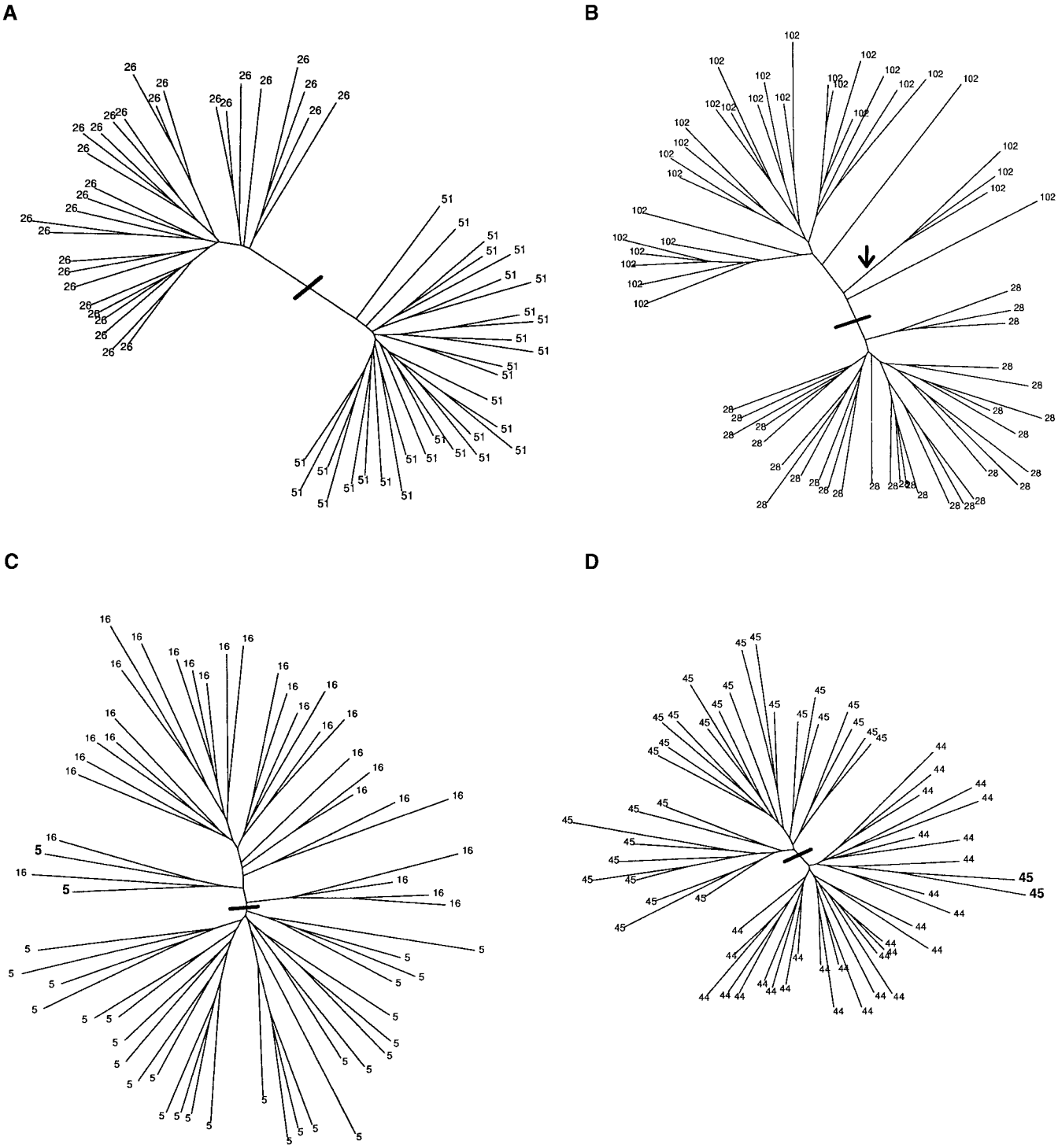

FIGURE 3.—Neighbor-joining trees of individuals taken from two breeds at a time. (A) One of 170 pairwise neighbor-joining trees that was *strongly* consistent with breed affiliations of individuals. (B) One of 18 neighbor-joining trees that was *weakly* consistent with breed affiliations but not strongly consistent. The arrow points to the longest internal edge of the tree. (C and D) The only two trees among 190 two-breed trees that were not consistent with breed affiliations of individuals. For all trees, individuals are labeled by their breed codes. The thick line is drawn on the edge that minimizes the number of individuals that do not group with other members of the same breed. Individuals in boldface type fall on the side of the thick line that does not include most members of their breeds.

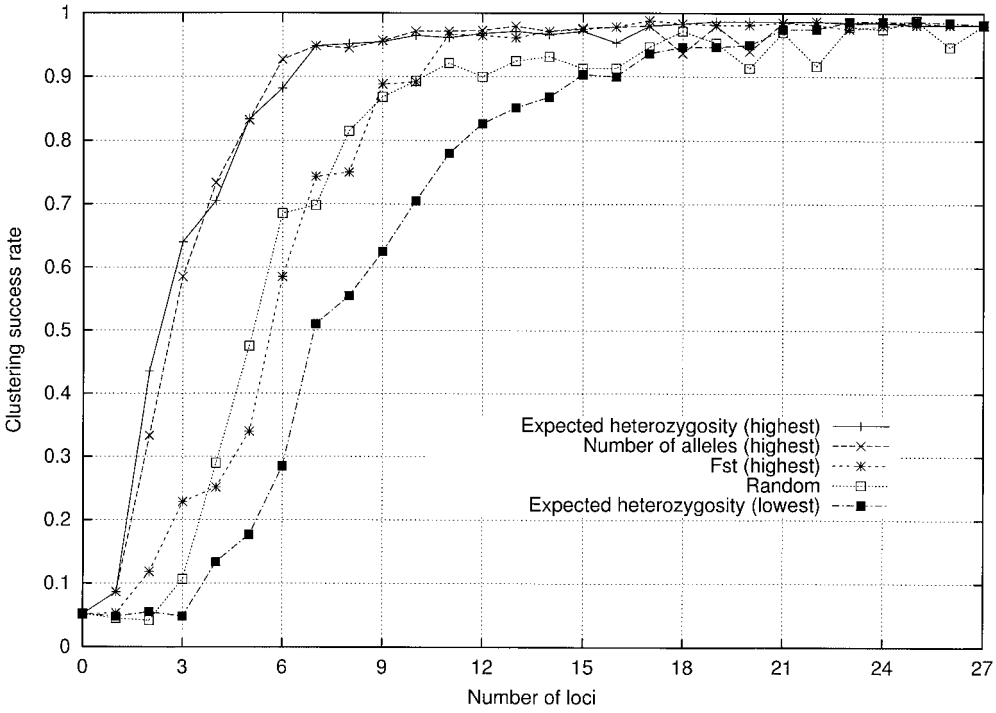

FIGURE 4.—Clustering success rate as a function of the number of markers used. Sets of markers were chosen in order of five criteria: highest number of alleles, highest expected heterozygosity, lowest expected heterozygosity, highest F_{st}, and a random ordering.

rate the tree into components; under these circumstances, the strong criterion of consistency must be applied. When this criterion is used, cluster analysis performs slightly better than neighbor-joining trees in separating populations. Clustering also offers the opportunity for significance testing using $R \times C$ tests of association (ROSENBERG *et al.* 2001). Permutation tests for significance of neighbor-joining clusterings are more cumbersome.

Strategies for successful clustering: Highest expected heterozygosity and highest number of alleles provided the best ways to select loci for clustering and were better than highest F_{st}. This result was surprising: The most useful genetic marker for clustering populations and assigning individuals is one that varies greatly across populations but little within them. A perfect locus for these purposes would be monomorphic within any given breed but polymorphic across breeds (REED 1973). In the absence of such loci, it seems that F_{st}, which quantifies the between-breed component of genetic variation, should be an appropriate criterion on which to rank the clustering potential of loci.

Several highly variable markers, the tetranucleotide loci LEI192 and LEI228 most dramatically, had many alleles that were specific to at most a few populations and frequent in those populations. These markers also had generally low F_{st} values, since more common alleles did not greatly differ in frequency across breeds. It seems likely that these "diagnostic" alleles were partly responsible for the extremely successful clustering with the number of alleles and expected heterozygosity statistics. For 27 markers, we observed 101 alleles private to a single breed out of 326 total alleles. For 62 of the private alleles, two or more copies were observed, and, thus, these alleles are unlikely to result from genotyping errors. Since so many alleles in this study were breed specific and since many more were found only in two or three breeds, it is possible that these alleles had a substantial effect on clustering success. However, the number of private alleles in a sample decreases as the sample size from a breed increases. A large number of private alleles is not a property of most data sets, and, thus, the highest number of private alleles cannot be recommended as a criterion method for choosing the best markers to use. In closely related populations, private alleles may be uncommon: for example, using data from a study of 11 human populations (JIN *et al.* 2000), only 137 of 792 alleles were private to 1 population (and two or more copies were only observed for 41 of these 137), even though sample sizes were much smaller than those used here.

The successful performances of all other criteria compared to the reverse ordering by expected heterozygosity demonstrate that a careful choice of markers increases the power to achieve accurate clustering. This idea that a careful choice of markers can improve statistical power has been employed in the estimation of population of origin for admixed individuals (REED 1973; SHRIVER *et al.* 1997). For admixture inference in populations that result from the combination of two ancestral groups, SHRIVER *et al.* (1997) suggested that markers for assignment should be those with maximal allele frequency differentials in the ancestral populations. In the more general situation of a large collection of populations, we suggest that expected heterozygosity, number of alleles, and, to a lesser extent, F_{st}, are suitable for

TABLE 4

Breed-specific numbers of correctly assigned individuals out of 30 individuals per breed

	Breed no.																				
No. of loci	4	5	13	16	18	19	21	26	27	28	32	33	37	42	44	45	50	51	102	3402	Total
0	1	0	1	1	0	2	4	1	1	2	2	0	1	2	3	0	0	3	2	5	31
1	1	0	1	1	1	5	0	0	0	1	30	2	0	3	1	1	1	1	2	1	52
2	27	0	1	0	6	29	15	27	23	1	30	1	14	27	11	6	0	13	0	30	261
3	28	0	15	1	8	29	14	28	28	18	30	28	16	29	18	16	20	7	21	30	384
4	30	16	21	0	20	30	11	29	29	30	30	27	20	30	16	11	22	16	26	9	423
5	30	18	27	18	20	30	30	30	29	30	30	29	29	30	12	19	25	8	26	30	500
6	30	23	28	20	26	30	30	30	29	30	30	30	28	30	12	15	26	26	26	30	529
7	30	26	30	22	26	30	30	30	29	30	30	30	28	30	30	28	28	26	26	30	569
8	30	25	29	21	27	30	30	30	30	30	30	30	29	30	30	28	28	28	26	30	571
9	30	26	28	20	29	30	30	30	30	30	30	30	29	30	30	28	28	28	27	30	573
10	30	25	29	26	29	30	30	30	30	30	30	30	30	30	30	28	29	26	27	30	579
11	30	25	29	26	29	30	30	30	30	30	30	25	30	30	30	30	29	27	27	30	577
12	30	27	30	23	30	30	30	30	30	30	30	30	30	30	30	30	28	27	26	30	581
13	30	26	30	26	30	30	30	30	30	30	30	30	30	30	30	28	28	28	27	30	583
14	30	26	30	26	29	30	30	30	30	30	30	30	30	30	29	28	28	29	25	30	580
15	30	25	30	27	30	30	30	30	30	30	30	30	30	30	30	29	28	29	26	30	583
16	30	11	30	29	30	30	30	30	30	30	30	30	30	30	30	29	29	29	25	30	572
17	30	29	30	27	30	30	30	30	30	30	30	30	30	30	30	29	29	29	25	30	588
18	30	29	30	26	30	30	30	30	30	30	30	30	30	30	30	29	30	29	27	30	590
19	30	28	30	28	30	30	30	30	30	30	30	30	30	30	30	29	30	29	28	30	592
20	30	29	30	29	30	30	30	30	30	30	30	30	30	30	30	29	30	30	25	30	592
21	30	29	30	27	30	30	30	30	30	30	30	30	30	30	30	29	30	30	27	30	592
22	30	29	30	27	30	30	30	30	30	30	30	30	30	30	30	29	30	30	27	30	592
23	30	29	30	27	30	27	30	30	30	30	30	30	30	30	30	29	30	30	26	30	591
24	30	29	30	27	30	30	30	30	30	30	30	30	30	30	30	29	30	30	26	30	591
25	30	29	30	25	30	30	30	30	30	30	30	30	30	30	30	29	30	30	26	30	589
26	30	29	30	25	30	30	30	30	30	30	30	30	30	30	30	29	30	30	26	30	589
27	30	29	30	25	30	30	30	30	30	30	30	30	30	30	30	29	30	30	26	30	589

Sets of markers include the most variable markers according to the highest expected heterozygosity.

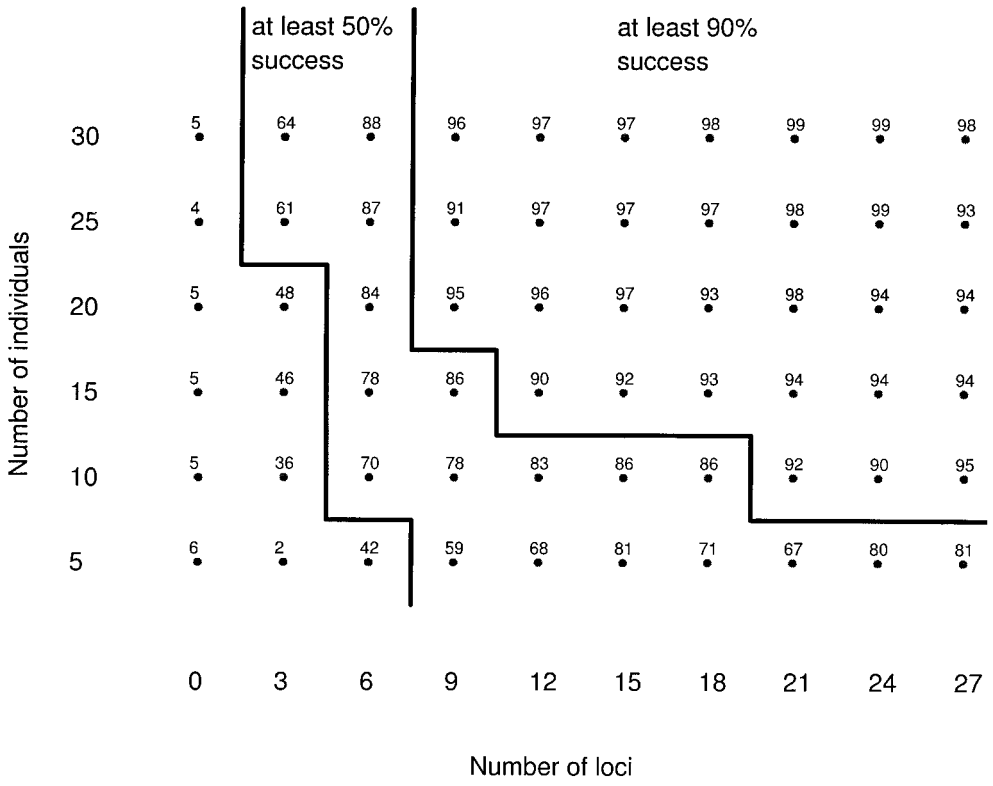

FIGURE 5.—Clustering success rate as a function of the number of markers and the number of individuals used. The most variable markers were chosen according to the highest expected heterozygosity in the full data set. Random sets of 5, 10, 15, 20, and 25 individuals per breed were chosen from among the 30 individuals genotyped in each breed.

choosing markers for cluster analysis and individual assignment. All three criteria are applicable to microsatellite data; however, expected heterozygosity is the most practical general criterion. For some types of markers, such as single nucleotide polymorphisms, the number of alleles is often the same for all markers. The calculation of F_{st} presumes knowledge of population structure, so that for initial cluster analysis of a collection of individuals of unknown origin, the statistic cannot be calculated. This recommendation of the expected heterozygosity criterion is empirical and potentially of limited generality; the true property of a marker that causes it to produce successful clusterings is unlikely to be any of the properties mentioned here. A detailed simulation-based approach will be needed for understanding this true property.

Depending on the species under consideration, the relative cost of genotyping more individuals and genotyping more markers will vary. We did not achieve 90% success in clustering when only 5 individuals were used per breed, but with 10 or more individuals per breed, clustering was highly successful when enough markers were used. Similarly, we did not achieve 90% success when fewer than 6–7 markers were used, even when all individuals were included. Thus, as a minimum, for similarly diverged populations to those in our study, at least 12–15 highly variable markers should be genotyped in at least 15–20 individuals per hypothesized population to achieve accurate clustering. In species for which genetic research is still preliminary, genotyping can be done sequentially: A small number of individuals can be genotyped for many markers. The most variable markers can then be selected for future study and then genotyped in a large sample.

While using the most variable markers allows researchers to minimize genotyping effort, we caution against using this type of marker set with statistical methods that assume a random set of loci and that make inferences based on mean variabilities across loci. A marker set selected for maximal variability will inflate estimates of divergence times estimated using the genetic distance $(\delta\mu)^2$ (GOLDSTEIN *et al.* 1995) and it will bias population growth statistics (*e.g.*, ZHIVOTOVSKY *et al.* 2000).

We note that the 20 breeds genotyped here were chosen from among many breeds used in an earlier study (HILLEL *et al.* 1999) to maximize diversity within the larger collection of breeds. Thus, it is possible that the populations we used are generally more divergent than populations in other species of interest and even more divergent than most chicken groups. This is evidenced by a high F_{st} value of 0.313 for these breeds. In less diverged populations, including human groups, the number of markers necessary for maximal clustering will certainly be >12–15. A more comprehensive analysis of *structure*, using simulated data sets of different F_{st} values and different numbers of populations, will be needed to determine the generality of our results. Rather than attempting such an analysis here, we chose to use a single large data set for developing recommen-

dations. This has allowed us to explore issues that will be of interest in future applications of genetic cluster analysis to data sets of unknown structure, such as the variation of clustering solutions across runs, the difference in ease of clustering across populations, and the placement of problematic individuals.

Problematic individuals: In most runs, the clustering success rate remained $< \sim 98\%$, though this level could be obtained when 15–20 highly variable markers were used. Given this observation, it is surprising that the full set of 27 markers did not achieve 100% accuracy. Errors in the clustering algorithm seem to be an unlikely explanation, since roughly the same sets of individuals were placed in the wrong clusters in runs that used different sets of loci or that produced different clustering solutions. Since all breeds were sampled from populations maintained in different locations, it is unlikely that recent admixture or labeling errors explain the improper placements. We suspect that the inability to achieve perfect clustering results from the fact that some individuals were genetically atypical of their breeds, and the algorithm could not recognize breeds of origin for these individuals. The frequently misplaced individuals 102_19, 102_20, and 102_21 derived from a flock of zoo animals that may have undergone considerable genetic drift. Individuals 16_21, 16_22, and 16_23 came from a single flock, one of many that was incorporated into the breed 16 sample; this flock may have been managed differently from the others. Interestingly, only one individual was misplaced from the closely related breeds 44 and 45: This suggests that *structure* may be useful for distinguishing lines from different breeding companies, in spite of common origin and similar selection objectives.

Cluster analysis and population assignment: Placement of individuals into clusters is related, but not identical, to assignment of unknown individuals to populations. Assignment tests assume the existence of distinct populations and use properties of those groups, such as allele frequencies, to infer the source populations of unknown individuals (Buchanan *et al.* 1994; Paetkau *et al.* 1995; Rannala and Mountain 1997; Cornuet *et al.* 1999; Davies *et al.* 1999; Ciampolini *et al.* 2000; Pritchard *et al.* 2000). Properties of the potential source populations are ideally known, but in practice they are generally inferred in such a way that any specific individual is not assigned using information that derived from knowledge of its genotype. In our evaluation procedure, individuals are first used to create clusters, and the same individuals are assigned to clusters and then to breeds. While this approach allows us to precisely define the clustering success rate, the fact that information from any given individual is used in inferring its origin prevents us from interpreting its inferred population of origin as a proper assignment.

Our results are best interpreted as verification that these individuals indeed form genetic clusters that correspond to their breed designations and that they can be used in a training sample for assignment of future unknowns. This training sample can be utilized differently by various assignment algorithms. For example, the method of Paetkau *et al.* (1995) estimates allele frequencies from a training sample and assigns each unknown individual to the population in which its genotype is most likely, assuming Hardy-Weinberg and linkage equilibrium. The method of Cornuet *et al.* (1999) also estimates population allele frequencies from a training sample and uses the smallest genetic distance of the unknown individual to the various populations for assignment. With the *structure* algorithm of Pritchard *et al.* (2000), a genetic clustering solution is obtained that includes the training sample and the unknowns. Next, each unknown is assigned to a cluster. Finally, depending on which individuals from the training sample are also assigned to that cluster, the unknown is eventually assigned to a breed. During this process, prior knowledge about individuals in the training sample can be incorporated—individuals in the training sample can be flexibly treated as being of known, unknown, or probabilistically known origin.

The importance of training samples for population assignment suggests a strategy by which future assignment studies can be optimized. For any species of interest, the most variable markers according to expected heterozygosity, number of alleles, or F_{st} should be genotyped on a large scale. New markers could be tested by the criterion, and highly variable new markers could potentially be included in the set of most variable markers, reducing the number of markers needed for clustering studies below the current recommendation of 12–15. A database of individual genotypes at these most variable loci could then be made publicly available. New individuals could be genotyped for the most variable markers and could then be added to the database. Individuals in the database who are known to represent certain breeds could be used as a training sample for assignment tests. Individuals who were misassigned or who were difficult to assign correctly could be excluded from the database, so that only the individuals who can confidently be assigned to the correct breeds would be included.

Such a database might be extremely useful to researchers who may only have one or a few unknown individuals that they wish to identify (*e.g.*, Primmer *et al.* 2000). This type of database could also serve as a repository of individuals with known origin for testing new statistical algorithms. Using such a database, genetic variation quantified in different studies can be made commensurable, and large numbers of individual multilocus genotypes can be combined into a single framework. The importance of training samples to assignment makes it necessary that the same microsatellite allele sizes be used by different laboratories. DNA from several individuals in our study is available from M. Tixier-

Boichard to calibrate allele size measurements, and genotypes are available at http://charles.stanford.edu.

Cluster analysis and genetic distinctiveness: We observed that some breeds were easier to separate into clusters than others, in the sense that all individuals in some breeds were correctly placed with only a small number of markers. This likely derives from the presence of distinctive multilocus genetic combinations in the breeds that were easiest to separate. Thus, we suggest that the relative number of loci required for the correct clustering of several breeds can be used as a way of identifying populations that are genetically distinctive with respect to a collection.

In addition to resolving questions about population histories (Rosenberg *et al.* 2001), the characterization of genetically distinctive populations can assist in conservation of within-species diversity (Moritz 1994; Paetkau 1999). For species in which relative conservation value of different populations is of interest, the ease with which a population can be separated from other groups by cluster analysis can be incorporated into assessments of its conservation potential, along with relevant ecological, economic, and evolutionary criteria. More generally, cluster analysis has great potential to help identify populations with different allele frequencies and different multilocus genetic combinations. Although genetical divergence among populations may not reflect adaptive diversity (Crandall *et al.* 2000), conservation programs that wish to maintain genetic diversity of endangered species could benefit from a purely genetical method by which individuals can be clustered without regard to their sampling location. For agricultural species, although the preservation of specific phenotypes is perhaps of primary interest, conservation of genetic diversity is of great importance toward ensuring that future breeding programs will have a large base on which to perform artificial selection (Notter 1999).

Relationships of chicken breeds: Considerable attention has been devoted to the study of genetic diversity and relationships of chickens. Some studies focused on commercial breeds (Dunnington *et al.* 1994; Crooijmans *et al.* 1996; Kaiser *et al.* 2000), others mainly considered local breeds (Mafeni *et al.* 1997; Takahashi *et al.* 1998; Wimmers *et al.* 1999, 2000), and some studied mixed collections (Ponsuksili *et al.* 1996, 1998, 1999; Vanhala *et al.* 1998; Hillel *et al.* 1999; Zhou and Lamont 1999). We suggest here that the frequent clustering of pairs of breeds into the same clusters illuminates a new approach toward determining genetic similarity. Populations can be considered similar if they are frequently placed in the same genetic cluster. Several of the frequently obtained groupings of populations are expected from knowledge of chicken population histories. The two most difficult populations to separate were the two commercial brown-egg layers (44 and 45). Other frequent groupings involved the two broiler populations (42 and 50), as well as other collections of populations that were of the same general category. The combinations (33, 44, 45), (37, 3402), and (44, 45, 51) all included selected populations. The grouping (18, 37) may reflect the fact that breeds 18 and 37 are unselected and selected descendants, respectively, of ancestral white leghorn populations. Groupings of traditional breeds, such as (5, 16) and (16, 18), are more surprising. It seems particularly strange that the Icelandic landrace (16), probably isolated for several hundred years, would sometimes group with the Middle Eastern Bedouin breed (5). One hypothesis is that populations 5, 16, and 18 represent unselected groups similar to ancestral Mediterranean chickens, which may have colonized many European countries along maritime trading routes. A more detailed historical analysis of chicken breeds will be required to explain such surprising relationships.

Conclusions: We have discussed the application of genetic cluster analysis to 600 individuals from 20 chicken breeds, demonstrating that the technique has great potential to correctly identify population structure. We have argued that individual clustering provides a more appropriate characterization of population structure in these groups than does a neighbor-joining tree. Last, we have proposed recommendations on future uses of genetic cluster analysis and individual assignment tests in similarly diverged collections of populations: (1) At least 12–15 highly variable loci should be genotyped in at least 15–20 individuals per hypothesized population; (2) markers with the highest expected heterozygosity, number of alleles, and F_{st} can be used in genetic cluster analysis to minimize genotyping costs; (3) databases of multilocus genotypes obtained at highly variable markers in individuals of known origins can be established to provide training samples for assignment algorithms; (4) genetically distinctive populations can be identified on the basis of how difficult it is to separate them from other breeds when cluster analysis is used; and (5) cluster analysis can provide an additional tool for identification of population relationships, history, and within-species genetic units for conservation.

The authors thank Nina Dudnik and Jonathan Pritchard for helpful comments. This study arose during a visit by N.A.R. to the laboratory of J.H. N.A.R. is supported by a Program in Mathematics and Molecular Biology graduate fellowship. This research was supported by the European Community-funded project AVIANDIV (Development of Strategy and Application of Molecular Tools to Assess Biodiversity in Chicken Genetic Resources, BIO4CT980342) and by National Institutes of Health grant GM28428 to M.W.F.

LITERATURE CITED

Beaumont, M., E. M. Barratt, D. Gottelli, A. C. Kitchener, M. J. Daniels *et al.*, 2001 Genetic diversity and introgression in the Scottish wildcat. Mol. Ecol. **10:** 319–336.

Bowcock, A. M., A. Ruiz Linares, J. Tomfohrde, E. Minch, J. R. Kidd *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

Buchanan, F. C., L. J. Adams, R. P. Littlejohn, J. F. Maddox and A. M. Crawford, 1994 Determination of evolutionary relationships among sheep breeds using microsatellites. Genomics **22:** 397–403.

Ciampolini, R., H. Leveziel, E. Mazzani, C. Grohs and D. Cianci, 2000 Genomic identification of the breed of an individual or its tissue. Meat Sci. **54:** 35–40.

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.

Crandall, K. A., O. R. P. Bininda-Emonds, G. M. Mace and R. K. Wayne, 2000 Considering evolutionary processes in conservation biology. Trends Ecol. Evol. **15:** 290–295.

Crooijmans, R. P. M. A., A. F. Groen, A. J. A. Van Kampen, S. Van Der Beek, J. J. Van Der Poel *et al.*, 1996 Microsatellite polymorphism in commercial broiler and layer lines estimated using pooled blood samples. Poult. Sci. **75:** 904–909.

Davies, N., F. X. Villablanca and G. K. Roderick, 1999 Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. Trends Ecol. Evol. **14:** 17–21.

Devlin, B., and K. Roeder, 1999 Genomic control for association studies. Biometrics **55:** 997–1004.

Dunnington, E. A., L. C. Stallard, J. Hillel and P. B. Siegel, 1994 Genetic diversity among commercial chicken populations estimated from DNA fingerprints. Poult. Sci. **73:** 1218–1225.

Felsenstein, J., 1993 *PHYLIP (Phylogeny Inference Package)*. Department of Genetics, University of Washington, Seattle.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman, 1995 Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

Groenen, M. A. M., H. H. Cheng, N. Bumstead, B. F. Benkel, W. E. Briles *et al.*, 2000 A consensus linkage map of the chicken genome. Genome Res. **10:** 137–147.

Hillel, J., A. Korol, V. Kirzner, P. Freidlin, S. Weigend *et al.*, 1999 Biodiversity of chickens based on DNA pools: first results of the EC funded project AVIANDIV, pp. 22–29 in *Poultry Genetics Symposium, Proceedings*, edited by R. Preisinger. Lohmann Tierzucht, Cuxhaven, Germany.

Jin, L., M. L. Baskett, L. L. Cavalli-Sforza, L. A. Zhivotovsky, M. W. Feldman *et al.*, 2000 Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. Ann. Hum. Genet. **64:** 117–134.

Kaiser, M. G., N. Yonash, A. Cahaner and S. J. Lamont, 2000 Microsatellite polymorphism between and within broiler populations. Poult. Sci. **79:** 626–628.

Mafeni, M. J., K. Wimmers and P. Horst, 1997 Genetic diversity in indigenous Cameroon and German Dahlem Red fowl populations estimated from DNA fingerprints. Arch. Tierz. **40:** 581–589.

Minch, E., A. Ruiz-Linares, D. B. Goldstein, M. W. Feldman and L. L. Cavalli-Sforza, 1998 *Microsat2: A Computer Program for Calculating Various Statistics on Microsatellite Allele Data*. Department of Genetics, Stanford University, Stanford, CA.

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.

Moritz, C., 1994 Defining 'evolutionarily significant units' for conservation. Trends Ecol. Evol. **9:** 373–375.

Mountain, J., and L. L. Cavalli-Sforza, 1997 Multilocus genotypes, a tree of individuals, and human evolutionary history. Am. J. Hum. Genet. **61:** 705–718.

National Research Council, 1996 *The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, DC.

Notter, D. R., 1999 The importance of genetic diversity in livestock populations of the future. J. Anim. Sci. **77:** 61–69.

Paetkau, D., 1999 Using genetics to identify intraspecific conservation units: a critique of current methods. Conserv. Biol. **13:** 1507–1509.

Paetkau, D., W. Calvert, I. Stirling and C. Strobeck, 1995 Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. **4:** 347–354.

Ponsuksili, S., K. Wimmers and P. Horst, 1996 Genetic variability in chickens using polymorphic microsatellite markers. Thai J. Agric. Sci. **29:** 571–580.

Ponsuksili, S., K. Wimmers and P. Horst, 1998 Evaluation of genetic variation within and between different chicken lines by DNA fingerprinting. J. Hered. **89:** 17–23.

Ponsuksili, S., K. Wimmers, F. Schmoll, P. Horst and K. Schellander, 1999 Comparison of multilocus DNA fingerprints and microsatellites in an estimate of genetic distance in chicken. J. Hered. **90:** 656–659.

Primmer, C. R., M. T. Koskinen and J. Piironen, 2000 The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proc. R. Soc. Lond. Ser. B **267:** 1699–1704.

Pritchard, J. K., M. Stephens and P. J. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA **94:** 9197–9201.

Reed, T. E., 1973 Number of gene loci required for accurate estimation of ancestral population proportions in individual human hybrids. Nature **244:** 575–576.

Rosenberg, N. A., E. Woolf, J. K. Pritchard, T. Schaap, D. Gefel *et al.*, 2001 Distinctive genetic signatures in the Libyan Jews. Proc. Natl. Acad. Sci. USA **98:** 858–863.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Shriver, M. D., M. W. Smith, L. Jin, A. Marcini, J. M. Akey *et al.*, 1997 Ethnic-affiliation estimation by use of population-specific DNA markers. Am. J. Hum. Genet. **60:** 957–964.

Takahashi, H., K. Nirasawa, Y. Nagamine, M. Tsudzuki and Y. Yamamoto, 1998 Genetic relationships among Japanese native breeds of chicken based on microsatellite DNA polymorphisms. J. Hered. **89:** 543–546.

Tixier-Boichard, M., G. Coquerelle and C. Vilela-Lamego, 1999 Contribution of data on history, management and phenotype to the description of the diversity between chicken populations sampled within the AVIANDIV project, pp. 15–21 in *Poultry Genetics Symposium, Proceedings*, edited by R. Preisinger. Lohmann Tierzucht, Cuxhaven, Germany.

Vanhala, T., M. Tuiskula-Haavisto, K. Elo, J. Vilkki and A. Mäki-Tanila, 1998 Evaluation of genetic variability and genetic distances between eight chicken lines using microsatellite markers. Poult. Sci. **77:** 783–790.

Weigend, S., 1999 Assessment of biodiversity in poultry with DNA markers, pp. 7–14 in *Poultry Genetics Symposium, Proceedings*, edited by R. Preisinger. Lohmann Tierzucht, Cuxhaven, Germany.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Wimmers, K., S. Ponsuksili, F. Schmoll, T. Hardge, E. B. Sonaiya *et al.*, 1999 Application of microsatellite analysis to group chicken according to their genetic similarity. Arch. Tierz. **42:** 629–639.

Wimmers, K., S. Ponsuksili, T. Hardge, A. Valle-Zarate, P. K. Mathur *et al.*, 2000 Genetic distinctness of African, Asian and South American local chickens. Anim. Genet. **31:** 159–165.

Zhivotovsky, L. A., L. Bennett, A. M. Bowcock and M. W. Feldman, 2000 Human population expansion and microsatellite variation. Mol. Biol. Evol. **17:** 757–767.

Zhou, H., and S. J. Lamont, 1999 Genetic characterization of biodiversity in highly inbred chicken lines by microsatellite markers. Anim. Genet. **30:** 256–264.

Communicating editor: G. B. Golding