*Adam Connor,[1] Ph.D. and Mark Stoneking,[2] Ph.D.*

# Assessing Ethnicity from Human Mitochondrial DNA Types Determined by Hybridization with Sequence-Specific Oligonucleotides

**REFERENCE:** Connor, A. and Stoneking, M., "**Assessing Ethnicity from Human Mitochondrial DNA Types Determined by Hybridization with Sequence-Specific Oligonucleotides,**" *Journal of Forensic Sciences,* JFSCA, Vol. 39, No. 6, November 1994, pp. 1360–1371.

**ABSTRACT:** A logistic regression model was developed to predict ethnic group from mitochondrial DNA (mtDNA) types determined by hybridization with sequence-specific oligonucleotide (SSO) probes of the two hypervariable segments of the mtDNA control region. The model was developed with, and tested against, a previously reported data set of 525 individuals from five ethnic groups (African-American, Southeast Asian, Caucasian, Japanese, and Mexican) involving 23 probes at nine regions within the two hypervariable segments [*1*]. The model correctly predicted the ethnic group of 65.3% of the overall sample; however, the success rate varied substantially among ethnic groups, with the most success obtained with Caucasians (81% correctly classified). A discriminant analysis yielded similar results. An example is given of using the model to predict the ethnic group of an SSO-type from a forensic case. Such models provide alternatives to traditional skeletal-based methods of predicting ethnicity, especially in cases where skeletal material is absent or incomplete.

**KEYWORDS:** pathology and biology, DNA, mtDNA, ethnicity

Hybridization of sequence-specific oligonucleotide (SSO) probes to DNA sequences has become a useful technique for detecting molecular genetic variation in human populations [*1–4*]. One potential use of SSO-typing is assisting identification of samples in forensic cases. Typically, a sample of forensic interest is compared with the mtDNA of an individual thought potentially to be either the source of the mtDNA or to be maternally related to the source. Matching SSO mitotypes (hereafter, *SSO-types*, or where no confusion may result, simply *mitotypes*) are taken to be evidence for identity or maternal relatedness of the samples, as detailed elsewhere [*1*]. Such comparisons can be very informative, but only where a potential match for the case sample has already been identified. It is possible, however, to use the information in an SSO-typing to predict which ethnic group the sample came from; these predictions may serve as a useful first step in narrowing the pool of potential matches. In this paper, we examine the strengths and limitations of using logistic

Received for publication 20 Oct. 1993; revised manuscript received 7 March and 22 April 1994; accepted for publication 25 April 1994.

[1] Postdoctoral Associate, Population Research Institute, The Pennsylvania State University, University Park, Pennsylvania.

[2] Assistant Professor, Department of Anthropology, The Pennsylvania State University, University Park, Pennsylvania.

This work was supported by NIJ grant 92-IJ-CX-K040 to M.S.

regression models on SSO-typings of hypervariable segments of the mtDNA control region to predict ethnic classification.

Analysis of the mtDNA control region has a number of advantages with regard to forensic cases [5]. First, mtDNA is present in high copy number. Typing mtDNA thus presents a greater probability of success in analyzing samples with either minute amounts of DNA, or in which the DNA may be highly degraded, than would typing single-copy nuclear genes. Second, the apparent haploid inheritance of human mtDNA simplifies the analysis of mixed samples. And lastly, mtDNA, and the control region in particular, show an extremely high level of polymorphism in humans [6–10].

Our strategy was to build probability models that used the binding or non-binding of various SSO probes to predict the ethnic classification of each sample. The probability models were constructed using stepwise logistic regression on the mitotypes of a sample of 525 individuals from five ethnic groups (Caucasians, African-Americans, Southeast Asians, Japanese, and Mexicans); mitotypes were defined by a series of 23 SSO probes at 9 distinct sites within the mtDNA control region [1]. The emphasis was not on interpreting the interrelationships among the probes (for example, their high-order interactions), but on building a model that could successfully predict ethnic classification. We apply the model to a forensic case that was previously analyzed [1], and discuss the prospects and limitations of this approach.

## Methods and Materials

The sample consisted of 525 unrelated individuals from five ethnic groups (142 Caucasians, 129 African-Americans, 74 Southeast Asians, 86 Japanese, and 94 Mexicans), described in more detail elsewhere [1]. MtDNA variation was previously characterized by hybridization with a series of 23 SSO probes at 9 distinct sites [1]; the frequency of each SSO variant is given in Table 1. For the purposes of this study, the data were reduced to a series of 23 dummy variables representing the binding or non-binding of each probe, and a categorical variable indicating ethnic group membership. These dummy variables were then used to predict the probability of membership in the various ethnic groups. Discriminant analysis [11] is traditionally used for this purpose, but it assumes an underlying multivariate normal error function, which is violated by the categorical nature of the data. We therefore focused on logistic regression models [12], although, for the purpose of comparison, the results of a matched discriminant analysis will be briefly discussed.

The outcomes at a given site were converted into a set of binary dummy variables, where a '1' for a particular dummy indicated the binding of the particular probe, while a '$-1$' indicated non-binding. For example, the results at the IA site were summarized by the binary dummies IA1, IA2, and IA3, with a blank at IA indicated by a $-1$ for each dummy. This resulted in a set of 23 dummy variables; the separation of sites into dummy variables allowed the selection procedure to discard the least useful probes. The IIA site was not used in the analysis because the Japanese were monomorphic for the IIA2 variant, which caused the logistic regression procedure to fail to converge.

The resulting set of 21 dummy variables possesses 210 possible first order interactions, although the 18 within-site interactions are not meaningful. Two separate strategies were used to construct models involving main effects and first order interactions. The first strategy was to use stepwise logistic regression to build a hierarchical logistic regression model. The stepwise analysis was first used to produce a main effects model with the first order interactions being added in a subsequent round of stepwise logistic regression [12]. To enter the model, a variable had to have a p-value of less than 0.15; variables whose p-values rose past 0.20 were dropped. This procedure retained 15 dummy variables as main effects; in order to maintain a satisfactory number of degrees of freedom (d.f.) per parameter

TABLE 1—*Frequencies (percent) of sequence variants at nine mtDNA SSO-defined regions for five populations. Data from* [1]. n ≡ *sample size,* B ≡ *blank.*

| Population | | | African- | Southeast | | | |
|---|---|---|---|---|---|---|---|
| *n* | | Caucasian | American | Asian | Japanese | Mexican | Total |
| Region | Variant | 142 | 129 | 74 | 86 | 94 | 525 |
| IA | 1 | 69.7 | 55.8 | 68.9 | 80.2 | 88.3 | 71.2 |
| | 2 | 18.3 | 8.5 | 4.1 | 2.3 | 1.1 | 8.2 |
| | 3 | 8.5 | 17.8 | 22.9 | 17.4 | 9.5 | 14.5 |
| | B | 3.5 | 17.8 | 4.1 | 0.0 | 1.1 | 6.1 |
| IB | 1 | 76.0 | 10.1 | 32.4 | 11.6 | 3.2 | 30.1 |
| | 2 | 10.6 | 77.5 | 44.6 | 75.6 | 67.0 | 52.6 |
| | 3 | 1.4 | 4.7 | 13.5 | 5.8 | 22.3 | 8.4 |
| | B | 12.0 | 7.7 | 9.5 | 7.0 | 7.4 | 9.0 |
| IC | 1 | 74.6 | 58.1 | 58.1 | 81.4 | 83.0 | 70.9 |
| | 2 | 9.9 | 0.8 | 17.6 | 7.0 | 1.1 | 6.7 |
| | 3 | 12.7 | 27.1 | 18.9 | 9.3 | 12.7 | 16.5 |
| | B | 2.8 | 14.0 | 5.4 | 2.3 | 3.2 | 5.9 |
| ID | 1 | 86.6 | 69.8 | 77.0 | 53.5 | 50.0 | 69.1 |
| | 2 | 9.9 | 18.6 | 23.0 | 46.5 | 47.9 | 26.7 |
| | B | 3.5 | 11.6 | 0.0 | 0.0 | 2.1 | 4.2 |
| IIA | 1 | 40.1 | 3.9 | 2.7 | 0.0 | 8.5 | 13.7 |
| | 2 | 59.9 | 95.3 | 97.3 | 100.0 | 91.5 | 86.1 |
| | B | 0.0 | 0.8 | 0.0 | 0.0 | 0.0 | 0.2 |
| IIB | 1 | 56.3 | 27.9 | 48.6 | 51.2 | 39.4 | 44.4 |
| | 2 | 8.5 | 1.6 | 9.5 | 7.0 | 10.6 | 7.0 |
| | 3 | 15.5 | 11.6 | 8.1 | 17.4 | 7.5 | 12.4 |
| | B | 19.7 | 58.9 | 33.8 | 24.4 | 42.5 | 36.2 |
| IIC | 1 | 67.6 | 35.7 | 74.3 | 73.3 | 88.3 | 65.3 |
| | 2 | 17.6 | 19.4 | 4.0 | 3.5 | 4.3 | 11.4 |
| | 3 | 2.8 | 1.5 | 14.9 | 8.1 | 2.1 | 5.0 |
| | B | 12.0 | 43.4 | 6.8 | 15.1 | 5.3 | 18.3 |
| IID | 1 | 95.1 | 74.4 | 82.4 | 83.7 | 79.8 | 83.6 |
| | 2 | 1.4 | 19.4 | 1.4 | 0.0 | 1.1 | 5.5 |
| | B | 3.5 | 6.2 | 16.2 | 16.3 | 19.1 | 10.9 |
| IIE | 1 | 38.0 | 45.7 | 28.4 | 43.0 | 28.7 | 37.7 |
| | 2 | 57.8 | 44.1 | 71.6 | 53.5 | 66.0 | 57.2 |
| | B | 4.2 | 10.1 | 0.0 | 3.5 | 5.3 | 5.1 |

estimated, only the first (most important) five interactions, as determined by the stepwise procedure, were retained in the model.

Non-hierarchical models, in which an interaction could be included without including the associated main effects, are usually avoided in logistic regression—not because the models are necessarily inappropriate, but because significant interactions in a model without main effects are impossible to interpret, and interpretation of the coefficients is usually the goal of modeling the data. In this case, however, the emphasis was on building a parsimonious description of the data, rather than on evaluating the theoretical significance of the coefficients of the model. The model coefficients have no known theoretical significance for the SSO data, beyond their descriptive value. The second strategy was therefore to consider non-hierarchical logistic regression models.

In a non-hierarchical analysis, main effects and interactions are considered simultaneously, rather than sequentially; the interaction between two sites can be added to a model that does not include the corresponding main effects. This would require dummy variables for all possible interactions. To reduce the number of first order interactions to a smaller set that could be feasibly included in the stepwise analysis, the 192 potentially meaningful interactions were first evaluated by means of the likelihood ratio statistic $G^2$ [*13*]; for example, to test the IA1*ID1 interaction, a full model containing both the main effects and the interaction was compared to a reduced model containing only the main effects in question. Since the $G^2$ for an interaction seemed likely to decline when the interaction was included in a more complete model that included other main effects and interactions, the relatively conservative $G^2$ cut-off for inclusion was set at 14, with a p-value of 0.007 for 4 d.f. The cut-off was selected by inspection of the results to provide a reasonable number of probably significant interactions for testing. In point of fact, no interaction with a $G^2$ of less than 16 was included in a final stepwise model.

The preceding cut-off yielded 11 interactions for testing. Because including interactions involving both ID1 and ID2 in a model resulted in non-convergence of the logit model, two possible sets of interactions were considered: the set containing ID1 interactions, and the set containing ID2 interactions. Forward stepwise logistic regression was implemented in a Gauss program [*14*] that made calls to the Gauss *Quantal Response* application [*15*]; variables whose inclusion resulted in a chi-squared improvement with an approximate *P*-value of 0.05 were retained. The addition of variables ceased when none resulted in a significant improvement of the model. At that point, the program switched to stepwise deletion of any variables with *P*-values greater than 0.05. When no variables in the model had *P*-values greater than 0.05, the remaining variables constituted the final model. Although in theory such a procedure may not detect all important variables or interactions, in this instance it seemed to produce an equally effective and slightly more parsimonious model than the hierarchical procedure.

## Results

### *Models*

The hierarchical stepwise model resulted in a $G^2$ value of 646.20, with 80 d.f. The two non-hierarchical models (Model 1 and Model 2) differed only slightly, with overall $G^2$ values of 612.512 (68 d.f.) and 603.16 (64 d.f.), respectively. The $G^2$ values of these models could not be simply compared, since no model was a subset of any other, but the differences appeared to be marginal, considering the difference in degrees of freedom. The hierarchical model correctly predicted the ethnic group of 63.6% of the overall sample; both non-hierarchical models correctly predicted the ethnicity of 65–66% of the sample. Since the second non-hierarchical model (Model 2) was slightly more parsimonious than the others and provided about as good a fit to the data, only the variables included in Model 2 were considered in the remainder of the analysis.

The success of the logit model in classifying the various ethnic groups varied, as is shown in Table 2. The model correctly classified 81.0% of Caucasians, 73.4% of Mexicans, 69.0% of African-Americans, 54.7% of Japanese, and only 31.1% of Southeast Asians. To provide some perspective on these results, a linear discriminant function was fit to the same variables and interactions using SYSTAT [*16*]; the resulting classification was very similar to the logit model (Table 3). The two procedures agreed in 487/525 = 92.8% of the cases. Despite this close agreement, they were correct in only 327 of the 487 cases, a success rate of 67% vs. 65% for the procedures separately. This suggests the possibility that the remaining 35% of the sample may be inherently difficult to classify—a possibility that will be explored in the subsequent discussion.

### *An Example*

To give a concrete example of predicting the ethnic group of a sample from the SSO-typing, we re-analyze a case presented previously [*1*], in which the skeletal remains of a

TABLE 2—*Classification of ethnic groups by Model 2. Each row shows the group membership (in percent) predicted by the model for the actual group.*

| Actual | Caucasian | African-American | Southeast Asian | Japanese | Mexican |
|---|---|---|---|---|---|
| Caucasian | 81.0 | 4.2 | 4.9 | 6.3 | 3.5 |
| African-American | 10.9 | 69.0 | 2.3 | 6.2 | 11.6 |
| Southeast Asian | 28.4 | 8.1 | 31.1 | 16.2 | 16.2 |
| Japanese | 10.5 | 8.1 | 5.8 | 54.7 | 20.9 |
| Mexican | 8.5 | 6.4 | 2.1 | 9.6 | 73.4 |

TABLE 3—*Classification of ethnic groups by the discriminant function analysis. Each row shows the group membership (in percent) predicted by the discriminant function for the actual group.*

| Actual | Caucasian | African-American | Southeast Asian | Japanese | Mexican |
|---|---|---|---|---|---|
| Caucasian | 80.3 | 4.9 | 4.9 | 7.0 | 2.8 |
| African-American | 9.3 | 66.7 | 2.3 | 10.1 | 11.6 |
| Southeast Asian | 32.4 | 6.8 | 36.5 | 6.7 | 17.6 |
| Japanese | 11.6 | 9.3 | 9.3 | 48.8 | 20.9 |
| Mexican | 8.5 | 5.3 | 2.1 | 8.5 | 75.5 |

child were typed and found to exactly match the SSO mitotype of a Caucasian mother whose own child had disappeared a year and a half before. The match was subsequently confirmed by DNA sequencing. This mitotype differed from all 142 Caucasian mitotypes observed in the SSO-type database. Table 4 summarizes the calculations of the log odds that the unknown sample belonged to the various ethnic groups, compared to the probability that the unknown sample was of Mexican origin. (The ethnic group chosen as the basis for comparison—here, Mexicans—is arbitrary; the results would be identical if log odds were calculated versus, say, African-Americans.) If the probability that the sample comes from ethnic group $g$ is denoted $\pi_g$, the log odds that the sample belonged to group $g$ vs. $g'$ equal $\ln\left(\frac{\pi_g}{\pi_{g'}}\right)$, which we may denote $L_{g,g'}$. Thus, if $L_{g,5}$ denotes the log odds of group $g$ vs. group 5 (Mexicans),

$$\pi_5 = \Pr\{\text{Sample is of Mexican origin}\} = \frac{1}{1 + \sum_{g=1}^{4} \exp(L_{g,5})}.$$

Probabilities $\pi_1$ through $\pi_4$ can then be calculated via

$$\pi_g = \pi_5 \exp(L_{g,5}).$$

The result of these calculations, shown in Table 4, is that the probability that the sample was of Caucasian origin is estimated to be 70.6%.

## Discussion

### *Limitations of the Models*

The probability model employed has a number of potential limitations. The full additive logit model, which would include all possible probes at all nine sites, is formally equivalent to a log-linear model that includes first level interactions—that is, non-independence between ethnic group and typing site. Such a model (including the possibility of blanks) would involve fitting cell values to a table with 414,720 cells. Since there are only 525 observations, it would necessarily be a sparse table, with many sample zeros; the mean cell count would be only 0.00127.

The fitted model does not require knowing the binding status of as many probes, which reduces the number of cells in the table to 2,592. Nevertheless, the average cell frequency still falls well below 5 (often suggested as a rule-of-thumb), and it is reasonable to question the effects of sample size on the estimated $G^2$ values. There is some evidence that $G^2$ behaves conservatively when most expected frequencies are smaller than 0.5 [*13,17–19*]. It is recommended that

$$\frac{n}{N} > \sqrt{\frac{10}{N}}$$

where $n$ is the sample size and $N$ is the number of cells in the table [*17*]. For the full table, $\frac{n}{N} = 0.00127$, whereas $\sqrt{\frac{10}{N}} = 0.0049$, so this condition would not be fulfilled. For the fitted model, $\frac{n}{N} = 0.2025$, whereas $\sqrt{\frac{10}{N}} = 0.0621$, suggesting that the $G^2$ statistic will behave appropriately. In order to test the significance of the model more systematically, a permutation test was applied, in which the column containing ethnic classification was randomly permuted, and Model 2 was then fit to the permuted data. In 999 trials, the observed $G^2$ value of Model 2 of 603.16 and percentage correctly predicted of 65.33% were never exceeded, yielding an approximate significance level of 0.001 for each [*20,21*].

Another well-known limitation to building probability models for any particular data set is over-prediction: the model tends to over-adapt itself to fitting the sample, with the consequence that the ability of the model to make predictions is overstated. It is therefore essential to perform some sort of cross-validation. In order to test for allocation bias, the estimates for Model 2 were jackknifed: each case was removed in turn, the model parameters were re-estimated, probabilities of group membership for each case were recalculated, and the case was classified as belonging to the group for which the estimated probability of membership was highest. In this way, no sampled mitotype was used to estimate the model parameters or predict its own classification. The results were substantially the same, with 63.2% correctly classified overall, suggesting that the allocation bias was not high.

TABLE 4—*Probabilities of group membership for the case analysis, based on Model 2.*

| Group | LOD (vs.<br>Group 5) | Odds (vs.<br>Group 5) | Probability of<br>Group |
|---|---|---|---|
| 1 | 5.205 | 182.215 | 70.6% |
| 2 | 3.824 | 45.791 | 17.7% |
| 3 | 3.314 | 27.499 | 10.7% |
| 4 | 0.481 | 1.617 | 0.6% |
| 5 | — | — | 0.4% |

The limited success of the logit model in classifying certain ethnic groups probably has several causes. The most obvious possibility is that the sequences used to design the SSO probes were in large part (19/52) from Caucasians [1]. Thus, despite the fact that all five populations showed statistically significant differences in probe frequencies at each of the nine sites [1], the success of the logit model in classifying Caucasians (81.0% were correctly classified) may suggest that predicting ethnic classification from SSO probe data would work better if the probes were designed using DNA sequences from each of the targeted ethnic groups. However, the pilot sample also included 23 Asians, so it is not at all clear that this is the reason the models did so well with Caucasians and so poorly with Southeast Asians. The fact that a number of sites did not make it into the final logit models does suggest that some probes could be deleted, while other probes might be added to detect variants unique to Southeast Asians.

A second possible cause of the limited ability of the logit to classify some ethnic groups is that socially defined ethnic groupings do not exactly correspond to mitochondrial cladistic history; the ethnic groups include "migrants" from other lineages. This pattern of evolutionary sharing of some mitotypes across ethnic groups is also observed in phylogenetic reconstruction of mtDNA sequences [8–10]. The logit and the discriminant function succeed and fail in classifying almost identical cases, and when they fail, they often fail in the same ways. This suggests the possibility of heterogeneity within ethnic groups. Examination of the logit probability estimate that a mitotype does indeed belong to its nominal ethnic group—that is, the estimated probability of correct classification—reveals that the distribution of these estimates is bimodal, as illustrated in Fig. 1. The discriminant probabilities show the same pattern. The bimodality suggests that many of the 'misclassified' mitotypes may indeed belong to other mitochondrial lineages, whatever their socially defined ethnicity. To the extent that this is true, there may be an upper limit to the success of logistic probability models for predicting ethnicity from mitotypes. This may be particularly true when the ethnic groupings are themselves not very specific, but it also suggests that there may be inherent limitations to the ability of mtDNA to accurately classify ethnic groups, especially since such classification relies on morphological traits that are not influenced by mtDNA.

Finally, it was probably unfortunate that the IIA site was dropped from the analysis, since it clearly provides some information as to the probability of Japanese ancestry. Failure to include the IIA site was an artifact of using a logit model, for the following reason. At the IIA site, the Japanese were monomorphic for the IIA2 variant. Thus, the SSO-type

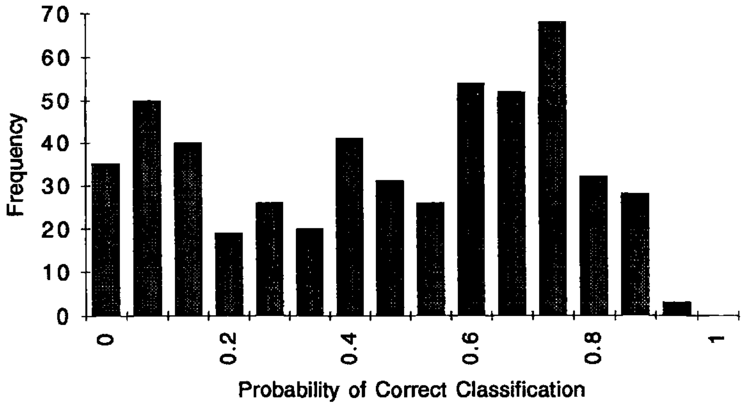

FIG. 1—*The distribution of the estimated probability of classifying each sampled mitotype into its nominal ethnic group, as calculated by Model 2.*

database implies that a mitotype with a IIA1 variant is certain not to be Japanese; that is, Pr{Mitotype is Non-Japanese Binding of the IIA1 probe} = 1. But in a logit model,

$$\Pr\{\text{Mitotype is Non-Japanese Binding of the IIA1 probe}\} = \frac{e^{c+bX}}{1+e^{c+bX}},$$

where $b$ and $c$ are constants and $X$ is the dummy variable representing binding or nonbinding of the IIA1 probe. For this expression to equal one, $b$ must approach infinity, and the logistic regression procedure fails to converge. One way to force convergence would be to arbitrarily pick a Japanese mitotype and change the IIA binding status from IIA1 to IIA2. Another would be to use discriminant analysis; in the discriminant analog to the univariate example presented above, the Japanese monomorphism would be represented by a sample variance of zero at the IIA site. But since discriminant analyses assume identical covariance matrices across populations, the sample variances would be pooled into a common estimate, and the difficulty would not be so apparent. In order to evaluate the effect of neglecting the IIA site, we tested a discriminant function model that included all main effects and all 11 of the "significant" interactions. The model correctly classified 68.0% of the overall sample, and both Japanese (58.1%) and Southeast Asians (40.5%) did show some improvement in the accuracy of their classification. This suggests that neglecting the IIA site did lead to a modest reduction in the predictive power of the model.

Another alternative that allows inclusion of a monomorphic site is to use metric space models, in which mitotypes are assigned to groups on the basis of their distance from group centroids, with the distance being calculated according to some metric. Various metrics were tried, for example, Euclidean, city block, and weighted variants of same. All of the metric models correctly classified 53–56% of the sample, and none was more successful than Model 2 at classifying any of the ethnic groups.

### *Typing Strategies*

An important practical question concerns whether it is better to cover many sites with a few probes each or a few sites with many probes each. Considering blanks suggests that the former approach is superior. Two probes at one site yield three possibilities: probe 1, probe 2, or a blank; while two probes at two sites yield four possibilities: probe 1/probe 2, probe 1/blank; blank/probe 2, or blank/blank. This pattern is exhibited by the contribution of probes to measures of diversity: Figures 2 and 3 show the pattern of declining diversity which occurs when probes are removed in a stepwise fashion, at each step preserving as much diversity as possible. The pattern in both these figures is of resisting the loss of sites: in Fig. 2, a site isn't completely lost until the removal of IIA1, on the fourteenth step; in Fig. 3, a site isn't lost until the removal of IID2, also on the fourteenth step. Since a site would have had to be lost by the fifteenth step, the empirical pattern comes very close to preserving the number of polymorphic sites for as long as is possible, suggesting that it may be better to spread any fixed number of probes over as many sites as possible. This pattern is also exhibited in the probes included in Model 2: with the exception of the IIA site (which was monomorphic in the Japanese and not considered in the analysis), at least one probe was included from every probe site. And at only one site, IB, were *all* the available probes included in the final model.

Such considerations are counterbalanced by the problems of sparsity, discussed above, which probably limits the number of probe sites that can be usefully included in a logistic model. For example, if we limit ourselves to a single probe at each site—guaranteeing the maximum number of sites for a given number of probes—the Koehler and Larntz criterion [*18*] implies that

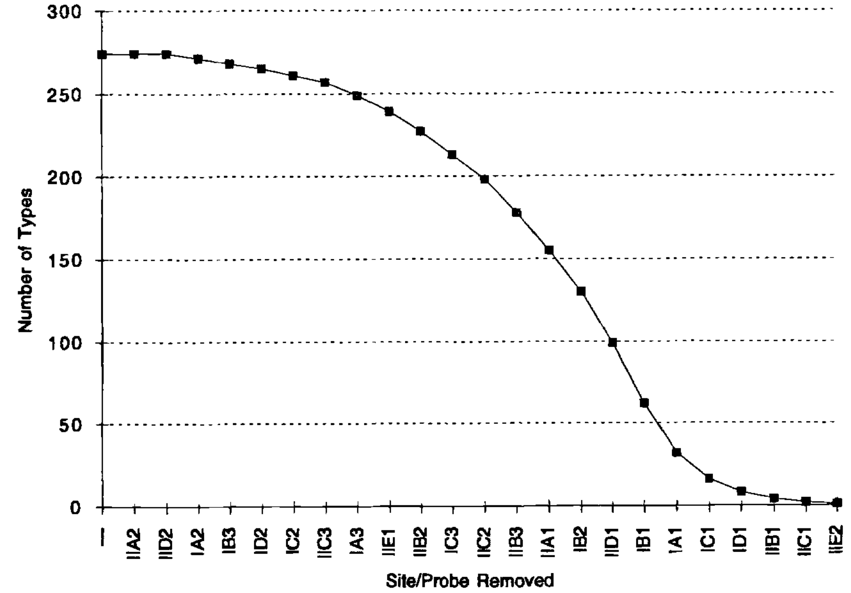

FIG. 2—*The number of distinct mitotypes present in the database of 525 SSO-types as probes were progressively removed from consideration. At each step, the probe removed was chosen so that the remaining number of distinct mitotypes was maximized.*

$$2^S < \frac{n^2}{10},$$

where $S$ is the number of probe sites. Thus,

$$S < \frac{2}{\ln(2)}\ln(n) - \frac{\ln(10)}{\ln(2)}.$$

When $n = 525$ (the size of the SSO-type database), the maximum number of sites that could be considered in any final model is fewer than 15. Because the limit goes up only as the log of the sample size, even very large sample sizes will suffer from this limitation; a sample size of 5000 would require the number of sites to be less than or equal to about 21.

### *Comparative Performance and Future Prospects*

There have been several studies in the past that sought to predict ethnicity (as well as other characteristics, such as sex) from forensic samples on the basis of skeletal biology, typically via a discriminant function analysis of selected traits. Giles and Elliot [22] studied the ability of eight cranial measurements to distinguish among American whites, American blacks, and American Indians, with sex as a known covariate; their discriminant analysis correctly predicted the ethnic group of 82.6% of the males and 88.1% of the females. İşçan [23] used discriminant function analysis on three measurements (selected via stepwise discriminant analysis) of the pelvis to distinguish between black and white Americans and had an overall success rate of 83% in males and 88% in females when age, which was known, was included in the analysis. Without age, the rates dropped to 79% and 83%, respectively. DiBennardo and Taylor [24] performed stepwise discriminant analysis on 32 measurements of the post-cranial skeleton to distinguish American whites and blacks; their final model retained 15 of the measurements and correctly identified the ethnic group of 95% of the sample. A thorough review of the identification of ethnicity from skeletal remains may be found in İşçan [25]. In general, the studies reviewed were successful in distinguishing American whites from American blacks 80–95% of the time.

Determining ethnicity from skeletons is therefore superior to the present mitotype analysis, although the correct classification of Caucasians from mitotype analysis does approach the success of discriminant analysis of skeletal traits. Additional SSO-probes, particularly designed from sequences from under-represented groups, should increase the classificatory power of mitotype analysis. Furthermore, skeletal discriminant analysis does suffer from the limitation of requiring a relatively intact skeleton, whereas mitotype analysis can be applied to limited or fragmented skeletal samples [1], as well as other biological material, including blood, teeth [26], and hair [27]. Moreover, mtDNA is a single segregating locus; it seems likely that analysis based on multi-locus models (which would be more comparable to analyzing the entire post-cranial skeleton) would be even more successful at ethnic classification.

### *Recommendations*

In cases where a relatively complete skeleton is available, discriminant analysis of various skeletal characters remains the most powerful method of predicting ethnicity. But when skeletal material is incomplete or absent, applying the logit model to mitotypes provides a reasonably accurate method of ethnic classification, particularly for Caucasians. It is anticipated that probes designed especially to detect the variation in other ethnic groups will improve the performance of the models, and designing and testing such probes is currently underway. In some cases it may be useful to construct models that address more specifically the ethnic groups of interest—for example, "Italian" instead of "Caucasian," or "Nigerian" instead of "African." The pattern with which probes generate diversity

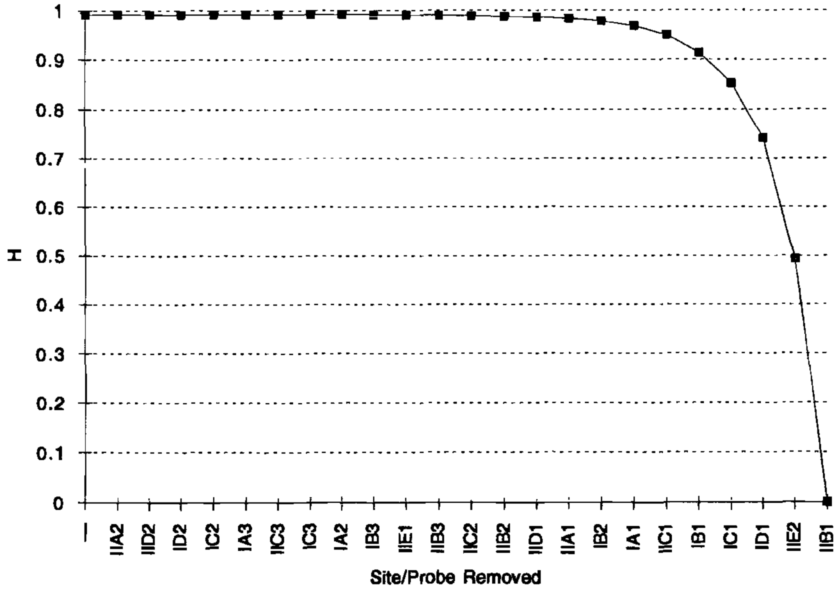

FIG. 3—*The genetic diversity, H, of the 525 SSO-types as probes were progressively removed from consideration. At each step, the probe removed was chosen so that the genetic diversity was maximized. Genetic diversity = H = 1 − Σ $X_i^2$, where $X_i$ is the frequency of the ith mitotype.*

suggests that investigators should try to maximize the number of distinct sites covered, and statistical considerations suggest that 15–20 probes is about the maximum that can be meaningfully analyzed in a logit model with feasible sample sizes. This implies that if logit models are to be used the best strategy is to look for 15 or so sites that are polymorphic in all of the populations of interest.

## References

[1] Stoneking, M., Hedgecock, D., Higuchi, R. G., Vigilant, L., and Erlich, H. A., "Population Variation of Human MtDNA Control Region Sequences Detected by Enzymatic Amplification and Sequence-Specific Oligonucleotide Probes," *American Journal of Human Genetics,* Vol. 48, 1991, pp. 370–382.

[2] Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B., and Erlich, H. A., "Analysis of Enzymatically Amplified β-globin and HLA-DQα DNA with Allele-Specific Oligonucleotide Probes," *Nature,* Vol. 324, 1986, pp. 163–166.

[3] Goedde, H. W., Singh, S., Agarwal, D. P., Fritze, G., Stapel, K., and Paik, Y. K., "Genotyping of Mitochondrial Aldehyde Dehydrogenase in Blood Samples Using Allele-Specific Oligonucleotides: Comparison with Phenotypic Hair Roots," *Human Genetics,* Vol. 81, 1989, pp. 305–307.

[4] Helmuth, R., Fildes, N., Blake, E., Luce, M. C., Chimera, J., Madej, R., Gorodezky, C., Stoneking, M., Schmill, N., Klitz, W., Higuchi, R., and Erlich, H. A., "HLA-DQα Allele and Genotype Frequencies in Various Human Populations, Determined by Using Enzymatic Amplification and Oligonucleotide Probes," *American Journal of Human Genetics,* Vol. 47, 1990, pp. 515–523.

[5] Wilson, M. R., Stoneking, M., Holland, M. M., DiZinno, J. A., and Budowle, B., "Guidelines for the Use of Mitochondrial DNA Sequencing in Forensic Science," *Crime Laboratory Digest,* Vol. 20, 1993, pp. 68–77.

[6] Aquadro, C. F. and Greenberg, B. D., "Human Mitochondrial DNA Variation and Evolution: Analysis of Nucleotide Sequences from Seven Individuals," *Genetics,* Vol. 103, 1983, pp. 287–312.

[7] Cann, R. L., Brown, W. M., and Wilson, A. C., "Polymorphic Sites and the Mechanism of Evolution in Human Mitochondrial DNA," *Genetics,* Vol. 106, 1984, pp. 479–499.

[8] Cann, R. L., Stoneking, M., and Wilson, A. C., "Mitochondrial DNA and Human Evolution," *Nature,* Vol. 325, 1987, pp. 31–36.

[9] Horai, S. and Hayasaka, K., "Intraspecific Nucleotide Sequence Differences in the Major Noncoding Region of Human Mitochondrial DNA," *American Journal of Human Genetics,* Vol. 46, 1990, pp. 828–842.

[10] Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C., "African Populations and the Evolution of Human Mitochondrial DNA," *Science,* Vol. 253, 1991, pp. 1503–1507.

[11] Mardia, K. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis,* Academic Press, New York, 1979.

[12] Hosmer, D. W. and Lemeshow, S., *Applied Logistic Regression,* John Wiley & Sons, New York, 1989.

[13] Agresti, A., *Categorical Data Analysis,* John Wiley & Sons, New York, 1990.

[14] Aptech Systems Inc., *Gauss Manual,* Aptech Systems, Inc., Kent, Ohio, 1992.

[15] Long, J. S., "Quantal Response Models," Gauss Applications Manual, V. 2.01. Aptech Systems, Inc., Kent, Ohio, 1990.

[16] Wilkinson, L., *SYSTAT: The System for Statistics,* SYSTAT, Inc., Evanston, Illinois, 1989.

[17] Larntz, K., "Small-sample Comparison of Exact Levels for Chi-squared Goodness-of-Fit Statistics" *Journal of the American Statistical Association,* Vol. 73, 1978, pp. 253–263.

[18] Koehler, K. and Larntz, K., "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *Journal of the American Statistical Association,* Vol. 75, 1980, pp. 336–344.

[19] Koehler, K., "Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables," *Journal of the American Statistical Association,* Vol. 81, 1986, pp. 483–493.

[20] Edgington, E. S., *Randomization Tests,* Marcel Dekker, Inc., New York, 1987.

[21] Noreen, E. W., *Computer-Intensive Methods for Testing Hypotheses: An Introduction,* John Wiley & Sons, Inc., New York, 1989.

[22] Giles, E. and Elliot, O., "Race Identification from Cranial Measurements," *Journal of Forensic Sciences,* Vol. 7, 1962, pp. 147–157.

[23] İşcan, M. Y., "Assessment of Race from the Pelvis," *American Journal of Physical Anthropology,* Vol. 62, 1983, pp. 205–208.

[24] DiBennardo, R. and Taylor, J. V., "Multiple Discriminant Function Analysis of Sex and Race in the Postcranial Skeleton," *American Journal of Physical Anthropology,* Vol. 61, 1983, pp. 305–314.

[25] İsçan, M. Y., "Rise of Forensic Anthropology," *Yearbook of Physical Anthropology,* Vol. 31, 1988, pp. 203–230.

[26] Ginther, C., Issel-Tarver, L., and King, M.-C., "Identifying Individuals by Sequencing Mitochondrial DNA from Teeth," *Nature Genetics,* Vol. 2, 1992, pp. 135–138.

[27] Vigilant, L., Pennington, R., Harpending, H., Kocher, T. D, and Wilson, A. C., "Mitochondrial DNA Sequences in Single Hairs from a Southern African Population," *Proceedings of the National Academy of Sciences USA,* Vol. 86, 1989, pp. 9350–9354.

Address requests for reprints on additional information to
Mark Stoneking, Ph.D.
Dept. of Anthropology
409 Carpenter Bldg
PSU
University Park, PA 16802