Authors’ Response

Sir:

Kranke et al. (1) contest the conclusion of our study of the “CODIS STR Loci Data from 41 Sample Populations” (2), that “there was little evidence for departures from Hardy-Weinberg expectations (HWE) in any of the populations” on the grounds that: (a) our application of the Bonferroni adjustment for multiple testing to our data on 12 to 13 loci, studied in each of the 41 populations, is inappropriate; (b) we disregarded the “significant” clustering of departures from HWE in two populations (Salishan and Navajo); and (c) we failed to pay attention to the distinctiveness of the Native American populations to explain the clusters of deviant test results in these populations. Further, Kranke et al. (1) contend that our data “provides significant evidence that at least three loci in Navajos (FGA, D7S820, and TH01) as well as Salishans (D3S1358, FGA, and D7S820) do not adhere to HWE” and hence, “the product rule should not be used to estimate the rarity of genotypes involving those loci in those populations unless corrective factors are involved.”

We demonstrate that their contentions are based on flawed statistical as well as population genetics logic. Moreover, even if their specifically chosen observations on our reported test results are taken in isolation (of the rest of the study results) as they did (1), their comments are of no consequence. The practices employed by the United States forensic community for the last decade generally do not use the strict product and, instead, allow for departures from HWE to estimate the rarity of genotypes in forensic computations (3).

The first statistical flaw of Kranke et al.’s arguments is in regard to the Bonferroni adjustment. It is true that the Bonferroni adjustment for multiple tests applies when independent sets of data are used to test the same hypothesis. In the context of DNA Forensics, however, the issue is: “Do the genotype frequencies at the CODIS STR loci conform to HWE in population samples”? Our study (2) involves the examination of this specific hypothesis. What would differ from one locus-population combination to another in the dataset we analyzed is the magnitude of departure from HWE, should the null hypothesis be not true. In the terminology of meta-analysis, this is called the variation of size effects (4), and of the various alternative forms of meta-analytic multiple test correction of p-values (see Table 15.2 of Ref 5), the Bonferroni adjustment pays less attention to the varying size effects of multiple sets of data, though it is the simplest and most widely used (6). Notably, applications of Bonferroni-type of adjustment of significance levels are also common in the current genome-scan studies of mapping complex disease traits by multilocus linkage and association analyses, where the markers vary from one test to another (7,8). Thus, its application is not unique to forensic database analyses.

Kranke et al. (1) committed another statistical error by implying that the Bonferroni adjustment assumes “the loci are equivalent with regard to HWE and discriminating power.” This simply is not true, because the Bonferroni adjustment is based on the null distribution of the minimum p-values of all of the locus-population combination of tests (see Table 15.2 of Ref 5), each of which is, in turn, dependent on the product of effect size and sample size of the respective dataset (4). Thus, even though the order statistic (i.e., the minimum of all observed p-values), against which the revised significance level of the Bonferroni adjustment is compared, does not utilize the full spectrum of distribution of the p-values, it is incorrect to say that the variations of size effects and discriminatory power in the individual locus-population combinations are ignored in applications of the Bonferroni adjustment.

Nonetheless, even if the Bonferroni adjustment is regarded as crude, the approach taken by Kranke et al. (1) to take clusters of deviant test results in isolation of the others, is statistically flawed. By concluding that “3 of the 31 noted departures from HWE would randomly be found in one of the 41 populations . . . is itself unlikely” they made two fundamental errors. First, their conclusion of unlikeliness is based on a p-value (the test procedure of which is not mentioned by Kranke et al.) of 0.078, not supportive of the level of significance generally employed in data analysis. Second, and more importantly, in order to test the true randomness, they should have considered the entire array of our test results (including making distinctions of possible size effects of departures in different population-locus combinations), not merely the most clustered occurrence of significant results. Their approach is effectively a “vote-counting” method of meta-analysis (9), which has been known to be inappropriate for synthesis of multiple test results when the test statistic values (individual p-values of the exact test in our case) are reported for each study (10).

Table 1 shows the relevant data (directly extracted from Ref 2) for a true test of randomness considering the entire array of our tests results. With this tabulation, the randomness of occurrences of significant (at 5% level) or non-significant departures from HWE is tested by $2 \times c$ contingency table analysis for each of the five major groups of populations, as well as for the pooled data, taking into account each individual locus-population test result. The significance test for randomness is done by following the permutation algorithm suggested by Roff and Bentzen (11), which does not assume any large sample property of the (Chi-square) test statistic. The last two columns of Table 1 clearly show that the occurrences of significant deviation from HWE in individual locus-specific tests are random between population samples within each of the five major

<table>
<thead>
<tr>
<th>Population Group</th>
<th>Number of Significant Departures*</th>
<th>Total Number of Population Samples</th>
<th>χ²</th>
<th>p-Value¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africans</td>
<td>11</td>
<td>7</td>
<td>7.41</td>
<td>0.888</td>
</tr>
<tr>
<td>US Caucasians</td>
<td>9</td>
<td>8</td>
<td>10.36</td>
<td>0.267</td>
</tr>
<tr>
<td>Hispanics</td>
<td>8</td>
<td>4</td>
<td>4.08</td>
<td>1.0</td>
</tr>
<tr>
<td>Asians</td>
<td>6</td>
<td>2</td>
<td>4.11</td>
<td>1.0</td>
</tr>
<tr>
<td>Native Americans</td>
<td>7</td>
<td>10</td>
<td>6.09</td>
<td>0.426</td>
</tr>
<tr>
<td>Pooled</td>
<td>41</td>
<td>31</td>
<td>41.79</td>
<td>0.376</td>
</tr>
</tbody>
</table>

* Extracted from exact tests of departure from the data presented in Ref 2, the significant departures from HWE at 5% level are: in African Americans: 0 of 13 (0/13) of the locus-specific tests for the FBI sample, 2/13 for Bahama, 1/13 for Jamaica, 1/13 for Trinidad, 1/13 for California, 0/13 for Alabama, 1/13 for Florida, 0/13 for Virginia, 0/13 for New York, 1/13 for Illinois, and 0/12 for Minnesota; in US Caucasians: 1/13 for FBI, 0/13 for California, 2/13 for Alabama, 1/13 for Florida, 0/13 for Virginia, 0/12 for New York, 3/13 for Michigan, 1/12 for Minnesota, and 0/12 for Canada; for Hispanics: 0/13 for FBI, 0/13 for California, 1/13 for Florida, 1/13 for New York, 1/13 for Michigan, 0/12 for Minnesota, 0/13 for Arizona, and 1/13 for Mexico; in Asians: 0/13 for Chinese, 0/13 for Japanese-1, 0/13 for Japanese-2, 0/13 for Koreans, 1/13 for Vietnamese, and 1/13 for General Asians; and in Native Americans: 0/13 for Michigan, 1/13 for Minnesota, 1/13 for Apache, 3/13 for Navajo, 1/12 for Northern Ontario, 3/12 for Salishan, and 1/12 for Saskatchewan samples.

† The significance levels (p-values) were obtained by permutation tests of the respective $2 \times c$ contingency table chi-square (with algorithm of ref. 11) with 10,000 replications of permutations.
population groups, as well as in the pooled sample of 41 populations. This randomness of deviations from HWE is in direct contradiction with the result obtained using the inappropriate method by Krane et al. (1).

To further illustrate that our observed 31 deviant test results out of the 524 tests performed are in accordance with the expectation of multiple testing, one can also use Fisher’s (12) original theory; namely, under the null hypothesis that HWE applies for all of the 524 tests performed, the distribution of the \( p \)-values (of exact tests) should follow a uniform distribution. Thus, the array of individual locus-population specific \( p \)-values can be tabulated for each group of populations to check for conformity with the uniform distribution. In the total data, the 524 \( p \)-values do not deviate from a uniform distribution by the Kolmogorov-Smirnoff non-parametric test (\( Z = 1.154 \), with 2-sided \( p \)-value of 0.134 with 10,000 replications of simulation). Further, the observed distributions of \( p \)-values across the five groups of populations (utilizing the 142 test results in African Americans, 114 in U.S. Caucasians, 103 in Hispanics, 78 in Asians, and 87 in the Native Americans) are seen to be homogenous (by a \( 5 \times 20 \) contingency table Chi-square test, obtained by grouping the \( p \)-values into 20 classes of equal interval length), since this test resulted in a Chi-square value of 57.99 whose empirical level of significance is 0.948 (with 10,000 replications of permutations). Thus, even if our original Bonferroni adjustment were to be questioned, when the data represented in (2) are re-analyzed by at least two other meta-analytic methods, each of which recognizes locus-population specific differences of discriminating power and size effects of possible departure from HWE, we still find no over-all significance of deviations from HWE (i.e., the hypothesis that at least one of the locus-population combination of genotype frequencies deviate from HWE is rejected) when the array of all 524 test results are considered simultaneously. In addition, homogeneity of distribution of \( p \)-values as well as randomness of occurrence of \( p \)-values below the nominal level of significance of 5% (in individual tests) suggest that there is no significant clustering of deviant test results in our data reported in (2).

The population genetic errors committed by Krane et al. (1) are more fundamental than the relatively obvious statistical errors. For example, in response to our comment that the majority of the initially found departures from HWE are due to genotypes consisting of rare alleles (2), they assert that in the Navajo population, “not even one of the homozygotes observed” at the deviant loci (FGA, D7S820, and TH01) “were homozygous for rare alleles.” However, they failed to recognize that for this population all of the rare (below five counts) alleles (FGA-28, D7S820-7, and TH01-9) occur in heterozygote forms. This should have been obvious to Krane et al. (1). Moreover, the frequencies of the heterozygotes with rare alleles are in excess of their respective HWE expectations, contributing to the overall significant departures from HWE at these loci. Incidentally, 11 of the 31 initial deviations from HWE, noted in our data analysis, are due to excess overall heterozygosity at the respective loci, and hence our statement with regard to rare alleles contributing to deviations from HWE cannot be equated to excess homozygosity. Of particular note, of the clusters of deviations from HWE in the Navajo and Salishan samples (three per population), erroneously claimed by Krane et al. (1), three (D7S820 in Navajo, and D3S1358 and FGA in Salishan) show heterozygosity excess (see Table 5 of Ref 2); thus population substructure within these populations, or genetic drift are not likely to account for these departures (because both of these factors would have caused heterozygote deficiency and not excess as observed).

The second population genetic error committed by Krane et al. (1) relates to their statement that the clusters of HWE deviations in the Navajo and Salishan samples is consistent with the relatively large \( F_{ST} \) values for the Native Americans (average \( F_{ST} \) of 0.0282) that we reported. This is clearly wrong; \( F_{ST} \), computed in our work, reflects the standardized allele frequency variation across the seven Native American populations sampled. This has no direct relationship with deviations from HWE within each population, which would have been reflected if they had estimated \( F_{IS} \) (13,14). Of course, as noted in our prior work (15), loci/populations with smaller within-population gene diversity generally exhibit larger \( F_{ST} \) in substructured populations. Nevertheless, as noted above, since at least three of the six HWE deviations in the Navajo and Salishan samples cannot be ascribed to hidden population substructure within these populations, there is no foundation to suggest a consistency of the observed deviations from HWE (within each sample) with their group-level \( F_{ST} \).

By claiming that “the Navajos have significantly higher rates of allele sharing than any other population, which suggests a greater degree of substructure within the Navajo population,” Krane et al. (1) made an even more egregious population genetic error. The data presented in their Table 1 are not at all surprising, and not novel either, as our group already has shown that with respect to VNTR as well as STR markers, individuals of smaller populations generally share more alleles, compared with those of larger populations (16,17). However, Krane et al. (1) apparently are not aware that the larger rate of allele sharing is a consequence of a reduced level of genetic diversity in smaller sized populations, and not necessarily due to their inherently higher hidden substructuring (17). This is so, because in each of these populations, one can compute the distribution of allele sharing based on allelic independence within and between loci (18), and as shown in (17), in spite of the larger rates of allele sharing between individuals of populations of reduced genetic diversity, the distribution of allele sharing between individuals can be in accordance with their expectations based on the assumption of mutual independence of alleles within and across loci. Data presented in Table 2 on the 23 population samples in which the initial 31 deviations from HWE had been observed in our work (2) exhibit this phenomenon. Clearly this shows that in each of the 23 populations where at least one initial deviation from HWE was observed (2), the observed rates of allele sharing are in accordance with their respective expectations under the hypothesis of mutual independence of alleles within and across loci (i.e., HWE for each locus, and pairwise as well as higher-order linkage equilibria (LE) between loci). This is true, in spite of the fact that the populations with lower average heterozygosity exhibit a higher rate of allele sharing. Thus, the contention that a higher degree of allele sharing in a population is reflective of its greater degree of substructuring is clearly incorrect. In fact, the analysis shown in Table 2 addresses another issue raised by Krane et al. (1), which is that LE tests should be carefully evaluated. The distribution of allele sharing is consistent with the assumption of LE (and more strongly, that of mutual independence of alleles within and across loci), even in the samples where some initial departures of HWE in locus-specific tests were observed in our original investigation (2).

Finally, we note that, irrespective of the results of tests of independence of alleles within and across loci in database analyses, the current forensic calculations of multilocus genotype profiles do not use the strict product rule and always incorporate conservative features (3). These include: (a) application of population substructure adjustment for homozygotes, with levels of \( \theta \) generally larger than the \( F_{ST} \) found in empirical data; (b) invoking a minimum threshold frequency for rare alleles; and (c) using the upper confidence limit on the point estimate of the multilocus profile frequency. Some laboratories even include computing conditional probabilities that
allow for substructure adjustments at the individual locus level for homozygotes as well as heterozygotes. With these protocols in place, corrective actions are always imposed in forensic computations, whether or not the population data show any sporadic departure from HWE and LE. In conclusion, the entire commentary by Krane et al. (1) is scientifically incorrect, both in the use of statistics and population genetics. There is no basis for their assertions that question the forensic practices for estimating DNA profile frequencies. Their commentary also is a useless exercise for homozygotes as well as heterozygotes. With these protocols in place, corrective actions are always imposed in forensic computations, whether or not the population data show any sporadic departure from HWE and LE. In conclusion, the entire commentary by Krane et al. (1) is scientifically incorrect, both in the use of statistics and population genetics. There is no basis for their assertions that question the forensic practices for estimating DNA profile frequencies. Their commentary also is a useless exercise for homogenizing the allele sharing statistics because \( n(n-1)/2 \), shown in the next column.

**References**


Ranajit Chakraborty, Ph.D.
Center for Genome Information
Department of Environmental Health
University of Cincinnati College of Medicine
Cincinnati, OH 45267-0056

Hee S. Lee, M.D.
Center for Genome Information
Department of Environmental Health
University of Cincinnati College of Medicine
Cincinnati, OH 45267-0056

Bruce Budowle, Ph.D.
Federal Bureau of Investigation Laboratory
Quantico, VA 22135