The 1996 NRC Report: Another Look

James F. Crow, Ph.D.
Genetics Department, University of Wisconsin, Madison, WI 53706, US

It as been a dozen years since DNA technology was first employed to solve a crime. This year the FBI celebrates the 10th anniversary of its activity in the forensic use of DNA. In response to numerous questions and doubts about the new technology, in 1989 the National Research Council formed a committee to study the issues. It produced a report in 1992 (1). The report immediately met with severe criticism, particularly because of the "interim ceiling principle". Although the aim was laudable, namely to remove the necessity for separate racial and ethnic databases, the procedure seemed arbitrary, didn't use all the relevant data, and was unnecessarily conservative. So a second committee was set up in the fall of 1994.

The second committee report (2) was issued shortly after the Seventh Annual Promega Symposium in the fall of 1996. By this time, a preliminary version of the report was available and the final report, which was issued a few months later, differed in only very minor ways. At this Promega Symposium I presented a summary of the report (3). Now, two years later, some things have changed and I should like to give an update with some comments. Although I served as chair of the Committee, these views are my own, although I believe that the Committee members would generally concur.

MAJOR REVISIONS

The major revision that I would make, were the report being written today, would be less emphasis on VNTRs and more on STRs. STRs, which permit PCR amplification, are a great improvement in most ways. The trend toward more use of STRs was already started at the time the report was published, especially since the report was largely finished more than a year before it appeared in print. We devoted a great deal of attention to problems of matching and binning of VNTRs. These conclusions are still correct, I think, but with each passing year they become less relevant as STRs become the rule. With STRs there is usually a unique genotype associated with a gel pattern. So there is therefore no need to group several alleles into one bin, with the statistical uncertainties that this leads to. There is also no need to introduce a correction for single bands that might really be from a heterozygote and the second band for some reason is not visible.

At the time of NRC2 there were abundant VNTR frequency data from a large number of populations. We recommended that the measure of population subdivision, θ, be set at 0.01 when used with our equations 4.4 and 4.10. (Let me note, for those who are reading my earlier paper (3), that every Greek θ appears as a q -- a computer's failure to translate from Roman to Greek letters. Fortunately this needn't cause any confusion.)

Because databases for STRs were less extensive and representative of fewer populations, we suggested using θ=0.03 as an interim measure. Now, with extensive data (R. Chakraborty and B. Weir, personal communications), it is clear that population subdivision is no greater when measured by STRs than when VNTRs are used . I would have expected this on theoretical grounds, but the Committee did not feel certain enough to act on this expectation. I would now recommend that θ be taken as 0.01.

Following the report, I would recommend using θ=0.01 in equation 4.4a and $\theta = 0$ in 4.4b. This is intended to adjust for different allele frequencies in subpopulations. This produces a conservative adjustment for homozygotes, but not for heterozygotes since the unadjusted H.W. frequencies are usually higher. Thus the recommendation is tilted slightly in favor of the defendant.

We suggested using the conditional match-probability equations 4.10 when the person leaving the evidence DNA and the suspect both came from the same subpopulation. Some reviewers have suggested that these formulae always be used as a hedge against uncertainty about population substructure. This is not unreasonable. With $\theta = 0.01$ the numbers are not greatly changed. For example, with 13 STR loci the average match probability in whites is about 2×10^{-13}. With $\theta = 0.01$ this value is approximately doubled.

I see no reason not to use convenience samples in databases. VNTR and STR markers are surely not correlated with anything that would affect criminal behavior, nor are there demonstrable differences among those data gathered from different social or occupational groups. A strict random sampling procedure is impractical, and probably would be no better.

We called for high standards of laboratory practice, including regular proficiency tests. We deliberately

avoided too much detail in recommendations for quality control and quality assurance. There were two reasons. One was that techniques and standards are continually improving. The other was that we knew of the existence of the DNA Advisory Board, DAB, a continuing body that can keep up with changes in the technology.

Our treatment of mixed samples was correct as far as it went. We could have gone further, with more complicated situations. Perhaps we should have. My excuse is that we were trying to keep the report simple, and a proper treatment of this subject would have had to be much more extensive.

POPULATION SUBDIVISION AND RELATIVES

The number of STR loci is almost unlimited. The FBI has now certified 13 loci and individual laboratories can use some or all, or add others. This means that the time will soon be here, if it is not already, when we no longer need pay attention to the troublesome problem of racial differences and population subdivision. Our classification of races was the common sense one. We chose as racial groups those that the average person can identify on sight; for Hispanics the identification is also partly linguistic. We were aware, of course, that these groupings are fuzzy at the borders; in fact some anthropologists go so far as to regard racial differences as meaningless. Yet there are clearly average differences, not only in external phenotype but in molecular markers. The databases for the major groups are well established. But there are problems when one is dealing with tribal populations like those of many Native Americans or isolated groups such as Pakistani in Britain. Furthermore, the presence of relatives (e.g. unknown half-sibs) complicates the analysis.

Ideally, the number of marker loci would be large enough that even close relatives or members of a small isolate could be distinguished. The most difficult relationship is that of full sibs, where the conditional probability of identity at a locus is always at least 1/4, no matter how rare the alleles (2, eq. 4.9). So, I ask, how close are we to the point where even sibs would almost always have different profiles.

Since the value 1/4 dominates the conditional probability, it is rather insensitive to allele frequency differences. With 15 typical STR loci, the match probability of sibs is about 1 in a million. For half sibs the probability is much less. We now find chance matches between close relatives with the same probability that only a short time ago applied to unrelated persons. So we are close to the time when, for most purposes, we can ignore population subdivision and close relatives for most simple cases. STRs represent a major advance over VNTRs for this situation, since the random match probability for brothers depends mostly on the number of loci and much less on the number of alleles per locus.

Interestingly, it was to avoid having to make racial and ethnic distinctions that the writers of NRC1 chose the ceiling principle (1). Soon we shall accomplish the same end by simply using more loci, and we shall have come full circle.

CRITICISMS

Our report has been reviewed a number of times. I have seen many of the reviews, but by no means all. For the most part, the critics have been favorable. But not all have been, and I would like to comment to what seem to me to be the most important criticisms. Several of these, some quite severe, appeared in a single journal issue (4 and following articles).

Some of the reviewers still advocated calculating error rates based on past proficiency tests and essentially adding these to match probabilities to give one composite probability. Our Committee rejected this suggestion for a number of reasons. For example it would require an inordinate number of tests to determine the relevant error rate for a laboratory with any acceptable precision, and pooling estimates for all labs would tell you little about the particular one that is involved. I see no reason to change our view. The argument is made that this combined probability is fairer to the defendant. I continue to believe that the best protection for a wrongly accused innocent person is an independent retest, as NRC2 recommended. Our arguments are given in Chapter III of the report. For additional arguments from a more legal perspective see (5). Of course, we favored high standards, including regular proficiency tests. Our recommendation 3.2 says: Laboratories should participate regularly in proficiency tests, and the results should be available for court proceedings. We also said that some of the tests should be blind.

One critic, Newton Morton (6), believes we too often erred on the side of the defendant. He argues that one should always make the best possible estimate and let the courts make whatever adjustments they see fit. There is much to say for this view. We, too, rejected extreme conservative calculations, such as the interim ceiling principle or a direct count from the database. Our recommendation for single bands in VNTRs, using $2p$ rather than p^2 as particularly criticized. We chose it because, while other less extreme corrections are usually conservative, this one always is. In any case, as STRs replace VNTRs, this argument becomes moot. In general, we tried to give unbiased estimates. When there was

uncertainty, as there always is about population structure, we opted for a conservative adjustment. Although I concede the force of Morton's criticism, I still believe our procedure was good. Err we must, and I prefer to err on the side of the defendant.

Another criticism for which I have great respect comes from several people, especially Bruce Weir (7). In several places in the report we did not use optimum statistical procedures. For example, there is an exact test for departures from Hardy-Weinberg ratios. We chose instead a good approximation, that tests especially the direction of the departure, since an increase of homozygotes is what might be expected. So we opted for a test answering approximately the question that seemed to be of most interest rather than an exact test that answered a slightly different question. In several places in the report (an example is Table 4.5) we chose for illustration a calculation of θ that we thought would be easy to follow, although it is not the procedure one would use for optimum statistical efficiency. The Committee was repeatedly urged to keep the report simple, and we tried to do so. Often simple procedures are not the best (although for the ones we used, the difference is very slight). We were writing for users of the data, not for those who generate the numbers. But I think that we could have been more explicit in emphasizing this in the report.

We made relatively little use of formal statistical procedures, even such techniques as bootstrapping. We relied instead on empirical comparison of different databases in the belief that this was a way to encompass all the errors in the system, or at least most of them. Hence our reliance on rather inelegant graphs, such as appear on page 150, rather than calculations. It is true, as Weir (7) has pointed out, that the 10 fold error may be exceeded for very small and very large probabilities. But for probabilities like 10^{-12} I am not concerned with an error larger than 10 fold, and probabilities less than 10^{-3} would not be used as evidence.

The strongest and most persistent criticism has come from Bayesians. We used a hypothesis-testing approach, which was certain to elicit Bayesian criticism. We tried to give procedures acceptable to the courts. Hence our emphasis was on match probabilities or simple likelihood ratios. We did suggest that a range of prior probabilities might be presented in order to generate a posterior probability, but I am not aware of this being accepted by the criminal courts. We deliberately chose not to use of Bayesian analyses, mainly on the ground that American courts have not seemed to be accepting of them.

DATABASE SEARCHES

Our most controversial recommendation concerned making calculations when the suspect was found through a database search. We argued that, on the hypothesis that the contributor of the evidence DNA was not in the database, a Bonferroni correction was in order. The probability of finding at least one match in a homogeneous database of size N is $1 - (1-p)^N$ where p is the frequency of the profile in the population. This is approximately Np if N is much less than the reciprocal of p. (Let me take this opportunity to correct an unfortunate typo in the Report (2). See P 40 and also P161, two lines below Recommendation 5.1. The statement should say ". . . the source of the evidence sample is <u>not</u> someone in the database."

We were thinking of relatively small databases, not those large enough to yield more than one match. Some critics say that we answered the wrong question. Some of them simply ignore the ascertainment problem and say that the fact that the suspect was found through a database search is irrelevant, or largely so (7, 8). Others say that the likelihood ratio has meaning only when associated with prior odds, which in a large database would be small. These are not numerically unimportant disagreements, for the differences among these various procedures can be several orders of magnitude.

The 1992 Committee (1) recommended that the loci used to discover the suspect not be used to calculate the probabilities to be used in court, but that independent markers should be employed. Our Committee approved of this, but with a small number of VNTR markers, this seemed impractical. With an increasing number of STR loci this is no longer so impractical. Morton (6) in particular advocates going back to the NRC1 recommendation. He thinks the ascertainment procedure should not be ignored. It is clear, I think, that the NRC1 recommendation is unbiased. The question is, will it be accepted by the forensic community and statisticians? And which database should be used, that of convicted felons or the general population (if they should turn out to differ)?

Morton (6) says, in recommending the NRC1 procedure, says: "To the delight of scientists and judges and the disappointment of mathematicians, this solution puts an end to the controversy about interpretation of suspect trawls in a very large database." I doubt that this will end the controversy. We shall have to wait and see. But as databases increase in size the problem will continue to arise.

CONCLUSION

It is clear that our report is not the last word on the subject. Yet, I believe that the report has a number of solid accomplishments and has brought considerable order and clarification. Let me end by quoting myself, from the Preface to the Report (2).

"I have no illusion that our report will eliminate the controversy; remaining uncertainties and the adversary system in the courts guarantee its continuance. But I hope that we have substantially narrowed the range of acceptable differences.

REFERENCES

1. NRC. (1992) DNA Technology in Forensic Science. Washington (DC): *National Academy Press*.
2. NRC. (1996) The Evaluation of Forensic DNA Evidence. Washington, (DC); *National Academy Press*.
3. Crow J.F. (1997) The 1996 NAS Report. Proceedings from the Seventh International Symposium on Human Identification,1996 Promega Corporation, pp. 1-11.
4. Kaye D.H. (1997) DNA. MAS. NRC, DAB, RFLP, PCR, and More: An Introduction to the Symposium on the 1996 NRC Report on Forensic DNA Evidence. *Jurimetrics J* **37**:395-404.
5. Berger M.A. (1997) Laboratory Error Seen Through the Lens of Science and Policy. *U. C. Davis Law Rev* **30**: 1081-1111.
6. Morton NE. (1997) The Forensic End Game. *Jurimetrics J*, **37**: 477-494.
7. Weir B.S. (1996) The Second National Research Council Report on Forensic DNA Evidence. *Amer J Hum Genet* **59**:497-500.
8. Evett I.W., Weir B.S. (1998) Interpreting DNA Evidence. Sunderland (MA): *Sinauer Associates*.