B24 Guessing the Race From an STR Profile

Charles H. Brenner, PhD, Consulting in Forensic Mathematics, 6568 Sobrante Road, Oakland, CA 94611*

After attending this presentation, attendees will understand the routine DNA profile – the CODIS STR loci for example – properly evaluated, usually provides strong information about the ancestry of the contributor.

Calculation of probable racial origin of a crime stain can occasionally be a helpful hint and is extremely easy and cheap to compute. Therefore, this presentation will impact the forensic community and/or humanity by providing a tool that the investigator may as well have.

The routine DNA profile – the CODIS STR loci for example – properly evaluated, usually provides strong information about the ancestry of the contributor.

It can be useful to know the race or population group of origin of a crime stain. For example, in April 2003, the Louisiana serial killer Derrick Todd Lee was arrested based on a tip from a citizen who had, according to the newspapers, long been inclined to be suspicious of him. However, the informant decided his hunch was worth acting on only when and because the authorities, in about March, issued a statement contradicting the earlier incorrect public assumption that the killer was Caucasian. Lee is an African-American.

The March announcement (in substance, it was that no race should be eliminated from consideration) came shortly after a racial analysis by DNAPrint based on their specially developed set of markers. However, such a high-tech approach was not essential in this case. In fact I had been consulted on the case in January and on the basis of the STR locus profile obtained by the Lafayette and Baton Rouge crime labs provided essentially the same information. A dragnet, operating on a racial assumption based perhaps on a profiler's analysis or perhaps on some vague witness account, had methodically extracted DNA samples from 800 Caucasians in an attempt to match the crime stain. After calculating likelihoods for several races – Caucasian, African-American, Japanese, Korean, Chinese, Hispanic, Vietnamese – I said, "Caucasian is the last race I'd look at." (This was a little glib – on the numbers it is really second last, before Vietnamese.)

The DNA criminalist commonly calculates several "frequencies" corresponding to the Caucasian, African-American, etc., population groups. "Frequency" is the wrong word though, since the number is invariably so small that if it were really interpreted as a frequency it would imply a fraction of a person. What the number actually represents is expected frequency, or probability: The probability that a random e.g., Caucasian person unrelated to the crime stain would have the crime stain type. That is, it would if instead of making the customary conservative computation, we try to be as accurate as possible. To that end, the right procedure is simply to include (temporarily) the crime stain in the population sample – note that this avoids ever having a probability of zero – and then use the consequent sample frequency for each allele as a good estimate of its random match probability. The product rule, perhaps elaborated by the NRC homozygote and heterozygote theta correction (but with a realistic, i.e., very small, value of theta), is then used to combine the individual allele probabilities to form a profile probability.

The profile probability assuming Caucasian origin and the profile probability assuming African-American origin are two probabilities of the same thing under different hypotheses. As such, their ratio is by definition the likelihood ratio supporting Caucasian over African-American origin for the source sample. In the Baton-Rouge case, the likelihood ratio favored African-American over Caucasian by about 6:1. Testing Caucasians is thus quite inefficient. Curiously, the likelihood for Japanese, Chinese, and especially Korean populations was even higher, the likelihood for Koreans exceeding that for Caucasians by a factor of 200. That does not, of course, mean that a Korean was probably the culprit, only that any single random Korean would be a more plausible suspect than any single random Caucasian or anything else. Hence lacking some reason to the contrary (such as knowledge that Koreans are improbable serial killers), the most efficient dragnet would begin with Koreans. As few Koreans as there are in Baton Rouge – about 0.1% of the population – the net chance that a Korean would be the killer was 6%. Disproportionately large, but still small. Caucasians and AfricanAmericans, each of which represent about half the people in Baton Rouge, figured as 11% and 67% respectively to include the killer.

The 6:1 likelihood ratio favoring African-American over Caucasian seems like a strong clue, but in fact it was a little unlucky it was so small. On average we can expect to do better. This table shows the typical likelihood ratio that can be expected when comparing various population groups:

Cau	His	AA	Jap	Vietnamese
Caucasian	5	40	30	300
Hispanic		30	10	200
African American			300	5000
Japanese				8

In the long run forensic STRs will certainly not be the most accurate tool for assessing racial origin. However, they do offer a number of advantages. The typing is routine; it doesn't cost extra. Extensive population data is available, for a large number of populations, whereas for a specialized test the data will need to be gathered at considerable expense just for that test. Once it is appreciated that we are dealing with

 ** Presenting Author*

probabilities and not frequencies, it is apparent that large population studies are not necessary. Small ones generally are somewhat less informative, but their smallness does not inhibit inferring a probability. Probability is, after all, a summary of whatever information we may posess. Finally, the method of analysis of STR data conveniently gives a likelihood ratio which means that though it may not always give a definite answer it will rarely be misleading; when it is not sure, it tells you it is not sure.

Racial Attribution, DNA Typing, Likelihood Ratio

 ** Presenting Author*