Frederick R. Bieber

Genetic diversity in the human population
Use of DNA for race/ethnicity prediction

Describe yourself in one word without using a name. So just do that. What comes to mind? And once you've done that I am going to come back to that after I am through with my talk. One word only. In English. Who are you or, if you would rather, describe yourself in one word without using a name. Just do it. And then I will tell you what happened when I asked my Harvard undergraduates to do that. Because I think it relates to some of the topics today. How we think about ourselves and describe ourselves.

Before we get into the meat of my talk I want to - reminding ourselves of this caveat. Someone asked me – "Aren't there", yesterday, "Fred, you didn't talk about the possible sources of error in DNA testing or in forensic testing in general, and I think that was a major oversight" Well it was an oversight but not intentional. I just didn't get to that slide which was two slides later so in the interest in of being complete, I have a non-definitive list of some of the potential problems and pitfalls that we all hopefully recognize that are potentially present in any investigation or case. And it would be sort of a laundry list that I would go over with Frank Gaziano before helping him prepare for a case involving DNA or involving someone from the defense community who asked me to review a case with them. These are some of the issues that I would want them to think about.

Certainly we considered a number of these in the so-called Death Penalty Report to the Governor that we presented a week ago about some of the errors of commission or omission that could potentially lead to false arrest or false convictions and, God forbid, false executions. I won't say any more to that other than except in the interest in being complete.

I had another question after David's talk about this concept that David alluded to of ethnicity tracing, or excuse me, of pedigree searching by looking for more extended family members. And these are data just looking at the real frequencies in the human population of the STR loci and looking at the theoretical overlap in the allele distribution in - say a victim and a sib or suspect and a sib. And a victim and a control, or non-related person. How many alleles would be shared across the thirteen CODIS loci? Thirteen loci - so if you were heterozygous in each you would have 26 alleles in your profile in the CODIS system. So we'd expect that siblings, a victim and a sib or a suspect and a sib, a dependant and a sib would share on the average about fifteen or sixteen alleles in common, not necessarily in locus identity, but alleles in common and that would be roughly twice that of unrelated individuals. The important point about this projected data set is that it shows you that one could make predictions even if you don't have a complete hit in the database - that it's not him but it's probably his brother. And this is not higher mathematics, as David and others have alluded to this as well. But just to show you a pedigree, at a single locus with parents 11/14 and these numbers refer to the size of the repeat unit, the number of STR repeats, 11 on one chromosome and 14 on the other. Simple probability calculation, what's the chance that any brother or any sibling of this person who is a 9/14 at this locus would also be a 9/14. And if you happen to know the genotype of the parents you could say it's one-half times one-half, or one-quarter. One quarter of the time at that locus the sibs would be expected to be identical.

If the father was 11/11 and mom was 11/11, the chance that all sibs would be the same was 100 percent. So it's not, you are not just multiplying a quarter to the n – n being the number of loci you are testing, it would depend in fact on the genotypes of the parents. This becomes very relevant in consanguineous matings, which are very common in some cultures and populations. Arranged marriages between cousins or uncle-niece are very common and may be even the rule rather than the exception. So the amount of locus identity and allele sharing among sibling populations, brothers in particular, can be greater than you'd expect.

We took a look at this – "we" meaning Mike Bourke and Karl Ladd from the Connecticut State Police crime lab and I, we gathered, with the subjects' approval, the cheek swabs from siblings and non-sibs and did the STR profiling with the 377 system at Connecticut. And we collected about 100 sib pairs and some unrelated pairs as controls, the sibship sizes ranged from 2 to 8 and they were from 37 sibships and some 28 kinships where there were some multiple generations of families. And these are the data in which again, paralleled the mathematical expectations and are quite, I think, interesting in two respects.

Looking at the full sib data on the left, we see that we found some sibs shared locus identity, meaning the same genotype at up to nine loci. These are regular sibs, nonconsangineuous mating offspring. And the range of allele sharing in these sibs was from eleven alleles in common amongst the whole profile up to 23 as compared to the right hand column to the unrelated non-siblings. The data of Mike Brundle's experiments are so close in predictions that you wonder if I fudged the data. But I didn't. I didn't and we didn't.

And we presented these data at an American Academy of Forensic Science meeting a few years ago. And nobody who thinks about math and statistics really would be surprised and we hoped they wouldn't be. One of the bottom line issues is that 15% of the sibs matched at between seven and ten loci – before we said nine, now we say ten. I put that in there to remind me that we checked and double checked any matches that were more than seven loci. And we found one error, there was a clerical error and changed it back to nine. So we did our own internal QA on the data and found a human error that occurred in tabulating.

Now after I presented that data, Chris Thomsey, who is a terrific lady in Pennsylvania State Police said, "Fred, we beat you. We got a ten locus match in sibs." And I said I don't believe it. That is stretching it. It wouldn't be zero but it would be rare. She said, "Well, we were doing a QA/QC check on our CODIS system." And she decided - if I understand that phone conversation correctly - she decided to look at any ten-locus or more matches in her database thinking they could be duplicate samples, identical twins, or typographical errors of entry. And she found one. And she went back with the rules available in her state and found that the individuals – she'd gotten permission from all concerned - they had the same surname. And they came into CODIS at a different time and it turns out they are brothers. Not surprised.

There are a lot of sib pairs in prison and a fair number of identical twin pairs in the criminal justice system. And I said, "Well they are not just regular brothers. I bet you a dollar and a cup of tea that they are the offspring of consangineuous mating." And this is what she was able to find out, again using all the right methods of finding this out. The mother had taught her son sex education the old fashion way right at home. Education begins at home. So she had her own biological son whom she then had another son with so these are - so-called, they are socially, sort of uncle-nephew - but they are in fact biological half-sib and more than half-sib, they are father-son here.

So, and we have also heard but I cannot confirm and I haven't seen the data so I can't tell you that I have actually seen the electrotheragram. So this is personal communication. We have heard another personal communication from someone south of Georgia and north of Key West that they have an eleven-locus match in sibs. We don't know any details. So the chance for high amount of locus identity – let alone allele sharing - is quite real. And so one of the things that I always would ask Frank or Kerry Miles -are there any twins involved? 160, there are one in 260 -individuals is a monozygotic twin. You better be sure you exclude those brothers or twins before trial especially if you haven't the greatest DNA sample and you are only presenting data on seven-locus comparisons with the samples.

I wanted to say one final word about some of the areas that we haven't a lot of chance to talk about that are really, I think are ramping up very quickly. This is going to be the subject of the remainder of my comments as well as Pilar's. But non-human DNA evidence is an exploding area not only animal DNA mainly - dog and cat - but also some other farm animals and plants to trace the source of marijuana plants, for example. The Connecticut State Police are very involved in that and there is a major emphasis, coordinated in some ways by (Lewis ?) of the FBI to put together a team of individuals to study recognized pathogens in being able to trace the genetic origin of particular strains of microorganisms. And here is our friend the Y chromosome again, this is going to relate to our discussion of the concept of race or ethnicity. And there's a whole battery of STR markers on the Y that travel together as a haplotype and so the particular allelic profile of a single allele agent at this loci on the Y chromosome the single combination of the alleles is what we call a haplotype or a haplogroup and this moves from father to son with a low mutation rate.

These SNPs as I mentioned are coming on the scene. There are SNPs not only on the autosomes of the Y chromosome and X for that matter but also in the mitochondria. So the allelic variation at the level of the single nucleucotide has been is well known on the mitochondrial chromosome and the SNP typing – Kim this is important for your upcoming hearing — the SNIP profiling of mitochondria is showing that people who belong to the same group in the traditional sequencing of the region can be subclassified once you give SNP typing of the remainder of the mitochondrial chromosome. So it's sort of like the ABO blood group. People with high blood (?) can be sub categorized once you do STR profiling, and the same is true for mito.

(Question from audience): Can you just say how common or rare these variations are, I mean how many alleles there are in a particular…

Response: The problem with SNPs - one problem is that they are very few alleles, usually just two in the single base pair and that makes the analysis of mixtures a bit of a challenge because when you have a mixture you are very likely to get all of the alleles of the population at that locus and therefore no one can be excluded. So the mathematics of testing for mixtures and SNPs is still being worked out. And that's a bit of a challenge.

Did you fill out the one-word description? Could you now turn it over and ask…maybe I can show… when I asked my students this question, I got five categories of answers. Gender - some individuals just said "Man." OK. I got Occupation, typically "student." That's their occupation since I was polling students. Religion. I got one Rom Zen Buddhist, where he told me not only the type of the major category but the subcategory. Hobby: motorcyclist, someone said cat-breeder, a hyphenated word. Species: I have human and primate. I had one student call himself a primate. And I am glad he is. But by far away the most common one word response to 'who am I' or 'what am I,' one word other than their name was their country of birth, their perceived race or ethnicity. German. Swiss. Argentino, we have a lot of international students here. But the single most common response was their country of birth, their perceived race or ethnicity. So whether or not we can ever agree on what this concept means and I hope Pilar will help us understand this, clearly… Did any of you choose any of these categories? Gender? Did anyone pick gender? Okay. Occupation. A couple, good. Religion? Hobby? One finds himself... Any non-humans? How about country, race or ethnicity? What else did we choose? Oh…you're praising yourself. I'm glad we can have healthy self-concepts here.

Comment from audience: You're assuming that the adjective is a positive thing.

Fred Bieber: We're using the (unclear: maybe "Chatham rules here"?) Can you tell us what is on your card?
Audience: Dedicated
Fred Bieber: Dedicated.
Audience: To what? Audience: Not committed. Dedicated.
Fred Bieber: He's dedicated to expanding CODIS.

So regardless of how we answered that question, in the real world of medical genetics, one – Pilar –can bring us back to the other side of this coin – we clearly perceive race and ethnicity as an important attribute factor to consider in delivering medical care to patients and in these primary ways, and we can't go into detail, but you know that every newborn in the United States by law, a statute has a heel stick taken for the testing for treatable genetic diseases. The prototype for that is PKU, Phenylketonuria, a recessive condition, which if left untreated leads to a devastating neurological problem. Tay-Sachs Disease might be another that couples might seek genetic screening for based on their region of origin in Europe, because the prevalence or the incidences of Tay - Sachs or the prevalence of healthy heterozygous carriers is high. Cystic fibrosis is the most common autosomal recessive condition in Caucasians, about 1 in 25 to 1 in 30 US whites is a carrier for one of the 1000 or more CF mutations. Sickle cell affects those whose origins are around the Mediterranean basin and the equatorial regions. And so there is an ethnic, if you will, distribution of disease that isn't lost on the medical community and so certain individuals if they identify themselves in this way and usually they are asked by good nurses or docs or obstetricians, certainly, as potential risk factors because of the duty to inform, the duty to inform the patient of the risk or potential risk from which they may take some action mainly through genetic screening testing which would potentially prevent or alter the way they receive medical care. Allele frequency differences are well known throughout the world and sometimes they are dramatic and sometimes they are subtle. In the forensic community, these frequency differences are described in the end of every journal - in every issue of the *Journal of Forensic Sciences* and those differences of allele frequencies are used to cue(??) the point estimate of the match probabilities in different relevant populations for presentation in court and indeed many courts have ruled, most courts have ruled and have been asked to, that the presentation of the fact of a DNA match whether it's a one locus or thirteen cannot be admissible – it's prejudicial without a description of some measure of statistical interpretation: how common or how rare would that loci be in a randomly selected unrelated African American, Caucasian, and so on we describe whatever ethnic group.

Questioner: How important do you think that is?

Fred Bieber: How important do I think it is? I think it can be important in some inbred populations of genetic isolates.

Questioner: Let me give an example. I sit on the editorial board for the *Journal of Forensic Sciences*. And during our meeting this year to discuss what we will do and what we won't do, there is a distinct possibility that we are going to drop the report of the frequencies because it has come to the assumption that, regardless of what the race is that you are dealing with from a forensic perspective they are basically rare. It's a unique situation and there is not enough interest in supporting the use of that report anymore (too faint) I would argue that it's a fact of life that they are facing unique situations. So I'm not sure allele frequencies are that important.

Fred Bieber: Well, the problem that it creates when you separate individuals based on these somewhat arbitrary divisions, is when you have DNA mixtures. And I'm never really certain what the…how you should be doing this. Should you be looking at the subset of individuals in greater Detroit or greater Boston or looking for the (unclear) distribution there

Questioner: That’s a good point, because based on your previous comment on SNPs although, you know, I’m a technocrat apparently (too faint) robotics and everything else and I’ll applaud (too faint) when they come out, but the actual practicality from a forensic perspective is I think you have to make a convenience versus the actual research. So if we consider 70% of the crime scene are composed of a mixture, SNPs they are going to be our variation our version of the DQ Alpha of Ciro(??).

Fred Bieber: This is ah…I think we need to move to the break out sessions. You are addressing issues that I hope will come up there. But the way that SNPs will be relevant is in the question of ethnic and race prediction ….

Questioner: Not in a mixture

Fred Bieber: Not in a mixture, but there are lots of single source samples left at scenes.

Let me show you an example of the current use of SNPs and there are literally now, if you look through the literature, hundreds and hundreds of abstracts. In fact this is from the Human Genome Meeting in 2002 and Klaus Lindpainter, one of the former fellows at Brigham and Women’s Hospital is screening for cardiovascular disease associated genes and the heritable variation in the SNP profiles is being used as the length of(??) markers to look for disease susceptibility loci, if there are any. There are, I don’t know. I just put in bold some of the findings in this abstract a year and half ago whether we think SNPs will be used in forensics tomorrow or a year from tomorrow I can tell you that they’re being used actively by the medical scientists in every country that I know of that has the technology because there is an interest not only in looking at disease susceptibility but specifically drug sensitivity and many of the iatrogenic illnesses that patients suffer from are those of medical mishaps that relate to prescription pharmaceuticals where adverse drug reactions occurred and they could be avoided if you knew what drug one individual from a certain ethnic or racial category, however we define it might be in. So there is an active growth in number of chips being developed, these are oncology chips that have an array of several hundred SNP variants and other mutational events that are known to affect the retino blastoma gene, the colon cancer gene, and so on, that are involved in the pathway to colon cancer or other heritable diseases.

What is race? Ethnicity? I am not sure that I am the person to ask or answer that question, but in thinking about it as I have for awhile because this issue of race and ethnicity population databases comes up all the time in the Frye/Daubert hearings in forensics. I had a little chance to think about it especially having grown up in Canada and went to a high school in Buenos Ares, Argentina, to a boarding school where a lot of my pals in the school were from all over the Latin continent. And I came to realize very quickly that the only thing in common that any of us had really was linguistic commonality. We all spoke Castellana (?). We all spoke proper Spanish in Spanish and the proper dialects. And we were - along with Castellana - in Argentina. But I realized that my Argentine brother, Eduardo, was a direct descendant of an Italian family and had nothing in common culturally in any way with some of our Bolivian classmates who were indigenous people who pre-dated Spanish conquistadors by thousands and thousands of years. And the Spanish speaking population in Miami again is a polyglot, if you will, of different cultures. So the term 'Hispanic' has always bothered me because Argentines and Chileans as you know, may have European roots. And it makes me ask the rhetorical question 'are we not all Americans from Africa?' because the data from the mitochondria within the Y chromosome, Y chromosome (too faint).

It is believed by the archeologists and paleontologists, the people like Luka Cavalli, from Stanford and colleagues, who over the years have studied the ancestral migrations of mankind and humankind. It is believed by many that this breakup out of Africa occurred a hundred to a hundred and fifty thousand years ago. We may have reached Europe as a homosapien forty or fifty thousand years ago. A fair number of groups at Stanford and other universities around the world have mapped Y chromosome haplotypes and mitochondrial sequence data and have followed human migrations by tracking the mutation rates which differ within these populations. And this is a map that may or may not be a consensus map of human migrations but more or less it conforms to I think a reasonable theory of the migrations of the human species out of Africa as I just mentioned. And this is a long route to travel from Cairo to Lima, walking, and by considering the distance and mutation rates it appears that the peoples all have common origins and these are Y STR data that look at the distribution of a major Y chromosome group and the different clades(?) around the world. These data are more or less consistent with one another, between the Y data and the mito data.

David Lazer:… that - there…Is it based on like - for example – North American and South American? Is that based on indigenous population?

Fred Bieber: Usually not. They are having a hard time collecting from some of the indigenous groups now. There aren't really too many any more now who are truly indigenous. Many have a mixture.

Let's talk about the point of our afternoon and the (faint) break out session. There might be lots of reasons from a sweat band profile that comes from a bank robbery or a missing person who washed up on shore or a solider blown apart in Kabul to - identify him or her and bring him home. So the triaging of mass remains from disasters for family reunifications or the narrow focus on certain suspects.

Did he have red hair? Did he have blue eyes? How tall is he? - all the kinds of things that we would ask witnesses to a bank robbery now might be enhanced if you had some way of predicting phenotypes; physical appearance in some way based on the DNA helix. I am not proposing that this be done but I can tell you it would be a potential use. These are some of the things that have come up in discussion already.

There are two melanocortin receptors, or at least two melanocortin receptors, alpha and beta, that have been associated in a large percentage of cases with either premorbid obesity or some mutations in that locus, or alterations if you'd rather use that word, other allelic forms associated with red hair color. It is well known that the telomeres, the ends of our chromosomes, shorten with age either in tissue culture or in vivo. And while there is no perfect positive predictive value at this time with the length of those TAA3G sequences of the telomeres with age, it is at least plausible to consider knowing the age of an individual from their DNA sample. It's not here yet. Maybe it will never be here but at least it has been discussed.

Amelogenin is always used, as you know, to look for the two allelic forms of XY males or the single allelic form of XX females, so gender is already something of a phenotype that we predict from the DNA sample left at the bank robbery. I mentioned in passing yesterday the correlation between that rare last name Lazer and that particular Y STR haplotype. So one could potentially make correlations between clades or even surnames with a large array of Y SNPs.

And I'm not sure this is the real word but geoethnocultural SNP profiling. This is the real hot-button issue that we hope to discuss at some length in the break out session today. What predictions could we make from 10,000 or 50,000 SNPs on a chip? All sorts of predictions. Does the person have Type I Diabetes? Does the person carry the gene for sickle cell anemia? Is she or he a Tay-Sachs carrier? Does he have red hair? Or any other phenotype that we might find useful in a criminal investigation. And I am almost done.

But I just think that it is important that you know that the cases that David (dead spot) yesterday (dead spot) from the Louisiana serial murder case brought this company called DNAPrint Genomics in Sarasota, Florida to attention. They had previously offered a 71 SNP profile analysis for $158 US and they were primarily marketing this Ancestry by DNA. The website by the way is AncestryByDNA.com. They offered it to people searching their genealogy. You know, was my mother, a Macintosh, the same Macintosh from the Isle of Skye all the same root, the Cohen family family of the Askenazi who had special religious privileges. Those folks are tracking the YSNP and the YSTR haplotypes to determine whether they had a hereditary rights to perform certain religious duties and so on.

So this company has now got on the bandwagon a little bit and have offered their services to investigators in the Louisiana case and if you hear them talking about the story over tea at PROMEGA you get the impression that they really solved the case. I don't have first-hand knowledge of that case investigation so I can't comment but they are passing out their CDs with the box lunches at the PROMEGA Meeting. And there has been a fair amount of interest. I wouldn't say enthusiasm at this point because it's hard to know what they are actually doing and we haven't seen their data but this is Derrick Todd Lee who has been arrested and charged with several of these crimes, the serial killings in Louisiana. The traditional profiling as mentioned indicating a white male, that's largely because it was believed that most serial murderers were white loners and they lived with their mothers. It's true.

And this SNP profiling indicated otherwise – they predicted an amount of sub-Saharan African ancestry and indigenous ancestry. And these are the kinds of triangular plots. This is my last slide. These are the triangle plots that you will see if you order such testing on yourself. And Jack Valentine, our friend from Florida at the University of Central Florida actually ordered this test on himself and some of his students with their knowledge and they compared these triangle plot projections with what they thought they were. And basically to interpret this plot of the triangle you can put any group that you would define, however you would define African or white or Native American, you be the judge.

They used as their background known samples taken from several hundreds of sub-Saharan Africans and indigenous Native Americans, these are their terms not mine, European whites. I think they have a fourth group that I'm blanking on… Asian. Jeff might recall what part of Asia they are from. They claim to use a maximum likelihood analysis. I did my PhD thesis on maximum likelihood analytical techniques and I have asked them to see their data and I haven't seen it.

So I would like to see a peer review of their publication on the method that they used because it would possibly have value at least in this portion of the academic world. But I think before using it routinely the forensic community would certainly want to explore what data basis they are using for their comparisons. So they compare your SNP profile against the known frequency of the different SNP polymorphisms in these various groups. So they drop a line to the opposite point here on the triangle and where the dot is helps you - if you were 0% Native American your dot would be right down along this line somewhere. But since it is up it's around 15% or so, and depending on where it is here we draw this line here and it would tell you what proportion of your genetic makeup. Vis a vis the SNP profile - that is more common than the typical profile of Europeans and for the Africans. We draw the line here and we see where along this line your dot or my dot ….

Questioner: Is that Jack?

Fred Bieber: I don't know if it's Jack. But he said that they got it right. He's a Scotsman through and through and you can tell that from his accent …

Well you know my female students who know I am in forensics are always coming in, "Fred, should I get mace? What should I do to protect myself in this city?" Just don't park next to a white van. So this is the end of my formal comment.