DNA "FINGERPRINTS" AND THEIR STATISTICAL ANALYSIS IN HUMAN POPULATIONS

A.Marie Phillips
The University of Melbourne
Australia

In 1985 the concept of a "DNA fingerprint" was introduced as a means of evaluating human identity and relatednes. (Jeffreys, Wilson, & Thein, 1985). The possible forensic and legal applications of DNA evidence were quickly appreciated and such data are now frequently presented in court cases involving serious crimes such as murder and rape. DNA evidence is also used in establishing paternity, in determining relatedness in immigration and inheritance disputes, and in identifying disaster victims. Such cases, especially those involving famous people, are widely reported in the media and are of interest to the general population. Also, many people will be called to serve on juries in cases where DNA evidence is presented. As statistical concepts are involved in evaluating such evidence, "DNA fingerprinting" as a topic can be used to introduce statistical analysis to undergraduates. If a non-mathematical approach is taken many concepts can be taught to secondary school children, extending their understanding of statistics while holding their interest with practical " real-life" examples.

INTRODUCTION

Individuals of all ages are fascinated by stories involving people crime, and identity. This fascination with ourselves, and with the dark and tragic side of ourselves, can be used to teach principles of statistics to children of all ages. The use of DNA fingerprinting as a means of identification is well established throughout the world and a significant proportion of people will be involved in its use, many as members of juries. Others will be exposed to its concepts and complexities on an almost daily basis via the media. The question underlying the use of forensic data is *what is the probability that two or more people have the same genetic fingerprint/ profile*? A very useful reference book on this subject is *Interpreting DNA evidence* by I.W.Evett and B.S.Weir (2001).

If we could sequence and compare the DNA in each person we would find that we are all different. Even natural clones, such as identical twins, will have acquired a few differences, mutations in their DNA, during their growth from a single cell to an adult. Those of us who are not clones have many more differences. However we cannot sequence an entire individual's DNA and instead we rely on differences in length of short stretches of repeated DNA at the end of chromosomes. These pieces of DNA, called variable numbers of tandem repeats, VNTRs or short tandem repeats STRs**,** can be visualised as shown in Figure1.

SOME QUESTIONS ABOUT FIGURE 1.

If the picture shows DNA found at the scene of the crime, what are the implications for the people whose DNA profile is shown (lanes 1 to 7)? Can we say that any of these people did not commit this crime? That one of these people is probably guilty? Can we assign a numerical value to our answer?

Our answers to the last two questions depend on the probability of two or more people having the same pattern of *these* DNA fragments, the same sized STRs.

STRs do not code for any proteins or any known important cell function. They are part of our "junk" DNA and changes in length do not affect our health or likelihood of marrying and reproducing. The STRs chosen for forensic purposes are inherited independently of each other. They are on different chromosomes. Therefore, in theory, if STR1 of 40 bases is present in the population at a frequency of 1 in 50, and STR2 of 60 bases at a frequency of 1in1000, then only 1 in 50,000 people would be expected to have both STRs with this number of bases. *(Would the student be surprised if a survey showed 1 in 45,000 to have this combination? 1 in 1000?)* 10 different STRs are usually monitored in each "DNA fingerprint" and are chosen so that the probability of a false positive match is about 1 in a billion.

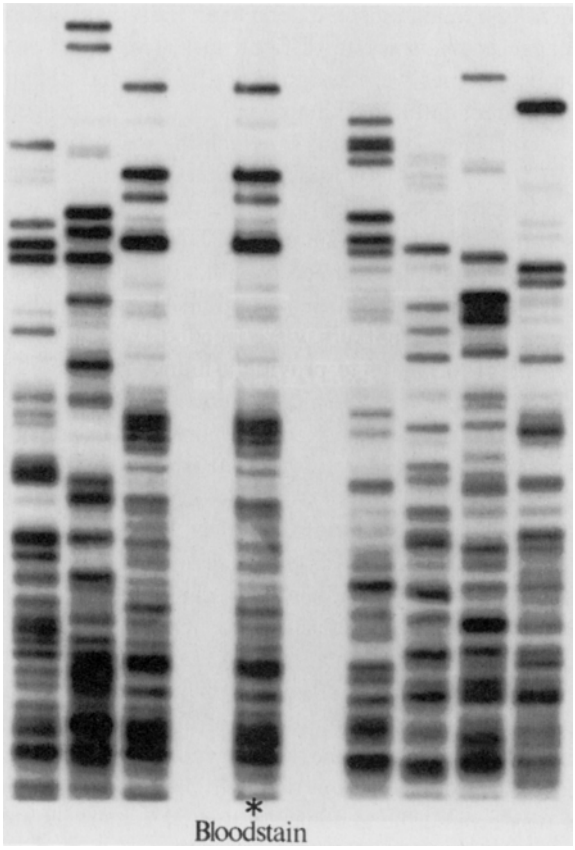

Figure 1. The STR profiles of seven suspects compared with
the STR profile of a bloodstain found at the crime scene
(Figure reproduced with permission of Orchid Cellmark Diagnostics.
Georgetown, MD. USA Feb. 2002)

That we will probably be different even from our brothers and sisters (but not our identical twin) can be demonstrated as follows. Each of us arises from an egg and a sperm and after fertilization each of our cells have two copies (one from the egg and one from the sperm) of each DNA sequence, including the STRs in our cells. For a female, STR1 from mother, STR1(m) will be passed on in one of two sizes, the one obtained from her father, STR1(mp), or the one obtained from her mother,STR1(mm). Similarly from our fathers we will get either STR1(pp) or STR1(pm) but not both. If the STRs from the parents are all of a different length the chance of two siblings having the same "DNA fingerprint" pattern is 2^{10} x 2^{10} or 1 in 1,048,576. However, as we shall see, this probability is likely to be less in siblings due to "identity by descent". *The Dolan DNA Learning Center, Cold Spring Harbor Laboratories, http://vector.cshl.org provides excellent coverage of Genetics for non-specialists, including programs on DNA fingerprinting. Another source of information is http://www.interactive-genetics.ucla.edu.*

In the early days of DNA fingerprinting the independent nature of STRs led to very large probabilities being quoted in trials (in one case it was asserted that there was 1 chance in 79 quadrillion of a false match) and consequently there were heated debates between population geneticists as to the real versus the expected probability, (see Roberts, 1991; Chakraborty & Kidd, 1991; Lewontin & Hartl, 1991). *Acting out or discussing these debates can be a useful learning exercise for senior secondary students and undergraduates.*

The quoted probability of 79 quadrillion was conditional on choosing the suspect from an infinite, random mating population and on independence of the chosen STR's. However, human populations are neither infinite nor random mating. It will have become apparent that our DNA fingerprint depends on the size of STRs passed to us through our parents, grandparents and

remote ancestors. This transmission of fragments unchanged through generations (fragments that are *identical by descent)* may affect the probability of a false positive match. There are a number of population concepts to consider that may affect the probability analysis of data.

NON RANDOM MATING

We indicated above the very low probability of two sibs having the same DNA fingerprint given that all the copies of each STR from their parents were different. However this is unlikely. People do not randomly mate with respect to their DNA e.g. tall people tend to marry tall people. As STRs do not encode any characteristic one might expect that people *would* mate randomly with respect to STRs *(Do you know the size of your partner's STRs? Did you choose to marry a person with very short STRs?)*. However, people frequently mate with someone from a similar culture. The city of Melbourne, Australia, is home to approximately three million people. These people are the population of Melbourne but this population consists of several sub-populations. We have, for example, large Italian, Greek and Vietnamese communities. When these people first came to Australia they chose to live close to people with a similar religion, with similar tastes in food, etc. People that came from the same part of the world and hence had the same ancestral gene pool. As they spend leisure time together they meet their mates and marry within the group. The population of Melbourne is therefore not a single population but a group of populations. In each of these populations the frequency of a particular STR may be very different. Using a data base that represents all Melbourne we may find that STR1 of 40 bases has a frequency of 1 in 6000, but in the Vietnamese community it may be 1 in 100,000 and in those of United Kingdom ancestry 1 in 50. The question can be asked *if an STR1 of 40 bases was found in DNA at the crime scene would it be more likely that the criminal would have a British accent or an Vietnamese accent*? The effect of population substructure affecting probability can be readily demonstrated to children using coloured buttons to represent the different sized STRs. Older children can manipulate the STR/button pool and observe changes in STR frequency**.** A computer simulation of a number of these concepts is available on www.handsongenetics.com.

To ensure that the likelihood of a false positive match is minimised the reference data base used must reflect the STR composition of the population to which the suspect belongs and a correction is introduced to account for the likely relatedness of individuals in the population. *How can you determine statistically that your reference data pool is adequate, remembering that the freedom of an individual will depend on your analysis? If you are the prosecuting group in the class how do convince the jury of classmates that your reference sample of 150 people is adequate when the population of the city is 800,000?*

FOUNDER POPULATIONS AND BOTTLENECKS

The frequency of a particular STR in a population may be very high if the population has passed through a recent bottleneck. An illustration of a rare DNA sequence becoming very common in a population following such a bottleneck is given in Oliver Sachs 's book *"The Island of the Colour Blind*"(1996). Monochromatic colour blindness is a very rare disorder but now approximates 10% on this island. The population was greatly reduced by a typhoon and the current population is the result of breeding from a very limited gene pool after the disaster. Particular STRs would also be expected to have increased in frequency by chance, depending on their presence or absence in the survivors of the typhoon.

Founder populations can be quite small, for example the island population established from the mutineers on the Bounty. Some excellent examples of Founder Effect are given in an article in Nature (Diamond & Rotter, 1987) including a discussion of some examples from the Afrikaner population of South Africa. More than 1 million living Afrikaners have the names, and the genes (and the STRs) of 20 original settlers. Although the genetic markers will have been shuffled and some altered by mutation there will be less variability between these people than in a population arising from a larger group of migrants as in the United States, or the criminals transported to Australia. Indigenous populations of very long standing are also likely to show greater variability. *How would the number of founders affect the number of STRs used for forensic analysis?*

IS THIS PERSON WHO THEY SAY THEY ARE? ARE THESE PEOPLE RELATED?

In criminal cases STRs are chosen to maximise the differences between people, however there are times when we will be most interested in the similarities between related individuals. Family specific STRs are very rare, but some combinations of STRs are more common in families due to "identity by descent". These can be used to identify people and body parts following disasters such as plane crashes. In the main this requires matching profiles from relatives and the deceased. Statistical analysis of such *DNA profiles* was used to determine that "Anastasia" was not the Romanov princess, and that Thomas Jefferson most probably fathered at least one child by a slave. *How would the need to use DNA from relatives rather than a reference database affect the analysis?*

DNA profiling can be used to resolve migration disputes involving questions about the relatedness of individuals to landed immigrants. A variation of this is the "missing heir", when someone unrecognized by the family appears to lay claim to an inheritance. If the "heir" is male, Y chromosome STRs can be very useful in establishing relatedness, as a man can only inherit the Y chromosome from his father. All direct male descendents will have the same Y chromosome i.e. uncles, nephews, paternal grandfathers and great-grandfathers. An interesting article relating to the topic of Y chromosome transmission is "*In the name of the father: surnames and genetics* (Jobling, 2001 and references therein).

DNA profiling and human population complexity offer a wealth of information and examples which relate statistical concepts and everyday life. More advanced students could look at the problems associated with probability calculations when mixtures of biological fluids from more than one individual are present. These students may find the software packages available at http://statgen.ncsu.edu/, for example the Genetic Data Analysis, GDA, package interesting and useful.

ACKNOWLEDGMENT

I would like to thank Henry Roberts, Acting manager, Biological Examination Branch, Victoria Forensic Science Centre, Forensic Drive, Macleod, Australia, for his fascinating and helpful discussion of the topic.

REFERENCES

Chakraborty, R., & Kidd, K.K. (1991). The utility of DNA typing in forensic work. *Science, 254*, 1735-1739.

Diamond, J.M., & Rotter, J.I. (1987). Observing the founder effect in human evolution. *Nature, 329*, 105-106.

Evett, I.W., & Weir, B.S. (2001). *Interpreting DNA evidence*. Sinauer Associates, Inc., MA, USA.

Jeffreys, A.J., Wilson, V., & Thein. S.L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, *314*, 65-73.

Jobling, M.A. (2001). In the name of the father: surnames and genetics. *Trends in Genetics, 16*, 353-356.

Lewontin, R.C., & Hartl, D.L. (1991). Population genetics in forensic DNA typing. *Science, 254*, 1745-1750.

Roberts, L. (1991). Fight erupts over DNA fingerprinting *Science 254,* 1721- 1723.

Sachs, O. (1996). *The island of the colour blind.* Pan Macmillan Australia Pty Ltd.