DNA Profiling in Forensic Science

Peter Gill, *Forensic Science Service, Birmingham, UK*
Rebecca Sparkes, *Forensic Science Service, Birmingham, UK*
Gillian Tully, *Forensic Science Service, Birmingham, UK*

Since 1985, DNA profiling in forensic science has rapidly evolved into an important technique that solves both major and minor crimes. A valuable innovation was the inception of the UK National DNA database that now comprises over 700 000 samples.

Introductory article

Article Contents

Introduction

The discovery of hypervariable minisatellite deoxyribonucleic acid (DNA) by Jeffreys and co-workers in 1985 provided forensic scientists with a means of discriminating between individuals that was many orders of magnitude greater than the traditional methods used before that date. Probably the most important forensic application of DNA profiling is in the identification of rapists. The original protein-based methods were difficult to interpret because both vaginal fluid and semen have enzyme and blood group activity, hence mixtures often occurred that were difficult to interpret. However, early experimentation with DNA demonstrated that it was relatively easy to separate sperm from other cellular material, effectively purifying the male from the female component and simplifying the interpretation.

Since the original introduction of DNA profiling in the late 1980s, the technique has undoubtedly revolutionized the way in which forensic science laboratories work. The evolution of DNA profiling has also been dramatic over the past decade. In particular, the discovery of the polymerase chain reaction (PCR) opened up new possibilities including analysis of minute crime-stains that were not amenable to the original methods. In addition, the latest techniques are significantly faster and cheaper, and this led to the inception of the UK National DNA database in 1995. It is envisaged that this will contain millions of DNA profiles within a few years. The speed of implementation of new technologies continues. Undoubtedly, within a few years, new techniques will enable even faster methods of analysis, with corresponding benefits of reduced costs. Inevitably the use of DNA profiling in forensic science will become even more widespread than it is today.

Historical, nonPolymerase Chain Reaction Methods

The original 'DNA fingerprinting probes' described by Jeffreys and co-workers were based on probes consisting of tandem repeats of the myoglobin locus which detected multiple hypervariable loci when hybridizations were carried out under low stringency. A number of different probes were originally isolated. Of these, two different probes were used, designated 33.15 and 33.6. Together, up to 36 independently inherited bands are detected. DNA profiles are visualized rather like a bar-code pattern on autoradiographic gels.

Minisatellite DNA consists of repeating sequences of bases. For example, Jeffreys and co-workers observed the core sequence GGAGGTGGGCAGGA. These sequences are repeated hundreds of times, such that the DNA fragment itself may be up to 20 kilobases (kb) in size.

Although the two multilocus probes 33.15 and 33.6 were used for a short period in the UK and some European countries, by 1990 their use had been superseded by new probes that could detect a single locus (single-locus probes; SLPs): instead of multiple bands being present, just two bands were visualized per probe. This was accomplished by the introduction of new probes that hybridized to Southern blots at high stringency. These new probes were developed in the laboratories of Alec Jeffreys and other scientists. Individually, the discriminating power was lower than for the multilocus probe approach, but this was compensated for by utilizing a battery of four different probes in sequential hybridizations of Southern blots.

The switch to SLPs had several advantages over MLP technology: they had increased sensitivity, were easier to interpret as the patterns are much simpler and, moreover, it was possible to convert the patterns into a computerized

digital code, based on the distance migrated during electrophoresis.

SLPs were used for several years, but in turn have now been completely superseded by the PCR.

The PCR method is an *in vitro* method for amplifying extracted DNA through successive cycles of duplication with specific DNA polymerase; it results in 50–100-fold increased sensitivity. Furthermore, PCR works on highly degraded samples – an important advantage for typical casework. The longer the DNA fragment, the less efficient the PCR. There were problems with the amplification of high molecular weight DNA, hence minisatellites became inappropriate. Short tandem repeat (STR) DNA is now the method of choice. These are typically highly variable sections of DNA that are short in length (250–400 base pairs) and therefore amenable to PCR. They consist of short repeating blocks of DNA, typically between two and five bases long.

Short Tandem Repeats: Structure and Nomenclature

STR repeats have been subdivided into various categories.

Simple repeats

Example: HUMFES/FPS $(ATTT)_{8–14}$

The nomenclature of simple repeats is straightforward. The notation is based upon the number of tandem repeats in the STR (**Figure 1**).

Simple with nonconsensus repeats

Example: HUMTH01 $(TCAT)_{5–11}$

Allele designated 9.3: $(TCAT)_4CAT(TCAT)_5$

If there is a variant such as the 173-bp variant of HUMTH01, then its notation is based on the number of complete repeats followed by a decimal point and the number of bases comprising the partial repeat (9.3).

Compound repeat sequences with nonconsensus repeats

Example: HUMVWFA31

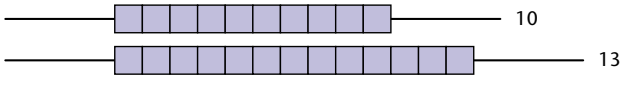

Figure 1 Schematic diagram of a short tandem repeat genotype. This individual has 10 and 13 repeats respectively on a homologous pair of chromosomes. The genotype is therefore 10,13 (the lower figure comes before the higher). The repeating blocks consist of ATTT in the HUMFES/FPS example.

Alleles designated 13, 15–22 $(TCTA)(TCTG)_{3–4}(TCTA)_{8,10–17}$

Allele designated 14 TCTA TCTG TCTA $(TCTG)_4(TCTA)_3TCCA(TCTA)_3$

The notation also works well for compound repeat sequences such as HUMVWFA31. This consists of two types of repeat (TCTA) and (TCTG). A variant consisting of 15 complete repeats of four bases and one partial repeat of two bases has also been discovered. The notation (15.2) follows naturally from before.

Complex repeat sequences

Example: D21S11

Class 1: $(TCTA)_{4–6}(TCTG)_{5–6}(TCTA)_3TA(TCTA)_3TCA(TCTA)_2TCCATA(TCTA)_{8–16}$——————————$TC$

Class 2: $(TCTA)_{4–6}(TCTG)_{5–6}(TCTA)_3TA(TCTA)_3TCA(TCTA)_2TCCATA(TCTA)_{8–16}TA.TCTA.TC$

In the complex repeat D21S11 there are several different types of repeat sequence: tetrameric (TCTA) and (TCTG), an invariant trimeric, an invariant dimeric (TA) and an invariant hexamer (TCCATA). In addition, the presence or absence of a hexamer (TATCTA) at the end of the repeat further complicates the nomenclature.

Complex hypervariable repeats

Example: Human β-actin related pseudogene (ACTBP2)

Complex hypervariable $(AAAG)_n$ repeats are much more difficult to accommodate using nomenclature based upon the number of tetrameric repeat sequences. This is because variant monomers, dimers, trimers and tetramers are scattered throughout the locus.

Nevertheless, in common with simpler STRs, regularly spaced allelic ladders can be constructed, and alleles designated by direct comparison. Since a notation based upon the number of repeats is not possible with this system, the alternative is to designate alleles based upon their size. The size of the STR will in turn be dependent upon the primers used.

It is possible for alleles of the same size to be different sequences (particularly with AAAG repeats). This does not affect the interpretation as population frequency estimates will be based on all of the alleles that comprise a given size class.

Allelic Ladder Markers

Allelic ladder markers are widely available in the forensic community, acting as controls to enable the designation of alleles in samples. In addition, it was recommended by the DNA Commission of the International Society of Forensic Haemogenetics that, ideally, alleles in the control ladder

should be sequenced. The HUMTH01 allelic ladder (**Figure 2**) consists of six common alleles and one rare allele (type 11).

If laboratories are to compare data, then a common standard is needed. It follows that for each locus a standard allelic ladder is needed; the nomenclature follows from a straightforward comparison against the allelic ladder. It is important to recognize that, in many loci investigated, rare size variants exist that will have mobilities intermediate to alleles represented in the allelic ladder; hence the rungs in the allelic ladder should be related to the repeat size of the locus. For example, for simple tetrameric repeats 4 bp is appropriate. However, it is possible to compare loci utilizing different allelic ladders (and nomenclatures) provided that proper comparisons are made so that one designation can be converted into another.

Multiplexing

Fluorescence-based automated sequencers have high throughput and allow the rapid analysis of sequence information. The availability of four distinguishable fluorescent dyes facilitates the development of STR multiplex PCR systems by allowing the labelling of loci that have overlapping allele size ranges with different fluorescent dyes. This means that several different PCR reactions can be carried out in the same tube. Fluorescent labels are attached to the 5′ end of one of the primers for each locus. Four differently coloured labels can be used, although one is normally reserved as an internal size standard. Amplified DNA is then loaded on to an electrophoretic gel. The labelled fragments move past a laser, which induces fluorescence. The resulting light is collected by a photomultiplier tube and converted into a signal. Software interprets the signal as peaks. Sizing relative to standards is automatic, hence multiplex systems can be employed to generate population databases rapidly. Because some loci overlap in size, it is necessary to tag primers with different dye labels; consequently the systems developed in the authors' laboratory can be utilized only with Applied Biosystem automated sequencers.

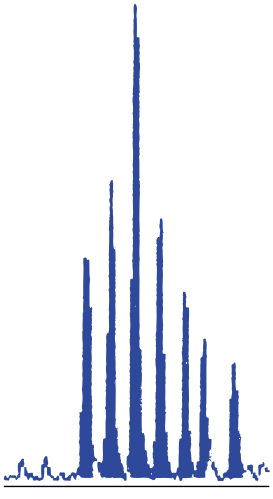

Figure 2 HUMTH01 allelic ladder showing seven alleles (from left to right – 5, 6, 7, 8, 9, 9.3, 11) resulting from an Applied Biosystems 377 automated sequencer.

Several factors are considered when choosing new loci for potential candidates in multiplexes:

1. High discriminating power (observed heterozygosity greater than 70%).
2. The predicted length of alleles must be approximately 90–500 bp (the higher the molecular weight, the lower the precision of measurement). Also, the lower the size of the STR locus, the less chance of locus or allelic drop-out because of degradation of the sample.
3. Chromosomal location (to ensure that closely linked loci are not chosen).
4. Robustness and reproducibility of results, showing minimal artefacts.

Building Multiplex Systems (Design and Validation)

To build a multiplex system, primers must be chosen so that annealing temperatures are similar and have low affinity either to one another or to regions of the DNA outside the specific target template; this is achieved with the help of computer programs. Once a system has been designed, primer concentrations must be optimized so that even signals are obtained after the PCR. The first multiplex used in the UK was a quadruplex consisting of the loci HUMTH01, HUMVWA31, HUMFES/FPS and HUMF13A1.

The full forensic validation of the quadruplex was described by Lygo and co-workers in 1994. No spurious bands or mistyping occurred with old stains, although loss of sensitivity may happen; locus drop-out was also recorded in some that had been stored at 100% humidity for up to 1 month at 56°C. High molecular weight loci (HUMF13A1 and HUMFES/FPS) are more likely to be affected when samples are heavily degraded, and the DNA profiles produce a characteristic pattern where the size of the signal increases linearly from high to low molecular weight.

The discriminating power of the first system was relatively low (1 in 10 000), hence it was apparent that a new system comprising more loci was needed. This resulted in the development of a new multiplex comprising six loci and the amelogenin sex test (**Figure 3**). This new system (second-generation multiplex; SGM) was subject to the

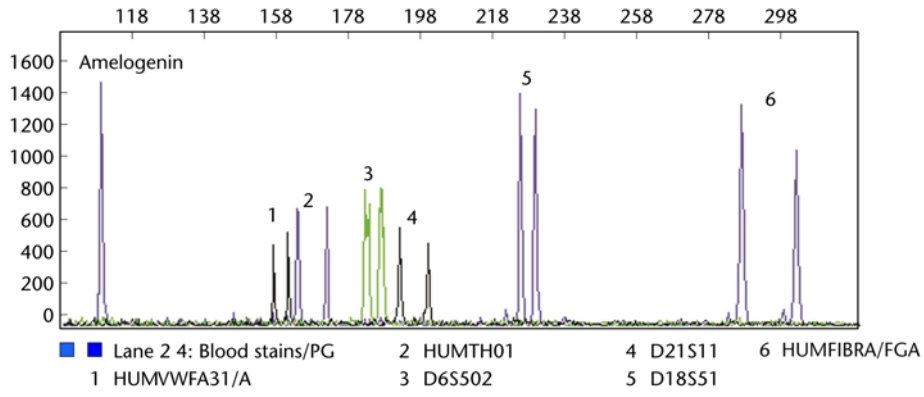

Figure 3 Second-generation multiplex used on the national deoxyribonucleic acid (DNA) database of England and Wales.

same validation procedures previously described. The advantage over the previous system is that the discriminating power is much greater (1 in 50 million). In 1999, a new validated system known as SGM plus was introduced. This system comprises 10 loci and the discriminating power is currently better than 1 in 1000 million.

DNA Databasing: the UK National DNA Database

In 1995, under the initiative of the Home Secretary, the Forensic Science Service, as custodian, was asked to set up the national DNA criminal intelligence database to cover England and Wales. It is projected that more than 200 000 samples per annum will be analysed, with a target database size of 5 million profiles. The system operates by analysis of buccal (mouth) scrapes or hair roots taken from any individual suspected or convicted of a recordable offence. The SGM results are held on computer in the form of a digital code based on the number of STR repeats for each locus. During normal casework, operational laboratories carry out analysis of crime material such as semen or blood stains. The DNA profiles derived from these samples are compared against the existing database. If a match is found, the police are informed to enable further investigations to be carried out. If suspects are subsequently found to be innocent, they are removed from the database.

To date, more than 700 000 samples have been stored on the DNA database. More than 65 000 matches between suspects and samples have been obtained. DNA profiling is usually thought of in terms of solving serious crimes, but it is interesting to note that the majority of matches arise from petty offences, such as burglary.

Probabilistic Considerations

If a crime is committed for which there is body fluid (e.g. bloodstain) evidence, and a suspect has been apprehended, then there are two possibilities to consider: either the suspect has shed the blood (C) or someone else has shed the blood ($\bar{C}$) . The strength of the evidence comprises a comparison of the two scenarios, and this is defined by the likelihood ratio ($LR = C/\bar{C}$). The probabilities of the two scenarios are conditional probabilities:

C: the probability of the evidence if the biological sample originated from the suspect

$\bar{C}$: the probability of the evidence if the biological sample originated from an unknown individual.

Suppose that a case comprises suspect (S) and victim (V), and the DNA profile matches that of the suspect. The probability of the evidence if C is true is 1. The probability of the evidence if the suspect is innocent ($\bar{C}$) is dependent on how common the DNA profile is in the population. Typically, this is a very small figure ($P < 1$ in 1000 million with current STR systems).

The rarity of a given DNA profile is dependent on the population from which the perpetrator came. The rarity will vary between different races. For example, the Forensic Science Service uses databases derived from three different racial groups: white Caucasian, Asian (Indian subcontinent) and African Caribbean (**Figure 4**). It is always necessary to condition the DNA profile on one of these race groups. In the absence of information about the race of the perpetrator, the scientist will default to the most conservative estimate derived from the three race groups. Additional calculations are carried out if the perpetrator could be related to the suspect.

Sometimes the case will demand special consideration, particularly if relatives are potential suspects. For example, if the suspect is innocent, could the perpetrator be related to the suspect? The likelihood ratio would be reduced in this scenario. Unless information is given to the contrary, the scientist will usually assume that the perpetrator is

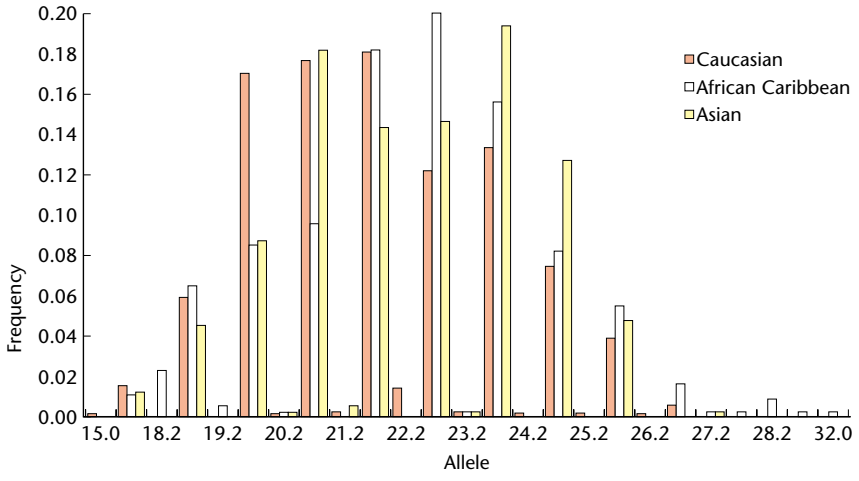

Figure 4 Example of a population frequency database. Three ethnic groups are compared for the frequency of HUMFIBRA/FGA. The allelic designation based on the number of repeats is given on the *x*-axis; respective frequencies are shown on the *y*-axis.

unrelated to the suspect, but the statement will make the assumption clear and will indicate the need to reconsider the evidence if it is believed that the perpetrator could be a close relative of the defendant.

The prosecutor's fallacy

It is important to realize that the forensic scientist can interpret evidence only in relation to the scenario presented by the court. A bloodstain found at a murder scene does not necessarily imply guilt. The DNA evidence must be combined with the other evidence in the case (by the jury) to reach the verdict. Phrases such as: 'The probability of the DNA profile coming from someone other than Mr X is less than 1 in 50 million' is known as the transposed conditional or the 'prosecutor's fallacy'. The statement must be rephrased: 'The probability of the DNA evidence, if it came from the suspect, is 1 in 50 million'. The emphasis is on the conditional probability.

Mitochondrial DNA

Mitochondria are cell organelles that contain a closed circular DNA molecule approximately 16.5 kb in size. The entire mitochondrial genome was sequenced by Anderson and co-workers in 1981, and this is used as the universal reference standard. The greatest variation is found within the control region; hence this is the region that is usually analysed by the forensic scientist (**Figure 5**). The control region consists of approximately 800 bases. When two randomly chosen individuals are compared, there are on average eight base differences between them.

Whereas there is only one nucleus per cell containing two copies of genomic DNA such as STRs, there are hundreds or thousands of mitochondria available for analysis. Forensic scientists typically deal with degraded material. Naturally occurring exonucleases gradually degrade the DNA and the average fragment size becomes progressively smaller so that it becomes more difficult to detect. Because mitochondrial DNA (mtDNA) molecules exist in such high copy number, sufficient longer-length DNA will remain long after the genomic DNA has completely degraded. In addition, mtDNA is inherited through the maternal line: the mother passes her mtDNA to all offspring and will share the same mtDNA with her grandmother, great grandmother, and so on. The mtDNA molecule acts rather like a time machine, passing from generation to generation relatively unchanged. It has been established that mutations can occur at a rate estimated at 1 per 33 generations and this usually leads to a condition known as heteroplasmy where a single base change may be observed, but the original and mutated forms coexist in admixture. This condition was prevalent in the Romanov family: compared with living relatives, the remains of Tsar Nicholas II demonstrated heteroplasmy at a single nucleotide position. Consequently, mtDNA is ideal for the analysis of highly degraded material which tends to be devoid of genomic DNA. Bone is one of the most difficult evidence types encountered. Here, there is also a strong overlap between forensic science and anthropology (ancient DNA analysis). In the forensic context, skeletal remains may be discovered; there may be other evidence, such as dental records, to suggest an identity of the body and, provided a (living) maternal relative (e.g. sibling) can be located, the hypothesis of identity can be tested. mtDNA has proven invaluable in solving historical mysteries such as the fate of the Romanov family and testing the claim of Anna Anderson to be the Duchess Anastasia.

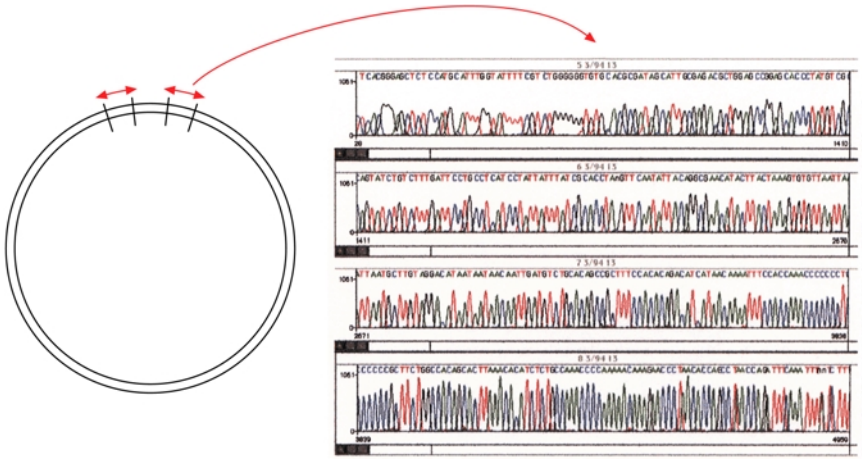

Figure 5 Schematic representation of a mitochondrial deoxyribonucleic acid molecule illustrating a 400-base sequence from the control region.

The second use of mtDNA is the analysis of hair shafts. Although it is not problematical to extract genomic DNA from the hair root, this is usually not an option unless the hairs have been physically plucked. Hairs that are shed naturally, because they are in the telogen phase, are usually devoid of roots and the shaft itself is devoid of genomic DNA. mtDNA can be routinely analysed, however.

Probabilistic Considerations of Mitochondrial DNA

Calculation of the frequency of occurrence of a mitochondrial sequence or 'mitotype' is simply the number of times that the sequence has been observed in the database. Caveats apply to mtDNA reporting; in particular, courts need to be aware that the same mitotype can be shared by maternal relatives (e.g. siblings). A typical discrimination power will be one in a few hundred. This contrasts with figures typically quoted for genomic DNA.

The Y Chromosome

Increasing interest is being shown in the analysis of Y-chromosome polymorphisms. The Y chromosome is haploid and paternally inherited. It is passed from father to son relatively unchanged, except by the gradual accumulation of mutations. It is therefore the direct antithesis of mitochondrial DNA in forensic terms, except that it is does not exist in multiple copies; hence the sensitivity of the test is equivalent to that for ordinary STRs. Several kinds of polymorphism are found in the Y chromosome, but only two are of interest to the forensic scientist:

1. Binary polymorphisms: single base substitutions are relatively common in the Y chromosome.
2. In addition to binary markers, there are also microsatellite markers.

These markers are used by geneticists to construct population phylogenies. There is little doubt that, as more markers are discovered, the potential for generating discriminating race markers becomes a realistic possibility. Y-chromosome markers may also be useful in the interpretation of male–female mixtures; it is unlikely that Y-chromosome markers could be used on their own.

Casework Examples

STRs are used predominantly in casework. Suspects are profiled by means of buccal scrapes, as this is the cheapest method available. However, comparisons will always be made with items received from a crime scene. These items will typically include blood stains from affrays or burglary scenes, and semen stains or semen-contaminated vaginal swabs from sexual crimes.

The example in **Figure 6** shows a typical example from a rape case where a mixed male–female DNA profile has been obtained. An interpretation of the mixed profile can be attributed to a combination of DNA from the suspect (male) and the female (victim).

STRs will work on very degraded samples. A good example is demonstrated by analysis of samples from the Waco, Texas disaster of 1993 where body tissues and bone damaged by fire and decomposition were taken from the

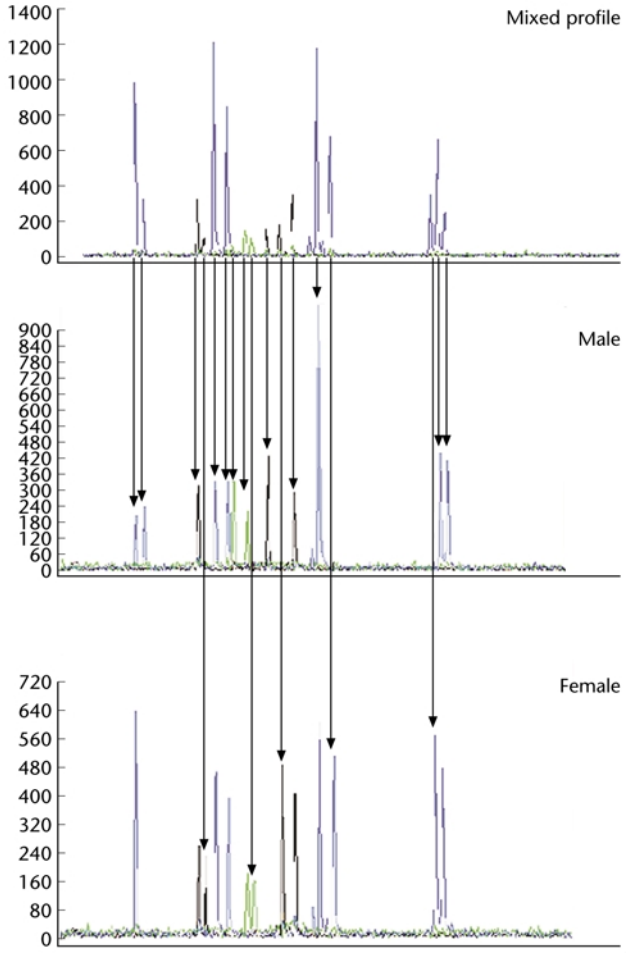

Figure 6 Typical case comprising a mixture (crime sample) from a male suspect and a female victim. The peaks that correspond to the male and female are shown by arrows. One explanation for the crime sample is that it comprises a suspect–victim mixture. See **Figure 3** for explanation of loci in these particular profiles.

remains of 61 people. A total of 50 bodies gave full STR results.

Nevertheless, there is a limitation to the sensitivity of STRs. When samples are very seriously degraded, such as aged bone, or there is simply too little sample to test, results are not achieved. Although research is underway to analyse ever smaller samples using STRs (indeed preliminary results have been reported from just a single cell), the technique of choice is currently mtDNA.

mtDNA is used on discrete evidence types including hairs, aged bone, teeth and faeces. The sequencing technique described above is labour intensive and expensive. However, Tully and co-workers have recently introduced a 'minisequencing' test (**Figure 7**), which considerably speeds the process. Rather than sequence 800 bases of the control region, just 12 regions containing a

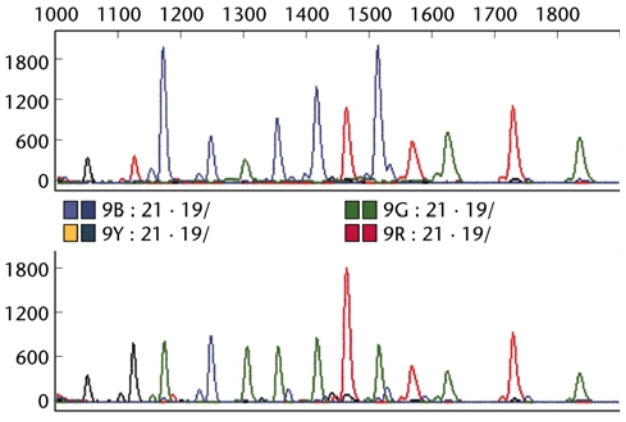

Figure 7 Example of a minisequencing test. A suspect's buccal scrape (a) was compared with a sample extracted from faeces (b). The profiles are different, so the suspect was excluded.

highly polymorphic base position are compared using a multiplex reaction. If two samples have the same origin, the profiles will be identical. However, it may still be necessary to confirm the result using full sequencing because this test is not as powerful. The minisequencing test is an excellent exclusion tool.

Further Reading

Anderson S, Bankier AT, Barrell BG *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**: 457–465.

Bar W, Brinkmann B, Budowle B *et al.* (1997) DNA recommendations. Further report of the DNA commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics. *International Journal of Legal Medicine* **110**: 175–176.

Clayton TM, Whitaker JP and Maguire CN (1995) Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Science International* **76**: 7–15.

Clayton TM, Whitaker JP and Fisher DL (1995) Further validation of a quadruplex STR DNA typing system: a collaborative effort to identify victims of a mass disaster. *Forensic Science International* **76**: 17–25.

Findlay I, Taylor A, Quirke P, Frazier R and Urquhart A (1997) DNA fingerprinting from single cells. *Nature* **389**: 555–556.

Gill P, Ivanov PL, Kimpton C *et al.* (1994) Identification of the remains of the Romanov family by DNA analysis. *Nature Genetics* **6**: 130–135.

Gill P, Kimpton C, Aliston-Griner R *et al.* (1995) Establishing the identity of Anna Anderson Manahan. *Nature Genetics* **9**: 9–10.

Gill P, Urquhart A, Millican E *et al.* (1995) A new method of STR interpretation using inferential logic – development of a criminal intelligence database. *International Journal of Legal Medicine* **109**: 14–22.

Ivanov PL, Wadhams MJ, Roby RK *et al.* (1996) Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georji Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nature Genetics* **12**: 417–420.

Lygo JE, Johnson PE, Holdaway DJ *et al.* (1994) The validation of short tandem repeat (STR) loci for use in forensic casework. *International Journal of Legal Medicine* **106**: 302–311.

Parsons TJ, Muniec DS, Sullivan K *et al*. (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nature Genetics* **15**: 363–368.

Sparkes R, Kimpton C, Watson S *et al*. (1996) The validation of a 7-locus multiplex STR test for use in forensic casework. (I) Mixtures, ageing, degradation and species studies. *International Journal of Legal Medicine* **109**: 186–194.

Sparkes R, Kimpton C, Watson S *et al*. (1996) The validation of a 7-locus mutliplex STR test for use in forensic casework. (II) Artefacts, casework studies and success rates. *International Journal of Legal Medicine* **109**: 195–204.

Tully G, Sullivan KM, Nixon P, Stones RE and Gill P (1996) Rapid detection of mitochondrial sequence polymorphisms using solid-phase fluorescent minisequencing. *Genomics* **34**: 107–113.

Whitaker JP, Clayton TM, Urquhart AJ *et al*. (1995) Short tandem repeat typing of bodies from a mass disaster: high success rate and characteristic amplification patterns in highly degraded samples. *Biotechniques* **18**: 670–677.