Criminal intelligence databases and interpretation of STRs

P. Gill, A. Urquhart, E. Millican, N. Oldroyd, S. Watson, R. Sparkes and C.P. Kimpton

Forensic Science Service, Priory House, Gooch Street North, Birmingham B65QQ, UK

INTRODUCTION

DNA profiling in forensic science in the UK is focussed on the analysis of short tandem repeat (STR) loci using PCR. It is the technique of choice for the national strategy to create criminal intelligence databases. Apart from the increased sensitivity inherent with any PCR technique, with STRs there is also the advantage of definitive allelic identification. This is a consequence of lower measurement errors associated with the use of polyacrylamide gel electrophoresis to detect DNA fragments ranging between 200-400bp in size (Ziegle et al. 1992).Because of their small sizes STRs are more likely to be successful on old or badly degraded material (Gill et al. 1994; Hagelberg et al. 1991; Jeffreys et al. 1992; Weigand et al. 1993) an important aspect of forensic casework.

The National DNA database Unit

Recently, a change in the UK legislation allowed the formation of a national DNA database. The purpose is to store DNA profiles derived from individuals either suspected or convicted of crimes. DNA is sampled from either buccal scrapes or from hair-roots. The aim is to store 135,000 DNA profiles per year and it is envisaged that the database may eventually contain 5 million profiles. This is a significant proportion of the UK population of 60 million people. As the custodian of the DNA database, the Forensic Science Service (FSS) has constructed a unit at the FSS headquarters, Birmingham, UK, consisting of c.125 scientists, eight 373A ABD automated sequencers and six 377 ABD automated sequencers.

METHOD

To decide the method of choice, the following requirements must be fulfilled:

- It must be reliable (the quality of results must be high).
- Throughput must be high.
- The process must be cost-effective.

Automation is the key to achieving all three requirements.

Choice of loci

Several factors are considered when choosing candidate loci:

- Discriminating power (Jones 1972) of >0.9 (Observed heterozygosity >70%).
- The predicted length of alleles must be approximately between 90-500bp (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.

- Chromosomal location (to ensure that closely linked loci are not chosen).
- Robustness and reproducibility of results, low stuttering characteristics.

Dimeric loci cannot be used because slippage during amplification results in spurious bands that are difficult to interpret whereas trimeric and tetrameric loci are less prone to this problem. Complex hypervariable loci such as HUMACTBP2 have more than 30 alleles. Although originally described (Polymeropoulos et al. 1992; Warne et al. 1991) as a tetrameric repeat, sequencing by Urquhart et al. (1993) and Weigand et al. (1993) has revealed that repeats differ by 1,2, or 3 bps; in addition, different alleles the same size, may have different sequences. To designate alleles to these types of STRs is difficult, although there is no reason why interpretation cannot be carried out based using size estimates (bp), but the system is not discrete

To achieve higher discriminating powers required for large criminal intelligence DNA databases, we have chosen to utilise the advantages of complex tetrameric repeat loci described by Urquhart et al. (1993). Because these have alleles that differ in size by 2bp, the DPs are greater compared to those previously utilised by Kimpton et al. (1994) (Table 1). The loci used in the National DNA database are listed; there are 3 complex STRs included - D18S51 (15 common alleles); D21S11 (21 common alleles); HUMFIBRA/FGA (21 common alleles).

Table 1. List of loci used in the National DNA database; chromosomal locations and Pms for 3 different ethnic groups.

Locus	Chromosomal	Probability of Match (Pm)		
	location	Caucasian	Afro-Caribbeans	Indo-Pakistani
AMELOGENIN	Xp22, Yp11.2	X,Y	X,Y	X,Y
HUMVWFA31/A	12p12-pter	0.064	0.057	0.075
HUMTH01	11p15-15.5	0.086	0.100	0.084
HUMFIBRA*	4q28	0.044	0.027	0.031
D21S11*	21	0.051	0.042	0.046
D18S51*	18q21.3	0.029	0.024	0.042
D6S502	6	0.047	0.061	0.054
	Total Pm	**1.7×10^{-6}**	**1×10^{-9}**	**1.2×10^{-8}**

* Display alleles differing by 2bps

MULTIPLEXING

The availability of 4 distinguishable fluorescent dyes facilitates the development of STR multiplexes (i.e. single tube reactions) enables loci that have overlapping allele size ranges to be labelled with different colours.

To build a multiplex system, primers must be chosen so that annealing temperatures are similar and have low affinity either to each other or to regions of the DNA outside the specific target template; this is achieved with the help of computer programs such as Oligo™. Once a system has been designed, primer concentrations must be optimised so that even signals are obtained after PCR .

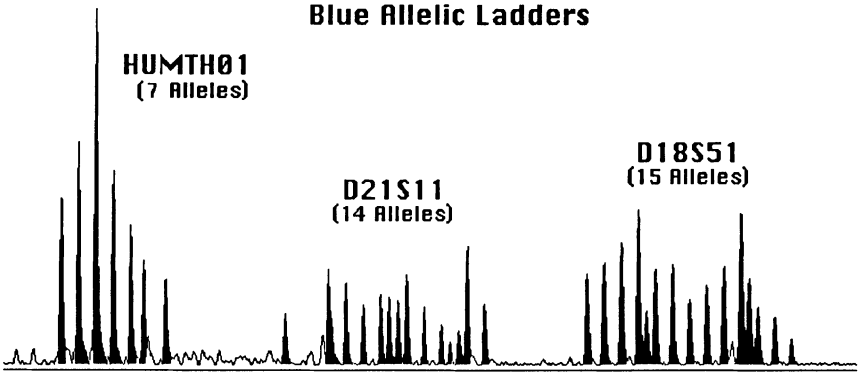
Accordingly, we have developed a quadruplex consisting of the loci HUMTH01, HUMVWFA31, HUMFES/FPS and HUMF13A1 for use in routine casework throughout the UK (Kimpton et al. 1994; Lygo et al. 1994). All four loci have discrete classes (Kimpton et al. 1993). An allelic class may comprise 2 or more alleles, for example we do not distinguish between the HUMTH01 9.3 allele and the rarer 10 allele which differs in size by 1bp. The combined frequency of the two alleles is therefore used for interpretation purposes.

The main purpose of using multiplexes, is to speed the process of analysis. Inevitably, there may be some loss of efficiency of amplification since the conditions used are a compromise. This has no significant implications for the database, since the operator has large quantities of undegraded DNA available for analysis. Furthermore, the DNA is never a mixture. In casework, where the sample is less predictable, singleplexes may sometimes be used to identify difficult (e.g. degraded) samples (i.e. singleplexing and multiplexing are not mutually exclusive techniques).

DETERMINING THE SIZE OF STR ALLELES

To determine the sizes of DNA fragments, standard marker ladders consisting of all the common alleles of a given locus can be used for comparison on an electrophoretic gel (Puers et al. 1993; Puers et al. 1994). They are made by mixing together DNA from different individuals displaying the entire range of alleles for comparison and carrying out PCR on the mixture. Ladder markers are widely utilised and available throughout the forensic community.

Fig. 1. Multiplexed allelic ladders for 3 STR loci.



However, the major advantage of fluorescence is that internal size markers can be incorporated and this improves sizing accuracy both within and between gels (Mayrand et al. 1992; Ziegle et al. 1992; Kimpton et al. 1993). Whereas the use of allelic ladders is indispensable with non-automated systems, fluorescently tagging STR loci enables inclusion of size standards as an internal size marker within each lane. This reduces measurement errors and allows automatic sizing of STR-PCR products with GENESCAN™ 672 analysis software. For the majority of loci, reliable sizing can be achieved by direct comparison with a lambda *Pst I* restriction digests (ABD GS2500 or GS350). For calibration purposes allelic ladders are used in our laboratory (if new software is installed, for example).

The consistency of automatic size calling against the GS2500 or GS350 ladder markers was evaluated for each locus studied by examination of the distribution of computer-generated band sizes of allelic ladders for a large number of samples (72 allelic ladders run across 6 different gels). These experiments are used to set windows based on the range of observed sizes of any given allele. Accuracy has been demonstrated to be within a range of approximately 1.2 bp (Table 2 of Kimpton et al. 1993 ; Fig. 2). Ranges based on observations are programmed into the ABD GENOTYPER™ software, enabling automatic allele designations to be made.

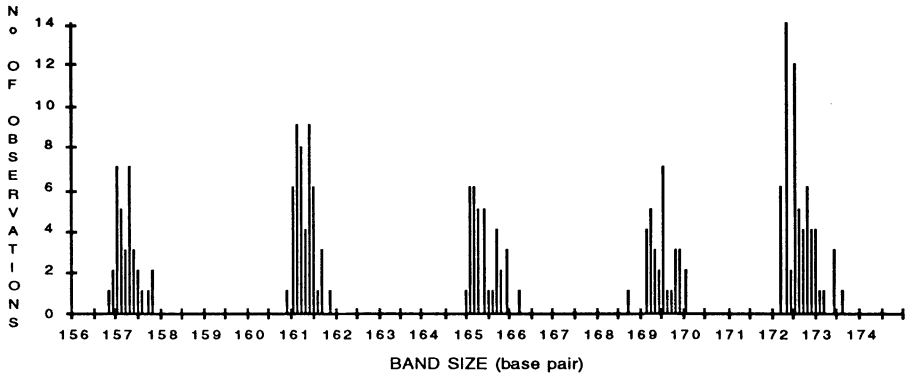


Fig. 3: Distribution of HUMTH01 alleles sized against the GS2500 standard (the allele at the far right comprises 9.3 and 10 alleles, hence the range is relatively large).

A NEW METHOD OF RULE-BASED INTERPRETATION USING ALLELIC LADDER CONTROLS

HUMFIBRA/FGA (Mills et al. 1992) is a complex repeat locus with 21 common alleles differing by increments of just 2bp. If complex loci were perfect 2bp repeats, then the window established to encompass measurement error for any given allele would be just ±1bp of its mean. However, intermediate alleles differing by 1bp may also occur. For example, we have recently discovered a 22.3 allele in HUMFIBRA/FGA. This means that to identify these very rare ($p<0.001$) alleles, windows must be no greater than ±0.5bp. The difficulty is that measurement errors may exceed this value, hence a different approach is needed to use complex STRs as discrete systems.

An alternative approach to that described above is to calculate the relative difference in size between a questioned sample and an allelic ladder marker control on the same gel. All sizes, in base pairs, are calculated by reference to the internal GS350 standard. If the distance is within a predetermined range, then the allele can be designated. This is a fundamentally different approach to that previously described. Whereas the former method calculates 'absolute' windows based upon repeated running of allelic standard markers, the new method calculates windows relative to ladder markers on the same gel - windows are therefore set for each individual gel.

Examination of the population distribution of HUMFIBRA/FGA (Fig. 3) reveals that the most common alleles such as 19,20,21 are complete tetramers, whereas intermediates (19.2,20.2,21.2) are much rarer ($p<0.02$). If α.0 represents a complete no. of repeats, we can generalise that α.0 repeats are common; α.1 and α.3 alleles are always very rare ($p<0.001$) and α.2 repeats are intermediate. In some systems we have examined e.g. (D21S11), some α.2 variants are relatively common, yet the extreme rarity of α.1 and α.3 alleles holds true for complex loci listed in Table 1.

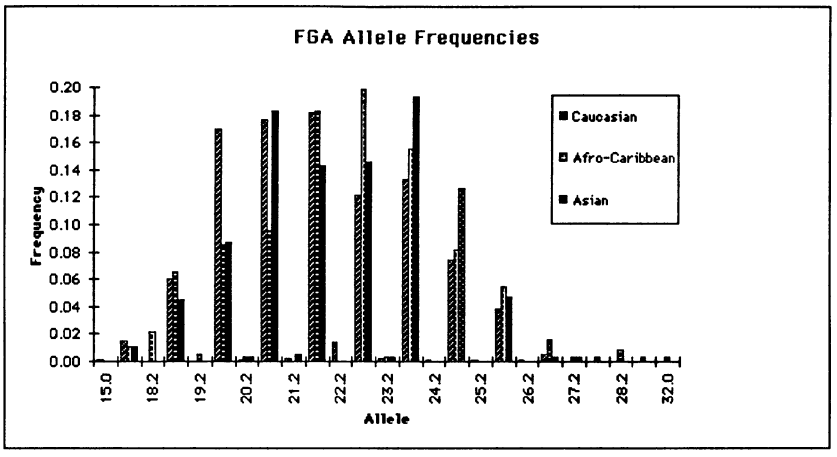


Fig. 3: HUMFIBRA/FGA population survey of 3 different ethnic groups

For HUMFIBRA/FGA, the following conditional logic applies:

- **If** an α.1 or α.3 allele is observed then this is an extremely rare occurrence.
- Given that the size distribution of individual STR alleles may be in the region of ±0.6bp, an explanation may be that an α.0 or α.2 variant is in a tail of its measurement error distribution such that it now resides in an adjacent window, normally occupied by an α.1 or α.3 variant.
- To test this possibility, re-run the sample to determine if the result is reproducible.
- **If** the allele is a true α.1 or α.3 variant then the locus would normally be heterozygous (unless, the population from which the sample is derived is atypical, or inbred). It follows that the partner allele must also normally be a common α.0 or α.2 variant.
- **If** a heterozygous sample is observed where both alleles are apparently α.1 or α.3 variants then that result suggests the strong possibility that 2 common variants are shifted into the tails of their error distributions. In fact band shifts are strongly correlated (Gill, unpublished observations).

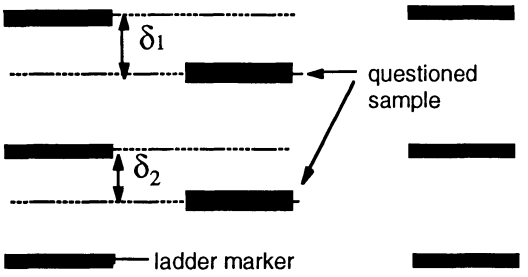


Fig. 4: Exaggerated diagrammatic representation of band shift showing correlation. The questioned sample is compared to ladder markers on the gel.

Measuring the correlation effect is a useful diagnostic tool. To do this the following procedure is used (fig 4):

- **If** both alleles are <0.5bp from the closest α.0 or α.2 ladder marker then continue to the next test.
- Measure the correlation effect (c= δ1-δ2). **If** c<0.5 then the alleles may be designated.
- Extremely rare α.1 or α.3 variants can only be designated **if** the sample has been separately analysed and the same results obtained.

The utility of this procedure is best illustrated by reference to an actual example

Identification of a rare HUMFIBRA/FGA 22.3 allele

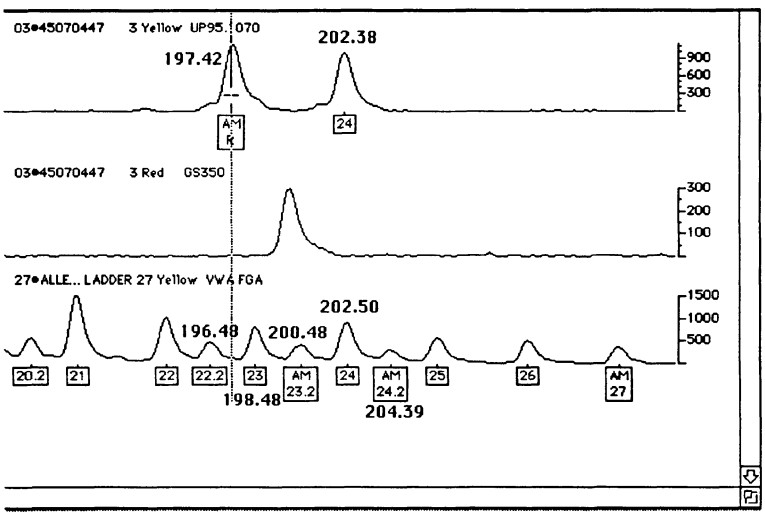


Fig 5: Interpretation of an extremely rare variant allele in HUMFIBRA/FGA - showing designations made by GENOTYPER. The top lane is the questioned sample; the bottom lane is the allelic ladder.

Table 2: Compilation of allele sizes illustrated in Fig 5.

Allele Designation	Size in questioned sample (bp)	Size in allelic ladder (bp)	Difference (δ)
22.3	197.42	197.48 [t]	-0.06 (δ1)
22.2		196.48	+0.98*
23		198.48	-0.98*
24	202.38	202.50	-0.12 (δ 2)

* **If** the questioned allele (designated 22.3) is truly a common allele it must be either 22.2 or 23. The test fails this condition since δ >0.5 (ignore sign).
[t] The 22.3 allele is not in the allelic ladder, but its size can be estimated as the mean of alleles 22.2 and 23. The difference passes the condition d <0.5 (ignore sign), **if** the allele is 22.3.

Correlation measurements

- **If** the questioned allele is 22.3 then the observed correlation with allele 24 (c=δ1-δ2) is +0.06bp, passing the condition since c<0.5
- **If** the questioned allele is 22, then the observed correlation with allele 24 is +1.1bp, failing the condition since c>0.5.
- **If** the questioned allele is 22.2, then the observed correlation with allele 24 is -0.86bp, failing the condition since c>0.5 (ignore sign) .
- Because the putative allele is an α.3 variant, the sample was separately reanalysed, confirming the above interpretation.

Modification of the above procedure may be needed for different loci but the same principles can be applied, given knowledge of rare and common variants; for example in HUMTH01, both 9.3 and 10 alleles are common, whereas 9.2 and 10.1 alleles are not observed (or extremely rare). Otherwise all remaining common alleles are α.0 variants.

Universal application of the relative measurement method

These principles, incorporating use of the 0.5bp rule, can be universally applied to interpretation of STRs, regardless of the platform or method used, provided that allele sizes are always cross-referenced to allelic ladders. The number of inconclusive results (apparent extremely rare alleles observed) is dependent upon the resolving power of the system used, but it is not a pre-requisite that the true measurement error <0.5bp.

In our hands the 'absolute' method of window estimation suffers from the drawback that windows must be reset when new acrylamide batches are prepared - other factors may necessitate the use of different windows in different laboratories. On the other hand, the relative measurement approach has a universal applicability because the windows are constant for each locus and independent of the method used. The principle of rule-based interpretation packages built upon experimental observations can be extended to enable construction of algorithms or computer based expert systems. This will automate the interpretation procedure, guiding the scientist to interpret results in the presence of artefacts, such as stutters, and to distinguish components of mixtures.

REFERENCES

DNA recommendations (1994) Int. J. Leg. Med. 107 159-160

Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K, (1994) Identification of the remains of the Romanov family by DNA analysis. Nature Genetics 6: 130-135.

Hagelburg E. Gray IC. Jeffreys AJ (1991) Identification of the skeletal remains of a murder victim. Nature 352: 427-429.

Jeffreys AJ, Allen MJ, Hagelburg E, Sonnberg A (1992) Identification of the skeletal remains of Josef Mengele by DNA analysis. For. Sci Int. 56: 65-76.

Jones D.A. (1972) Blood samples: Probability of discrimination. J. Forens. Sci. Soc. 12: 335-359.

Kimpton CP. Gill P. Walton A. Urquhart A. Millican ES. and Adams M (1993) Automated PCR profiling employing 'multiplex' amplification of short tandem repeat loci. PCR Methods and Applications 3: 13-22

Kimpton C. Fisher D. Watson S. Adams M. Urquhart A. Lygo. J.E. & Gill P (1994) Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. Int. J. Leg. Med. 106: 302-311.

Lygo JE, Johnson PE, Holdaway DJ, Woodroffe S, Whitaker JP, Clayton TM, Kimpton CP, Gill P (1994) validation of of short tandem repeat (STR) loci for use in forensic casework. Int. J. Leg. Med. 107: 77-89

Mills KA, Even D, Murray JC (1992) Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA) Hum Molec. Genet. 1992 1 779

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1992) Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene H-beta-Ac-psi-2 (ACTBP2). Nucleic Acids Res. 20 1432.

Puers C, Hammond H,A, Jin L, Caskey C,T, and Schumm J.W. (1993) Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01 $[AATG]_n$ and reassignment of alleles in population analysis by using a locus-specific allelic ladder. Am J Hum Genet. 53: 953-958.

Puers C, Hammond HA, Caskey CT, Lins AM, Sprecher CJ, Brinkmann B, Schumm JW (1994) Allelic ladder characterization of the Short Tandem Repeat polymorphism located in the 5' flanking region to the human coagulation factor XIII A subunit gene. Genomics 23: 260-264.

Urquhart A, Kimpton CP, Gill P (1993) Hypervariability of the tetranucleotide repeat of the human beta-actin related pseudogene H-beta-Ac-psi-2 (ACTBP2) locus. Hum Genet 96: 637-638

Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in short tandem repeat sequences - a survey of twelve microsatellite loci for the use as forensic identification markers. Int J Leg Med. 107 13-20.

Warne D, Warkins C, Bodfish P, Nyberg K, Spurr NK (1991) Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene 2 (ACTBP2) detected using the polymerase chain reaction. Nucleic Acids Res. 1991 20 1432.

Weigand P, Budowle B, Rand S, Brinkmann B (1993) Forensic validation of the STR systems SE33 and TC11. Int J Leg Med. 105: 315-320.

Ziegle JS, Su Y, Corcoran KP, Nie L, Mayrand PE, Hoff LB, McBride LJ, Kronick MN, Diehl SR, (1992) Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics 14: 1026-1031.