Selection of STR loci for forensic identification systems.

Urquhart, A.J., Oldroyd, N.J., Downes, T., *Barber, M., Alliston-Greiner, R., Kimpton, C.P. and Gill, P.D.

The Forensic Science Service, Birmingham, UK
*Metropolitan Police Forensic Science Laboratory, London, UK

INTRODUCTION

Short Tandem Repeat (STR) profiling is rapidly growing as a method of individual identification for forensic and other purposes. Several multiplex STR systems are available (e.g. Kimpton et al, 1993), offering matching probabilities of about 10^{-4}. The quadruplex STR system developed in our laboratory and presently in use in forensic casework gives a matching probability of this order in three British populations (Kimpton et al, 1993). We have investigated numerous STR loci for use in further multiplex STR systems. Here we discuss the criteria for locus selection, with particular reference to the repeat sequence at the loci under investigation.

SELECTION CRITERIA

STR loci were initially selected on the basis of their reported heterozygosity (over 70%, or a Discriminating Power of >0.8). The next considerations are their size range, optimal annealing temperature and ability to co-amplify with other selected loci. Once a prototype multiplex has been designed, we investigate the propensity to form stutter bands, the ratio of n to n+1 peaks (Robertson et al., 1995).

SEQUENCE VARIATION

We have divided STR loci into 3 categories on the basis of their variable sequence (Urquhart et al, 1994). *Simple repeats* comprise an invariant unit of 3, 4 or 5 base pairs repeated a variable number of times; *compound repeats* contain 2 or more adjacent simple repeats; *complex repeats* consist of several repeat blocks of variable unit length along with more or less variable intervening sequence (Urquhart et al, 1994). Each type of repeat may exhibit alleles which differ from the consensus sequence at that locus, such as the well-characterised 9.3 allele at HUMTHO1 (Puers et al, 1993). Some loci exhibit alleles which show sequence variation between alleles of the same size, e.g. HUMVWFA31/A, HUMACTBP2 and D11S554 (Urquhart et al, 1993 and 1994; Adams et al, 1993).

NEW STR SEQUENCE DATA

The most comprehensively investigated tetranucleotide repeat loci are those which comprise TCTA/TAGA and AAAG/CTTT repeat units. This may reflect the primers used to search for repeats rather than their relative frequency in the genome. A search of GenBank for loci containing $(TCTA)_5$ or its complement revealed several loci in which the majority repeat unit TCTA had apparently mutated (Urquhart et al, 1994). The most common variant repeat unit was TCTG, but TCCA, TCA and TA were also found. We have recently sequenced 3 more TCTA loci (Fig. 1). TCTG repeats are seen at D6S502, TCA at D3S1359 and TA at D19S253. In addition, we find a TGTA unit at D3S1358 and a CA unit at D19S253.

D6S502	$(TCTA)_{8-12}$ $(TCTA)_{1-2}(TCTG)_{1-2}(TCTA)_{9-15}$
D19S253	$TCTA.TA.TCTA.CA.TA(TCTA)_{5-14}$
D3S1359	$(TCTA)_{11-18}$ $(TCTA)_{7-9}TGTA(TCTA)_{5-12}$ $(TCTA)_{7-9}TGTA.TCTA.TCA(TCTA)_{8-17}$

Figure 1. Sequences found at the D6S502, D19S253 and D3S1359 loci.

We have also sequenced the primarily AAAG repeats at the D18S51, HUMFIBRA and HUMAPOAI1 loci (Barber et al, to be published) (Fig. 2). D18S51 is a compound repeat of AAAG and AG. Most alleles contain an AG dinucleotide followed by 4 AG dinucleotides immediately 3' to the AAAG tract. A minority of alleles contain 6 AG dinucleotides in this position. This difference causes the occurrence of intermediate alleles (designated X.2). The repeat at the HUMFIBRA locus is complex, comprising polymorphic GAAG and AAAG tracts as well as other variants on the AAAG basic unit. The difference between X and X.2 alleles is the absence in the X.2 alleles of one AG dinucleotide immediately 3' to the large AAAG tract. One X allele and 2 X.2 alleles contain insertions immediately 5' to the large AAAG tract. One (shown on the left in Fig. 2) comprises $(AAAG)_5GAAG$. Of the other two (on the right in Fig. 2), one has a sequece of $(AAAG)_3(GAAG)_3$, while the other has this sequence followed by $(AAAG)_{13}(ACAG)_5$. Other alleles close in size to the latter have been seen, bit not yet sequenced. The alleles containing these insertions are the largest seen at the locus. We speculate that these insertions are stabilized by the presence of variant (non-AAAG) sequence within the repeat area, as has been proposed at other repeat loci. Alleles at HUMAPOAI1 are extemely complex both by our narrow definition and in the general sense of the word. Insertions containing odd numbers of base pairs are common, leading to alleles differing by 1 bp accross almost the entire alllelic size range.

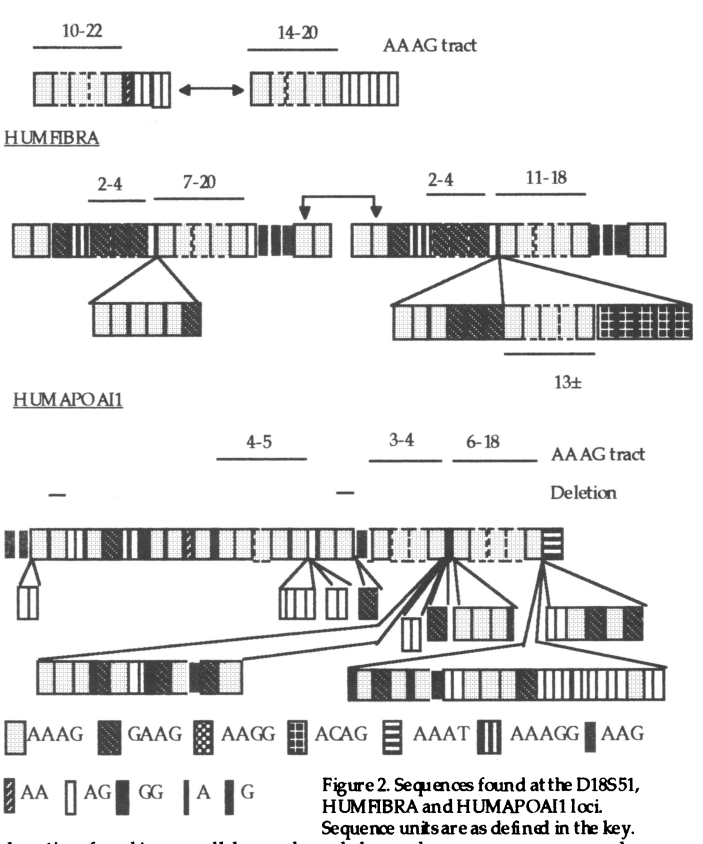

Figure 2. Sequences found at the D18S51, HUMFIBRA and HUMAPOAI1 loci. Sequence units are as defined in the key. Insertions found in some alleles are shown below each consensus sequence, and deletions and variable-length polymorphic tracts are above. Arrows show possible mutations between two types of allele at a locus.

STR EVOLUTION

Inspetion of the sequence within and adjacent to repeat sequences suggests some mechanisms that may be involved in STR evolution. The most common repeat units associated with AAAG repeats are AG and AAGG, while those at TCTA repeats are TCTG, TA and TCA (Urquhart et al, 1994). As might be expected, substitutions normally retain purine or pyrimidine at the affected positon, while the commonest deletions preserve adjacent bases. Mutations to di- or tetranucleotides, particularly AAGG, AG and TCTG appear prone to further expansion. Interestingly, both AAAG and TCTA appear to mutate to ACAG or its complement TCTG (AAAG at one HUMFIBRA allele, TCTA at many loci).

ALLELES DIFFERING BY ONE BASE PAIR

There is a well-chracterised 1 bp allele difference at the HUMTHO1 locus (Puers et al, 1993). We have recently seen an 8.3 allele at this locus. Since the launch of the UK's National DNA Database, we have typed a huge number of individuals at six STR loci. We have found two apparent instances of 1 bp allele differences at other loci. The alleles concerned are a 64.1 allele at D21S11 and a 16.1 allele at HUMFIBRA. We have developed an interpretation system which will reliably type alleles differing by 1 bp (Gill et al, to be published). Sequencing of these alleles is under way.

REFERENCES

Adams M, Urquhart A, Kimpton C, Gill P (1993) The human D11S554 locus: four distinct families of repeat pattern alleles at one locus. Hum. Molec. Genet. 2: 1373-1376

Barber M, McKeown B, Parkin B (to be published) Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus (HumFGA). Forensic Sci. Int.

Gill P, Urquhart A, Millican E, Oldroyd N, Watson S, Kimpton C (to be published) Criminal intelligence databases and interpretation of STRs. Adv. Forensic Haemogenetics

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

Puers C, Hammond HA, Jin L, Caskey CT, Schumm JW (1993) Identification of repeat sequece homogeneity at the polymorphic short tandem repeat locus HUMTHO1[AATG]$_n$ and reassignment of alleles in population analysis by using a locis-specific allelic ladder. Am. J. Hum. Genet. 53: 953-958

Robertson JM, Badger CA, Buoncristiani MR (1995) Design of short tandem repeat systems suitable for human identification. Proc. 5th Intl. Symposium on Human Identification 1994. Promega Corp.

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

Urquhart A, Kimpton CP, Gill P (1993) Sequence variability of the tetranucleotide repeat of the human beta-actin related pseudogene H-beta-Ac-psi-2 (ACTBP2) locus. Hum. Genet. 92: 637-638