ORIGINAL ARTICLE

A. Urquhart · C. P. Kimpton · T. J. Downes · P. Gill

Variation in Short Tandem Repeat sequences – a survey of twelve microsatellite loci for use as forensic identification markers

Received: 1 February 1994 / Received in revised form: 28 April 1994

Abstract Alleles at 12 Short Tandem Repeat loci have been sequenced to investigate candidate loci for a multiplex Short Tandem Repeat system for forensic identification, and for single-locus amplification of Short Tandem Repeat loci. Variation from the consensus sequence was found at 6 loci, while one locus, D21S11, was found to be complex in sequence. The presence of non-consensus alleles does not rule out loci for inclusion as forensic identification markers, but size differences between alleles of 1 base pair require very precise sizing. We suggest criteria for the suitability of Short Tandem Repeat loci as forensic identification markers, and propose a universal allele nomenclature for simple and compound Short Tandem Repeats. The effect of the repeat unit sequence of the evolution of Short Tandem Repeats is discussed.

Key words Short Tandem Repeats · Microsatellites DNA sequencing · Polymerase Chain Reaction · Forensic DNA typing

Zusammenfassung Allele an 12 Short-Tandem-Repeat Loci wurden sequenziert, um Kandidaten für ein Multiplex Short Tandem Repeat System für forensische Identifikationen und für Single-Locus Amplifikationen von Short-Tandem-Repeat Loci zu untersuchen. Abweichungen von der Konsensus-Sequenz wurden an 6 Loci gefunden, während ein Locus, D21S11, als Komplex in der Sequenz gefunden wurde. Die Anwesenheit von Non-Konsensus-Allelen schließt solche Loci nicht aus für die Einbeziehung als forensische Identifikationsmarker. Aber Größendifferenzen von einem Basenpaar zwischen Allelen erfordern eine sehr genaue Größenbestimmung. Wir empfehlen Kriterien für die Eignung von Short-Tandem-Repeat Loci als forensische Identifikationsmarker und schlagen eine universale Allelnomenklatur für einfache und komplexe Short-Tandem-Repeats vor. Die Auswirkung der Sequenz der Repeateinheit auf die Entwicklung von Short-Tandem-Repeats wird diskutiert.

Schlüsselwörter Short-Tandem-Repeats Mikrosatelliten · DNA-Sequenzierung · Polymerase Kettenreaktion · Forensische DNA-Typisierung

Introduction

Analysis of Short Tandem Repeat (STR) sequences by the polymerase chain reaction (PCR) is becoming the method of choice for the forensic identification of body fluids (Kimpton et al. 1993, 1994; Fregeau and Fourney 1993; Wiegand et al. 1993). Because of problems caused by 'shadow bands' when analysing dinucleotide repeats (Hauge and Litt 1993), the less common tri-, tetra- and pentanucleotide repeats are preferred.

STR sequences vary in the length of repeat unit, the number of repeats and the rigour with which they conform to an incremental repeat pattern. 'Simple' repeats contain units of identical length and sequence, 'compound' repeats comprise 2 or more adjacent simple repeats, 'complex' repeats may contain several repeat blocks of variable unit length, along with more or less variable intervening sequences.

We have recently studied sequence variation at 2 complex STR loci, HUMACTBP2 (SE33) and D11S554 (Urquhart et al. 1993; Adams et at. 1993). Both these loci were originally reported to have allele sizes which differed by 4 base pair increments (Warne et al. 1991; Phromchotikul et al. 1992). However, our sequence data showed that at both loci allele size differences of 1, 2 or 3 base pairs also exist.

Since allele designation of STR PCR products depends on accurate sizing, we investigated a range of simple, compound and complex STR loci which were being screened in this laboratory for use as forensic identification markers. The markers used in our quadruplex STR system (Kimpton et al. 1993, 1994), a *Pst*I digest of bacteriophage lambda

A. Urquhart (✉) · C. P. Kimpton · T. J. Downes · P. Gill
Central Research and Support Establishment,
The Forensic Science Service, Birmingham B5 6QQ, UK

Table 1 PCR primers used

Locus	Primers	Reference
HUMVWFA31	5' CCCTAGTGGATGATAAGAATAATCAGTATG 3' GGACAGATGATAAATACATAGGATGGATGG	Kimpton et al. 1992, GenBank M25858
HUMTH01	5' GTGATTCCCATTGGCCTGTTCCTC 3' GTGGGCTGAAAAGCTCCCGATTAT	Polymeropoulos et al. 1991f, GenBank D00269
HUMF13A01	5' GAGGTTGCACTCCAGCCTTT 3' ATGCCATGCAGATTAGAAA	Polymeropoulos et al. 1991cm, GenBank M21986
HUMFES/FPS	5' GGGATTTCCCTATGGATTGG 3' GCGAAAGAATGAGACTACAT	Polymeropoulos et al. 1991b, GenBank X06292
HUMCD4	5' TTGGAGTCGCAAGCTGAACTAGC 3' TCATGCGTCCATGGTCCGGAGCCTGAGTGACAGAGTGAGAACC	Edwards et al. 1991, GenBank M86525
HUMPLA2A1	5' CCCACTAGGTTGTAAGCTCCATGA 3' TACTATGTGCCAGGCTCTGTCCTA	Polymeropoulos et al. 1990b, GenBank M22970
HUMFOLP23	5' ATTGTAAGACTTTTGGAGCCATTT 3' TTCAGGGAGAATGAGATGGGC	Polymeropoulos et al. 1991d, GenBank J00145
HUMCYAR04	5' CTCTGGAAAACAACTCGACCCTTC 3' TGGGTGATAGAGTCAGAGCCTGTC	Polymeropoulos et al. 1991a, GenBank M30798
HUMTFIIDA	5' GCCTATTCAGAACACCAATA 3' TGGGACGTTGACTGCTGAAC	Polymeropoulos et al. 1991e, GenBank M34960
HUMFABP	5' GTAGTATCAGTTTCATAGGGTCACC 3' TTACGCGTCTCGGACAGTATTCAGTTCGTTTC	Polymeropoulos et al. 1990a, GenBank M18079
HUMGABRB15	5' CTAGAAAGCTAGCAAGGTGGAT 3' GCTCATTAAACACTGTGTTCCT	Dean et al. 1991, GenBank M59216
HUMD21S11	5' ATATGTGAGTCAATTCCCCAAG 3' TGTATTAGTCAATGTTCTCCAG	Sharma and Litt 1992, GenBank M84567

DNA labelled with the fluorescent dye ROX, sized alleles precisely but not accurately, and in our hands sizing alleles differing by only 1 bp could not be performed without the use of an allelic ladder. Since many laboratories are now involved in STR analysis, and before STR data becomes widespread as courtroom evidence, it would be convenient if a universal system of allele designation and nomenclature were adopted. To this end, we have investigated 12 prospective human STR identification markers to ascertain convenient, easily understood and scientifically accurate methods of allele nomenclature.

Materials and methods

The 12 loci studied and their respective PCR primer sequences are shown in Table 1. All primers were derived from the published or GenBank sequences, but the HUMFABP 3′ primer was lengthened to incorporate and *MluI* restriction site, and the HUMCD4 5′ primer had a 20 bp extension, designed by Jeffreys and co-workers (Jeffreys et al. 1991), to produce allele sizes compatible with one of our multiplex STR systems (Kimpton et al. 1993). DNA was prepared from whole blood as described previously (Gill et al. 1990). Allele designations at each locus had been made previously (Kimpton et al. 1993). For sequencing, heterozygotes with allele sizes differing by at least 12, and ideally 20, base pairs, were selected, and between 8 and 22 alleles were sequenced at each locus. This was intended to be a representative sample, rather than an exhaustive survey, of alleles at each locus.

PCR amplification was performed using 10 ng of genomic DNA in a 50 µl reaction volume. Reactions included 1 × Parr-Excellence buffer (10 m*M* Tris-HCl pH 8.3, 50 m*M* KCl, 1.5 m*M* $MgCl_2$, 1% Triton X-100; Cambio Labs, Cambridge, UK), 1.25 U Taq polymerase (Perkin Elmer Cetus, Norwalk, CT, USA), 200 µ*M* each deoxyribonucleotide triphosphate and 0.5 µ*M* each of 2 primers for each locus: 35 cycles of PCR (1 min 95°C, 1 min 54°C, 1 min 72°C) were performed for the loci HUMVWFA31, HUMF13A01 and HUMFES/FPS. For the other 9 loci the annealing temperature was 60°C; other conditions were identical and 35 cycles were again performed.

PCR products were electrophoresed in agarose gels, excised and purified as described previously (Urquhart 1991). Purified PCR products were sequenced from both ends with a Taq Dye-Deoxy Terminator Cycle Sequencing kit (Applied Biosystems, Foster City, Calif., USA) using the PCR primers as sequencing primers. Sequence analysis was performed on a 373A Sequencer (Applied Biosystems, Foster City, Calif., USA) using 373 Data Collection, 373 Analysis and SeqEd software (Applied Biosystems, Foster City, Calif., USA).

Results and discussion

The consensus sequences of repeat regions at the 12 loci are shown in Figs. 1–3. The repeat unit at each locus was defined as the first in-frame repeat unit on the strand listed in the GenBank database. Where necessary, ambiguity codes were used, in accordance with the recommendations of the Nomenclature Committee of the International Union

LOCUS	
Simple	
HUMFES/FPS	$(ATTT)_{8-14}$
HUMPLA2A1	$(ATT)_{8-17}$
HUMFABP	$(ATT)_{8-15}$
Simple with non-consensus alleles	
HUMTH01	$(TCAT)_{5-11}$
173bp allele	$(TCAT)_4\underline{CAT}(TCAT)_5$
HUMFOLP23	$(AAAC)_{4-8}AAAA.AAAC$
174bp allele	$(AAAC)_{10}\ \underline{---------}$
HUMCD4	$(TTTTC)_{3-13}$
141bp allele	$(TTTTC)_3\underline{CTTTC}(TTTTC)_8$
HUMF13A01	$(GAAA)_{4-17}GAGTAAAA$
181bp allele	$(GAAA)_5\ \ \underline{------}AA$
HUMCYAR04	$AT(CTT)_2TTTTGTCTATGAATGTGCCTTTTTTGAAATCATATTTTTAAAATAT(TTTA)_{7-13}$
166bp allele	$AT(CTT)_{\underline{1}}TTTTGTCTATGAATGTGCCTTTTTTGAAATCATATTTTTAAAATAT(TTTA)_7$

Fig. 1 Simple repeat sequences. The variable repeat regions are shown along with their flanking sequence where relevant. Differences from the consensus sequence for each locus are underlined

of Biochemistry (1985). Hence, M signifies A or C, Y signifies C or T, K signifies G or T, R signifies A or G and V signifies A, C or G.

Of the 12, 8 loci (HUMFES/FPS, HUMPLA2A1, HUMFABP, HUMTH01, HUMF13A01, HUMCYAR04, HUMFOLP23 [formerly HUMDHFRP2] and HUMCD4) were classified as simple repeats (Fig. 1), although HUMTH01, HUMF13A01 and HUMCYAR04 each had one common allele which differed from the consensus, while in HUMFOLP23 and HUMCD4 there was variation in individual repeats units. For HUMF13A01 and HUMCYAR04 the non-consensus allele was the smallest allele found, and each involved a deletion outside the repeat region (Fig. 1). The deletion in HUMCYAR04 is 1 of 2 CTT trinucleotides 51 bp 5′ to the repeat region, while that in HUMF13A01 was a GAGTAA hexanucleotide immediately 3′ to the repeat. This deletion occurred in all 6 of the 181 bp alleles sequenced. At the HUMTH01 locus, the largest common alleles, although a 178 bp allele is found very occasionally (Edwards et al. 1992; our unpublished data), are sized at 173–174 bp. We have sequenced 11 alleles sized at 173/174 bp, of which 8 were 173 bp and 3 were 174 bp. All the 173 bp alleles had the same single-base deletion of a thymidine residue in the fifth of 10 TCAT repeats. These observations have recently been reported elsewhere (Puers et al. 1993).

The HUMFOLP23 locus has been called a simple repeat in this study, although most alleles contain the octamer AAAA.AAAC following the run of AAAC repeats. Nine alleles (including one 174 bp allele, the largest allele size found at this locus) had 4–8 AAAC repeats followed by AAAA.AAAC. One 174 bp allele had 10 AAAC repeats without the following 8 bp. Since the two 174 bp alleles are indistinguishable by band-sizing methods, we decided to designate both as alleles with 10 AAAM repeats. It is not known whether smaller alleles consisting solely of AAAC repeats exist. Similarly, the HUMCD4 locus has been called a simple repeat although both the GenBank sequence (Edwards et al. GenBank M86525) and one allele (out of 11) that we sequenced had CTTTC as the fourth repeat instead of the consensus TTTTC. Alleles were designated as YTTTC repeats.

Three loci (HUMGABRB15, HUMTFIIDA and HUMVWFA31) were classified as compound repeats (Fig. 2).

LOCUS	
Compound	
HUMGABRB15	$(GATA)_{5-12}(GATC)_{2-4}(TATC)_{1-2}$
HUMTFIIDA	$(CAG)_3(CAA)_3(CAG)_{9-11}CAA(CAG.CAA)_{0-1}(CAG)_{9-24}CAA.CAG$
Compound with non-consensus alleles	
HUMVWFA31	$(ATCT)_2(GTCT)_{3-4}(ATCT)_{9-13}$
144bp allele	$(ATCT)_2(GTCT)_4\ \ (ATCT)_5\underline{AT}(ATCT)_4$

Fig. 2 Compound repeat sequences. The variable repeat regions are shown. Differences from the consensus sequence for the HUMVWFA31 locus are underlined

	x		y
GenBank	$\texttt{(TCTA)}_4\ \texttt{(TCTG)}_6\ \texttt{(TCTA)}_3\texttt{TA(TCTA)}_3$	$\texttt{TCA}$	$\texttt{(TCTA)}_2\texttt{------(TCTA)}_8\texttt{-----------TC}$
Odd	$\texttt{(TCTA)}_{4-6}\texttt{(TCTG)}_{5-6}\texttt{(TCTA)}_3\texttt{TA(TCTA)}_3$	$\texttt{TCA}$	$\texttt{(TCTA)}_2\underline{\texttt{TCCATA}}\texttt{(TCTA)}_{8-16}\texttt{--------TC}$
Even	$\texttt{(TCTA)}_{4-6}\texttt{(TCTG)}_{5-6}\texttt{(TCTA)}_3\texttt{TA(TCTA)}_3$	$\texttt{TCA}$	$\texttt{(TCTA)}_2\underline{\texttt{TCCATA}}\texttt{(TCTA)}_{8-16}\underline{\underline{\texttt{TA.TCTA.}}}\texttt{TC}$

GenBank	$\texttt{GTCTATCTATCCAGTCTATCTACNTCCTATNNAG}$
Odd	$\texttt{GTCTATCTATCCAGTCTATCTAC}\underline{\texttt{C}}\texttt{TCCTAT}\underline{\texttt{T-}}\texttt{AG}$
Even	$\texttt{GTCTATCTATCCAGTCTATCTAC}\underline{\texttt{C}}\texttt{TCCTAT}\underline{\texttt{T-}}\texttt{AG}$

Fig. 3 Complex repeat sequence at the D21S11 locus. The variable repeat region is shown along with 33 bp of 3′ sequence for the GenBank sequence (Sharma and Litt 1992) and our consensus sequences for the odd-numbered and even-numbered alleles. The invariant hexanucleotide which we found in all alleles and the 2 amendments we made to the GenBank sequence are singly underlined. The hexanucleotide found in all even-numbered alleles is doubly underlined. The segments marded x and y are those used for allele designation (see text)

The HUMGABRB15 locus consists of blocks of GATA, GATC and TATC repeats, all of which vary in number between alleles. The aggregate number of the 3 repeat types (i.e. the number of KATM repeats) was used for allele designation. Similarly, the HUMTFIIDA locus contains 7 or 9 blocks of CAG or CAA repeats, and allele designation was for CAR.

Apart from one allele in one individual, HUMVWFA31 was a straightforward compound repeat with the sequence $(ATCT)_2(GTCT)_m(ATCT)_n$, and allele designation is for RTCT. However, one non-consensus allele of 144 bp was observed in which the 3′ATCT tract contained an AT dinucleotide (Fig. 2). This could have arisen either by deletion of TC from a 146 bp 16 allele or by duplication of TA in a 142 bp 15 allele. This allele was the only HUMVWFA31 allele seen which differed from the 4 bp repeat pattern in over 1500 alleles sized at the locus (unpublished data).

The repeat at the D21S11 locus was classified as a complex repeat. Although originally reported to have 4 bp increments between alleles (Sharma and Litt 1992), later work (Kimpton et al. 1993) revealed that alleles differing by 2 bp were common. The consensus sequences of the repeat at the D21S11 locus are shown in Fig. 3, aligned with the sequence from GenBank (Accession number M84567; Sharma and Litt 1992). A total of 16 alleles was sequenced, including the largest and smallest, and 2 pairs of identically sized alleles, one pair or which had identical sequence. The variable sequence consisted largely of TCTA and TCTG repeat blocks, although an invariant TA dinucleotide and an invariant TCA trinucleotide were also present. Alleles which were given even-numbered allele designations (see ‘Allele designation and nomenclature’ below) had a TATCTA hexanucleotide after the final block of TCTA repeats. Alternatively, this can be viewed as a TA insertion before the last TCTA repeat. In all 16 alleles sequenced, a TCCATA hexanucleotide was found which does not appear in the GenBank sequence (Sharma and Litt 1992). If the absence of this hexanucleotide genuinely occurs, this is a further mechanism for 2 bp allele differences. Sequencing the D21S11 locus also allowed us to make 2 amendments to the GenBank sequence. In all 16 alleles sequenced, the N at position 288 in the GenBank sequence was a C, and the NN at positions 295-6 was a single T residue (Fig. 3).

Allele designation and nomenclature

Designation and nomenclature of alleles at STR and VNTR loci has been fairly haphazard in the past. For presentation of forensic evidence in the courts, and for transfer of data between different laboratories, some standardisation of allele designation is necessary. The most widely applicable method would be to call each allele by its length in basepairs. This method would be suitable for VNTRs, normal STRs and hypervariable STRs such as HUMACTBP2 (Urquhart et al. 1993) and D11S554 (Adams et al. 1993). However, the allele size is dependent on the primers used, and requires a precise and accurate sizing method. An alternative is to call alleles by the number of repeat units they contain. This is easy for simple repeats and some VNTRs, and can be applied to compound repeats with the use of ambiguity codes, but is too cumbersome for complex repeats. Problems also arise when intermediate alleles occur, as with the HUMVWFA31 144 bp allele in this study and the various anodic and cathodic allele variants in some VNTR systems. Both allele designation methods discussed above may involve loss of informativeness, since the repeat pattern at any individual allele is not specified. However, this is inevitable with all methods which distinguish alleles solely by size, and the increase in informativeness gained by sequencing every allele is more than offset by the increased cost. In line with the recommendations of the DNA Commission of the International Society for Forensic Haemogenetics (1992), we have called alleles at all simple and compound repeat loci by their repeat number, using redundancy codes for compound repeats. For intermediate alleles and other alleles that fail to align with the incremental ‘ladder’ at each locus, digits after a decimal point were used to indicate the number of basepairs by which the allele exceeded the previous ‘rung’ of the ladder. Thus, the 144 bp allele at the HUMVWFA31 locus was designated 15.2, the 166 bp allele at the HUMCYAR04 locus was designated 6.1, the 173 bp allele at the HUMTH01 locus was designated 9.3, and the 181 bp allele at the HUMF13A01 locus was designated 3.2. It should be noted that the use of the number after the decimal point does not necessarily imply the presence of a partial repeat (cf. HUMF13A01, HUMCYAR04), but many indicate variation outside the repeat region.

Table 2 Characteristics of $(TCTA)_5$-containing human loci in GenBank

Locus	Accession number	Reference	Sequence motif			
			Alternative repeat (longest block)		Shortened repeat[a] (number present)	
			TCTG	TCCA	TCA	TA
D21S11	M84567	This study	5–6	1	1	2/3
HUMVWFA31(A)[b]	M25858	This study	3–4	2	–	0/1
HUMVWFA31(B)[b]	"	Mancuso et al. 1989	2	–	1	–
HUMVWFA31(C)[b]	"	Mancuso et al. 1989	1	–	2	–
HUMGABRB15	M59216	This study	–	–	–	–
T03428	T03428	Khan et al.1992	–	–	–	–
D18S53	Z16461	Weissenbach et al. 1992	–	1	–	–
D2S121	Z16545	Weissenbach et al. 1992	–	–	–	–
D7S513	Z16989	Weissenbach et al. 1992	1	–	–	–
D10S225	Z17156	Weissenbach et al. 1992	1	2	1	1
HS7SKP41	X04237	Murphy et al. 1984	1	–	–	–
HSCSF1PO	X14720	Hampe et al. 1989	–	–	–	–
HSIGK6	Z00004	Jaenichen et al. 1984	–	–	–	–
HUMBKM	M35828	Erickson et al. 1988	1	–	–	2
HUMCD4(A)[b]	M86525	Edwards et al. GenBank	–	–	1	–
HUMCD4(B)[b]	"	Edwards et al. GenBank	1	–	–	1
HUMCD4(D)[b]	"	Edwards et al. GenBank	2	2	1	–
HUMHPRTB	M26434	Ansorge et al. 1990	1	–	–	–
HUMRPTPOLF	L02066	Weber and May 1989	–	–	1	–
HUMSIRPOBD	M87698	Hudson et al. 1992	–	–	–	–
HUMSIRPOCP	M87736	Hudson et al. 1992	1	–	1	–
HUMPGK1	S75476	Riley et al. 1991	1	–	–	–
HUMPAH	L10305	Goltsov et al. 1993	1	–	3	–
DXS981	M38419	Mahtani and Willard 1993	1	–	0/1	–
D12S66	–	Roewer et al. 1992	1	?[c]	–	–
D12S67	–	Roewer et al. 1992	6	?[c]	–	–
DYS19	–	Roewer et al. 1992	1	?[c]	–	1

[a] TCA trinucleotides and TA dinucleotides were only included if they had at least one TCTA, TCTG or TCCA tetranucleotide repeat immediately adjacent on each side

[b] HUMVWFA31 A region: bases 1683–1770; B region: bases 1889–2063; C region: bases 2084–2343. HUMCD4 A region: bases 5624–5686; B region: bases 5944–6043; D region: bases 7101–7340

[c] Full sequence not given in reference

Allele designation for the complex repeat at D21S11 is more problematical. Each allele contains a mixture of di-, tri-, tetra- and hexanucleotides (Fig. 3). Three options were considered: naming alleles by length in base pairs, using an arbitrary system of allele designation, and naming alleles by the number of TV dinucleotide repeats. We decided on the third option largely because it was consistent with nomenclature at the other loci in this system. However, there were 2 minor inconsistencies. The system of nomenclature excludes the invariant TCA trinucleotide, and treats the CA in the centre of the TCCATA hexanucleotide as a TV dinucleotide. The allele designation at the D21S11 locus thus indicates the aggregate number of TV dinucleotides (plus one CA dinucleotide) in the two regions labelled x and y in Fig. 3.

TCTA and related repeats

We noted several similarities between the sequences of D21S11 and HUMVWFA31. When written as TCTA repeats, both contain compound $(TCTG)_p(TCTA)_q$ regions and the sequence motif $(TCTA)_rTA(TCTA)_s$ appears (either once or twice) in D21S11 alleles and in the 144 bp non-consensus HUMVWFA31 allele. A search of GenBank human sequences for $(TCTA)_5$ and its complement $(TAGA)_5$ produced 22 matches, including D21S11, HUMGABRB15 and 3 regions of HUMVWFA31 (Table 2). These sequences, plus 5 others recently published, DXS981 (Mahtani and Willard 1993), phenylalanine hydroxylase (Goltsov et al. 1993), D12S66, D12S67 and DYS19 (Roewer et al. 1992), were examined for sequence motifs common to several sequences. These are summarised in Table 2. Many of the TCTA repeat blocks had single repeat units that differed by one base from the consensus repeat, the commonest being TCTG and TCCA (occurring in 67% and 21% of sequences respectively). At some loci, notably D21S11, HUMVWFA31 and D12S67, longer blocks of TCTG were present. Truncated repeats were also present at some loci. The commonest trinucleotide, TCA, occurred at 37% of the loci, while the dinucleotide TA was present in 22% of loci. The frequency of these truncated repeats may well be an underestimate, since many of the reported TCTA repeats were discovered by hybridisation to $(TAGA)_n$ or similar oligonucleotides, the binding of which would be decreased by imperfect repeats.

Searches were also performed for $(TCTG)_5$ and $(TCCA)_5$, but both failed to find sequences associated with extensive TCTA repeats. However, the TCCA repeat in the HUMIGCAAA locus (Yu et al. 1990) included 2 separate TCA trinucleotides. Interestingly, the non-consensus 9.3 allele at the HUMTH01 locus (see above) can be regarded as a TCA trinucleotide in the middle of a $(TTCA)_n$ tract.

The sequence at the HUMGABRB15 locus (Fig. 2) can be written as $(TAGA)_{5-12}(TCGA)_{1-3}(TCTA)_{1-2}$, i.e. 2 mutually palindromic TCTA tracts surrounding a TCGA tract. It is possible that the central tract is formed by limited gene conversion, the 2 TCTA tracts in opposite orientation acting on each other.

These observations suggest a scenario for evolution of TCTA repeats. Since the TCA trinucleotide could not be generated by duplication from $(TCTA)_n$, we would suggest that this unit originates by deletion of a thymidine residue from TCTA. By analogy, TA dinucleotides may arise by deletion of C from TCA or, alternatively, by deletion of TC or CT from TCTA. Of course, they could also arise by TA duplication. In a simple TCTA repeat (e.g. D2S121) individual TCTA units may mutate to TCTG (or TCCA), giving imperfect but essentially simple repeats (e.g. D7S513). These imperfections may then expand, either by genuine size expansion or by gene conversion (Jackson and Fink 1981; Slightom et al. 1988), producing a compound repeat such as HUMVWFA31. TCTA (and TCCA) units may also undergo deletion to TCA or TA, producing complex repeats such as D21S11. Variation at the D21S11 locus suggests that repeat expansion can continue after event such as TCA generation. These events would lead to degeneration of simple TCTA repeats with time into complex repeats, containing TCTA, TCTG and TCCA blocks interspersed with dinucleotide and trinucleotide truncated repeats. Indeed, some degnerate TCTA repeats such as those at the HUMVWFA31 and HUMCD4 loci (Mancuso et al. 1989; Edwards et al. Gen Bank M86525) can reach thousands of bp in length.

Other repeats

As discussed above, the most common mutation in TCTA repeats is to TCTG, i.e. and A > G transition. The HUMTFIIDA and HUMCD4 loci appear to have developed by similar events, respectively by CAG > CAA and AAAAG > GAAAG. Only at the HUMFOLP23 locus is there an apparent transversion (A > C); in this case, the mutational event could also be a deletion. The predominance of transitions is also seen at AAAG repeats (Urquhart et al. 1993; Adams et al. 1993) where the usual nonconsensus repeat is AAGG.

Non-consensus alleles

Of the 12 loci studied, 6 showed non-consensus alleles, and 2 of these, HUMFOLP23 and HUMCD4, only differed from other alleles in sequence and could not therefore be distinguished by sizing. Nonconsensus alleles showed a distinct tendency towards the ends of allele size ranges. That at HUMCD4 was the second largest allele, while at HUMTH01 the 9.3 allele was 1 bp smaller than the largest common allele. The HUMFOLP23 non-consensus allele was the same size as the largest alleles found, and those at HUMF13A01 and HUMCYAR04 are the smallest alleles found at the loci. Only the extremely rare HUMVWFA31 15.2 allele falls towards the middle of the allele size range.

It is possible that mutation to non-consensus allele is a mechanism which prevents both extreme expansion to high repeat numbers and extreme contraction to low-number, non-polymorphic, repeats. For the 2 loci where the non-consensus allele is at the low end of the size range, HUMF13A01 and HUMCYAR04, deletions outside the repeat sequence cause the non-consensus allele, while those at the top end of their range are caused by either deletion within a repeat (HUMTH01) or substitution within a repeat (HUMFOLP23 and HUMCD4).

Alleles at the D21S11 locus show a bimodal distribution, with odd-numbered and even-numbered alleles showing distributions over different size ranges (Fig. 4). Presumably this is due to the effect of the TA dinucleotide on increase or decrease of repeat number over time.

Alleles at other polymorphic STR loci require investigation to determine the extent to which sequence effect evolution at these loci.

Implications for forensic use

We have surveyed 12 STR loci to investigate candidate loci for an STR system for forensic identification. The major considerations for selection of loci were discriminating power, absence of linkage, agreement with Hardy-Weinberg equilibrium, low levels of 'shadow bands' (Hauge and Litt 1993), compatibility with other loci (for a multiplex STR system) and accurate sizing of alleles. Where alleles differ by 2 bp or more, sizing using the Pst I digest of bacteriophage lambda as a marker consistently distinguished alleles, but alleles differing by 1 bp required sizing by an allelic ladder. Hence, the non-consensus alleles at HUMF13A01 and HUMVWFA31 were sized and designated accurately. However, at the HUMTHO1 locus, the 9.3 and 10 alleles were treated as a single pooled allele group, since it was not possible to consistently distinguish

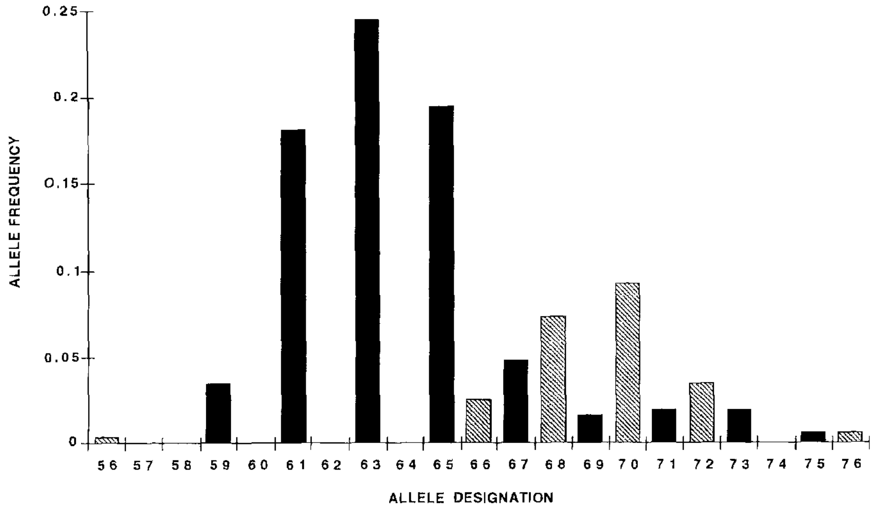

Fig. 4 Allele distribution at the D21S11 locus. Odd-numbered alleles are solid, even-numbered are hatched, showing two overlapping normal distributions. Data are pooled from 157 British individuals (i.e. 314 chromosomes) or various racial origin

Table 3 Allele size ranges and non-consensus alleles at the 12 STR loci. Sizes are as determined by sequencing the alleles

Locus	Repeat	Consensus alleles		Non-consensus allele
		Smallest allele	Largest allele	
HUMFES/fPS	ATTT	8 (211 bp)	14 (235 bp)	–
HUMPLA2A1	ATT	8 (110 bp)	17 (137 bp)	–
HUMFABP	ATT	8 (213 bp)	15 (234 bp)	–
HUMTH01	TCAT	5 (154 bp)	11 (178 bp)	9.3 (173 bp)
HUMF13A01	GAAA	4 (183 bp)	17 (235 bp)	3.2 (181 bp)
HUMCYAR04	TTTA	7 (169 bp)	13 (193 bp)	6.1 (166 bp)
HUMFOLP23	AAAM	6 (158 bp)	10 (174 bp)	10 (174 bp)
HUMCD4	YTTTC	3 (96 bp)	13 (146 bp)	12 (141 bp)
HUMGABRB15	KATM	9 (124 bp)	15 (148 bp)	–
HUMTFIIDA	CAR	27 (168 bp)	40 (207 bp)	–
HUMVWFA31	RTCT	12 (130 bp)	21 (166 bp)	15.2 (144 bp)
D21S11	TV[a]	56 (209 bp)	76 (249 bp)	–

[a] Excluding TCA trinucleotide (see text)

between them. This led to a slight loss of informativity at this locus.

The sequence data presented here are of less relevance to non-fluorescent STR analysis in which allele designation is by comparison with an allelic ladder. However, the allele designations suggested here are relevant whichever method of analysis is used. Small (1 or 2 bp) differences in allele sizes can cause problems using non-fluorescent detection methods, particularly where there is an appreciable difference in motility between denatured DNA strands.

In the future, an ideal multiplex STR system would consist of loci in which alleles differ by a minimum of 2 bp. The presence of non-consensus alleles does not rule out loci for inclusion as forensic identification markers, but size differences between alleles of 1 bp require very precise sizing. With the use of allelic ladders, more discriminating hypervariable loci such as HUMACTBP2 (Urquhart et al. 1993) and D11S554 (Adams et al. 1993) could be used. D21S11, though complex in sequence, can be sized in our system and, as a highly discriminating locus, would be a useful component of a multiplex STR system.

From the loci investigated in this study, we have developed a quadruplex STR system including the loci HUMFES/FPS, HUMTH01, HUMF13A01, and HUMVWFA31 (Kimpton et al. 1993, 1994), and further, more discriminating, systems are under investigation.

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 Mol Genet 2: 1373–1376

Ansorge W, Caskey CT, Erfle H, Zimmermann J, Schwager C, Stegemann J, Civitello A, et al. (1990) Automated DNA sequencing of the human HPRT locus. Genomics 6: 593–608

Dean M, Lucas-Derse S, Bolos A, O'Brien SJ, Kirkness EF, Fraser CM, Goldmann D (1991) Genetic mapping of the b1 GABA receptor gene to human chromosome 4, using a tetranucleotide repeat polymorphism. Am J Hum Genet 49: 621–626

DNA Commission of the International Society for Forensic Haemogenetics (1992) DNA recommendations – 1992 report concerning recommendations of the DNA Commission of the International Society for Forensic Haemogenetics relating to the use of PCR-based polymorphisms. Int J Leg Med 105: 63–64

Edwards MC, Clemens PR, Tristan M, Pizzuti A, Gibbs RA (1991) Pentanucleotide repeat length polymorphism at the human CD4 locus. Nucleic Acids Res 19:4791

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12: 241–253

Erickson RP, Ross CE, Gorski JL, Stalvey JR, Drumm MM (1988) Bkm sequences from the human X chromosome certain large clusters of GATA/GACA repeats. Ann Hum Genet 52:167–176

Fregeau CJ, Fourney RM (1993) DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification. Bio Techniques 15:100–109

Gill P, Sullivan KM, Werrett DJ (1990) The analysis of hypervariable DNA profiles: problems associated with the objective determination of the probability of a match. Hum Genet 85:75–79

Goltsov AA, Eisensmith RC, Naughton EF, Jin L, Chakraborty R, Woo SLC (1993) A single polymorphic STR system in the human phenylalanine hydroxylase gene permits rapid prenatal diagnosis and carrier screening for phenylketonuria. Hum Mol Genet 2:577–581

Hampe A, Shamoon BM, Gobet M, Sherr C, Galibert F (1989) Nucleotide sequence and structural organisation of the human FMS proto-oncogene. Oncogene Res 4:9–17

Hauge XY, Litt M (1993) A study of the origin of 'shadow' bands seen when typing dinucleotide repeat polymorphisms by the PCR. Hum Mol Genet 2:411–415

Hundson TJ, Engelstein M, Lee MK, Ho EC, Rubenfield MJ, Adams CP, Housman DE, et al. (1992) Isolation and chromosomal assignment of 100 highly informative human simple sequence repeat polymorphisms. Genomics 113:622–629

Jackson JA, Fink GR (1981) Gene conversion between duplicated genetic elements in yeast. Nature 292:306–311

Jaenichen HR, Pech M, Lindenmaier W, Wildgruber N, Zachau HG (1984) Composite human V kappa genes and a model of their evolution. Nucleic Acids Res 12:5249–5263

Jeffreys AJ, Macleod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204–209

Khan AS, Wilcox AS, Polymeropoulos MH, Hopkins JA, Stevens TJ, Robinson M, Orpana AK, et al. (1992) Single pass sequencing and physical and genetic mapping of human cDNAs. Nature Genet 2:180–185

Kimpton CP, Walton A, Gill P (1992) A further tetranucleotide repeat polymorphism in the vWF gene. Hum Mol Genet 1:287

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 Applic 3: 13–22

Kimpton CP, Fisher D, Watson S, Adams M, Urquhart A, Lygo J, Gill P (1994) Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. Int J Leg Med (in press)

Mahtani MM, Willard HF (1993) A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci. Hum Mol Genet 2:431–437

Mancuso DJ, Tuley EA, Westfield LA, Worrall NK, Shelton-Inloes BB, Sorace JM, Alevy YG, et al (1989) Structure of the gene for human von Willebrand factor. J Biol Chem 264:19514–19527

Murphy S, Altruda F, Ullu E, Tripodi M, Silengo L, Melli M (1984) DNA sequence complementary to human 7 SK RNA show structural similarities to the short mobile elements of the mammalian genome. J Mol Biol 177:575–590

Nomenclature Committee of the International Union of Biochemistry (1985) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Eur J Biochem 150:1–5

Phromchotikul T, Browne D, Litt M (1992) Microsatellite polymorphisms at the D11S554 and D11S569 loci. Hum Mol Genet 1:214

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1990a) Trinucleotide repeat polymorphism at the human intestinal fatty acid binding protein gene (FABP2). Nucleic Acids Res 18:7198

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1990b) Trinucleotide repeat polymorphism at the human pancreatic phospholipase A-2 gene (PLA2). Nucleic Acids Res 18:7468

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1991a) Tetranucleotide repeat polymorphism at the human aromatase cytochrome P-450 gene (CYP19). Nucleic Acids Res 19:195

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1991b) Tetranucleotide repeat polymorphism at the human c-fes/fps proto-oncogene (FES). Nucleic Acids Res 19:4018

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1991c) Tetranucleotide repeat polymorphism at the human coagulation factor XIII A subunit gene (F13A1). Nucleic Acids Res 19: 4036

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1991d) Tetranucleotide repeat polymorphism at the human dihydrofolate reductase psi-2 pseudogene (DHFRP2). Nucleic Acids Res 19:4792

Polymeropoulos MH, Rath DS, Xiao H, Merril CR (1991e) Trinucleotide repeat polymorphism at the human transcription factor IID gene. Nucleic Acids Res 19:4037

Polymeropoulos MH, Xiao H, Rath DS, Merril CR (1991f) Tetranucleotide repeat polymorphism at the human tyrosine hydrolase gene (TH). Nucleic Acids Res 19:3753

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

Riley DE, Goldman MA, Gartler SM (1991) Nucleotide sequence of the 3′ nuclease-sensitive region of the human phosphoglycerate kinase (PGD1) gene. Genomics 11:212–214

Roewer L, Arnemann J, Spurr NK, Grzeschik K-H, Epplen JT (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet 89:389–394

Sharma V, Litt M (1992) Tetranucleotide repeat polymorphism at the D21S11 locus. Hum Mol Genet 1:67

Slightom JL, Koop BF, Xu P, Goodman M (1988) Rhesus fetal globin genes. J Biol Chem 263:12427–12438

Urquhart AJ (1991) Human tyrosinase-like protein (TYRL) carboxy terminus: closer homology with the mouse protein than previously reported. Nucleic Acids Res 19:5803

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

Warne D, Watkins C, Bodfish P, Nyberg K, Spurr N (1991) Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene 2 (ACTBP2) detected using the polymerase chain reaction. Nucleic Acids Res 19:6980

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, et al. (1992) A second-generation linkage map of the human genome. Nature 359:794–801

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

Yu L, Peng C, Starnes SM, Liou RS, Chang T-W (1990) Two isoforms of human membrane-bound alpha Ig resulting from alternative mRNA splicing in the membrane segment. J Immunol 145:3932–3936