Development of Highly Polymorphic Pentanucleotide Tandem Repeat Loci with Low Stutter

By Jeff Bacher, Ph.D., and James W. Schumm, Ph.D.
Promega Corporation

INTRODUCTION

All eukaryotic genomes contain regions of simple repetitive DNA, called short tandem repeats (STR*) or microsatellites, which consist of tandem repeats of a small number of bases (1-3). The number of repeats at a STR locus can be highly variable among individuals, resulting in length polymorphisms that can be detected by relatively simple PCR-based assays. Thousands of these highly informative STR loci have been identified in the human genome. Consequently, STRs provide an abundant class of polymorphic markers that have gained popularity for use in genetic linkage analysis, mapping, identity testing and many other applications (3-6).

The combined properties of high power of discrimination, few microvariants and low levels of stutter make the pentanucleotide repeat STR loci ideal markers for forensic analysis.

Despite all the positive characteristics of STRs, two significant drawbacks complicate their use in forensic analysis. First, stutter artifacts are often seen following amplification of STR loci (7,8). These stutter products are minor fragments that differ in size from the major allele by multiples of the core repeat. The amount of stutter observed for STR loci tends to be inversely correlated with the length of the core repeat unit. Thus, stutter is most severely displayed with mono- and dinucleotide repeat loci, to a lesser extent with tri- and tetranucleotide repeat markers, and is nearly undetectable in much longer tandem repeats found in VNTR loci (2,3,8,9). The presence of stutter artifacts, presumed to result from a DNA polymerase slippage event during DNA replication (7,10), complicates the unambiguous assignment of alleles and automation of the genotyping procedure. Stutter products can be particularly problematic in mixed DNA samples since they are the same size as the actual alleles (e.g., generally 4bp smaller than the actual allele in tetranucleotide repeat loci). It is not always possible to distinguish a faint band in a mixed sample as a real allele or stutter if its position is four bases shorter than the more prominent allele band.

The second major drawback of current STR marker systems relates to the difficulty in resolving 2-4bp differences in larger DNA fragments, limiting the size of PCR products that can be analyzed easily and the number of loci that can be multiplexed. Difficulty in separation of larger DNA fragments is due to spatial compression in the upper regions of gels or diffusion during separation in other matrices. Alleles that differ by increments larger than 4bp extend the useful fragment sizes that can be separated, allowing resolution of larger DNA fragments and/or multiplexing of extra loci.

To overcome limitations in the current STR systems, we set out to identify and characterize a class of polymorphic markers that contained tandem repeat units greater than 4bp, but were small enough to allow PCR amplification and multiplexing. We predicted that STRs with larger core repeats could be separated and detected more easily and precisely, would allow multiplexing of more STR markers and exhibit minimal stutter.

STRATEGIES TO ISOLATE NEW STR MARKERS

Two strategies were used to isolate new STR markers with longer core repeat units. STR loci with 5-9bp long core repeats were identified, first, by searching human databases and, second, by screening of small-insert genomic libraries enriched for commonly occurring pentanucleotide repeat motifs.

GenBank® database searches offered the advantage that sequences surrounding the STR loci were immediately available, allowing rapid development of PCR-based analysis. However, more than 20,000 unique 5-9bp repeat motifs exist, making this approach tedious and time-consuming unless automated. Our searches of the GenBank® database revealed that the frequency

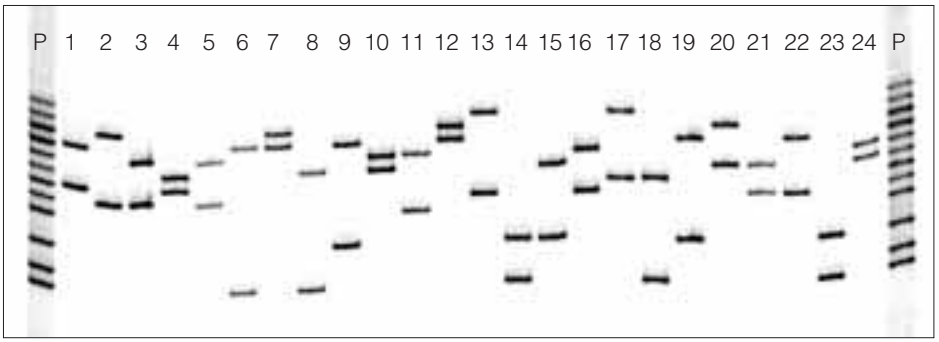

Figure 1. FMBIO® image of pentanucleotide short tandem repeat locus D. Twenty-four DNA samples from African-American individuals were amplified using fluorescein-labeled primers; the PCR products were separated on 4% denaturing polyacrylamide gels and visualized by scanning on an FMBIO® II fluorescent scanner. The first and last lanes include pooled (P) DNA samples.

of occurrence of STR loci was inversely proportional to repeat unit length and number of tandem repeats. Over 200 loci containing 5-9bp short tandem repeats (with ≥5 perfect tandem repeats) were identified in GenBank® searches and evaluated for polymorphism levels. Based on the number of polymorphic loci identified in databases, we estimate that over 1,000 STRs with 5-9bp repeat units exist in the human genome, most being pentanucleotides. This translates into about 1 every 3 million base pairs in the human genome compared to about 1 in 20,000bp for all tri- and tetranucleotide repeats (11). Despite the efficiency of evaluating STR loci for which sequence data is available, this approach alone did not produce a large number of polymorphic STR loci.

A second approach we used to isolate pentanucleotide repeat loci relied on screening of genomic libraries for the presence of clones containing targeted repeats. To increase recovery rates of pentanucleotide STR loci from genomic libraries, an enrichment strategy was used (12). Briefly, size-selected (250-600bp) human *Mbo* I fragments were ligated to linkers to give a whole-genome library. These *Mbo* I fragments were amplified using primers complementary to the linkers and the PCR products were denatured and hybridized to small nylon filters, each containing complementary pentanucleotide repeat sequences. The filters were washed to remove unbound DNA, and fragments containing targeted repeats were recovered. The hybridization-selected DNA was reamplified, cloned into the pGEM®-3Z Vector**, and transformed into JM109 Competent Cells. Next, colony hybridization was performed and clones containing targeted repeats were detected with the aid of alkaline phosphatase-labeled probes. All colonies that were positive in these colony hybridizations were reassayed to confirm results.

PENTANUCLEOTIDE REPEATS

Recombinants selected by colony hybridization were sequenced and inspected for the presence of tandem repeats. Unique sequences not found in the GenBank® database and having a minimum of five perfect tandem repeats were chosen and primers designed for amplification of the locus to determine polymorphism levels. The initial screen for polymorphism was performed using pooled DNA samples from 15 individuals to get a rough estimate of the number of alleles that exists for each locus. Pentanucleotide loci displaying four or more alleles were tested with individual DNAs to determine preliminary heterozygosity values (Figure 1).

Over 50 of the polymorphic pentanucleotide STR loci were mapped to determine their chromosomal location (see Table 1). This was achieved using both radiation hybrid mapping techniques to determine physical map location and by standard meiotic linkage mapping techniques using the CEPH kindred reference panel and the CRIMAP multipoint linkage program (13).

For markers to have value in forensic analysis and paternity determination, they must display a significant degree of polymorphism within each major racial/ethnic group. Development of population statistics for the new pentanucleotide markers is focused on several of the more informative pentanucleotide repeat markers (i.e., Penta B, C, D and E). This work is being done by Promega Corporation in collaboration with Bode Technologies (Springfield, VA) and involves genotyping and analysis of over 200 individuals of Caucasian-American, African-American, Hispanic-American and Asian-American descent for each locus. A summary of the preliminary results for this study is shown in Table 1. The analysis is not completed at this time, but preliminary results show that information content and polymorphism levels of some of the new pentanucleotide loci are comparable to or exceed those of the STR loci commonly used in the forensic analysis and paternity testing communities.

Unfortunately, STR loci that are highly polymorphic often have correspondingly high numbers of undesirable microvariant alleles (14,15). The presence of microvariant alleles (alleles differing from one another by lengths other than the repeat length) complicates separation, interpretation and assignment of alleles. However, genotyping of over 800 individuals in the Promega/Bode Technologies population study revealed few or no microvariant alleles at several highly polymorphic pentanucleotide STR loci (e.g., Penta B, C, D and E).

To evaluate whether pentanucleotide STR loci display lower stutter, selected loci (Penta A-E) were amplified, run on an ABI PRISM® 377 DNA sequencer, and the peak heights of main allele bands and stutter bands were determined. Figure 2 shows the average and range of percent stutter observed for pentanucleotide STR loci A through E and the 13 CODIS (Combined DNA Index System) tetranucleotide STR loci. With the exception of TH01, all five of the pentanucleotides tested revealed stutter rates below commonly used tetranucleotide tandem repeat loci, supporting the hypothesis that STRs with larger core repeat units have less stutter than those with smaller repeat units. However, it is important to realize that repeat unit size is not the only factor affecting the amount of stutter found at a particular locus. Other major factors affecting the level of stutter are the total length of the repeat region (larger alleles tend to display more stutter) and whether the repeat is interrupted (STR loci

Table 1. Preliminary population statistics and chromosomal location of loci Penta A though G†.

Locus	Population	Probability of Match	Power of Discrimination	Typical PI	Power of Exclusion	Total # Analyzed	Fraction Heterozygotes	Chromosome Location	Allele Range††
A	African-American	0.07	0.93	2.10	0.53	42	0.76	8p	7-18
	Asian-American	0.22	0.78	2.08	0.53	25	0.76		
	Caucasian-American	0.09	0.91	1.39	0.34	39	0.64		
	Hispanic-American	nd	nd	nd	nd	nd	nd		
B	African-American	0.02	0.98	1.89	0.49	208	0.74	7q	5-31
	Asian-American	0.07	0.93	3.11	0.67	205	0.84		
	Caucasian-American	0.03	0.97	3.40	0.70	211	0.85		
	Hispanic-American	0.03	0.97	2.43	0.59	209	0.79		
C	African-American	0.06	0.94	1.85	0.48	207	0.73	9p	3-15
	Asian-American	0.10	0.90	2.14	0.54	205	0.77		
	Caucasian-American	0.10	0.90	1.98	0.51	210	0.75		
	Hispanic-American	0.08	0.92	2.54	0.60	208	0.80		
D	African-American	0.03	0.97	4.18	0.76	209	0.88	21q	2.2-17
	Asian-American	0.06	0.94	2.03	0.52	207	0.75		
	Caucasian-American	0.06	0.94	3.64	0.72	211	0.86		
	Hispanic-American	0.05	0.95	3.21	0.68	212	0.84		
E	African-American	0.02	0.98	4.50	0.77	207	0.89	15q	5-24
	Asian-American	0.02	0.98	5.18	0.80	207	0.90		
	Caucasian-American	0.03	0.97	4.22	0.76	211	0.88		
	Hispanic-American	0.02	0.98	3.00	0.66	210	0.83		
F	African-American	0.10	0.90	6.25	0.84	25	0.92	6q	5-20
	Asian-American	0.23	0.77	1.63	0.42	26	0.69		
	Caucasian-American	0.12	0.88	1.79	0.46	25	0.72		
	Hispanic-American	nd	nd	nd	nd	nd	nd		
G	African-American	0.08	0.92	3.00	0.66	24	0.83	22q	6-17
	Asian-American	0.12	0.88	1.09	0.23	24	0.54		
	Caucasian-American	0.10	0.90	1.35	0.33	35	0.63		
	Hispanic-American	nd	nd	nd	nd	nd	nd		

†Preliminary population data from collaboration between Promega Corporation and Bode Technologies (Springfield, VA).
††Allele ranges may change as more sequence data becomes available.

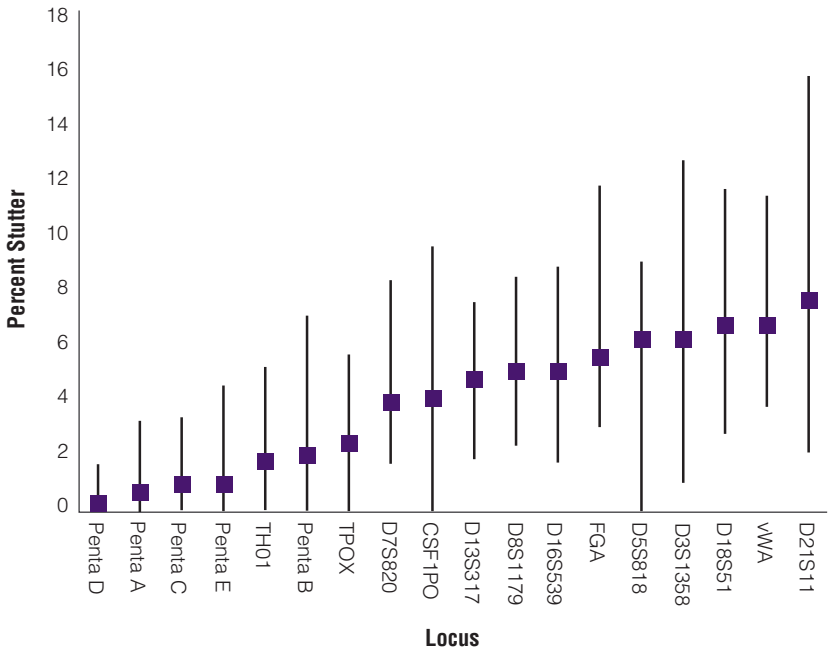

Figure 2. Percent stutter for pentanucleotide tandem repeat loci A through E and the 13 CODIS tetranucleotide repeat loci. Boxes represent the average percent stutter of the minor N-4 or N-5 band relative to the main allele for at least 20 DNA samples. The bars indicate high and low percent stutter range observed for all alleles. Analysis was performed on the ABI PRISM® 377 DNA sequencer using accurately quantified DNA samples (1ng DNA per PCR reaction) for locus-to-locus uniformity.

containing one or more repeat units with a different DNA sequence than the core repeat unit generally display less stutter; 8).

The work presented here represents the culmination of an extensive search for ideal markers for DNA forensic analysis. The combined properties of high power of discrimination, few microvariants and low levels of stutter make the pentanucleotide STR loci described here ideal markers for forensic analysis. Efforts are currently in progress to incorporate the best pentanucleotide STR loci into STR multiplexes. This includes next generation multiplexes, like the *GenePrint*™ PowerPlex™ 2 System, which will contain one pentanucleotide locus, and the *GenePrint*™ PowerPlex™ 16 System, which will include two pentanucleotide loci in addition to the 13 CODIS loci and Amelogenin.

ACKNOWLEDGEMENTS

Linkage mapping was performed in collaboration with Dr. H. Doris Keller, Ph.D. (Washington University School of Medicine, St. Louis, MO). The work was supported by grant #1-R43-MH52940-01 from the NIH.

This research was supported in part by grant #1-R43-MH52940-01 from the NIH. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

REFERENCES

1. Tautz, D. (1989) *Nucl. Acids Res.* **17**, 6463.
2. Weber, J. and May, P. (1989) *Am. J. Hum. Genet.* **44**, 388.
3. Edwards, A. *et al.* (1991) *Am. J. Hum. Genet.* **49**, 746.
4. Hammond, H. *et al.* (1994) *Am. J. Hum. Genet.* **55**, 175.
5. Murray, J. *et al.* (1994) *Science* **265**, 2049.
6. Sheffield, V. *et al.* (1995) *Hum. Mol. Genet.* **4**, 1837.
7. Levinson, G. and Gutman, G. (1987) *Mol. Biol. Evol.* **4**, 203.
8. Walsh, S. *et al.* (1996) *Nucl. Acids Res.* **24**, 2807.
9. Xiao-Ping, Z. *et al.* (1998) *Genes Chromosomes Cancer* **21**, 101.
10. Schlotterer, C. and Tautz, D. (1992) *Nucl. Acids Res.* **20**, 211.
11. Weber, J. (1990) *Genomics* **7**, 524.
12. Armour, J. *et al.* (1994) *Hum. Mol. Gen.* **3**, 599.
13. Lander, E. and Green, P. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2363.
14. Moller, A. *et al.* (1994) *Int. J. Leg. Med.* **106**, 319.
15. Brinkman, B. *et al.* (1995) *Int. J. Leg. Med.* **107**, 201.

*See patent statement on page 2.