Pentanucleotide Repeats: Highly Polymorphic Genetic Markers Displaying Minimal Stutter Artifact

Jeffery W. Bacher, Laura F. Hennes, Trent Gu, Allan Tereba, Katherine A. Micka, Cynthia J. Sprecher, Ann M. Lins, Elizabeth A. Amiott, Dawn R. Rabbach, and Jennifer A. Taylor, Cindy Helms, Hellen Donis-Keller and James W. Schumm

INTRODUCTION

Short tandem repeat (STR) polymorphisms are becoming the standard genetic markers used throughout the world for forensic DNA analysis and development of forensic databases (Hammond *et al.*, 1994; Urquhart *et al.*, 1994; Oldroyd *et al.*, 1995; Lins *et al.*, 1998). Early work with STR polymorphisms based on dinucleotide short tandem repeats immediately revealed the presence of artifact products, called stutter, which were associated with the PCR amplification of these loci (Levinson & Gutman, 1987; Schlotterer and Tautz, 1992). Stutter products are minor fragments that differ in size from the authentic alleles by one or more core repeat lengths. While problems with stutter has been much reduced with the adoption of current tetranucleotide STRs, stutter bands are still observed at intensity levels 2 – 15% of the authentic alleles. Additional reduction of the level of this artifact has been a goal of the forensic community since the first STR systems were identified. The motivation for reducing stutter is to simplify interpreting results, especially in DNA samples that are mixtures of two or more individuals or when two alleles from a single individual are close in size.

Our approach to achieving this goal has been the discovery and development of new genetic markers that offer a high degree of polymorphism, but do not reveal the typical level of stutter artifact displayed in existing STR systems. Based on the observation that the amount of stutter associated with tetranucleotide STR loci is considerably less than with dinucleotide repeat loci, we reasoned that genetic markers that contained longer repeats might display even less stutter. This work describes our search for short tandem repeat markers with this characteristic, their isolation and initial characterization, and how they can be included into multiplex STR systems for efficient use in forensic science, paternity determination, and other forms of genetic analysis.

BACKGROUND ON THE USE OF GENETIC MARKERS FOR HUMAN IDENTIFICATION

No two individuals, except identical twins, have exactly the same sequence of DNA in their cells. These differences in the estimated three billion base pairs (bp) of DNA in the human genome result in unique DNA profiles that can be used to distinguish individuals. Determining the sequence of the entire genome of an individual is impractical, however, analysis of highly variable regions of the human genome provides sufficient genetic information in most cases to identify an individual using DNA. Therefore, polymorphic markers from these highly variable regions have long been sought for use in genetic analysis and identity applications.

The discovery and development of a series of new types of polymorphic markers has contributed to the evolution of modern day forensic DNA analysis. The first DNA polymorphic markers, called restriction fragment length polymorphisms (RFLP), were proposed in 1980 by Botstein and co-workers as the solution to the limited number of markers available for genetic analysis (Botstein *et al.*, 1980). RFLPs are DNA-based polymorphisms that result from base substitutions, insertions or deletions that cause a modification in the length of fragments generated by digestion with a particular restriction endonuclease. These loci are simple Mendelian co-dominant genetic markers that are detected by Southern hybridization assays of restriction endonuclease-digested DNA (Sambrook *et al.*, 1989). Although first discovered in yeast, systematic screening revealed that RFLPs were not uncommon in humans, although the majority showed a low degree of polymorphism. By 1987, hundreds of RFLPs had been identified (Willard *et al.*, 1985; Nakamura *et al.*, 1987a; Schumm *et al.*, 1988) and the first human linkage map based on 393 RFLPs was constructed (Donis-Keller *et al.*, 1989). RFLP markers that distinguish individuals from one another have also been known since 1980 (Wyman & White, 1980). However, the application of RFLP markers is limited by the requirement of large amounts of non-degraded DNA, the need for tedious and time consuming Southern hybridization assays, and their relatively low level of informativeness.

The next generation of markers, described in the mid 1980s by Jeffreys, White and Nakamura, was a subset of RFLP markers called variable number of tandem repeats[1] or VNTRs (Jeffreys *et al.*, 1985a; Wyman & White, 1980;

Nakamuar *et al.*, 1987a,b). VNTR markers are abundant and highly polymorphic regions of DNA containing nearly identical sequences, 14 to 80 bases in length, repeated in tandem fashion. Different homologous chromosomes in a human population contain different numbers of these repeats. These markers are highly polymorphic, sometimes displaying up to forty or more alleles at a single genetic locus. The first forensic use of DNA took place in England and made use of Alec Jeffrey's original method of "DNA fingerprinting" with VNTR markers (Inman & Rudin, 1997). Many VNTR systems have since been widely adapted for forensic DNA analysis because of their high power of discrimination. However, most limitations described for RFLP markers also apply to VNTR markers.

The next advance involved the joining of the polymerase chain reaction (PCR) technology with the analysis of VNTR loci (Kasai *et al.*, 1990). VNTR loci small enough to be PCR-amplified were discovered (amplifiable VNTR). These can be analyzed without the need for Southern transfer. The amplified products are separated through agarose or polyacrylamide gels and detected by incorporation of radioactivity during the amplification or by post-staining with silver or ethidium bromide. The most widely used of the amplifiable VNTRs is D1S80, which contains a 16 bp repeat with at least 29 alleles ranging in size from 350 to 1000 bp corresponding to 14 to 41 repeats (Kasai *et al.*, 1990; Demers *et al.*, 1995). Problems with preferential amplification of smaller alleles, length limitations related to PCR, and the paucity of high quality markers of this type have limited the use of amplifiable VNTRs in forensic DNA analysis (Demers *et al.*, 1995).

The discovery of the polymorphic nature of microsatellites or short tandem repeat[2] (STR) loci in 1989 (Tautz *et al.*, 1989; Weber & May, 1989; Litt & Luty, 1989) has overcome several of the deficiencies of previous methods. Currently considered to be the best available markers, these loci are similar to the amplifiable VNTRs except they usually contain tandem repeat sequences two (dinucleotide), three (trinucleotide), or four (tetranucleotide) bases long. STR loci are highly abundant and polymorphic in the human genome. It is estimated that there are between 50,000 and 100,000 (CA)n repeats in the genome with an average spacing of approximately 30 kb (Hamada *et al.*, 1982; Tautz & Renz, 1984; Stallings *et al.*, 1991). Informative trinucleotide and tetranucleotide repeats are 10 fold less common than CA repeats, with the estimated 10,000 loci occurring on average once every 300-500 kb (Weber, 1990; Edwards *et al.*, 1991a).

Amplification results of tetranucleotide repeat loci are easier to interpret than di- and trinucleotide repeats because they typically produce only a single stutter band for each true allele (Walsh *et al.*, 1996). Because of this, the forensic DNA community has moved primarily towards tetranucleotide repeats. Genotyping of tetranucleotide STRs is also straightforward because discrete alleles are obtained due to small amplimer size (typically 60-400bp) that allows resolution of single base pair differences between alleles. Production of discrete alleles facilitates accurate allele determination by direct comparison to allelic ladders run on the same gel and permits accurate comparison of results among laboratories (Puers *et al.*, 1993; Xiao *et al.*, 1998). The smaller size of STR amplimers also allows simultaneous electrophoretic analysis of several systems on the same gel (i.e., multiplexing). Since STRs analysis is PCR based, only small quantities of DNA are needed, and the technique is amenable to use of degraded DNA. Thus, the quantity and integrity of the DNA sample is less of an issue with STRs than with earlier RFLP and VNTR methods.

RATIONALE FOR USE OF STRS WITH LONGER REPEAT LENGTHS

Despite all the positive characteristics of STRs, there are still drawbacks that relate to their use in forensics. First, stutter artifacts are generally seen following amplification of STR loci (Levinson & Gutman, 1987). These are minor fragments, often one repeat length shorter than the majority product. 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 mononucleotide 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 (Weber & May, 1989; Edwards *et al.*, 1991a; Walsh *et al.*, 1996; Xiao-Ping *et al.*, 1998). The presence of stutter artifacts, presumed to result from a DNA polymerase slippage event during DNA replication (Levinson & Gutman, 1987; Schlotterer and Tautz, 1992), complicates the unambiguous assignment of alleles and automation of the genotyping procedure. They can be particularly problematic in mixed DNA samples since they are the same size as the actual alleles. Therefore, in mixed samples it is not always possible to distinguish a fainter stutter band from a real allele if its position is four bases shorter than a more intense main allele band.

Another drawback to 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 easily analyzed and number of loci that can be multiplexed. Difficulty in separation of larger DNA

fragments is due to the spatial compression in the upper regions of the gel or greater diffusion in longer capillary electrophoresis runs. Alleles that differ by increments larger than 4 bp extends the useful separation region allowing resolution of larger DNA fragments and/or multiplexing of more loci.

To provide STR systems with enhanced properties we set out to identify and characterize a class of polymorphic markers which contained intermediate sized tandem repeat units that are larger than those commonly found in STRs (2-4bp), but small enough to allow PCR amplification and multiplexing (i.e., smaller than 14-80bp repeats found in VNTRs). We predicted that STRs with larger core repeats (5-9bp) could be separated and detected more easily and precisely, allow multiplexing of more STR makers and exhibit minimal stutter.

STRATEGIES TO ISOLATE NEW STR MARKERS

The number of published reports of polymorphic STR loci containing core repeat lengths between 5 and 9 base pairs is minimal (Edwards *et al.*, 1991b; Chen *et al.*, 1993; Harada *et al.*, 1994; Comings *et al.*, 1995; Utah Marker Development Group, 1995; Jurka & Pethiyagoda, 1995). Therefore, two strategies were used to identify STR markers with these longer core repeat units. The first approach was to search human DNA sequence databases for entries containing short tandem repeats. The alternate strategy followed was to screen small-insert genomic libraries enriched for commonly occurring STR 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, there are over 20,000 unique 5-9 bp repeat motifs that exist, making this approach tedious and time consuming unless automated. Our searches of the GenBank database revealed that the frequency of occurrence of STR loci was inversely proportional to repeat unit length and number of tandem repeats. Over 200 loci containing 5-9 bp short tandem repeats (with 5 or more perfect tandem repeats) were identified in GenBank searches and evaluated for polymorphism levels. Based on the number of polymorphic STRs identified from screening GenBank, we roughly estimate that only around 1000 polymorphic STRs (with 3 or more alleles) containing 5-9 bp repeat units exist in the human genome, most being pentanucleotides. This translates into about 1 every 3 million bp in the human genome, about 10 times less common than all tri- and tetranucleotide repeats (Edwards *et al.*, 1991a). Despite the efficiency of evaluating STR loci for which sequence data is available, this approach alone did not produce a large number of polymorphic ITR loci.

The second approach we used to isolate STR repeat loci involved screening of small insert genomic libraries enriched for the presence selected repeats using hybridization selection (Armor *et al.*, 1994). Based on GenBank human database searches we decided to focus our efforts on the most common repeat lengths and motifs, that is, the pentanucleotide (AAAAC)n, (AAAAG)n, and (AAAAT)n repeats. To increase recovery rates of pentanucleotide repeat loci from genomic libraries an enrichment strategy was employed (Armor *et al.*, 1994). The procedure, summarized in Figure 1, started with size selected (250-600bp) human MboI fragments ligated to linkers to give a whole-genome library. These MboI fragments were amplified using primers complementary to linkers and the PCR products denatured and hybridized to small nylon filters each containing fixed complementary pentanucleotide repeat sequences. Filters were washed to remove unbound DNA fragments and fragments containing targeted repeats were recovered. The hybridization selected DNA was reamplified, cloned into a plasmid vector, and transformed into competent bacterial cells. Next, colony hybridization was performed and clones containing targeted repeats were identified with chemiluminescent detection (Bronstein & McGrath, 1989; Tizard *et al.*, 1990). All colonies positive by colony hybridizations were re-assayed to confirm results. This approach produced many more pentanucleotide repeat candidates than the GenBank search, but required significantly more work to evaluate.

INITIAL CHARACTERIZATION OF PENTANUCLEOTIDE REPEAT LOCI

Recombinant clones selected by colony hybridization were DNA sequenced and inspected for presence of tandem repeats. To simplify the sequencing of over 1,500 clones, a method of preparing sequencing templates utilizing crude bacterial cell lysates was substituted for standard plasmid preparation[3]. This essentially allowed us to by-pass the time consuming step of plasmid preparations that was a major bottleneck in the process. DNA sequencing was then performed using ABI dye terminator sequencing chemistry and the ABI 377 PRISM® DNA Sequencer following manufacturer's protocol.

Unique sequences not found in GenBank database and having a minimum of five perfect tandem repeats were identified and selected. Next, primers were designed for amplification of the locus to determine polymorphism levels using the Oligo Primer Design Software program

(National Biosciences, Inc., Plymouth, MN). The Initial screen for polymorphisms was performed using two pooled DNA samples, one containing a mixture of human genomic DNAs from 15 individuals (mostly of Caucasian-American origin) and the other containing 54 CEPH individuals from the NIGMS Human Genetic Mutant Cell Repository (Coriell Cell Repositories, Camden, NJ). Figure 2 illustrates a typical polymorphism screen using pooled DNA. Those loci with four or more alleles were re-evaluated using individual DNAs to determine preliminary number of alleles, allele frequencies and heterozygosity values for each locus. For example, Figure 3 displays the DNA profiles of 24 African-American individuals following amplification, separation, and detection of alleles at the Penta D locus. A summary of the preliminary screening results for pentanucleotide repeats Penta A through G is shown in Table 1. In general, we found that less than half of the loci tested were polymorphic (having more than one allele), close to 10% displayed four or more alleles, and only a small number (about 2%) exhibited levels of heterozygosity (exceeding 75%) typically required for forensic DNA analysis.

CHROMOSOME LOCALIZATION OF PENTANUCLEOTIDE REPEAT LOCI

For human identification applications, the chromosomal location of markers is important primarily because loss of discrimination power in paternity determination occurs when two loci are genetically closely linked. Therefore, fifty of the polymorphic pentanucleotide repeat sequences with more than four alleles and greater than 50% heterozygosity were mapped to determine their chromosomal location. Both physical and genetic means were employed to localize loci on human chromosomes in collaboration with Cindy Helms and Helen Donis-Keller at the University of Washington (St. Louis, MO). Table 1 gives chromosomal location for the loci Penta A through G.

Physical mapping was performed in two stages. First, markers were assigned to a specific chromosome by somatic cell hybrid (SCH) mapping (Jones *et al.*, 1997; Washington *et al.*, 1993). Next, sub-chromosomal localization was determined by radiation hybrid (RH) mapping (Boehnke *et al.*, 1991; Walter *et al.*, 1994).

Somatic cell hybrid mapping was performed with the NIGMS Mapping Panel #2 (Coriell Cell Repositories, Camden, NJ) consisting of 24 somatic cell hybrids, each retaining a single intact human chromosome. Primer sets flanking each pentanucleotide locus to be mapped were used to amplify DNA from the 24 somatic cell hybrids along with control human and rodent genomic DNAs. The PCR products were separated by gel electrophoresis, stained and photographed. Localization of pentanucleotide loci to a chromosome was determined by presence of equal sized PCR products from both a specific SCH line and the human genomic control, along with the absence of a PCR product from the remaining SCH DNAs.

To determine sub-chromosomal location of pentancleotide loci, radiation hybrid (RH) mapping was used. Radiation hybrid mapping is a somatic cell hybrid approach based on the fusion of lethally irradiated human donor cells to non-irradiated hamster cells. Twenty to thirty percent of the human donor genome is retained in the hamster cells. Unlike the SCH lines described earlier, the RH cell lines contain several smaller fragments from multiple human chromosomes. RH mapping uses the frequency of X-ray breakage between two markers as a statistical measure of the physical distance between markers. Mapping was performed with the GeneBride 4 panel (Research Genetics, Huntsville, AL) consisting of a set of 93 radiation hybrid lines that provides a mapping resolution of approximately 300 kb (Hudson *et al.*, 1995; Gyapay *et al.*, 1996). Primers flanking each of the 50 pentanucleotide loci were used to amplify DNA from the 93 radiation hybrid lines along with controls, separated on acrylamide gels, stained and photographed. RH lanes were compared with the human genomic control and scored for the presence or absence of the human band. Greater accuracy was obtained through duplicate typing of the RH panel. Results were sent to Whitehead Center for Genomic Research (WICGR) web based server which returned results of significant physical linkages, that is, links within 20 cR (equivalent to approximately 6 cM) of a WICGR framework marker.

The genetic approach to chromosome localization used in this study was standard meiotic linkage mapping. Basically, DNA samples from family members of large human pedigrees were used for this analysis. The inheritance of particular alleles was studied for each locus. If two loci are "closely linked", then alleles of each locus which are present together in a parent are often inherited together in children. If loci are far apart or randomly associated (e.g. on different chromosomes), then regardless of the allele content of the particular two loci in the parent, the alleles are randomly associated in the children.

Four families (K102, K884, K1347, 1362) from the CEPH kindred reference panel were chosen for the initial linkage mapping. Since these four families were among the eight families used by Genethon, there is abundant marker genotype data for linkage comparisons. The other four families used by Genthon (1331, 1332, 1413, 1416)

were used if no localization was observed with the first four families. Genotyping procedures included using P32 labeled primers in the PCR reactions, separating the PCR products on sequencing gels, drying the gels, exposing X-ray film, and scoring the developed films for segregation of the alleles in the families. The data were evaluated using the CRI-MAP multi-point linkage program (Lander & Green, 1987) to identify linkages and to place the pentanucleotide loci into intervals between mapped markers. Two-point analyses compared segregation of pentancleotide loci to each of the 371 markers in the 1996 Genome Screen Map from Washington University, MO. The odds for linkage were set at 1000:1 (lod 3), a level that suggests significant linkage. Final analysis and convergence of mapping data from all three methods is in progress.

POLYMORPHISM AND MICROVARIANT ANALYSIS OF SELECTED PENTANUCLEOTIDE REPEAT MARKERS

For new markers to have value in forensic analysis and paternity determination, it is necessary that they display a significant degree of polymorphism within each major racial/ethnic group. Pentanucleotide repeat loci displaying a significant amount of polymorphism in preliminary screens were included in a larger multi-racial group population screen. This work was done in collaboration with the BODE Technology Group (Springfield, VA) and involved genotyping and analysis of over 400 alleles in each of four racial groups (Caucasian-American, African-American, Hispanic-American, and Asian-American). A summary of the preliminary results for four specific pentanucleotide loci are shown in Table 1. The high degree of polymorphism contained in these systems is illustrated by comparison with the heterozygosity of the least polymorphic (TPOX) and most polymorphic (FGA) loci selected as core STR loci for the FBI Laboratory's Combined DNA Index System (CODIS) database (Evett *et al.*, 1996; Lins *et al.*, 1996). The CODIS database holds "DNA fingerprints" from individuals convicted of violent crimes allowing crime laboratories to exchange and compare DNA profiles electronically.

Previously, it has been observed that STR loci which are highly polymorphic often have correspondingly high numbers of undesirable microvariant alleles, that is, alleles differing from one another by lengths other than the core repeat length (Moeller *et al.*, 1994; Evett *et al.*, 1996; Rolf *et al.*, 1997). The presence of microvariant alleles complicates separation, interpretation, and assignment of alleles. However, in genotype determination of particular pentanucleotide loci, few or no microvariant alleles were revealed despite their highly polymorphic nature. The range of alleles identified and any microvariants observed at a frequency greater than one in a thousand are listed in Table 1 for Penta B, C, D and E loci. Sequence analysis of Penta D alleles 2.2 and 3.2 indicated that they contained deletions (13 bp and 8 bp, respectively) just upstream of the repeat region (AAAGA)5. These deletions caused the 2.2 and 3.2 alleles to size smaller than expected for the actual number of repeats they contained. The next larger known Penta D allele is allele 5, which is 8bp larger than allele 3.2. Therefore, Penta D allele 13.3 is the only allele that differs from other alleles by less than five bases. This low frequency of microvariant alleles observed in the highly polymorphic pentanucleotide repeat loci is uncommon for other STR loci. This point is illustrated by comparison of the Penta E locus and the most polymorphic CODIS locus, FGA. (Table 1).

CHARACTERIZATION OF STUTTER IN PENTANUCLEOTIDE REPEAT MARKERS

To evaluate whether pentanucleotide repeat loci did display lower stutter, five pentanucleotide loci were selected for a controlled evaluation of stutter fragments associated with these loci. In each case, one nanogram of twenty separate DNA samples was amplified at the individual locus. The resulting products were separated and analyzed for peak heights of the main product and the stutter fragment product for each allele using the ABI PRISM® 377 DNA Sequencer and GeneScan® software (ABI, Foster City, CA). Results of this study were compared with tetranucleotide repeat loci commonly used in forensic and paternity analyses. Figure 4 shows that the average amount of stutter observed with selected pentanucleotide systems was less than 2%, while stutter in typical tetranucleotide repeats averaged 2 to 8%, with some alleles exceeding 15%.

The improved performance characteristic of pentanucleotides will assist in interpretation of mixtures of DNA samples and supports the hypothesis that STRs with larger core repeat units have less stutter then those with smaller repeat units. The clean nature of the amplification product typical of pentanucleotides is illustrated with Penta D (Figure 3), where despite the intense signal of the amplified authentic alleles, no stutter at any allele is visible in the image. A comparison of STR loci with repeat sizes ranging from one to five base pairs further illustrates the point that stutter is less pronounced for loci containing longer core repeats[4] (Figure 5).

INTRODUCTION OF PENTANUCLEOTIDE REPEAT MARKERS INTO MULTIPLEX STR SYSTEMS

The combined properties of high power of discrimination, few microvariants and very low levels of stutter

make pentanucleotide STR loci ideal markers for forensic DNA analyses (see Table 1). Efforts are currently in progress to incorporate the best pentanucleotide repeat loci into STR multiplex systems (Amiott *et al.*, 1998). In particular, the *GenePrint*® PowerPlex™ 2 System and the *GenePrint*® PowerPlex™ 16 System will contain pentanucleotide repeat loci.

A prototype of the PowerPlex™ 2 System has been developed. It allows co-amplification of 9 STR loci (including all the INTERPOL loci selected for pan-European use and all the SGM loci developed by the Forensic Science Service, United Kingdom) plus Amelogenin, a gender identification locus. One of these loci is Penta E. Five STR loci (D3S1358, TH01, D21S11, D18S51 and Penta E) are labeled with fluorescein and four STR loci (vWA, D8S1179, TPOX and FGA) plus Amelogenin are labeled with carboxy-tetramethylrhodamine (TMR). The amplified product may be detected in combination with an internal lane standard with evenly space fragments labeled in a third dye, carboxy-X-rhodamine (CXR).

A prototype of the PowerPlex™ 16 System allows co-amplification of 15 STR loci (including all of the CODIS 13 core STR loci) plus Amelogenin, the gender identification locus. This multiplex system design is displayed in Figure 6. In addition to all the loci present in the same configuration and the same dye content as the PowerPlex™ 2 System, six additional loci (D5S818, D13S317, D7S820, D16S539, CSF1PO, Penta D) are present in a new dye. Two pentanucleotide loci are included in this system, that is, locus Penta E and locus Penta D.

These multiplex systems have been designed for optimum performance in demanding situations of forensic analysis and paternity determination. Amplification of so many loci simultaneously allows use of minimal amounts of sample material as well as efficient sample throughput. In addition, with a single amplification reaction, there is less chance of sample mix-up from multiple separate amplification reactions. Compatibility with use of degraded sample material has also been incorporated into the designs. For example, in the PowerPlex™ 16 System, twelve CODIS STR loci generate amplification products completely below 372 bases. Only the rare FGA alleles (less than 2% of those observed) are larger and these are all separated from one another by four bases allowing their easy assignment. Eight of the CODIS STR loci and five of the SGM loci produce amplification products completely contained below 261 bases, even for all rare alleles.

Additional power of discrimination is achieved by incorporation of the Penta E and Penta D loci. With the paucity of microvariants observed with these systems, the Penta E and Penta D alleles are all separated from one another by five bases except in extremely rare circumstances (less than 1 in 500 allele calls). This allows easy and confident allele separation and determination. The matching probability and power of exclusion for both multiplex systems (Amiott *et al.*, 1998) are summarized in Table 2. The corresponding values for the PowerPlex™ 1 and FFFL Systems (containing loci F13AO1, F13B, FESFPS, LPL) are included for comparison (Lins *et al.*, 1998). We are continuing to collaborate with the Bode Technology Group to evaluate additional population characteristics such as final genotype frequencies and Hardy-Weinberg Equilibrium calculations (Weir, 1996) of independence of all 19 STR loci represented in the FFFL and PowerPlex™ 16 Systems.

ACKNOWLEDGEMENTS

Linkage mapping was performed in collaboration with Dr. H. Donnis-Keller, Ph.D. (Washington University School of Medicine, St. Louis, MO). The work was supported by grant #1-R43-MH5240-01 from NIH.

This research was supported in part by grant #1-R43-MH5240-01 from NIH. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

FOOTNOTES

[1] The concept of detection of VNTR loci is protected and described by US Patent 4,963,663, "Genetic identification employing DNA probes of variable number tandem repeat loci" authored by Raymond L. White, Peter O'Connell, and Mark F. Leppert and issued October 16, 1990. A license to commercialize several significant aspects of this technology has been obtained by Promega Corporation.

[2] STR loci are subject to U.S. Pat. No. 5,766,847 and German Patent No. DE 38 34 636 C2, issued to Max-Planck-Gesellschaft zur Forderung der Wissenschaften, eV, Germany. Exclusive rights have been licensed to Promega Corporation for uses in human clinical research and diagnostics applications and all forms of human genetic identity. The development and use of STR loci is covered by U.S. Pat. No. 5,364,759 assigned to Baylor College of Medicine, Houston, Texas. Rights have been licensed to Promega Corporation for all applications.

[3] Cell lysates were made by taking 2 µl of overnight bacterial cultures and adding this to 100µl sterile nanopure

water in 96 well reaction plates and heating to 100oC for 4 minutes. Two microliters of the cell lysate were used as templates in PCR reactions using M13 forward and reverse primers (Promega, WI) to amplify insert region of pGEM vectors. PCR reaction products were cleaned-up with Qiagen QIAquick 96 PCR Purification plates (Hilden, Germany) following manufacturers protocol.

[4] 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 number of repeats present within the allele (i.e., larger alleles tend to display more stutter) and whether the repeat is interrupted (STR loci containing one or more repeat units having a different DNA sequence than core repeat unit generally display less stutter (Walsh *et al.*, 1996).

REFERENCES

1. Amiott, E.A., Micka, K.A., Sprecher, C.J., Lins, A.M., Rabbach, D.R., Nassif, N.A., Mandrekar, P.V., Bacher, J.W., Hennes, L.F., Gu, T., Tereba, A., Taylor, J.A., Schumm, J.W. (1998) Incorporating high quality genetic markers into forensically useful multiplexes. Proceedings from the 9th International Symposium on Human Identification.
2. Armor, J., Neumann, R., Gobert, S., and Jeffreys, A.J. (1994). Isolation of human simple repeat loci by hybridization selection. Hum. Mol. Genet. 3(4): 599-605.
3. Bär, W. *et al*. (1997) DNA Recommendations. Further report of the DNA commission of the ISFH regarding the use of short tandem repeat systems. Int. J. Leg. Med. 110:175.
4. Botstein D, White, R.L., Skolnick, M., and Davis, R.W. (1980) Construction of a genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. Am. J. Hum. Genet. 32:314
5. Boehnke, M., Lange, K., and Cox, D.R. (1991) Statistical methods for multipoint radiation hybrid mapping. Am J Hum Genet 49: 1174-1188.
6. Bronstein, I., and McGrath, P. (1989) Chemiluminescence lights up. Nature 338: 599-600.
7. Chen, H., Kalaitsidaki, M., Warren, A.C., Avramopoulos, D., Antonarakis, S.E., (1993). A novel zinc finger cDNA with a polymorphic pentanucleotide repeat (ATTTT)n maps on human chromsome 19p. Genomics 15(3): 621-5.
8. Comings, D.E., Muhleman, D., Dietz, G., Sherman, M., Forest, G.L. (1995) "Sequence of human tryptophan 2,3-dioxygenase (TDO2): presence of a glucocorticoid response-like element composed of a GTT repeat and an intronic CCCCT". Genomics, 29(2):390-6.
9. Demers, D., *et al*. (1995) Enhanced PCR amplification of VNTR locus D1S80 Using Peptide Nucleic Acid. Nucleic Acids Res. 23:3050-3055.
10. Donis-Keller, H., *et al*. (1989) A Genetic Linkage Map of the Human Genome. Cell 51:319-337.
11. Edwards, A., Civitello, A., Hammond, H.A., Caskey, T. (1991a) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am. J. Hum. Genet. 49:746-756.
12. Edwards, M.C., Clemens, P.R., Tristan, M., Pizzuti, A., Gibbs, R.A. (1991b) Pentanucleotide repeat length polymorphism at the human CD4 locus. Nucleic Acids Res. 19(17):4791.
13. Evett, I.W., Gill, P.D., Lambert, J.A., Oldroyd, N., Frazier, R., Watson, S., Pnachal, S. Connolly, A., Kimpton, C. (1996) Statistical analysis of data for three British ethnic groups from a new STR multiplex. Int. J. Legal Med. 110:5-9.
14. Gyapay G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'homme, J., Dib, Cl, Auffray, C., Morissette, J., Wiessenbach, J., and Goodfellow, P.N. (1996) A radiation hybrid map of the human genome. Hum Mol Genet 5(3): 339-346.
15. Hamada H., Petrino M.G., Kakunaga T. (1982) A novel repeated element with Z-DNA forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Nat. Acad. Sci. 79:6465-6469.
16. Hammond, H.A., Jin, L., Zhong, Y., Caskey, C.T. and Chakraborty, R. (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am. J. Hum. Genet. 55: 175-189.
17. Harada, S., T. Nakamura, S. Misawa. (1994). Polymorphism of pentanucleotide repeats in the 5' flanking region of glutathione S-transferase (GST) gene. Hum Genet 93:223-224.
18. Hudson T.J., *et al*. (1995) An STS-based map of the human genome. Science 270(5244):1945-54
19. Inman, K. and Rudin, N. (1997) An Introduction to Forensic DNA Analysis. CRC Press, New York, NY
20. Jeffreys, A.J., Wilson V., Thein S.L. (1985a) Hypervariable minisatellite regions in human DNA. Nature 314:67-73.
21. Jeffreys A.J., Wilson V., Thein S.L. (1985b). Individual-specific "fingerprints" of human DNA. Nature 316:76-79.
22. Jones MH *et al*., (1997) Chromosomal assignment of 311 sequences transcribed in human adult testis Genomics 40:155.
23. Jurka, J., and Pethiyagoda, C., (1995) Simple repetitive DNA sequences from primates: Compilation and analysis. J. Mol. Evol. 40:120-126.
24. Kasai, K., Nakamura, Y., White, R. (1990) Amplification of VNTR locus by PCR and its application forensic science. Journal of Forensic Sciences 35(5):1196-1200.
25. Lander, E., and Green, P. (1987) Construction of multi-locus genetic linkage maps in humans. Proc. Nat. Acad. Sci. 84:2363-2367.
26. Lins A., Sprecher C., Puers C., and Schumm J.W. (1966). Multiplex sets for the amplification of plymorphic short tandem repeat loci- Silver stain and fluorescent detection. BioTechniques 20(5):882-888.
27. Lins, A.M., Micka, K.A., Sprecher, C.J., Taylor, J.A., Bacher, J.W., Rabbach, D.R. Bever, R.A., Creacy, S.D., and Schumm, J.W. (1998) Development and Population Study of an Eight Locus Short Tandem Repeat (STR) Multiplex System. Journal of Forensic Sciences 43: 1178-1190.
28. Levinson, G. and Gutman, G. (1987). Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4(3):203-221.
29. Litt, M., and Luty, J.A. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within a cardiac muscle actin gene. Am J Hum Genet 44:397-401.
30. Moeller, A., Meyer, E., Brinkmann, B. (1994) Different types of structural variation in STRs: HumFES/FPS, HumVWA and HumD21S11. Int J Leg Med 106: 319-323.
31. Nakamura, Y., *et al*, (1987a). Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616-1622.
32. Nakamuar, Y., Carlson, M., Krapcho, K., Kanamori, M., White, R. (1987b) New approach for isolation of VNTR markers. Am. J. Hum. Genet. 43:854-859.
33. Oldroyd, N.J., Urquhart, A.J., Kimpton, C.P., Millican, E.S., Watson, S.K., Downes, T. and Gill, P.D. (1995) A highly discriminating octoplex short tandem repeat polymerase chain reaction system suitable for human individual identification. Electrophoresis 16: 334-337.
34. Puers, C., Lins, A.M., Sprecher, C.J., Brinkmann, B. and Schumm, J.W. (1993) Analysis of polymorphic short tandem repeat loci using well-characterized allelic ladders. Proceedings from the 4th International Symposium on Human Identification. pp. 161-172.

35. Rolf, B., Schuerenkamp, M., Junge, A., and Brinkmann, B. (1997) Sequence polymorphism at the tetranucleotide repeat of the human beta-actin related pseudogene H-beta-psi-2 (ACTBP2) locus. Int. J. Leg. Med. 110: 69-72.
36. Sambrook, J., Fritsch, E.F., and Maniatis, T. (eds.) (1989) Molecular cloning - A laboratory manual. 2nd edition, Cold Spring Harbor Laboratory Press. pp. 1.90-1.104.
37. Schlotterer, C., and Tautz, D. (1992) Slippage synthesis of simple sequence DNA. Nucl. Acids Res. 20(2):211-215.
38. Schumm J, *et al*. 1988. Identification of More Than 500 RFLPs by Screening Random Genomic Clones. Am. J. Hum. Genet. 42:143-159.
39. Stallings RL, Ford AF, Nelson D, Torney DC, Hilderbrand CE, Moyzis RK (1991) Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. Genomics 10:807-815.
40. Tautz, D. (1989) Hypervariablitiy of simple sequences as a general source for polymorphic DNA markers. Nuc. Acids Res. 17:6464-6471.
41. Tautz, D., and Renz, M. (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nuc. Acids Res. 12:4127-4138
42. Tizard, R. Cate., R.L., Ramachandran, L.K., Wysk, M., Voyta, J.C., Murphy, O.J., and Bronstein, I. (1990). Imaging of DNA sequences with chemiluminescence. Proc Natl Acad Sci, U.S.A. 87: 4514-4518.
43. Urquhart, A., Kimpton, C.P., Downes, T.J. and 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.
44. Utah Marker Development Group, 1995. A collection of ordered tetranucleotide repeat markers from the human genome. Am. J. Hum. Genet. 57:619-628.
45. Walsh, P.S., Fildes, N.J., and Teynolds, R. (1996) Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucl. Acids Res. 24 (14): 2807-2812.
46. Walter, M.A., Spillett, D.J., Thomas, P., Weissenbach, and Goodfellow, P. (1994). A method for constructing radiation hybrid maps of whole genomes. Nature Genet 7: 22-28.
47. Weir, B.S. Genetic data analysis II. Sunderland: Sinauer Associates, Inc., 1996.
48. Washington, S.S., *et al*. (1993) A somatic cell hybrid map of human chromosome 13. Genomics 18:486.
49. Weber, J.L., and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388-396.
50. Weber, J.L. (1990) Informativeness of human (dC-dA)n (dG-dT)n polymorphisms. Genomics 7:524-530.
51. Willard H, *et al*. (1985) Report of the committee on human gene mapping by recombinant DNA techniques. In Human Gene Mapping 8. Cytogenet. Cell. Genet. 40:360-489
52. Wyman, A., and White, R. (1980) A highly polymorphic locus in human DNA. Proc. Nat. Acad. Sci. 77:6754-6758.
53. Xiao, F.-X., Gilissen, A., Cassiman, J.-J., and Decorte, R. (1998) Quadraplex fluorescent STR typing system (HUMVWA, HUMTH01, D21S11 and HPTR) with sequence-defined allelic ladders: Identification of a new allele at D21S11. Forensic Sci. Int. 94(1):39-40.
54. Xiao-Ping, Z. *et al*. (1998) Determination of the replication error phenotype in human tumors without the need for matching normal DNA by analysis of mononucleotide repeat microsatellites. Genes, Chromosomes & Cancer 21:101-107.

Table 1. Preliminary population statistics and chromosomal location of loci Penta A through G (STR loci TPOX and FGA, the least and most polymorphic CODIS loci, are included for comparison).

Locus	Population	Matching Probability[1]	Typical PI[2]	Total # Analyzed	Percent Heterozgotes	Chromosome Location	Allele Range[3]	Average Percent Stutter	Micro-variants[4] (>1:1000)
A	African-American	1:15	2.10	42	76	8p	7-18	0.6	nd
	Caucasian-American	1:12	1.39	39	64				
	Asian-American	1:5	2.08	25	76				
	Hispanic-American	nd	nd	nd	nd				
B	African-American	1:41	1.89	208	74	7q	5-31	2.0	NONE
	Caucasian-American	1:35	3.40	211	85				
	Asian-American	1:15	3.11	205	84				
	Hispanic-American	1:30	2.43	209	79				
C	African-American	1:16	1.85	207	73	9p	3-15	0.9	NONE
	Caucasian-American	1:10	1.98	210	75				
	Asian-American	1:10	2.14	205	77				
	Hispanic-American	1:13	2.54	208	80				
D	African-American	1:33	4.18	209	88	21q	2.2-17	0.1	2.2, 3.2, 13.3
	Caucasian-American	1:17	3.64	211	86				
	Asian-American	1:17	2.03	207	75				
	Hispanic-American	1:18	3.21	212	84				
E	African-American	1:48	4.50	207	89	15q	5-24	0.9	NONE
	Caucasian-American	1:33	4.22	211	88				
	Asian-American	1:63	5.18	207	90				
	Hispanic-American	1:49	3.00	210	83				
F	African-American	1:10	6.25	25	92	6q	5-20	nd	nd
	Caucasian-American	1:9	1.79	25	72				
	Asian-American	1:4	1.63	26	69				
	Hispanic-American	nd	nd	nd	nd				
G	African-American	1:13	3.00	24	83	22q	6-17	nd	nd
	Caucasian-American	1:10	1.35	35	63				
	Asian-American	1:9	1.09	24	54				
	Hispanic-American	nd	nd	nd	nd				
TPOX	African-American	1:12	2.01	221	75	2p23-2pter	6-13	2.4	NONE
	Caucasian-American	1:5	1.41	215	65				
	Asian-American	nd	nd	nd	nd				
	Hispanic-American	1:6	1.53	220	67				
FGA	African-American	1:32	3.27	209	85	4q28	17-51.2	5.5	18.2, 19.2, 20.2,
	Caucasian-American	1:23	3.89	210	87				21.2, 22.2, 23.2,
	Asian-American	1:31	6.06	206	92				24.2, 25.2, 30.2,
	Hispanic-American	1:33	4.35	209	89				34.2, 43.2, 44.2,
									45.2, 46.2, 47.2,
									48.2, 50.2, 51.2

[1] Matching Probability: is the average number of people you would have to survey before you would find the same DNA pattern as a randomly selected individual.

[2] Typical paternity Index: is how many more times likely it is that the man being tested is the father than a randomly selected individual would be the father.

[3] Allele nomenclature follows the recommendations of the DNA Commission of the ISFH regarding the use of STR systems (Bär *et al.*, 1997).

[4] Sources of information about FGA microvariants were the Promega/Bode Technologies population study, Evett *et al.*, 1997, and personal communication from the Forensic Science Service, United Kingdom. The frequency of occurrence of some FGA microvariant alleles has not been well established and may be less than 1 in 1000.

Table 2. Matching probabilities for STR multiplex systems FFFL, PowerPlex™ 1 System, PowerPlex™ 2 System, and PowerPlex™ 16 Systems containing 4, 8, 8 and 15 STR loci, respectively.

Matching Probabilities			
	African-American	**Caucasian-American**	**Hispanic-American**
FFFL System	1.68×10^{4}	2.66×10^{3}	3.28×10^{3}
PowerPlex™ 1 System	2.74×10^{8}	1.14×10^{8}	1.45×10^{8}
PowerPlex™ 2 System	3.00×10^{11}	8.46×10^{10}	1.02×10^{11}
PowerPlex™ 16 System	1.40×10^{18}	1.83×10^{17}	2.94×10^{17}
Power of Exclusion			
	African-American	**Caucasian-American**	**Hispanic-American**
FFFL System	0.9459	0.9406	0.9019
PowerPlex™ 1 System	0.9982125	0.9968853	0.9973337
PowerPlex™ 2 System	0.9999219	0.9999242	0.9997134
PowerPlex™ 16 System	0.9999996	0.9999994	0.9999983

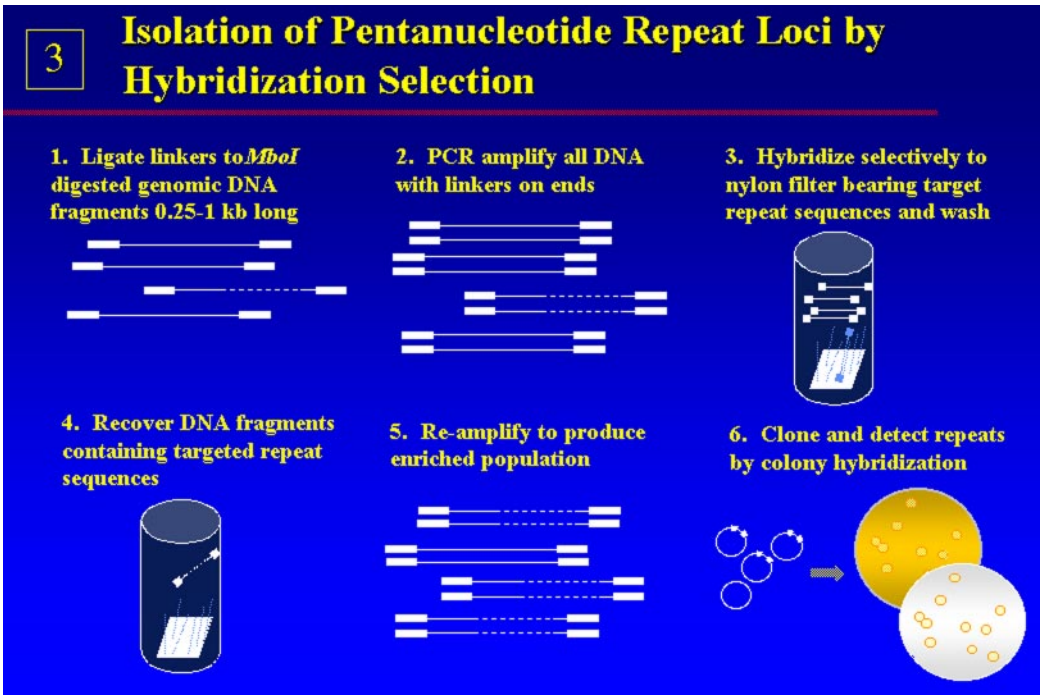

Figure 1. Isolation of pentanucleotide repeat loci by hybridization selection.

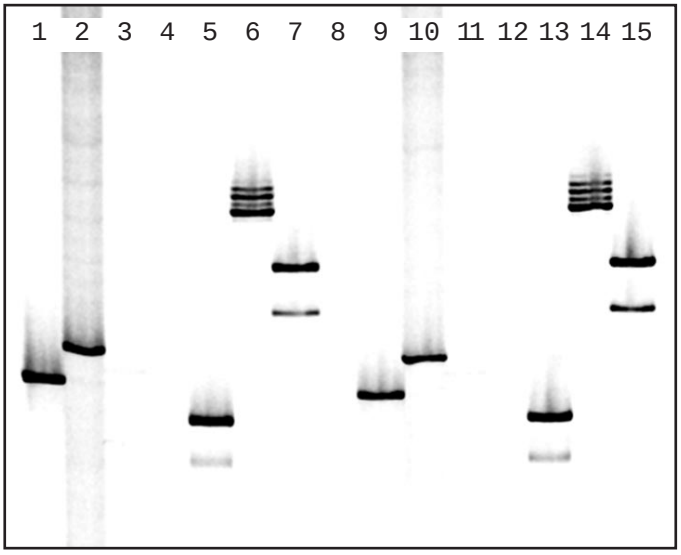

Figure 2. Initial polymorphism analysis of candidate pentanucleotide repeat loci using pooled DNA samples. Fluorescein labeled primers were used to amplify a mixture of DNA samples at one locus in a single PCR reaction. Each lane contains the amplified products from a different locus. Amplification products were separated using a 4% denaturing polyacryamide gel and detected using the FMBIO® II Fluorescent Scanner (Hitachi, San Francisco, CA). Polymorphic loci are identified by the presence of two or more bands per lane. Variation in band intensity is due to differences in frequency of alleles in the pooled DNA sample.

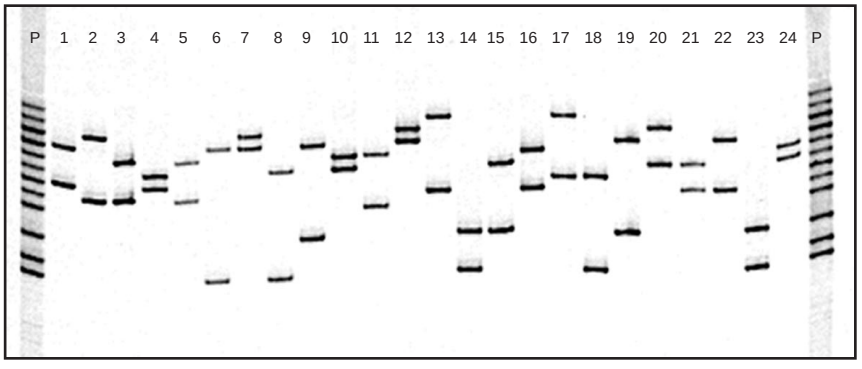

Figure 3. FMBIO® II image of pentanucleotide short tandem repeat locus D. Twenty four DNA samples from African-American individuals were amplified using fluorescein labeled primers, the PCR products separated on 4% denaturing polyacrylamide gels and visualized by scanning on FMBIO® II fluorescent scanner. The first and last lanes include pooled (P) DNA samples.

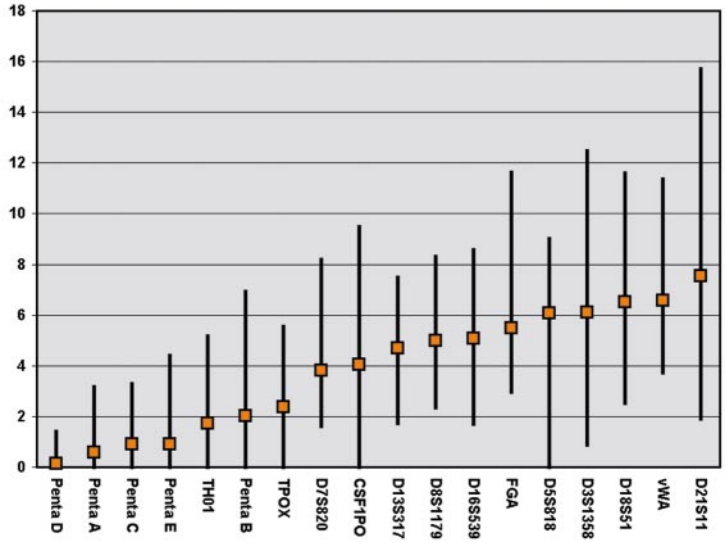

Figure 4. Percent stutter for pentanucleotide tandem repeat loci A through E and 13 tetranucleotide CODIS repeat loci. Boxes represent the average percent stutter of the minor N-4 or N-5 band relative to the main allele for at least 20 DNA samples. The bars indicate high and low percent stutter range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

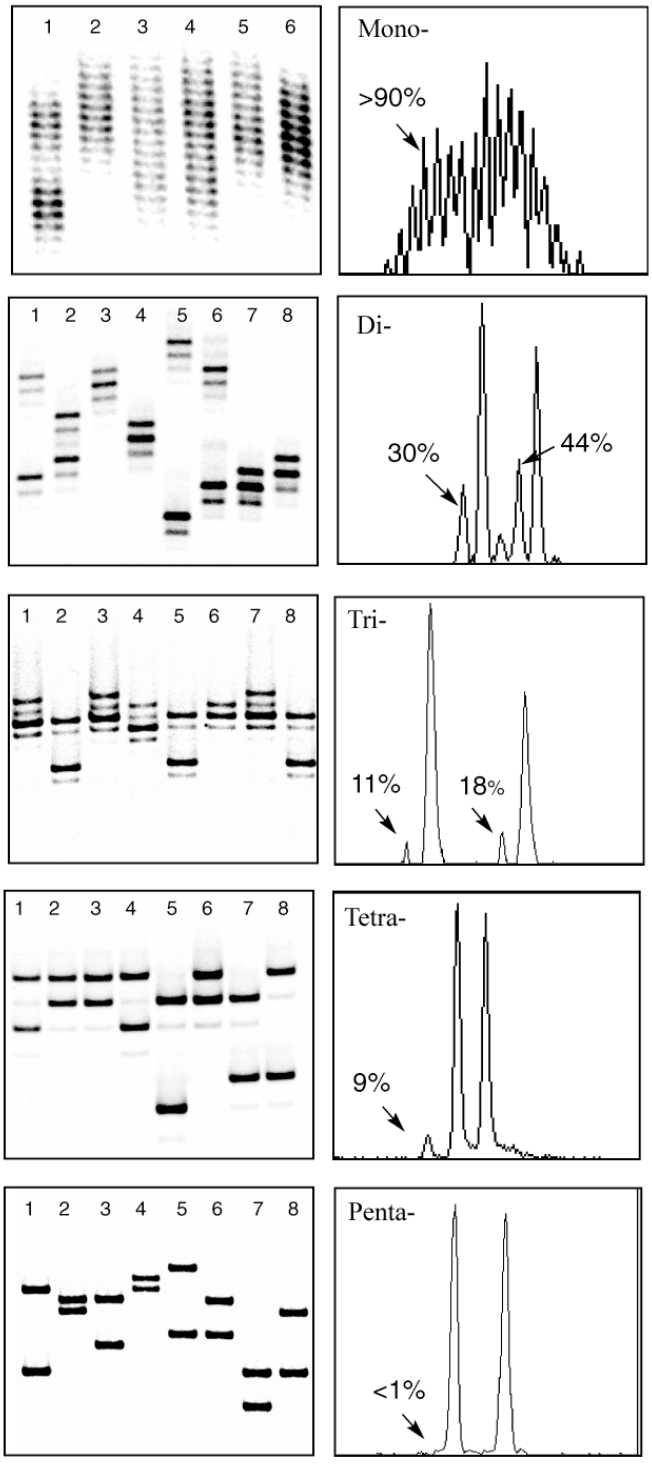

Figure 5. Stutter observed for STR repeat loci is more pronounced the shorter the repeat units. DNA samples were amplified using fluorescently labeled primers for mono- (BAT-40), di- (D5S346), tri- (TBP), tetra- (D5S818) and pentanucleotide STR repeat loci (Penta D). A gel image and lane trace of representative samples are shown for each repeat type, indicating stutter peaks and percent stutter.

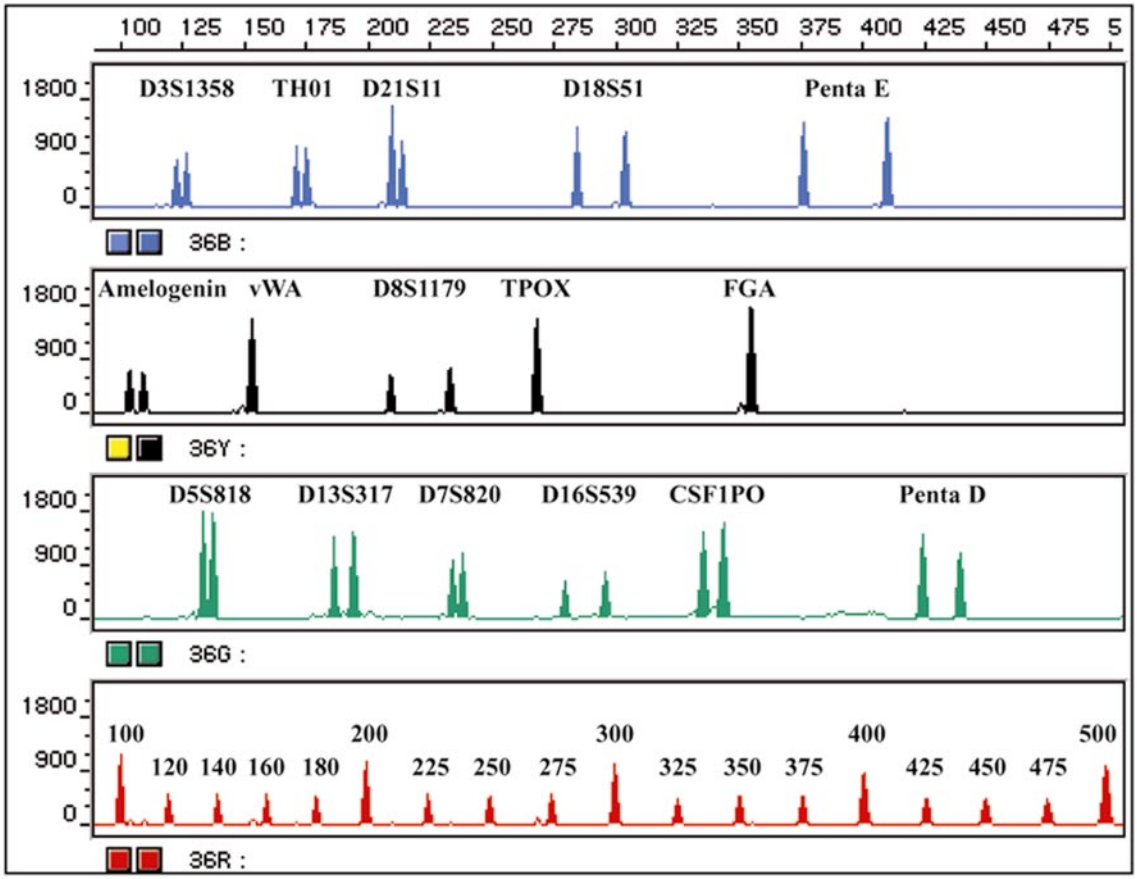

Figure 6. Electropherogram of single DNA sample amplified using the 16-locus prototype PowerPlex™ 16.2 System and detected with the ABI PRISM® 310 Genetic Analyzer. All 16 loci were amplified in a single reaction and detected in a single capillary. The multiplex contains pentanucleotide markers (Penta D and Penta E). Alleles were easily resolved even though some fragments were over 400 bp because length differences of at least five bases were observed.