High-Throughput Genotyping of Forensic STRs and SNPs Using Time-of-Flight Mass Spectrometry

John M. Butler and Christopher H. Becker
GeneTrace Systems Inc., 1401 Harbor Bay Parkway, Alameda, CA 94502

ABSTRACT

Time-of-flight mass spectrometry offers a rapid, accurate, and cost-effective means for high-throughput DNA analysis. Using proprietary chemistry and robotic sample preparation, GeneTrace has demonstrated that several thousand DNA samples may be processed daily with a single mass spectrometer. This mass spectrometry approach is a potential solution to the processing of samples currently being gathered for large criminal DNA databases around the world.

We present results from a number of tetranucleotide STR markers of forensic interest including the 13 CODIS core loci. We are also working on some new Y chromosome and mitochondrial DNA single nucleotide polymorphism (SNP) markers that may prove useful in screening forensic samples.

INTRODUCTION

Time-of-flight mass spectrometry is a rapid analysis technique capable of high-volume DNA processing that lends itself well to the formation of large DNA databases that are being developed around the world to aid forensic investigations. In this technique, DNA molecules are ionized with a brief laser pulse and separated in a matter of microseconds within the vacuum environment of a mass spectrometer. Accurate size measurements may be made with time-of-flight mass spectrometry in only a few seconds per sample with STR alleles (1,2) and single nucleotide polymorphisms of forensic interest (3).

GeneTrace Systems Inc. has developed an integrated high-throughput DNA analysis system involving the use of proprietary chemistry, robotic sample manipulation, and time-of-flight mass spectrometry. Due to the high-throughput capabilities of the mass spectrometer, which performs a rapid, serial analysis of each sample, large numbers of samples must be prepared in a highly parallel and automated fashion on a robotic workstation to utilize the technology to its fullest extent. Thermal cyclers are now commercially available for preparing 384 or even 768 PCR samples in parallel. The capability for constructing large DNA databases in a timely manner is becoming a reality.

FORENSIC STR LOCI

The FBI has designated 13 core STR loci for the nationwide Combined DNA Index System (CODIS) database. These STR loci include TH01, TPOX, CSF1PO, vWA, FGA, D3S1358, D5S818, D7S820, D13S317, D16S539, D8S1179, D18S51, and D21S11. The sex-typing marker amelogenin is typically included in STR multiplexes that cover the 13 core STR loci. Each sample must have these 14 markers tested in order to be entered into the national CODIS database. Therefore, the current national backlog of ~500,000 samples corresponds to at least 7 million genotypes. These samples are being stored in anticipation of future analysis and inclusion in CODIS.

To reduce analysis cost and sample consumption and to meet the demands of higher sample throughputs, multiplex STR analysis has become a standard technique in most forensic DNA laboratories. STR multiplexing is most commonly performed using spectrally distinguishable fluorescent tags and/or non-overlapping PCR product sizes (4,5). Multiplex STR amplification in one or two PCR reactions with fluorescently labeled primers and measurement with gel or capillary electrophoretic separation and laser-induced fluorescence detection is becoming a popular method among forensic laboratories for analysis of the 13 CODIS STR loci. The STR alleles from these multiplexed PCR products typically range in size from 100-350 bp with commercially available kits.

Due to the size constraints of mass spectrometry (1,2), we have adopted a different approach to multiplex analysis of multiple STR loci. Primers are designed such that the PCR product size ranges overlap between multiple loci but have alleles that interleave and are resolvable in the mass spectrometer (Figure 1). We have designed PCR primers that are closer to the STR repeat regions that those commonly used with electrophoretic systems. Higher resolution and sensitivity are possible in the mass spectrometer when PCR products are kept below 140 bp in size. The high accuracy, precision, and resolution of our mass spec approach permits multiplexing STR loci in such a manner.

MITOCHONDRIAL DNA AND Y CHROMOSOME SNPS

Single nucleotide polymorphisms (SNPs) are the most frequent form of DNA sequence variation in the human genome and as such are becoming increasing popular genetic markers for genome mapping studies, medical diagnostics, and identity testing (6-9). SNPs are typically biallelic with two possible nucleotides (alleles) having frequencies of >1% throughout the human population at a particular site in the genome. Although SNPs are less polymorphic than the currently used STR markers, SNPs are more abundant--occurring approximately every kilobase in the human genome (Table 1). The forensic community has used SNP markers such as HLA-DQA1 and PolyMarker. We are developing multiplexed SNP assays for rapid screening of variation in the mitochondrial DNA (mtDNA) control region and the Y-chromosome.

The extensive sequence variation in the control region of mitochondrial DNA has made it a useful system for identity testing particularly in situations where nuclear DNA markers are difficult to analyze (9). While full sequence analysis is the most comprehensive method for detecting mtDNA polymorphisms, it is relatively time-consuming and expensive. More rapid tests, such as the dot blot assay (10) or minisequencing (11), provide alternative methods for screening samples and can reduce the cost and time of analysis. For example, the Forensic Science Service has developed a multiplex solid-phase fluorescent minisequencing method capable of evaluating the nucleotide content at 12 polymorphic locations (10 substitution polymorphisms and 2 length polymorphisms) within the mtDNA control region (11). In this assay, poly(T) tails are used to aid electrophoretic separation of primers that probe the various polymorphic sites. The identity of each base is determined with fluorescently labeled dideoxynucleotides that terminate the extension reaction and label the SNP primer. An automated DNA sequencer capable of four-color fluorescent detection can be used to generate a readout that may be used as an effective exclusionary test in forensic analysis. We have adapted this mtDNA minisequencing assay to the more rapid readout format of mass spectrometry.

Genetic markers on the Y-chromosome have potential applications in human male identification, such as paternity testing and forensic rape cases. With the wealth of genetic information that is being uncovered as part of the Human Genome Project, new informative markers are constantly being discovered on the Y-chromosome. For example, 22 biallelic SNPs were reported this past year by a group from Stanford University and used to study human evolutionary history (12). In collaboration with Dr. Peter Oefner and Dr. Peter Underhill of Stanford University, we are beginning to examine some of these newly discovered Y-chromosome polymorphisms to determine their usefulness in human identity testing. These Y SNP markers should serve as a nice complement to the Y STR markers being developed by forensic laboratories around the world (13).

METHODS

STR Results

The expected masses for a triplex involving the STR loci CSF1PO, TPOX, and TH01 (commonly referred to as a CTT multiplex) are schematically displayed in Figure 1. All known alleles for these STR loci, as defined by STRBase (15), are fully resolvable and far enough apart to be accurately determined. For example, TH01 alleles 9.3 and 10 fall between CSF1PO alleles 10 and 11. For all three STR systems in this CTT multiplex, the AATG repeat strand is measured, which means that the alleles *within* the same STR system differ by 1260 Da. The smallest spread between alleles *across* multiple STR systems in this particular multiplex exists between the TPOX and TH01 alleles where the expected mass difference is 285 Da. TPOX and CSF1PO alleles differ by 314 Da while TH01 and CSF1PO alleles differ by 599 Da. By using the same repeat strand in the multiplex, the allele masses between STR systems all stay the same distance apart. Each STR has a unique flanking region and it is these sequence differences between STR systems that permit multiplexing in such a fashion as described here. An actual result with this CTT multiplex is shown in Figure 2.

It is also worth noting that this particular CTT multiplex was designed to account for possible, unexpected microvariants. For example, a CSF1PO allele 10.3 that appears to be a single base shorter than CSF1PO allele 11 was recently reported (5). With the CTT multiplex primer set described here, a CSF1PO 10.3 allele would have an expected mass of 21402 Da, which should be fully distinguishable from the nearest possible allele (i.e., TH01 allele 10) as these alleles would be 286 Da apart. Using our mass window of 100 Da as defined by our previous precision studies (2), all possible alleles including microvariants should be fully distinguishable. STR multiplexes are designed so that expected allele masses between STR systems are offset in a manner that possible microvariants, which are most commonly insertions or deletions of a partial repeat unit, may be distinguished from all other possible alleles.

Two possible multiplexing strategies for STR genotyping are illustrated in Figure 3. From a single punch of blood stained FTA™ paper, a multiplex PCR (simultaneously amplifying all STRs of interest) could be performed and then followed by a second-round PCR with primer sets that are closer to the repeat region to yield single or multiplexed STR products that are small enough for mass spec analysis. Alternatively, multiple punches could be made from a single blood stain on the FTA™ paper followed by singleplex or multiplex PCR with mass spec primers. After the genotype is determined for each STR locus in a sample, the information would be combined to form a single sample genotype for inclusion in CODIS or some other DNA database. This multiplexing approach permits flexibility for adding new STR loci or only processing a few STR markers across a large number of samples at a lower cost than processing extensive and inflexible STR multiplexes.

GeneTrace has developed automated genotyping software capable of processing mass spectral data and converting the mass information into a genotype at a rate of approximately one second per sample. This kind of sample processing speed is needed to maintain the capability of analyzing thousands of samples per day with robotic sample manipulation and time-of-flight mass spectrometry.

In addition to being amenable to automation and high volume sample processing, time-of-flight mass spectrometry is an excellent research tool for the development of new STR loci. Accurate allele calls may be made without allelic ladders meaning that genotype determination does not depend on the development of allelic ladders or the use of an internal sizing standard. The expected mass from a reference sequence is compared to the measured mass of an STR allele.

We have used a novel cleavable primer approach for multiplex SNP analysis of Y-chromosome SNP markers and mtDNA polymorphic sites. Multiplexing SNP markers may be achieved by using primers that are resolvable on a mass scale. With the approach described here, compatible primers with similar annealing temperatures may be used and cleavage sites may be placed at different positions in each primer. For example, one primer could have the cleavage site 5 bases from the 3'-end and another primer could have 8 bases between the cleavage site and the 3'-end. Thus, after performing the SNP extension reaction and cleaving the primers, each primer and extension product(s) can be easily resolved in the mass spectrometer. Figure 5 illustrates an SNP triplex from three Y SNP markers while Figure 6 demonstrates a 10plex SNP analysis from the mtDNA control region. These results have been confirmed by DNA sequence analysis (data not shown). In the mtDNA assay both strands of the PCR product are being probed and all four possible nucleotides are being detected simultaneously. As with multiplexed STR analysis, examining multiple SNP markers in the same reaction reduces time, labor, and cost compared to single reactions.

The cleavable primer approach for SNP analysis described here has a number of advantages over other existing technologies in terms of the molecular biology and the ability to be automated as well as the mass spectrometry detection. One feature of using a cleavable primer is the ability to probe both DNA strands simultaneously and still perform a solid-phase purification after a single PCR reaction. Because the 3'-end of the primer can be released following cleavage, the primer probe rather than the template strand may be biotinylated. Other primer extension methods, such as solid-phase minisequencing (11) and Genetic Bit Analysis™ (16), require that all of the primers used for probing SNP sites be on the same DNA strand because the template strand is captured and then made single-stranded through denaturation washes or digestion reactions. All four possible nucleotides can be probed simultaneously with our SNP approach, a fact that permits multiplexing any SNP marker as well as permitting clear determination of heterozygotes since all four bases are processed in the same reaction. Furthermore, as in all primer extension reactions, the specificity of the SNP determination is improved with primer hybridization and polymerase incorporation over just a hybridization event such as is performed with microchip arrays (8).

An advantage on the mass spectrometry side of the assay is the reduced size of the detected oligonucleotide, a fact that directly leads to higher sensitivity and resolution. Challenging heterozygotes that contain T and A and only differ by 9 Daltons can be readily resolved in the lower mass range (3). In addition, the reliability of making an SNP genotype determination is increased by having the primer present to act as an internal standard. Mass difference measurements are frequently more accurate in a mass spectrometer than absolute mass measurements by themselves. Most importantly, automated sample processing has been implemented on a robotic workstation so hundreds to thousands of samples may be processed in parallel each day per instrument.

ACKNOWLEDGMENTS

This work was supported by NIJ grant #97-LB-VX-0003. We thank Jia Li, Yuping Tan, Tom Shaler, Dan Pollart, Hua Lin, Christine Loehrlein, Kathy Stephens, Jon Marlowe, Vera Delgado, Wendy Lam, David Joo,

Gordon Haupt, Kevin Coopman, and Nathan Hunt for technical support and helpful discussions. Samples and sequence information from Dr. Peter Oefner and Dr. Peter Underhill of Stanford University are also greatly appreciated for the work with Y chromosome SNPs.

REFERENCES

1. Butler J.M., Li J., Monforte J., Becker C., Lee S. (1997) Rapid and Automated Analysis of Short Tandem Repeat Loci Using Time-of-Flight Mass Spectrometry. *Proceedings from the Eighth International Symposium on Human Identification* pp. 94-101.
2. Butler J.M., Li J., Shaler T.A., Monforte J.A., Becker C.H. (1998, in press) Reliable Genotyping of Short Tandem Repeat Loci Without an Allelic Ladder Using Time-of-Flight Mass Spectrometry. *Int. J. Legal Med*.
3. Li J., Butler J.M., Tan Y., Lin H., Royer S., Ohler L., Shaler T.A., Hunter J.M., Pollart D.J., Monforte J.A., Becker C.H. (1998, submitted) Single Nucleotide Polymorphism Determination Using Primer Extension and Time-of-Flight Mass Spectrometry. *Electrophoresis*.
4. Kimpton C.P., Gill P., Walton A., Urquhart A., Millican E.S., Adams M. (1993) Automated DNA Profiling Employing Multiplex Amplification of Short Tandem Repeat Loci. *PCR Meth. Appl.*, **3**:13-22.
5. Lazaruk K., Walsh P.S., Oaks F., Gilbert D., Rosenblum B.B., Menchen S., Scheibler D., Wenz H.M., Holt C., Wallin, J. (1998) Genotyping of Forensic Short Tandem Repeat (STR) Systems Based on Sizing Precision in a Capillary Electrophoresis Instrument. *Electrophoresis*, **19**: 86-93.
6. Landegren U., Nilsson M., Kwok P.-Y. (1998) Reading Bits of Genetic Information: Methods for Single-Nucleotide Polymorphism Analysis. *Genome Res.*, **8**:769-776.
7. Delahunty C., Ankener W., Deng Q., Eng J., Nickerson D.A. (1996) Testing the Feasibility of DNA Typing for Human Identification by PCR and an Oligonucleotide Ligation Assay. *Am. J. Hum. Genet.* **58**:1239-1246.
8. Wang D.G., et al. (1998) Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science*, **280**:1077-1082.
9. Butler J.M., Levin B.C. (1998) Forensic Applications of Mitochondrial DNA. *Trends Biotech*, **16**:158-162.
10. Stoneking M., Hedgecock D., Higuchi R.G., Vigilant L., Erlich H. (1991) Population Variation of Human mtDNA Control Region Sequences Detected by Enzymatic Amplification and Sequence-Specific Oligonucleotide Probes. *Am. J. Hum. Genet.*, **48**:370-82.
11. Tully G.., Sullivan K.M., Nixon P., Stones R.E., Gill P. (1996) Rapid Detection of Mitochondrial Sequence Polymorphisms Using Multiplex Solid-Phase Fluorescent Minisequencing. *Genomics* **34**:107-113.
12. Underhill P.A., Jin L., Lin A.A., Mehdi S.Q., Jenkins T., Vollrath D., Davis R.W., Cavalli-Sforza L.L., Oefner P.J. (1997) Detection of Numerous Y Chromosome Biallelic Polymorphisms by Denaturing High-Performance Liquid Chromatography. *Genome Res.* **7**:996-1005.
13. De Kniff P., et al. (1997) Chromosome Y Microsatellites: Population Genetic and Evolutionary Aspects. *Int. J. Legal Med.* **110**:134-140.
14. Haff L.A., Smirnov I.P. (1997) Single-Nucleotide Polymorphism Identification Assays Using a Thermostable DNA Polymerase and Delayed Extraction MALDI-TOF Mass Spectrometry. *Genome Res.* **7**:378-388.
15. Butler J.M., Ruitberg C.M., Reeder D.J. (1997) STRBase: a Short Tandem Repeat DNA Internet-Accessible Database. *Proceedings from the Eighth International Symposium on Human Identification* Promega Corporation, pp.38-47.
16. Nikiforov T.T., Rendle R.B., Goelet P., Roger Y.H., Kotewicz M.L., Anderson S., Trainor G.L., Knapp M.R. (1994) Genetic Bit Analysis: a Solid Phase Method for Typing Single Nucleotide Polymorphisms. *Nucleic Acids Res.*, **22**:4167-4175.

Table 1. Comparison of STRs and SNPs as Genetic Markers

Characteristics	**Short Tandem Repeats (STRs)**	**Single Nucleotide Polymorphisms (SNPs)**
Occurrence in Human Genome	~1 in every 15 kb	~1 in every 1 kb
General Informativeness	high	low (20-30% as informative as STRs)
Number of Alleles per Locus	typically >5	typically 2
Detection Methods	gel/capillary electrophoresis	Sequencing, PCR-RFLP, microchip hybridization, TaqMan probes, etc.
Multiplex Capability	>10 markers with multiple spectral channels	potential of 1000s on microchip
Mass Spectrometry Measurement	Measurement of PCR-amplified allele(s) mass	Mass difference measurement between primer and extension product(s)

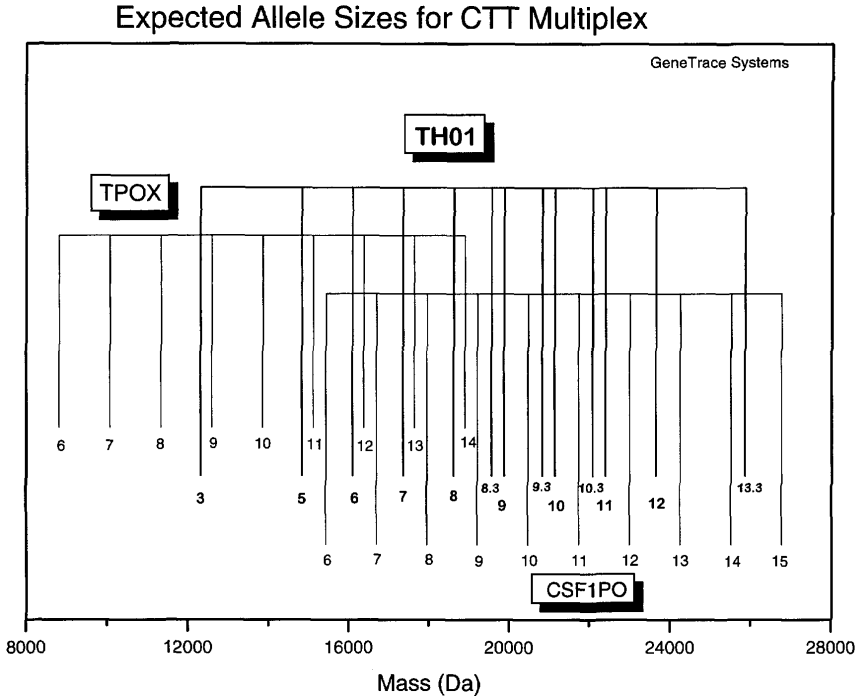

Figure 1. Schematic of expected allele masses for a CSF1PO-TPOX-TH01 (CTT) multiplex involving overlapping allele size ranges. All known alleles are fully distinguishable by mass with this interleaving approach.

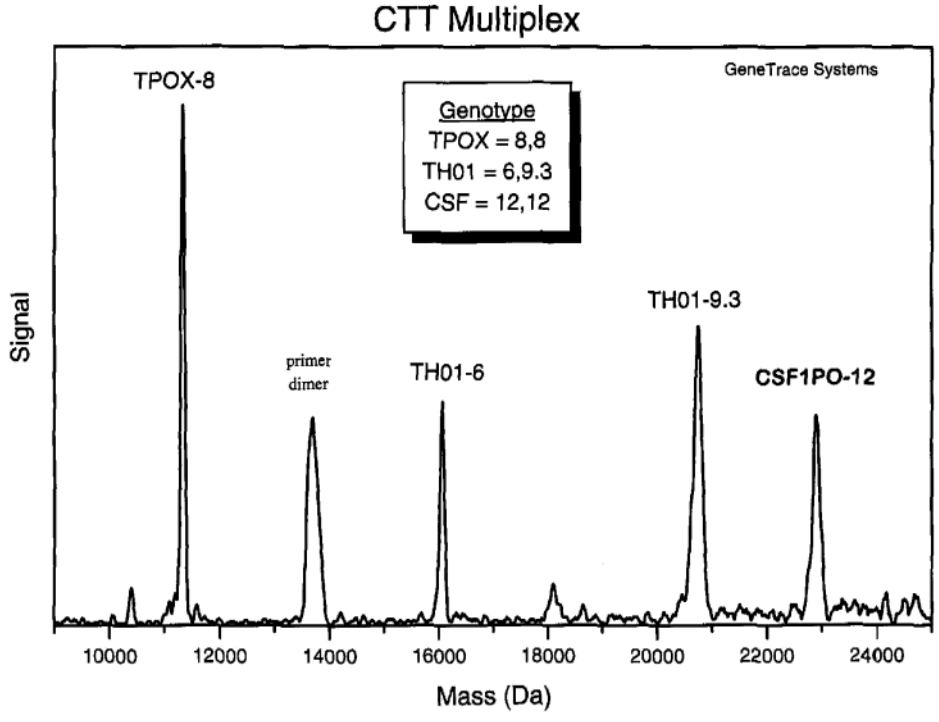

Figure 2. A mass spectrum of a CTT multiplex sample determined by our mass spectrometry approach.

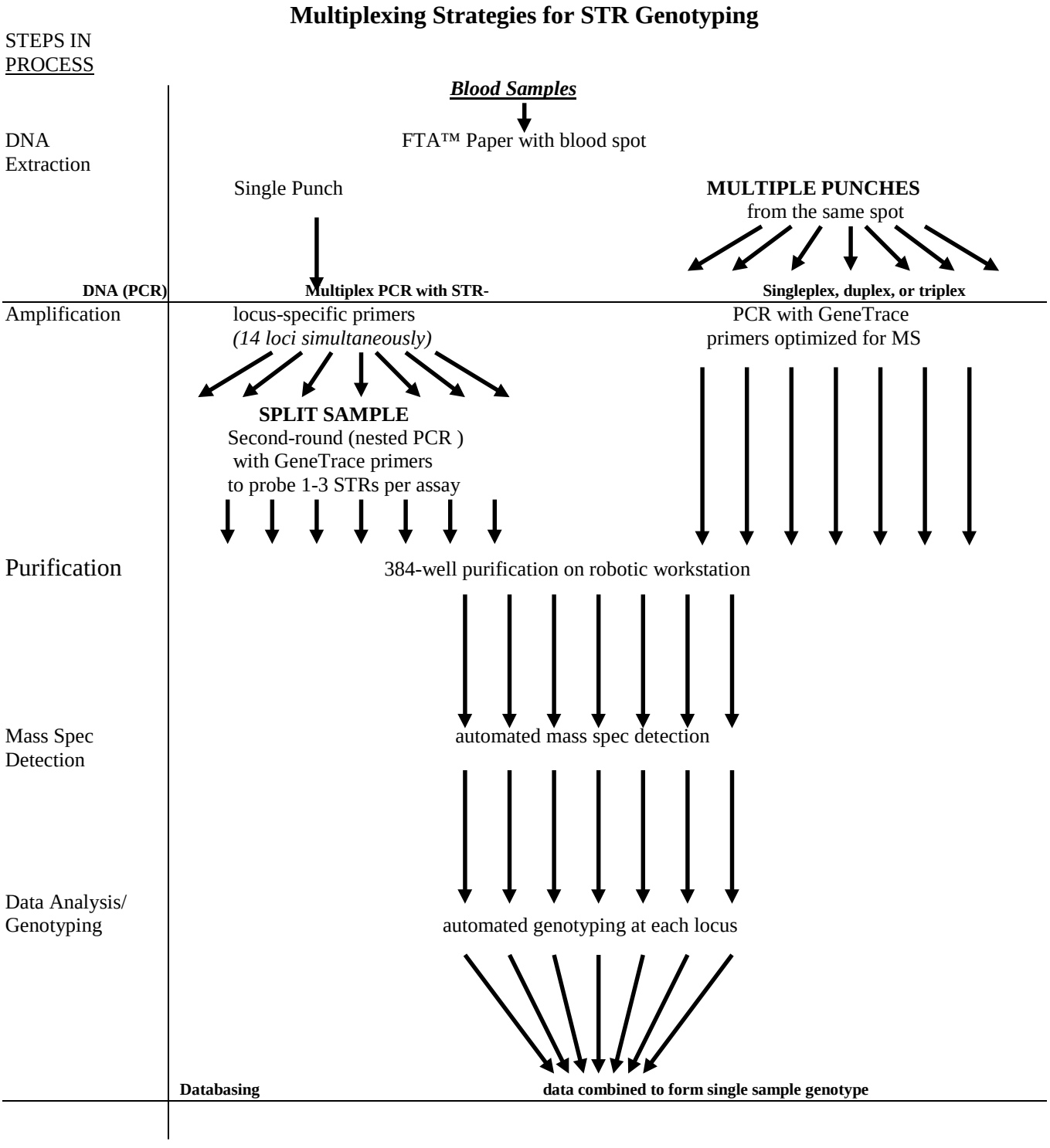

Figure 3. Multiplexing strategies for STR genotyping using a mass spectrometry approach.

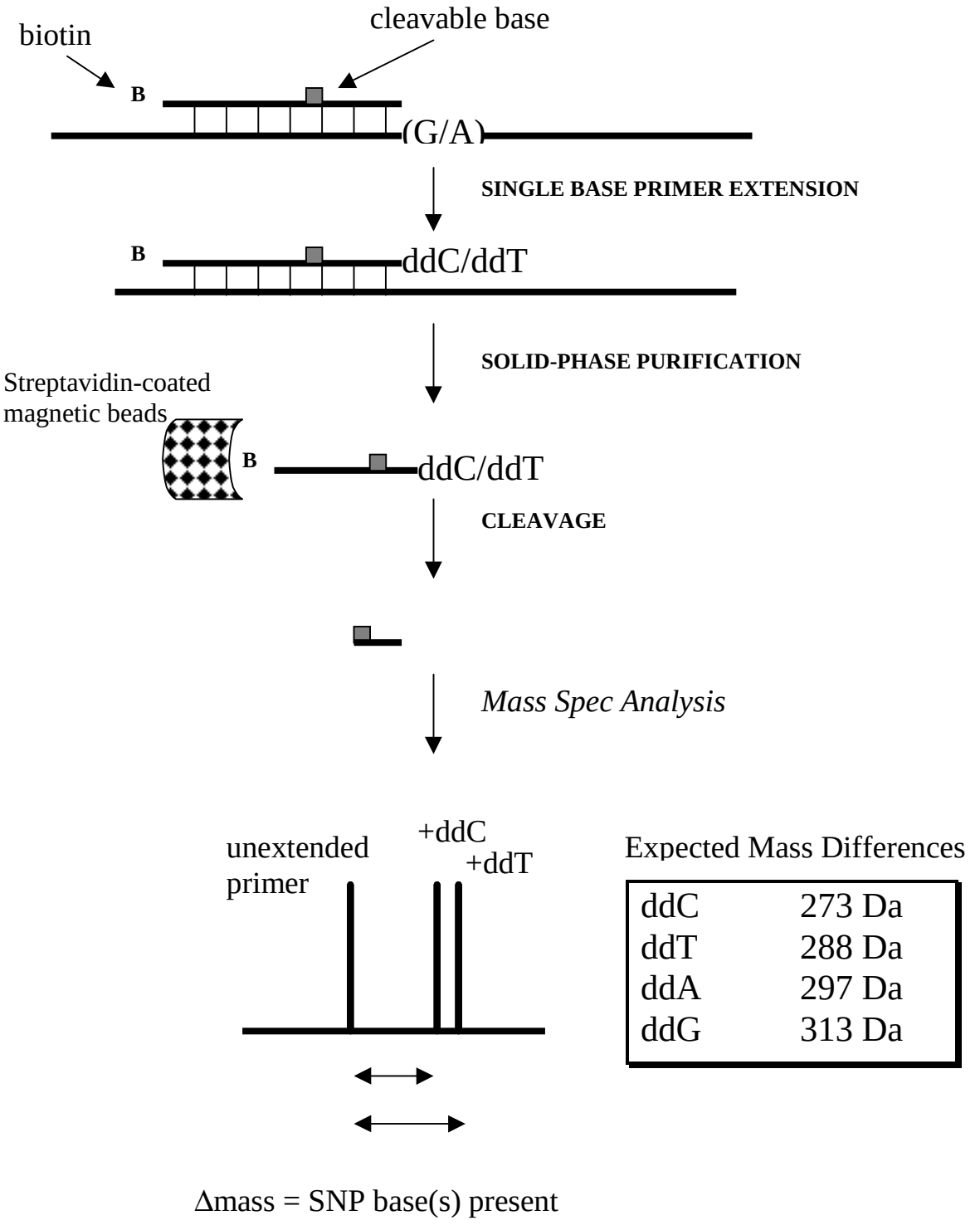

Figure 4. Schematic of SNP assay using a cleavable primer approach. An SNP primer hybridizes to the target DNA immediately upstream of the polymorphic site and a polymerase incorporates the complementary dideoxynucleotide in a single base extension reaction. The sample then undergoes a solid-phase purification via a capture and release protocol to prepare it for mass spec analysis. The mass difference between the primer and the extension product(s) indicates the nucleotide(s) that is present at the SNP site.

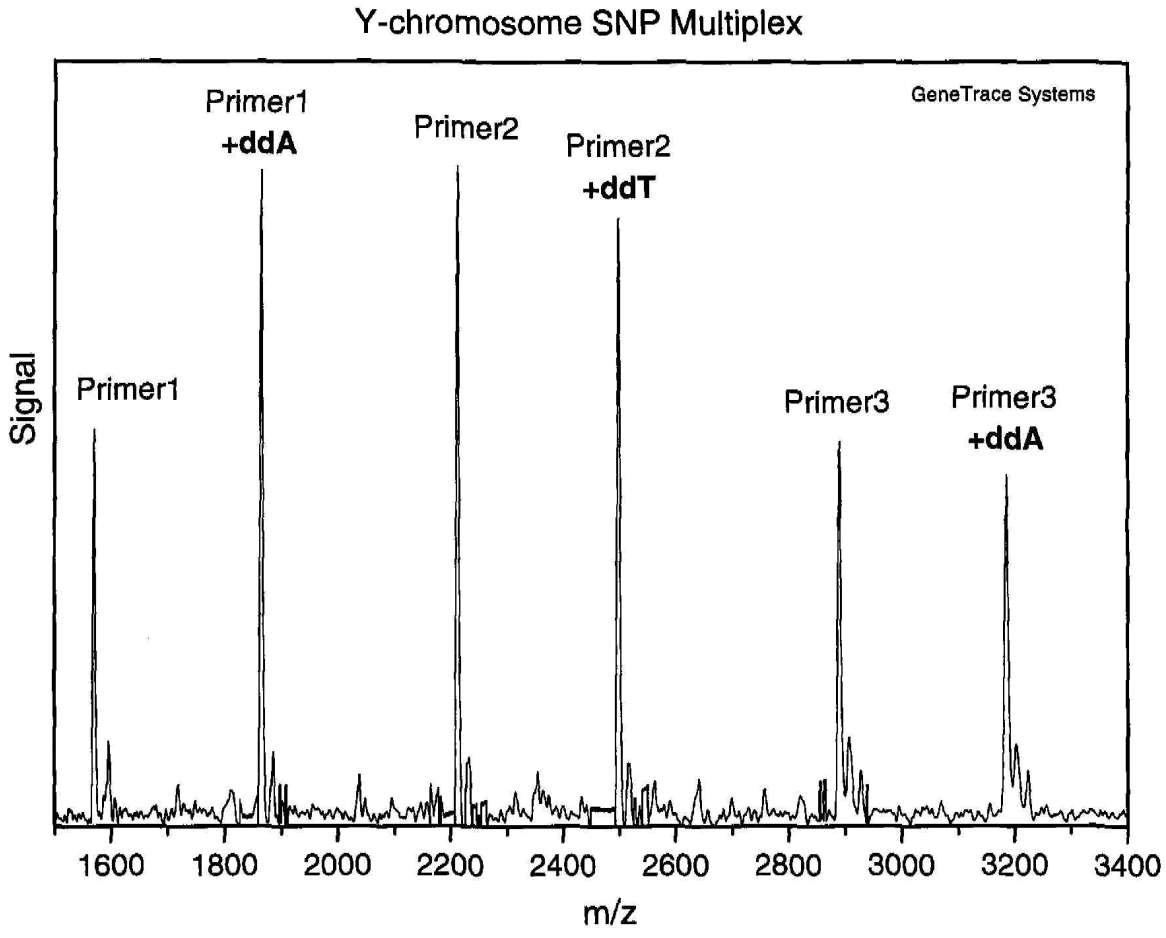

Figure 5. Simultaneously analysis of three Y chromosome SNP markers.

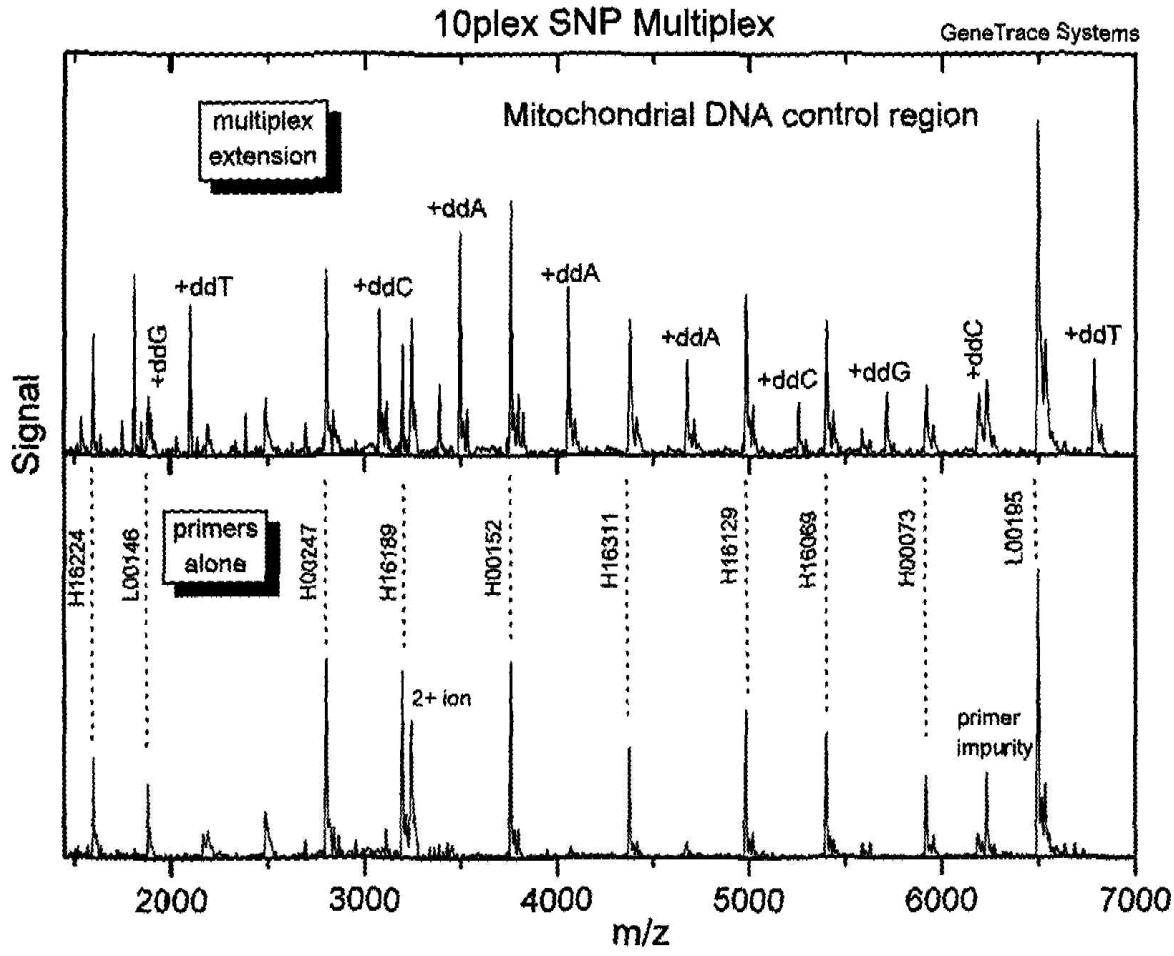

Figure 6. SNP 10plex for rapid determination of polymorphic sites in the mtDNA control region. The bottom panel shows the 10 primers by themselves. The top panel contains the multiplexed reaction products each labeled with the observed extension product. The results for this K562 PCR product are (in order across HV1 and HV2): H16069 (G), H16129 (C), H16189 (A), H16224 (G), H16311 (A), H00073 (C), L00146 (T), H00152 (A), L00195 (T), and H00247 (C). SNP nucleotide results have been confirmed by sequencing.