High-Throughput STR Analysis by Time-of-Flight Mass Spectrometry

John M. Butler, Kathryn M. Stephens, Joseph A. Monforte, Christopher H. Becker
GeneTrace Systems Inc., 1401 Harbor Bay Parkway, Alameda, CA 94502

ABSTRACT

Rapid, cost-effective methods for high-throughput DNA analysis are needed to process samples currently being gathered for large criminal DNA databases around the world. Within the U.S., several states have sample backlogs of over 50,000 samples with limited funds and manpower to analyze these samples. Currently available slab gel or capillary electrophoresis instruments can handle only a few dozen samples per day. Using time-of-flight mass spectrometry coupled with parallel sample preparation on a robotic workstation, we can process thousands of samples daily with a single mass spectrometer. Analysis times on the order of 10 seconds per sample with improved accuracy for sizing short tandem repeat (STR) alleles compared to electrophoresis methods have been demonstrated.

Following polymerase chain reaction (PCR) amplification of DNA templates, a proprietary solid-phase purification procedure is used to prepare the samples for mass analysis. We are currently working with over 20 tetranucleotide repeat DNA markers of forensic interest including HUMTH01, vWA, TPOX, CSF1PO and the sex-typing marker amelogenin. Mass difference measurements have been shown to benefit STR genotyping by permitting analysis of nontemplated addition, repeat content identification, and repeat spread measurements in heterozygotes. In short, time-of-flight mass spectrometry offers rapid and reliable genotyping of short tandem repeat DNA markers used in human identity testing.

INTRODUCTION

As of early 1998, over a half-million samples have been collected from convicted felons in 48 of the 50 states in the U.S. These samples are being stored in anticipation of future analysis and inclusion in the Combined DNA Index System (CODIS). The FBI has designated 13 core STR loci for the nationwide 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. Using current fluorescent technology, an estimated five to six years will be needed to process these samples at an enormous cost and commitment of labor resources.

Time-of-flight mass spectrometry (TOF-MS) offers a rapid, cost-effective alternative for genotyping large numbers of samples (1). Each DNA sample can be accurately measured in a few seconds. In this process, commonly referred to as MALDI (matrix-assisted laser desorption/ionization), DNA samples are mixed with an organic matrix and allowed to co-crystallize in a spatial array on a sample plate with each assay at a separate location. After the sample plate is placed in the mass spectrometer, which is under vacuum, a pulse of laser energy liberates a small portion of the DNA sample (Figure 1). While the generated ions travel to the detector in a matter of microseconds, multiple spectra are averaged for signal processing, which extends the measurement time to a few seconds. The DNA size is calculated by the time-of-flight to the detector in comparison to mass standards. Due to the increased accuracy with mass spectrometry, STR alleles may be reliably typed without comparison to allelic ladders (2). An absolute mass is measured with mass spectrometry rather than a relative mobility measurement (in comparison to DNA sizing standards) as in an electrophoretic analysis. GeneTrace-designed genotyping software then correlates the observed peak mass back to a genotype based on expected allele masses obtained from a reference sequence, the PCR primer positions, and the repeat unit mass. Each sample can be processed and genotyped in approximately one second using a standard desktop personal computer.

Two issues that impact mass spec results are DNA size and sample salts. Mass spec resolution and sensitivity are diminished when either the DNA size or the salt content of the sample is too large. By redesigning the PCR primers to bind close to the repeat region, the STR allele sizes are reduced so that resolution and sensitivity of the PCR products are benefited. Where possible, we design our primers to produce amplicons that are less than 100 bp although we have been able to resolve neighboring STR alleles that are as large as 140 bp in size. To overcome the sample salt problem, we use a patented solid-phase purification procedure that reduces the concentration of magnesium, potassium, and sodium

salts in the PCR products prior to their being introduced to the mass spectrometer (3). Without the reduction of the salts, resolution is diminished by the presence of adducts. Salt molecules bind to the DNA during the MALDI ionization process and give rise to peaks that have a mass of the DNA molecule plus the salt molecule. Adducts broaden peaks and thus reduce resolution. Our sample purification procedure, which has been entirely automated on a 96-tip robotic workstation, reduces the PCR buffer salts and yields "clean" DNA for the mass spectrometer. Using our robotic workstation in combination with a single high-throughput mass spectrometer, we have been able to purify and analyze over 2,000 samples in a single day.

MATERIALS AND METHODS

DNA Templates: Human genomic DNA from K562 cell line (Promega) and samples from several ethnic groups (Bios Laboratories, New Haven, CT). A set of 88 samples was provided by the California Department of Justice DNA Laboratory in Berkeley, CA.

Allelic Ladders: Profiler (PE Applied Biosystems) and PowerPlex™ (Promega) allelic ladders were diluted 1:1000 in deionized water and reamplified using GeneTrace-designed PCR primers

PCR Mix: 1X STR Buffer (Promega) or PCR buffer II for TaqGold polymerase (PE Applied Biosystems), 1 μM GeneTrace-designed primers, 1 U *Taq* (Promega) or Ampli*Taq* Gold (PE Applied Biosystems), 1-10 ng DNA template.

PCR Conditions: 2 min 94° C (11 min 95° C with TaqGold); 30-35 cycles of 30 sec each: 94° C, 55° C, 72° C; 5 min 72° C.

Sample Purification: A purification procedure involving solid-phase capture and release from streptavidin-coated magnetic beads was utilized (3) to purify the DNA. Parallel sample preparation was conducted on a robotic workdeck operated with a 96-pipet tip head designed by GeneTrace.

Mass Spectrometry: Each DNA sample was spotted with a 3-hydroxypicolinic acid matrix. A GeneTrace-designed linear time-of-flight instrument was used (4). The mass spectrometer was calibrated daily with two oligonucleotide mass markers (2).

Genotyping: Expected masses for alleles at each STR locus were calculated using GenBank sequences and information from STRBase (http://ibm4.carb.nist.gov:8800/dna/home.htm; ref 5). An additional mass of 313.2 Da was added to each allele to account for the nontemplate addition of adenine (6). Genotypes were assigned by comparing the measured mass to all expected allele masses. Typically mass bins of ~100 Da were used for each STR allele based on a standard deviation of 30 Da (2).

RESULTS AND DISCUSSION

Testing of new STR primers

New PCR primers were designed for each STR locus with sequence information from GenBank (http://www.ncbi.nlm.nih.gov). The PCR products produced from these primers are much smaller than commonly used from commercially available STR kits because the primers are closer to the repeat region (Table 1). We successfully designed and tested primers from the following commonly used STR loci: CD4, CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, DYS19, F13A1, F13B, FES/FPS, FGA, HPRTB, LPL, TH01, TPOX, and vWA as well as the sex-typing marker amelogenin. In addition we examined three new tetranucleotide STR loci: GATA132B04 (chromosome 5), D16S2622, and D22S445. To improve the sensitivity and resolution in the mass spectrometer, primers were placed close to the repeat region to make the PCR product size ranges under 140 bp in size when possible. In the case of CD4, LPL, and amelogenin, previously published primers were used. Primers were purchased from Biosource/Keystone (Menlo Park, CA) or synthesized in-house. One primer in each locus-specific set was biotinylated at the 5'-end for use in a solid-phase purification procedure (3).

The new primers were tested with 10 ng of K562 DNA template and the universal PCR conditions listed in the materials and methods section. In all cases, the expected K562 genotypes were obtained for the STR loci tested. We then proceeded to test samples from different ethnic groups to verify that the primers worked on a variety of DNA templates. Mass spectra from several samples for the D7S820 locus are shown in Figure 2. We have successfully detected PCR products from all 23 STR markers tested.

Validation Tests

A number of analytical tests have been performed to verify sensitivity, accuracy, precision, and resolution for STR analysis by mass spectrometry (2). For example, using our TPOX primers, we have been able to detect as little as 200 pg of genomic DNA from a serial dilution of

K562 DNA (no attempts were made to go below this quantity of DNA due to possible stochastic effects). The precision of measurements over multiple samples is approximately 30 Da or 0.1 nucleotides for each allele while we have obtained mass accuracies better than 6 Da or 0.02 nucleotides without using allelic ladders. However, we have found allelic ladders to be useful in demonstrating resolution with tetranucleotide STR alleles (Figure 3). In a previous publication, we demonstrated the ability to completely separate the peaks of a single base microvariant from a full allele with the separation of a TH01 allelic ladder containing alleles 9.3 and 10 (2).

As part of our validation testing for this new mass spec technology, we have compared our genotyping results with those obtained from an established, validated technique accepted in forensic DNA laboratories. The California Department of Justice Berkeley DNA Laboratory supplied 88 genomic DNA samples that had been previously typed with an ABI 310 capillary electrophoresis method and primers from the PE Applied Biosystems' Profiler STR multiplex kit. We have successfully typed these samples at several STR loci. In all cases where we obtained a result, we have made the correct genotype call. For example, with the TH01 locus, we obtained successful results in 82 out of the 88 samples on the first pass through the samples. The remaining six samples yielded a genotype on the second attempt for those samples. Most importantly, there was a 100% correlation for all of the genotypes obtained with the two methods.

Multiplex STR Analysis

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 (7). Due to the size constraints of mass spectrometry, 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. The high accuracy, precision, and resolution of our mass spec approach permit multiplexing STR loci in such a manner.

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 4a. All known alleles for these STR loci, as defined by STRBase (5), 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 4b.

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 (8). 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.

The Value of Mass Spectrometry for DNA Database Work

Our automated DNA sample preparation and time-of-flight mass spectrometry approach to STR analysis make rapid development of large DNA databases feasible. We recently demonstrated that over 2,000 samples could be analyzed in a single day using a single 96-tip robotic workstation and a single mass spectrometer. GeneTrace has built a state-of-the-art facility in Alameda, California, to process large numbers of DNA samples using this high-throughput approach. In the near future, we plan to offer a DNA typing service for development of large DNA databases such as will be required for the national CODIS system in the United States. The more quickly the samples can be analyzed and placed in the U.S. national CODIS DNA database or other forensic DNA databases, the more quickly repeat offenders may be brought to justice. Time-of-flight mass spectrometry

offers an effective and efficient solution to rapid DNA database development.

ACKNOWLEDGMENTS

We thank Jia Li, Tom Shaler, Dan Pollart, Joanna Hunter, Hua Lin, Yuping Tan, Mike Abbott, Gordon Haupt, and Christine Loehrlein for technical assistance and helpful discussions with this work. Steve Lee and John Tonkyn of the California Department of Justice were also helpful with providing samples for testing purposes. This research was supported in part by a grant from the National Institute of Justice (97-LB-VX-0003).

REFERENCES

1. Monforte J.A., Becker C.H. (1997) High-Throughput DNA Analysis by Time-of-Flight Mass Spectrometry. *Nature Med.* **3**:360-362.

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. Monforte J.A., Becker C.H., Shaler T.A., Pollart D.J. (1997) Oligonucleotide Sizing Using Immobilized Cleavable Primers. *US Patent* **5,700,642**.

4. Wu K.J., Shaler T.A., Becker C.H. (1994) Time-of-Flight Mass Spectrometry of Underivatized Single-Stranded DNA Oligomers by Matrix-Assisted Laser Desorption. *Anal. Chem.* **66**: 1637-1645.

5. Butler J.M., Ruitberg C.M., Reeder D.J. (1998) STRBase: a Short Tandem Repeat DNA Internet-Accessible Database. In: Proceedings from the Eighth International Symposium on Human Identification; 1997 Sept 17-20; Scotsdale (AZ), Promega Corporation, pp.38-47.

6. Clark J.M. (1988) Novel Non-Templated Nucleotide Addition Reactions Catalysed by Procaryotic and Eucaryotic DNA Polymerases. *Nucleic Acids Res.*, **16**:9677-9686.

7. 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.

8. 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.

Table 1. PCR product sizes for STR alleles using GeneTrace-designed primers compared to commercially available primers in STR multiplex sets.

STR Locus	Known Alleles	**GeneTrace Sizes**	Commercially Available Sizes*
Amelogenin	X, Y	106, 112 bp	107,113 bp; 212,218 bp
CD4	4—15	81-136 bp	Not available
CSF1PO	6—15	87-123 bp	291-327 bp
F13A1	3—17	112-168 bp	279-335 bp
F13B	6—12	110-134 bp	169-193 bp
FES/FPS	7—15	76-108 bp	222-254 bp
FGA	15—30	118-180 bp	206-267 bp
D3S1358	9—20	85-129 bp	114-142 bp
D5S818	7—15	89-121 bp	119-151 bp
D7S820	6—14	66-98 bp	212-244 bp
D8S1179	8—18	92-130 bp	128-168 bp
D13S317	7—15	98-130 bp	165-197 bp
D16S539	5,8—15	59-99 bp	264-304 bp
D18S51	9-27	120-192 bp	273-341 bp
D21S11	24-38	150-190 bp	189-243 bp
DYS19	8—16	76-108 bp	Not available
HPRTB	6—17	84-128 bp	259-303 bp
LPL	7—14	105-133 bp	105-133 bp
TH01	3--13.3	55-98 bp	171-214 bp
TPOX	6—14	69-101 bp	224-256 bp
vWA	11—22	126-170 bp	157-201 bp
Other STRs			
GATA132B04	10—14	99-115 bp	Not available
D22S445	10—16	110-130 bp	Not available
D16S2622	4—8	71-87 bp	Not available

* Commercial sources include PE Applied Biosystems and Promega Corporation

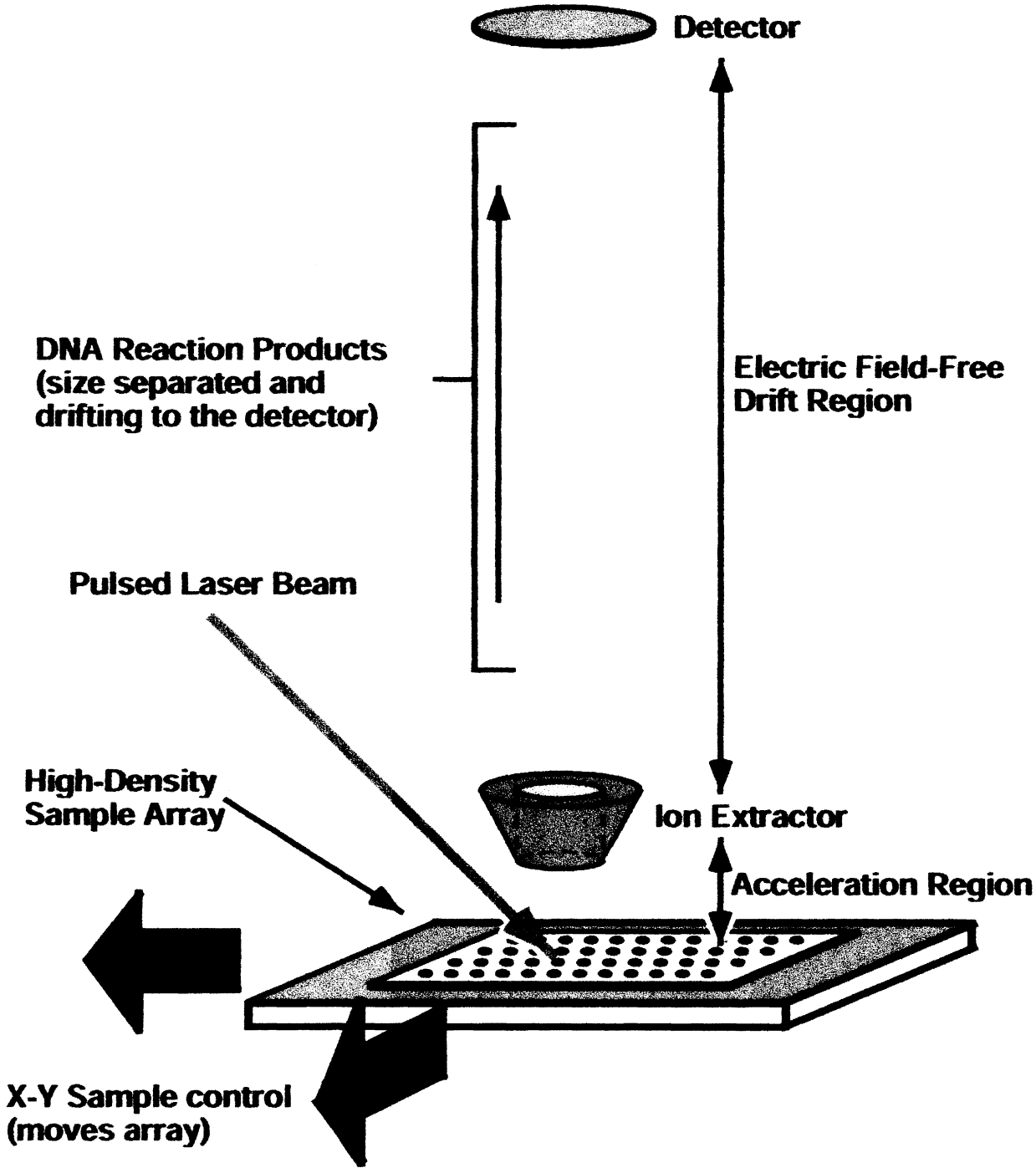

Figure 1. Schematic diagram illustrating the principal components of DNA analysis by time-of-flight mass spectrometry. Each STR sample is imbedded within a matrix compound in a ~1-mm diameter spot located on the sample plate. A pulsed laser beam liberates the DNA from the matrix for mass analysis. STR alleles are separated by their time-of-flight to the detector, a process that occurs in a few hundred microseconds.

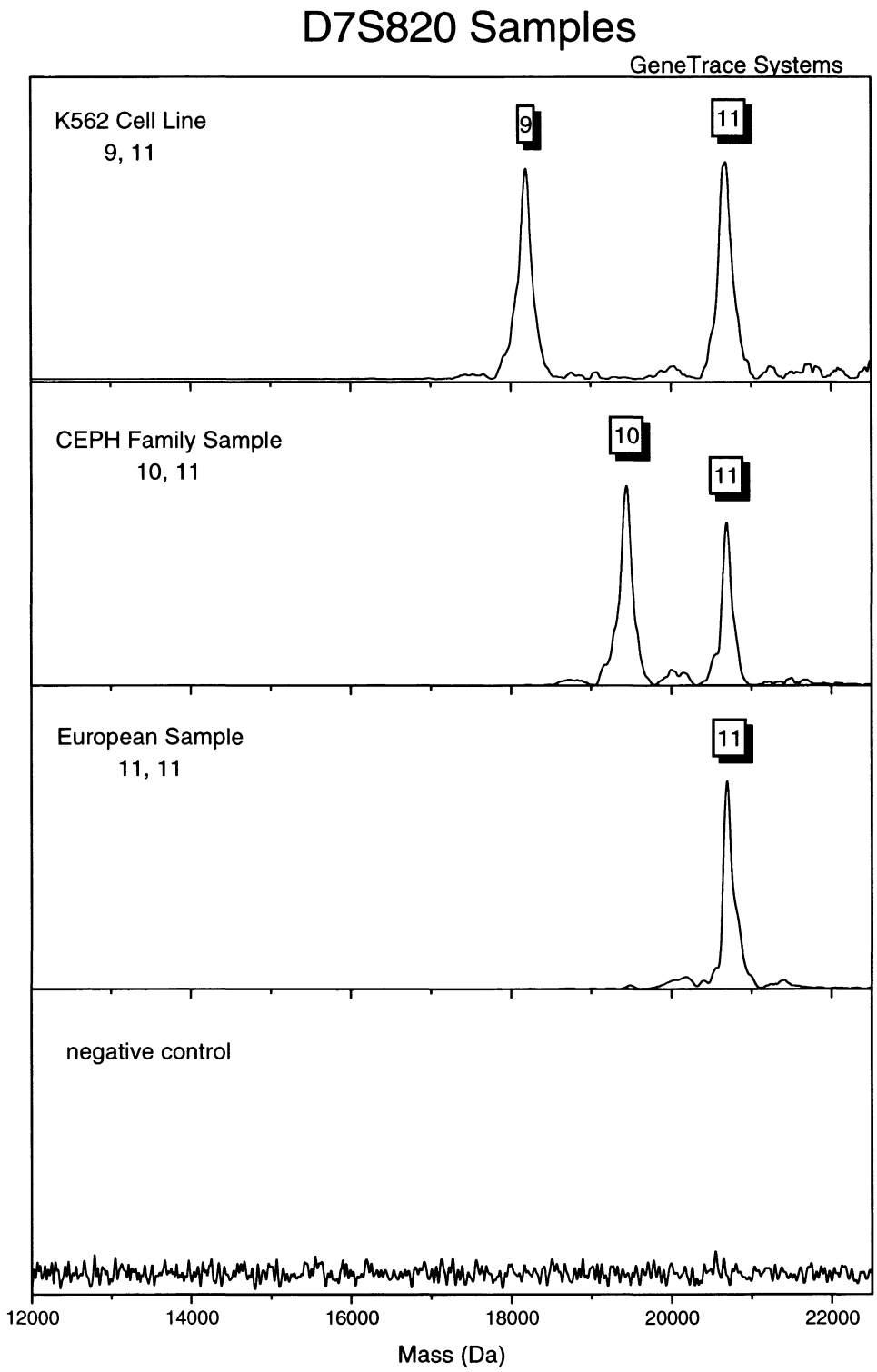

Figure 2. Mass spectra from several genomic DNA samples amplified at the D7S820 STR locus.

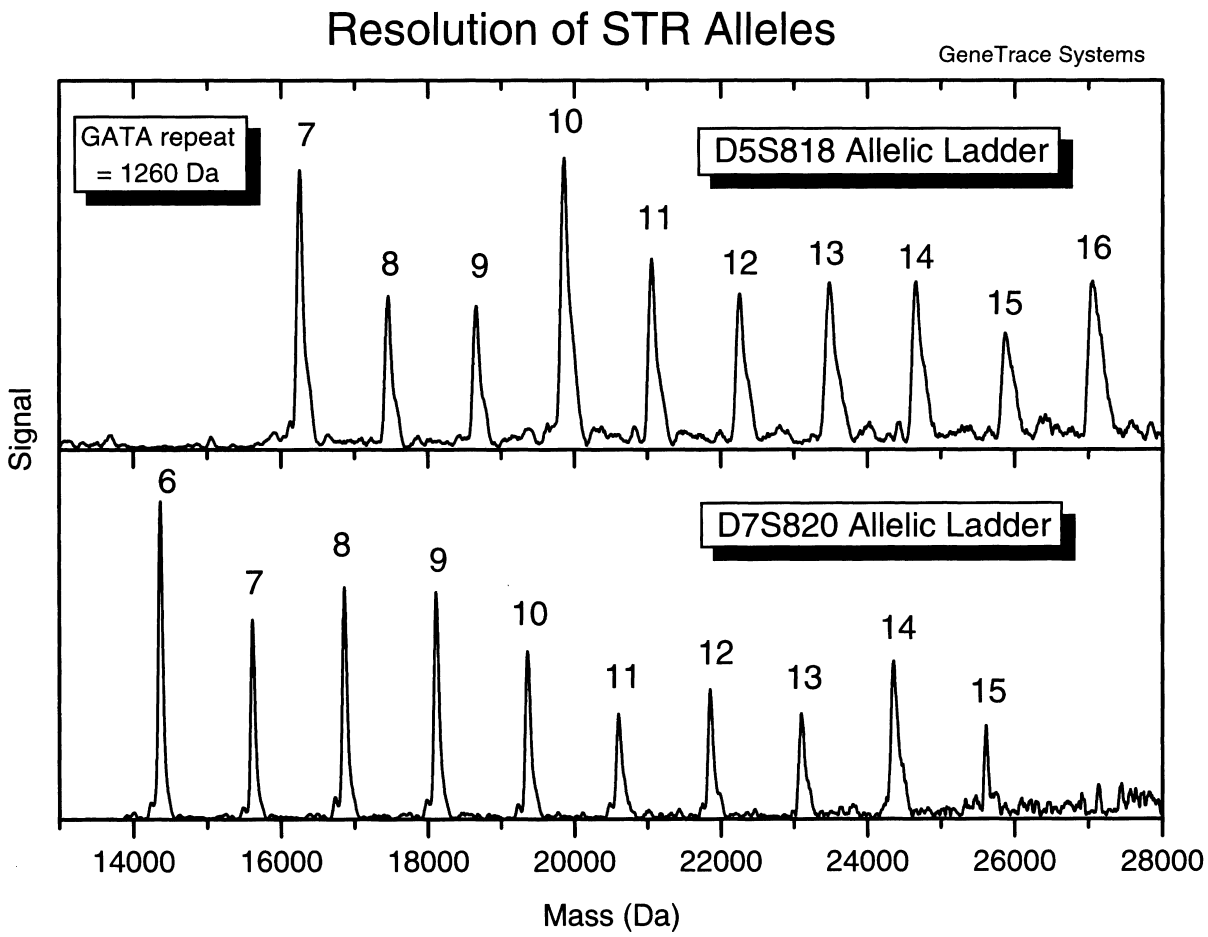

Figure 3. Resolution of tetranucleotide STR alleles demonstrated with D5S818 and D7S820 allelic ladders. The allelic ladders were reamplified from diluted PE Applied Biosystems AmpFlSTR Yellow allelic ladders using GeneTrace-designed primers and PCR conditions as described in the Materials and Methods section.

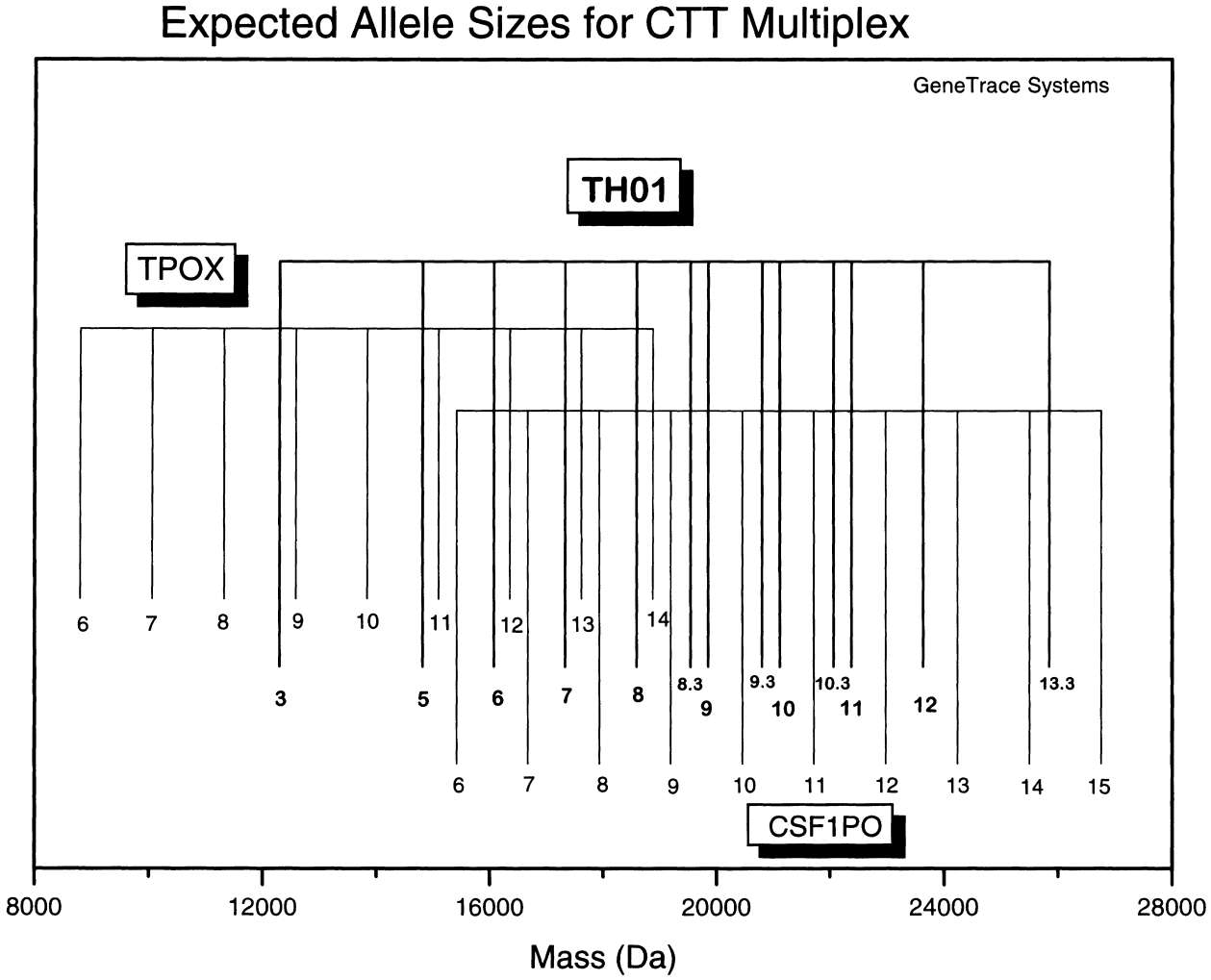

Figure 4. (a) Expected allele masses for a CSF1PO-TPOX-TH01 (CTT) multiplex involving overlapping allele size ranges. All known alleles are fully distinguishable by mass with this interleaving approach.

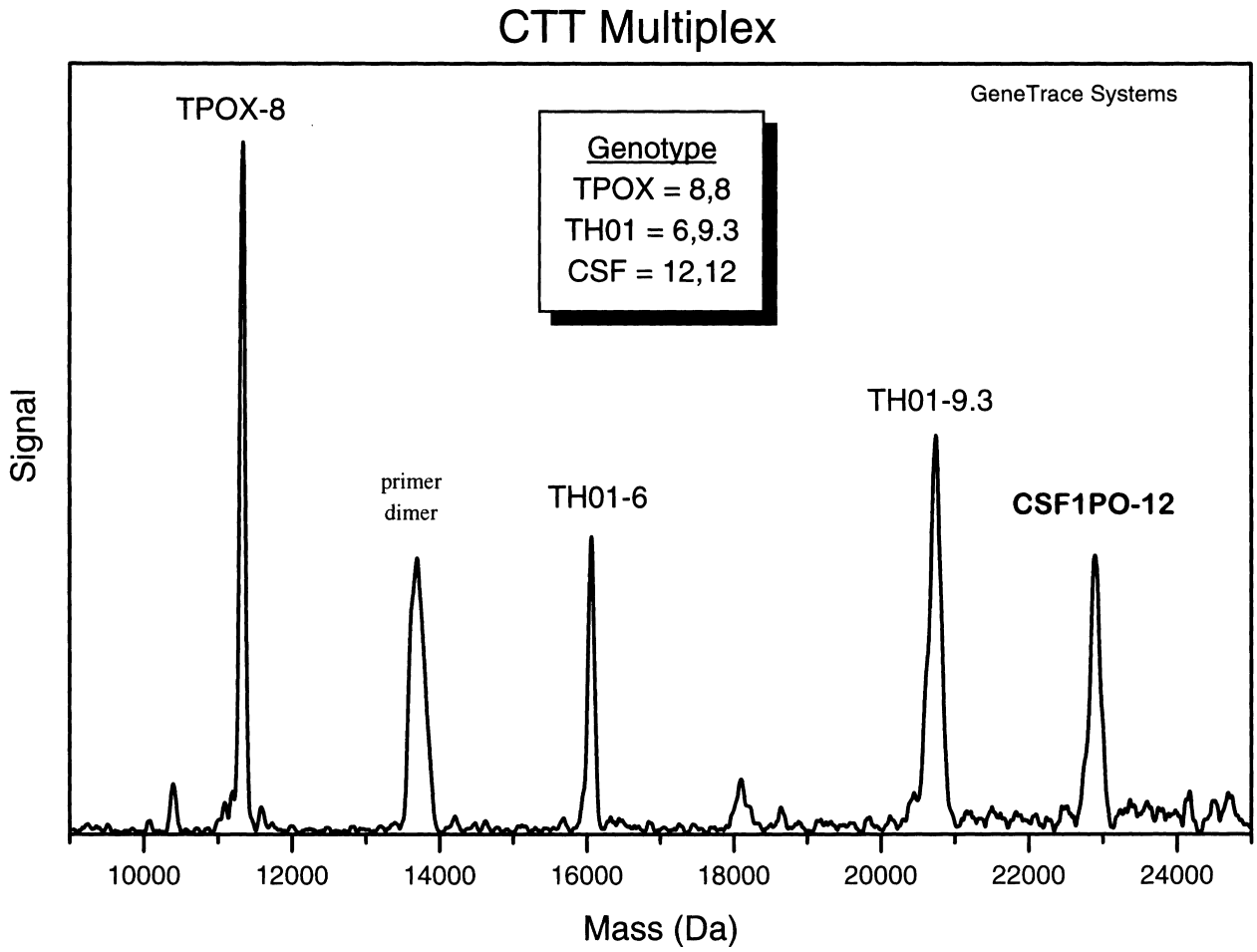

Figure 4 (b) A mass spectrum of a CTT multiplex sample.