High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing

Xiangpei Zeng [a], Jonathan L. King [a], Monika Stoljarova [a], David H. Warshauer [a], Bobby L. LaRue [a], Antti Sajantila [a,b], Jaynish Patel [c], Douglas R. Storts [c], Bruce Budowle [a,d,*]

[a] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[b] Department of Forensic Medicine, Hjelt Institute, University of Helsinki, P.O. Box 40, 00014 Helsinki, Finland
[c] Promega Corporation, 2800 Woods Hollow Road, Madison, WI 53711, USA
[d] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

A B S T R A C T

STR typing in forensic genetics has been performed traditionally using capillary electrophoresis (CE). However, CE-based method has some limitations: a small number of STR loci can be used; stutter products, dye artifacts and low level alleles. Massively parallel sequencing (MPS) has been considered a viable technology in recent years allowing high-throughput coverage at a relatively affordable price. Some of the CE-based limitations may be overcome with the application of MPS. In this study, a prototype multiplex STR System (Promega) was amplified and prepared using the TruSeq DNA LT Sample Preparation Kit (Illumina) in 24 samples. Results showed that the MinElute PCR Purification Kit (Qiagen) was a better size selection method compared with recommended diluted bead mixtures. The library input sensitivity study showed that a wide range of amplicon product (6–200 ng) could be used for library preparation without apparent differences in the STR profile. PCR sensitivity study indicated that 62 pg may be minimum input amount for generating complete profiles. Reliability study results on 24 different individuals showed that high depth of coverage (DoC) and balanced heterozygote allele coverage ratios (ACRs) could be obtained with 250 pg of input DNA, and 62 pg could generate complete or nearly complete profiles. These studies indicate that this STR multiplex system and the Illumina MiSeq can generate reliable STR profiles at a sensitivity level that competes with current widely used CE-based method.

© 2014 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Short tandem repeats (STRs) are the primary genetic markers used in forensic DNA human identification testing, due to their high discrimination power and relatively short amplicon size. For most analyses, a minimum of 15 loci are typed using commercially available kits. These kits enable multiplex amplification and generate amplicons less than 500 bp in size [1–4]. Although it would be desirable to have a multiplex system with similar and relatively short amplicon sizes, capillary electrophoresis (CE) detection limits that possibility. Those loci labeled with the same fluorescent tag must be separated by size to distinguish one from another. In addition, CE-based methods have other limitations, such as the number of STR loci (25–30 maximum) that can be typed in a multiplex system with current spectral capabilities and resolution, and output data cannot distinguish stutter products from alleles from a minor contributor(s) in mixtures [5,6]. Alternative approaches for sizing STR alleles include mass spectrometry with electrospray ionization or by matrix-associated laser desorption/ionization time-of-flight [7–9]. Advantages of mass spectrometry include no fluorescent dyes are required for detection, allelic ladders are not needed, and analysis is much faster than that of CE-based methods. However, mass spectrometry has not been widely implemented for forensic genetic analyses primarily due to its limited multiplex capacity and thus greater sample consumption to attain the same level of discrimination as CE-based methods.

Massively parallel sequencing (MPS) technology has the capability to sequence many targeted regions of multiple nucleic acid samples simultaneously with high coverage [10,11]. MPS

platforms allow for higher throughput sequencing when compared with Sanger sequencing at a substantially reduced cost per nucleotide. For example, 2–96 different samples can be sequenced simultaneously using commercial barcoding kits, such as Ion Xpress Barcode kit (Thermo Fisher) and Nextera XT Index kit (Illumina, San Diego, CA, USA) [12,13]. With the advent of MPS, some of the CE-based limitations may be overcome. Because locus assignment is not based on size, as is required by CE methods, STR amplicons can be engineered to be in general shorter and more equal in length, which makes them better suited for analyzing challenged samples. The amplicons can be more similar in size because sequence itself is used to distinguish a locus, not fluorescence and amplicon position after CE. MPS generates both allele lengths that are consistent with current STR data and the DNA sequences of alleles, which potentially can be used to increase discrimination power. Additionally, sequence variants may provide information to distinguish stutter products from minor contributor alleles in mixtures.

Several studies have been conducted to sequence STRs using MPS. Bornman et al. described a method using the Illumina GAIIx platform for accurately typing the 13 CODIS STR loci and Amelogenin locus from single individuals and one mixture sample [14]. They found that at least 18,500 reads were required to genotype an individual with 99% confidence for all CODIS loci and results were concordant with those generated using the PowerPlex 16 kit (Promega Corporation, Madison, WI). However, typing of larger amplicon alleles was somewhat limited as they employed a 150 base single end read length chemistry. Warshauer et al. [15] and Van Neste et al. [16] demonstrated that longer read lengths could overcome the allele drop out observed by Bornman et al. [14]. In addition, Fordyce et al. [17] and Scheible et al. [18] described STR typing was feasible on the 454 platform. All these studies indicate that multiplex STR typing by MPS holds promise for forensic applications. However, to date, a high sensitivity detection multiplex system with MPS had not been described. In this paper, a prototype multiplex STR System (Promega) was evaluated and demonstrated that low input target DNA can be sequenced and that the performance rivals that of current CE-based systems. The multiplex is a subset of the PowerPlex Fusion System (Promega), containing the CODIS 13 core loci, the Penta D, Penta E, D2S1338, and D19S433 loci, and the Amelogenin locus. The results described herein address amplicon size selection, library input sensitivity, and target input for sensitivity of detection and support that a sensitive STR typing system is achieved with MPS.

2. Materials and methods

2.1. Sample preparation

Whole blood was obtained from 24 volunteers with informed consent. All samples were anonymized to ensure the privacy of the contributing subjects in accordance with University of North Texas Health Science Center's Institutional Review Board. The DNA was extracted using the QIAamp® DNA Blood Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol [19]. The quantity of extracted DNA was estimated using the Quantifiler Human DNA Quantification Kit (Thermo Fisher, Foster City, CA, USA) on an ABI Prism 7500 Sequence Detection System (Thermo Fisher) following the manufacturer's protocol [20].

2.2. PCR amplification

Amplification of the template DNA was accomplished using the prototype multiplex STR system (Amelogenin, CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, FGA, Penta D, Penta E, TH01, TPOX, and vWA)

under the following conditions [21]. The PCR included Prototype NGS 5X Primer Pair Mix, PowerPlex® Fusion 5X Master Mix, Amplification Grade Water, and 16–500 pg genomic DNA. Amplification was performed on a GeneAmp® 9700 PCR System (Thermo Fisher) using the following thermal-cycling parameters: 1 min at 96 °C for polymerase activation; 30 cycles of 10 s at 94 °C for denaturation, 1 min at 59 °C for primer annealing, 30 s at 72 °C for primer extension; followed by a final extension of 10 min at 60 °C. Amplified products were purified using the MinElute PCR Purification Kit (Qiagen) and quantified using the Qubit dsDNA BR kit (Thermo Fisher) according the manufacturers' protocols [22,23].

2.3. Library preparation

Libraries were prepared using the TruSeq DNA LT Sample Preparation Kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol unless otherwise noted [24]. Two methods were used for size selection: the manufacturer-recommended diluted bead mixture (sample purification bead/sample purification bead + PCR grade water) and the MinElute PCR Purification Kit (Qiagen). The quantity of each indexed DNA library was determined using the Qubit dsDNA BR kit or Qubit dsDNA HS kit (Thermo Fisher) and normalized for sequencing according to the manufacturer's protocol [23,25]. Samples were indexed with up to 24 different indices provided in the TruSeq DNA LT kit.

2.4. MPS sequencing and data analysis

Each indexed DNA library (up to 24 unique indices) was normalized to 2 nM and then pooled. Pooled libraries were diluted to 10 pM and sequenced with the MiSeq v2 (2×250 bp) chemistry (Illumina). The MiSeq re-sequencing protocol for small genome sequencing was carried out according to the manufacturer's recommendation [26]. Base calls and quality scoring were performed using the on-board software – Real Time Analysis. The MiSeq output was converted automatically to the FASTQ format by the MiSeq Reporter software. Raw FASTQ files were exported from the MiSeq and analyzed using STRait Razor [15]. The depth of coverage (DoC) of each allele and allele coverage ratio (ACR), i.e., the lower coverage allele divided by the higher coverage allele, were calculated for each STR locus per sample.

2.5. Size selection

In the first phase of the size selection study, five diluted bead mixture ratios (BMRs) (0.60, 0.65, 0.70, 0.75, and 0.80) were used to compare the allele balance in three different samples. In the second phase of the size selection study, allele balance from 0.65 and 0.70 BMRs were compared with the results from the MinElute method for five different samples. The PCR input DNA was 500 pg and 100 ng of PCR product were used for library preparation. Indexed DNA libraries were normalized to 2 nM and then pooled for sequencing.

2.6. Library input sensitivity

One sample was amplified in six separate reactions (500 pg of DNA), purified, and the PCR products pooled to create a homogenous sample. From the pooled PCR products, libraries were prepared with inputs of 200, 100, 50, 25, 12.5, and 6 ng in triplicate.

2.7. PCR sensitivity

DNA samples from three individuals were amplified at six different amounts of input DNA: 500, 250, 125, 62, 31, and 16 pg.

After purification, 6 ng of amplified products were used for library preparation for all 18 reactions and 50 ng were used for 12 reactions (from the 16–125 pg of input DNA samples). For 6 ng amplicon products library yields were insufficient to attain a concentration of 2 nM; for those samples equal volumes (1 μl) were used when pooling libraries without normalization. For 50 ng of library input, libraries were normalized and then pooled.

2.8. Reliability

Two amounts of input DNA, 250 and 62 pg, were tested on 24 different samples to determine DoC and ACR variation. To assess run-to-run variation, the 24 individuals used in this study were divided into two separate runs of 12 individuals at two input amounts per run. 80 ng (first run 12 individuals) and 64 ng (second run 12 individuals) of amplified products were used for library preparation. The amount of product was based on sample with the lowest product yield per set.

2.9. CE STR analysis

Amplification of the template DNA (500 pg) was accomplished using the PowerPlex® Fusion System (Promega, Madison, WI, USA) according to the manufacturer's protocol [27]. Amplified product was separated in the 3130XL Genetic Analyzer (Thermo Fisher). Data analysis was conducted using GeneMapper® ID software v3.2 (Thermo Fisher).

3. Results and discussion

Initially, 500 pg of template DNA (optimum amount) were amplified using the Prototype NGS STR multiplex. The results (for 24 samples) were concordant with those of CE STR analysis (data not shown). The initial sample analysis suggested that this multiplex and the Illumina MiSeq system with a 2×250 bp sequencing chemistry can generate reliable STR profiles at similar quantities of input DNA used by current CE-based approaches. However, to be considered a viable alternative for CE-based STR typing, the multiplex would have to yield a sensitivity of detection of at least 100–200 pg. To determine the efficacy of this multiplex and MPS for a comparable sensitivity of detection and profile generation with that of CE-based analyses, several steps of the analysis were tested. These steps were: size selection purification, library input, and target input for initial PCR enrichment.

3.1. Size selection study

In theory, DoC should be dispersed evenly across alleles within and among STR loci. However, in the preliminary study, imbalanced ACRs (ACR < 0.5) were observed at those loci in which the alleles have a greater size differential, such as the Penta E and D2S1338 loci. These allele imbalances tend to be with those loci with the greatest allele spread and likely will vary among individuals with genotype specific allele ranges. At least two factors may affect locus DoC and ACR: amplification efficiency and the size selection method. Amplification efficiency is determined by the construct of the multiplex STR system and thus was not modified in this study. In this study, size selection was evaluated to improve ACRs at each STR locus.

In the TruSeq LT protocol, a BMR of 0.60 for a 350 bp size fragments and 0.50 for a 550 bp size fragments are recommended. The amplicon range used in the multiplex is not captured with these size ranges. The amplicon size range across the loci for common alleles is 176–332 base pairs. Preliminary results showed reduced coverage of smaller amplicons (i.e., those <200 bp; data not shown). Thus, an optimal BMR needed to be determined. The

BMR and recovered size length are correlated indirectly [28]. In the first phase of the size selection study, five different BMRs (0.60, 0.65, 0.70, 0.75, and 0.80) were used on three samples. The results confirmed that lower BMRs recovered larger amplicons better, while higher ratios were better for recovering smaller amplicons (data not shown). The average ACRs across all loci generally were similar among these five BMRs with the lowest average values observed in sample no. 1 (Supplemental Table 1). No one BMR yielded a superior ACR at all loci. Therefore, the least spread of ACRs per BMR was considered. Two out of three of the samples favored BMRs of 0.65 and 0.70. Samples treated with 0.75 and 0.80 BMRs tended to display more imbalanced ACRs for heterozygous loci with large allele spreads. Lower coverage was observed for the larger alleles. Therefore, BMRs of 0.65 and 0.70 were selected for the second phase of the size selection study. In this phase, 0.65 and 0.70 diluted bead mixtures were compared with the MinElute PCR Purification Kit method which was designed for purification of amplicons. Average ACRs of five samples for 0.65 and 0.70 BMRs and the MinElute method were comparable (Supplemental Fig. 1). The average ACR for locus D2S1338 decreased while at the Penta E locus it increased notably with the MinElute method compared with the bead approach. In the five samples tested, the alleles of the D2S1338 locus ranged in size from 225 to 257 bp with an allele spread within a sample of between 12 and 28 bp. The MinElute method tended to show a reduction in recovery of the larger heterozygous allele. The alleles of the Penta E locus ranged from 189 to 249 bp (a difference of 35–60 bp within samples). The ACR tended to increase for this locus with the MinElute method specifically by improving the recovery of the smaller heterozygous allele. These two loci suggest the MinElute method has a broader recovery range of amplicons during size selection and allows for better recovery of smaller amplicons. However, the locus-to-locus balance was substantially better with the MinElute method, and was especially notable at the Amelogenin locus (Supplemental Fig. 2). In addition, the MinElute method yielded greater amounts of library for sequencing. The range of values for the MinElute method was 46.8–186 ng, while the largest yield for both 0.65 and 0.70 BMR did not exceed 29.6 ng. Based on these results, the MinElute method was selected for size selection for the rest of the study.

Supplemental Figs. 1, 2 and Table 1 related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.11.022.

3.2. Library input sensitivity study

The recommended library input with the TruSeq LT protocol was 800–1000 ng. However, TruSeq LT was developed and optimized for genomic DNA. In the preliminary study, the amounts of library input of 24 samples were 50–200 ng. The protocol described herein substitutes this single copy target with amplified product. Thus, the amount of library input could be reduced considerably with the PCR-based enrichment of the STR loci.

The results using 500 pg of input DNA and various amounts of PCR product for library preparation showed that the average ACR was similar regardless of library input in the tested range of 6 to 200 ng (Supplemental Fig. 3). Of the six different library inputs, only the 6 ng input could not generate sufficient libraries (1.33, 1.29, and 1.79 nM) for normalization at the 2 nM target. Equal volumes of these samples (1 μl) were used when pooling libraries. All heterozygous loci had an ACR ≥ 0.6. While the library concentrations for the 6 ng input samples did not meet the 2 nM threshold for all samples, they still produced balanced heterozygote results. Normalization at 2 nM is recommended and was not modified from the prescribed protocol. Since a reduced library amount can yield the same result as a greater amount of library input, there is an indication that comparable results may be

obtained using less than the assumed optimum amount of target DNA (500 pg) placed into the PCR. While slight differences in DoC were observed, the locus-to-locus balance was similar among libraries (Supplemental Fig. 4). Overall, a wide range of amplified product (6–200 ng) could be used to build a library with little or no observable differences in the generated STR profiles. Therefore, as long the initial input DNA for the PCR can generate more than 6 ng of product, there is confidence that sequencing could yield a result.

Supplemental Figs. 3 and 4 related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.11.022.

3.3. PCR sensitivity study

Sensitivity of detection can indicate at what levels sufficient PCR product can be obtained and what degree stochastic effects occur during PCR. The recommended input DNA of the prototype multiplex STR protocol is 500 pg. In the previous size selection and library input sensitivity studies, STR profiles with ACRs typically ≥0.6 were obtained using 500 pg of input DNA. In addition, amplified product yield typically was greater than 50 ng. The library input sensitivity study above showed that as little as 6 ng of amplicon product could generate full STR profiles. Thus, 500 pg of input DNA was not considered as the only target amount, and not the minimal amount, of DNA for the PCR. In order to determine the sensitivity of the prototype multiplex STR system used in this study, DNA samples from three individuals were amplified at six different amounts of input DNA (ranging from 16 to 500 pg; 6 ng for library input). The results confirmed that 500 pg of input DNA generated high DoC for all alleles with ACRs ≥ 0.58 at all loci (Fig. 1, Supplemental Table 2). More imbalanced ACRs were observed for STR profiles generated from 62 pg of input DNA (Fig. 2, Supplemental Table 2). The ACRs for three loci were <0.15, for 22 loci ranged from >0.15 to 0.6; and 23 loci were >0.6; with an overall average ACR of 0.59 ± 0.25. At less than 62 pg of input DNA, allele imbalance was exacerbated as would be predicted for stochastic effects during PCR of low quantities of input DNA. In an attempt to determine whether the low ACRs at some loci could be increased, a second run was performed using 50 ng of amplicon products for library preparation. As expected, based on the library input study described above, the output of the second run was similar to the previous run (data not shown). These results generally are similar to those obtained with CE-based systems. Although there was increased imbalance for heterozygous types, complete loss of an allele was limited with 62 pg of input DNA. These results indicated that 62 pg may be an initial minimum input amount for analysis to be tested further.

Supplemental Table 2 related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.11.022.

3.4. Reliability study

In the PCR sensitivity study using 250 pg of input DNA, the majority of ACRs were ≥0.4 (only one example was <0.4) with an overall average ACR of 0.68 ± 0.15. STR profiles generated from 62 pg of input DNA had a lower average ACR of 0.59 ± 0.26 and a few examples of imbalance. Therefore two input amounts, 62 and 250 pg, appeared to be potential amounts for further evaluation of a minimal amount range. At 250 pg, full profiles could be obtained but there were a few examples of increased imbalance (compared with 500 pg input DNA) and at 62 pg of input DNA allele drop out began to occur on a more frequent basis. A total of 24 individuals were typed at these two input amounts. The PCR product input for library preparation was normalized to the sample per set that yielded the least amount of PCR product. Therefore, 80 ng (first run 12 individuals) and 64 ng (second 12 individuals) were used for library preparation. Varying the amount of input should not have a significant impact, as the library input

sensitivity study demonstrated that ACR and DoC per locus generally were similar for the range of library input. While a maximum amount of PCR product (based on the limiting individual sample) was used in this study, it is worth considering using half that amount. In this way, if an analysis fails for some practical reason the entire product is not consumed and a reanalysis can be performed.

The results showed that 250 pg of input DNA generated balanced ACR at all loci, all of the average ACR at 18 loci were above 0.75, while 62 pg generated average ACR above 0.5 except at the Amelogenin locus (Fig. 3, Supplemental Tables 3 and Table 4). However, one sample (no. 11) showed allele drop-out and low DoC with both the 62 and 250 pg of input DNA. The potential cause for this one sample's lower performance may be due to low sample quality, an incorrect quantitation value, or an initial pipetting error. Not considering this one sample, the average ACR increased at most loci for both 62 and 250 pg of input DNA (Supplemental Tables 3 and 4).

Supplemental Tables 3 and 4 related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.11.022.

The MiSeq generates approximately 7.5–8.5 Gigabases (Gb) of data from an optimal sequencing run using the MiSeq v2 (2×250 bp) chemistry. The allele lengths of the 18 STR loci in the prototype multiplex STR system range between 176 and 332 bp. Assuming equal DoC across all 18 STR loci, an individual allele would be expected to have over $34,700\times$ coverage for each allele (assuming a heterozygous type at all loci), with 24 indexed libraries sequenced simultaneously. However, DoC for the 23 samples with 250 pg of input DNA was between 1507–97246x and 0–89666x for the 23 samples with 62 pg of input DNA (excluding no. 11). Average locus coverage, which varied somewhat among loci for the 24 samples, was shown for both 250 and 62 pg of input DNA (Supplemental Figs. 5 and 6). The DoC generally was lower at the D2S1338, D8S1179 and vWA loci. The most likely explanation for this observation is a lower amplification yield during the PCR with the primer pairs used for these loci. Because the 24 samples were sequenced on separate runs, coverage differences occurred in part due to cluster density differences between these two runs. The TruSeq DNA LT Sample Preparation Kit provides only 24 indices, and therefore a run is limited to 24 samples. Determination of the maximum number of samples that can be sequenced per run (and in turn the minimum DoC needed for correct allele calling) will require additional indices and is underway. However, given the DoC observed in this study, the number of samples that can be run in a single analysis will be much greater than 24 samples. Overall, the reliability study showed that reasonably balanced ACRs and high coverage were generated at both 62 and 250 pg of input DNA. The average ACRs generated in the reliability study were noticeably higher than in the previous PCR sensitivity study. Several factors may contribute to this increase: such as improved experimental skills.

Supplemental Figs. 5 and 6 related to this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.11.022.

3.5. Sequence analysis

One advantage that MPS provides over CE is that intra-repeat (and possibly typeable flanking region) variation can be detected. Among the 24 individuals, there were 5 loci, D21S11, D2S1338, D3S1358, D8S1179 and vWA, where the same nominal allele (by number of repeats) could be differentiated by sequence (Table 1). These same sequences have been described by Brinkmann et al. [29], Heinrich et al. [30], Oberacher et al. [31], and within STRbase [32]. These same loci were described by Planz et al. [7] and Oberarcher et al. [31] to have alleles that differed by sequence by mass spectrometry analyses. However, these studies also detected variation at the D13S317, D7S820, and D5S818 loci. The most
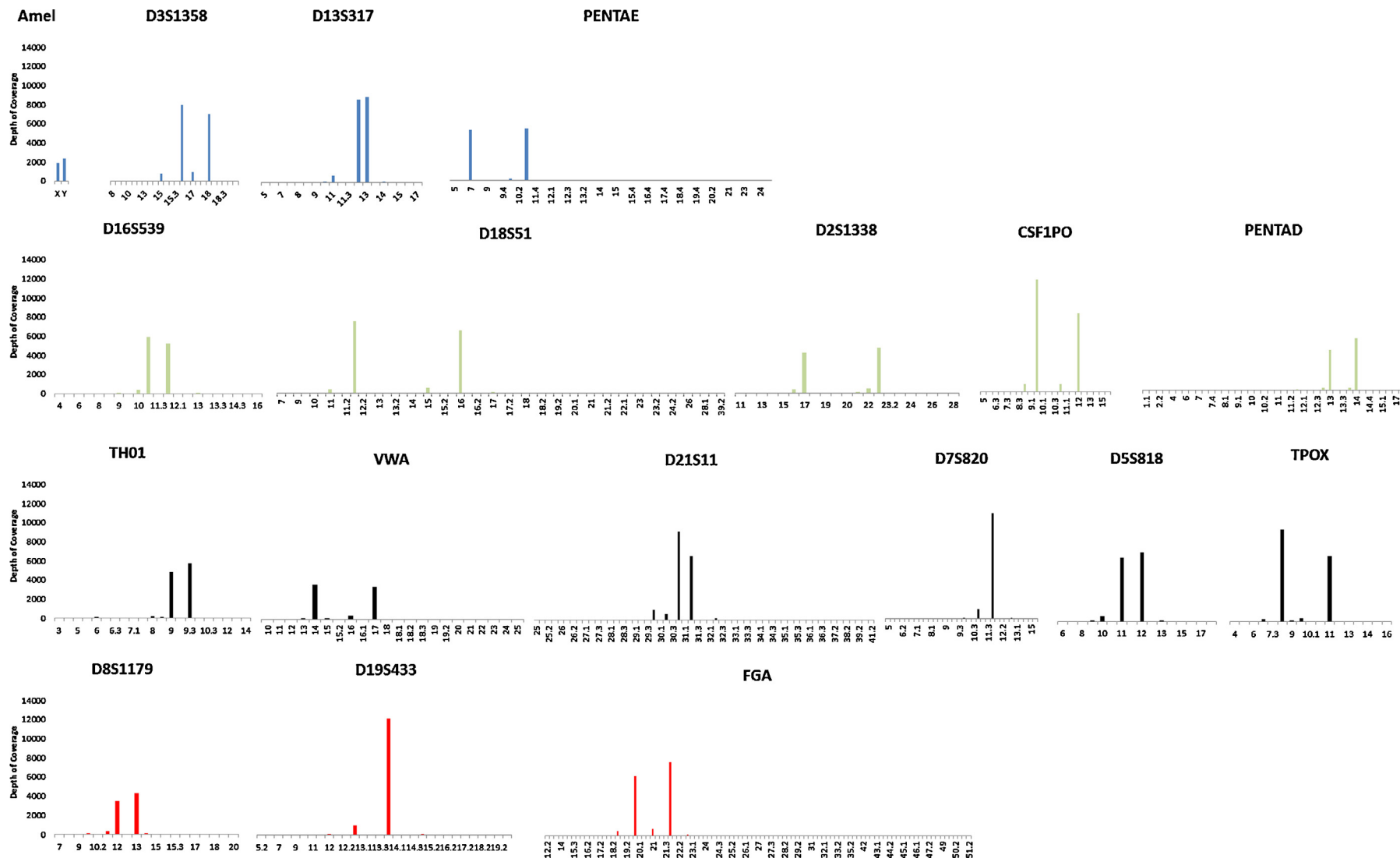
Fig. 1. A histogram portrayal of the depth of coverage by locus of one sample (no. 12) with 500 pg of input DNA.
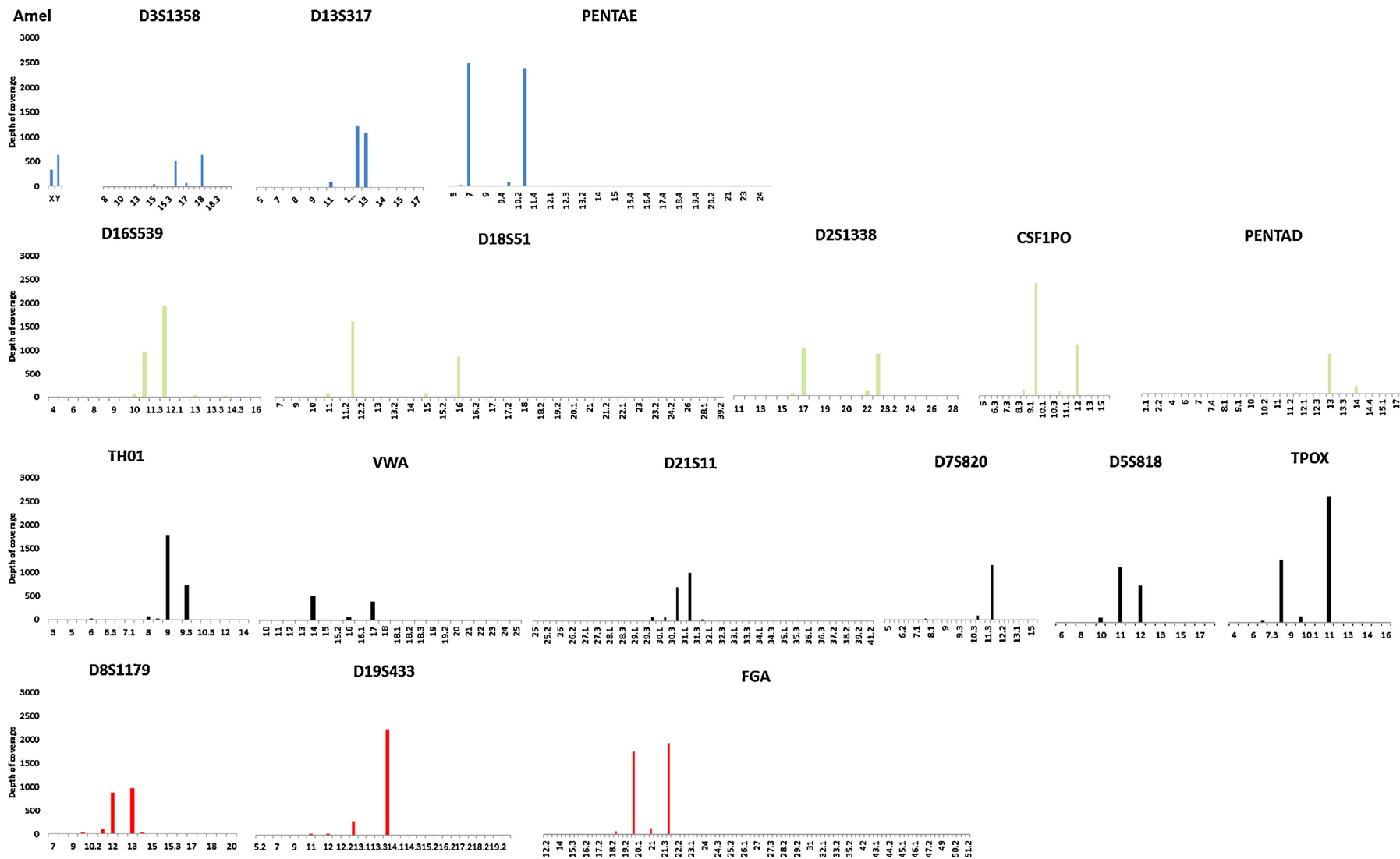
Fig. 2. A histogram result of depth of coverage by locus of one sample (no. 12) with 62 pg of input DNA.
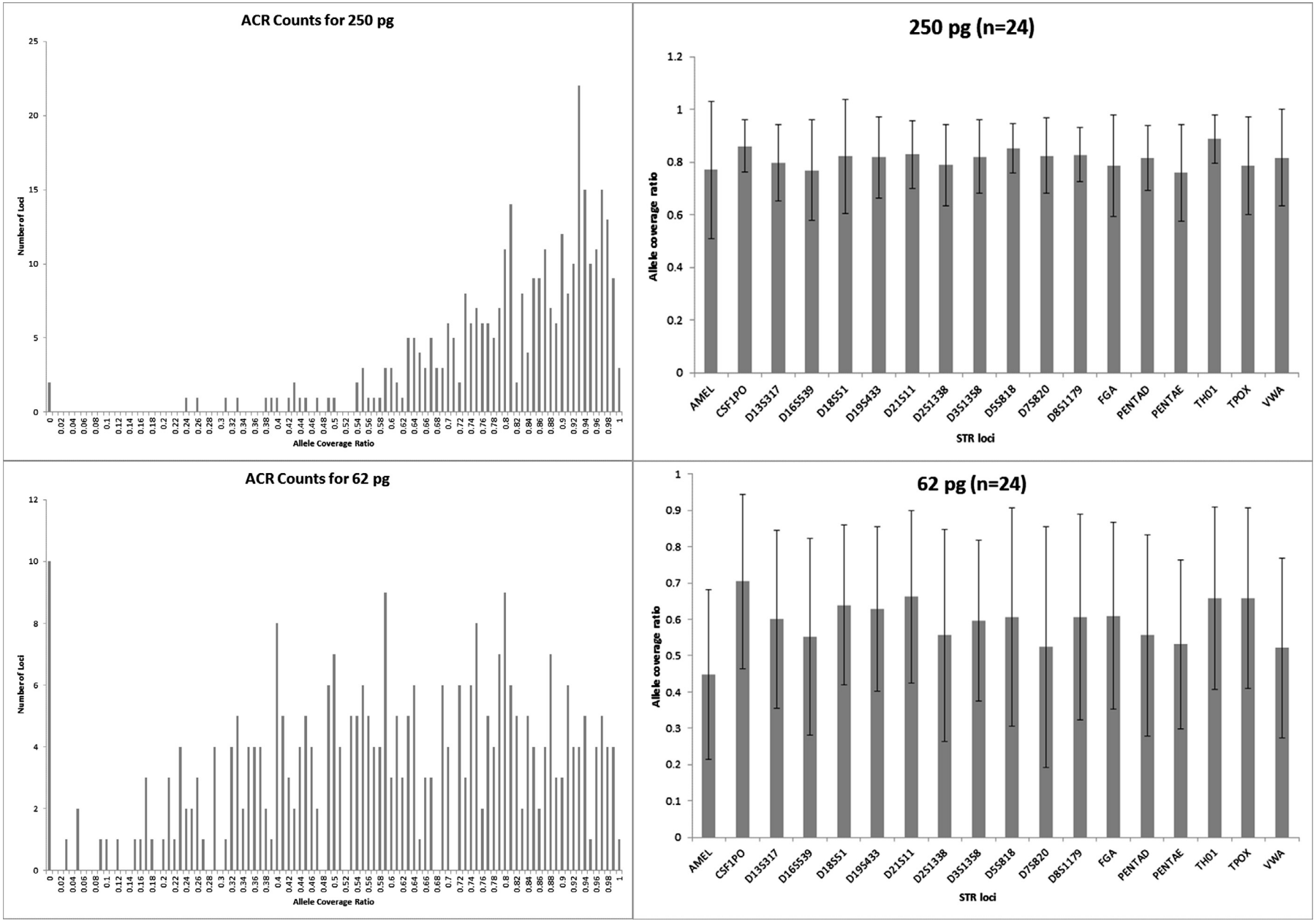
Fig. 3. The number of loci per allele coverage ratio (ACR) and ACRs at all loci of 24 samples. The top panels were results with 250 pg of input DNA, the bottom panels were results with 62 pg of input DNA. For the panels on the right, the bars are standard deviations.

Table 1

Loci, number of repeats, and number of observations by sequence among 24 individuals.

Loci	Repeats	Observations	Alleles
D21S11	29	8	TCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D21S11	29	1	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D21S11	30	3	TCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D21S11	30	2	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D21S11	30	5	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D21S11	31	5	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D21S11	31	2	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D21S11	32.2	4	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTA
D21S11	32.2	1	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTCTGTCTATCTATCTATATCTATCTATCTATCATCTATCTATCCATATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCTA
D2S1338	20	1	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCC
D2S1338	20	2	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTCCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCC
D2S1338	20	1	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC
D2S1338	22	1	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCC
D2S1338	22	2	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC
D2S1338	23	1	TGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC
D2S1338	23	5	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC
D2S1338	25	5	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC
D2S1338	25	1	TGCCTGCCTGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC
D3S1358	15	3	ATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
D3S1358	15	10	ATCTATCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
D3S1358	15	1	ATCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
D3S1358	16	10	ATCTATCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
D3S1358	16	4	ATCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
D3S1358	17	3	ATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
D3S1358	17	4	ATCTATCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
D8S1179	13	11	TCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D8S1179	13	5	TCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D8S1179	14	1	TCTATCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D8S1179	14	7	TCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D8S1179	14	1	TCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D8S1179	15	3	TCTATCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D8S1179	15	1	TCTATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
VWA	16	3	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGA
VWA	16	4	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGATAGA
VWA	17	1	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGA
VWA	17	15	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGATAGA
VWA	18	1	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGA
VWA	18	5	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGATAGA
VWA	19	1	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGATAGA
VWA	19	1	TAGATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGATAGA

Table 2
Selected examples of observed SNP patterns within STR loci of two samples. No. 2 was a homozygote at the D3S1358 locus (15/15), no. 6 was a heterzygote at the D3S1358 locus (15/15), no. 6 was a heterzygote at the D21S11 locus (30/31).

Locus: repeat	Sample name	Source	Sequence data
D3S1358: 14	No. 2	Stutter[a]	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGA**C**AGACAGACAGACAGATAGAT
D3S1358: 14	No. 2	Stutter[b]	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGA**T**AGACAGACAGACAGATAGAT
D3S1358: 15	No. 2	Allele 15	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGA**C**AGACAGACAGACAGATAGAT
D3S1358: 15	No. 2	Allele 15	AGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGA**T**AGACAGACAGACAGATAGAT
D21S11: 30	No. 6	Allele 30	TAGATAGATAGATAGATAGATAGATAGATAGATATGGATAGATAGATAGATAGATAGATAGATAGATAGA**C**AGACAGACAGACAGA**T**AGATAGATAGATAGATAGATAGA
D21S11: 30	No. 6	Stutter[c]	TAGATAGATAGATAGATAGATAGATAGATAGATATGGATAGATAGATAGATAGATAGATAGATAGATAGA**C**AGACAGACAGACAGA**C**AGATAGATAGATAGATAGATAGATAGA
D21S11: 31	No. 6	Allele 31	TAGATAGATAGATAGATAGATAGATAGATAGATATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGACACAGACAGACAGACAGA**C**AGATAGATAGATAGATAGATAGATAGA
D21S11: 31	No. 6	Stutter[d]	TAGATAGATAGATAGATAGATAGATAGATAGATATGGATAGATAGATAGATAGATAGATAGATAGATAGATAGACACAGACAGACAGACAGA**T**AGATAGATAGATAGATAGATAGATAGA

a Stutter of one of the 15 alleles at the D3S1358 locus in no. 2.
b Stutter of the other 15 allele at the D3S1358 locus in no. 2.
c Minus stutter of the 31 allele at the D21S11 locus in no. 6.
d Plus stutter of the 30 allele at the D21S11 locus in no. 6.

plausible explanation for this difference is that the variation at these loci tends to reside in the flanking region as opposed to within the repeat regions [31]. STRait Razor [15] currently detects variation only within the repeat region.

Table 2 showed examples of two different observed scenarios at the D3S1358 and D21S11 loci in different individuals. One example shows SNPs (C/T) distinguishing the two 15 homozygous alleles of the D3S1358 locus. The same SNP variants are seen in the stutter product (equivalent to '14' repeats). For the D21S11 locus, the nominal allele calls are a heterozygote 30 and 31. Typically, a heterozygote with two alleles that differ by one repeat is not used for calculating stutter percentages. However, because of the intra-repeat SNPs that are present, it is possible to distinguish the contributions of relevant plus and minus stutter contributions even with alleles and stutter that are nominally the same. These observations suggest that SNPs in true alleles also may reside in stutter products that could facilitate mixture deconvolution and thus allow for better source attribution. Further population studies are warranted to assess whether such observations hold or to what degree they hold.

While the sample size is small, the results demonstrate that variation can be detected readily and it was possible to distinguish alleles of the same nominal size in some loci. For these loci, mixture interpretation may be facilitated especially when a minor contributor's alleles are similar in signal intensity with that of stutter of a major contributor alleles. Future efforts will assess mixture interpretation and, if possible, whether slippage tends to occur more so in one area or another of the repeat region.

4. Conclusion

In this study, a new STR multiplex compatible with MPS was evaluated. This multiplex (17 STR loci + Amelogenin) system utilized PCR enrichment and library preparation with the TruSeq DNA LT Sample Preparation Kit. Sequencing was carried out on the MiSeq, and STRait Razor was used for allele calling. The MinElute PCR Purification Kit was demonstrated to be a better size selection method compared with diluted bead mixtures. The library input sensitivity study showed that a wide range of amplicon product could be used with no notable differences in the generated STR profile. Results on 24 different individuals showed that a high DoC and balanced heterozygote ACRs can be obtained with input DNA ($\geq$250 pg), which is similar to that typically observed in CE-based systems. Moreover, input DNA as little as 62 pg could generate complete or nearly complete (i.e., limited allele drop-out) profiles. These studies indicate that this STR multiplex and the Illumina MiSeq system can generate reliable DNA profiles at a sensitivity of detection level that rivals current CE-based approaches. The TruSeq DNA LT Sample Preparation Kit provides only 24 indices, and therefore a maximum of 24 samples can be sequenced per run. However, given the high DoC for the 24 samples, it is evident that many more samples can be multiplexed. Determination of the maximum number of samples that can be sequenced per run and the minimum DoC needed for correct allele calling is underway. Future studies will entail traditional validation analyses to determine the multiplex's capability of analyzing more challenged samples and mixtures compared with CE-based systems.

Acknowledgements

References

[1] P.J. Collins, L.K. Hennessy, C.S. Leibelt, R.K. Roby, D.J. Reeder, P.A. Foxall, Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR Identifiler PCR Amplification Kit, J. Forensic Sci. 49 (2004) 1265–1277.

[2] B.E. Krenke, A. Tereba, S.J. Anderson, E. Buel, S. Culhane, C.J. Finis, et al., Validation of a 16-locus fluorescent multiplex system, J. Forensic Sci. 47 (2002) 773–785.

[3] P. Martin, L.F. de Simon, G. Luque, M.J. Farfan, A. Alonso, Improving DNA data exchange: validation studies on a single 6 dye STR kit with 24 loci, Forensic Sci. Int. Genet. 13C (2014) 68–78.

[4] K. Oostdik, K. Lenz, J. Nye, K. Schelling, D. Yet, S. Bruski, et al., Developmental validation of the PowerPlex((R)) Fusion System for analysis of casework and reference samples: a 24-locus multiplex for new database standards, Forensic Sci. Int. Genet. 12 (2014) 69–76.

[5] K. Lazaruk, P.S. Walsh, F. Oaks, D. Gilbert, B.B. Rosenblum, S. Menchen, et al., Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument, Electrophoresis 19 (1998) 86–93.

[6] J.M. Butler, E. Buel, F. Crivellente, B.R. McCord, Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis, Electrophoresis 25 (2004) 1397–1412.

[7] J.V. Planz, B. Budowle, T. Hall, A.J. Eisenberg, K.A. Sannes-Lowery, S.A. Hofstadler, Enhancing resolution and statistical power by utilizing mass spectrometry for detection of SNPs within the short tandem repeats, Forensic Sci. Int. Genet. Suppl. Ser. 2 (2009) 529–531.

[8] J.V. Planz, K.A. Sannes-Lowery, D.D. Duncan, S. Manalili, B. Budowle, R. Chakraborty, et al., Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry, Forensic Sci. Int. Genet. 6 (2012) 594–606.

[9] F. Pitterl, K. Schmidt, G. Huber, B. Zimmermann, R. Delport, S. Amory, et al., Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations, Int. J. Legal Med. 124 (2010) 551–558.

[10] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, et al., A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 13 (2012) 341.

[11] S. Junemann, F.J. Sedlazeck, K. Prior, A. Albersmeier, U. John, J. Kalinowski, et al., Updating benchtop sequencing performance comparison, Nat. Biotechnol. 31 (2013) 294–296.

[12] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S.M. Gomes, L. Souto, et al., Evaluation of next generation mtGenome sequencing using the ion torrent personal genome machine (PGM), Forensic Sci. Int. Genet. 7 (2013) 543–549.

[13] J.L. King, B.L. LaRue, N.M. Novroski, M. Stoljarova, S.B. Seo, X. Zeng, et al., High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq, Forensic Sci. Int. Genet. 12 (2014) 128–135.

[14] D.M. Bornman, M.E. Hester, J.M. Schuetter, M.D. Kasoji, A. Minard-Smith, C.A. Barden, et al., Short-read, high-throughput sequencing technology for STR genotyping, BioTechniques (2012) 1–6.

[15] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, et al., STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (2013) 409–417.

[16] C. Van Neste, M. Vandewoestyne, W. Van Criekinge, D. Deforce, F. Van Nieuwerburgh, My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing, Forensic Sci. Int. Genet. 9 (2014) 1–8.

[17] S.L. Fordyce, M.C. Ávila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F.T. Petersen, et al., High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, BioTechniques 51 (2011) 127–133.

[18] M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat sequencing on the 454 platforms, Forensic Sci. Int. Genet. Suppl. Ser. 3 (2011) e357–e358.

[19] QIAamp® DNA Mini and Blood Mini Handbook, 2012 http://www.qiagen.com/resources/resourcedetail?id=67893a91-946f-49b5-8033-394fa5d752ea&lang=en.

[20] User guide: Quantifiler® Human and Y Human Male DNA Quantification Kits, 2014 http://tools.lifetechnologies.com/content/sfs/manuals/cms_041395.pdf.

[21] User guide. Prototype Multiplex STR System, 2013 Promega provided.

[22] MinElute® Handbook, 2008 http://www.qiagen.com/resources/resourcedetail?id=fa2ed17d-a5e8-4843-80c1-3d0ea6c2287d&lang=en.

[23] Qubit® dsDNA BR Assay Kits, 2011 https://tools.lifetechnologies.com/content/sfs/manuals/mp32850.pdf.

[24] TruSeq® DNA PCR-Free Sample Preparation Guide, 2013 http://supportres.illumina.com/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqdnapcrfree/truseq-dna-pcr-free-sample-prep-guide-15036187-b.pdf.

[25] Qubit™ dsDNA HS Assay Kits, 2010 http://tools.lifetechnologies.com/content/sfs/manuals/mp32851.pdf.

[26] MiSeq® System User Guide, 2013 http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq_system_user_guide_15027617_l.pdf.

[27] Technical Manual: PowerPlex® Fusion System, 2014 http://www.promega.com/~/media/files/resources/protocols/technical%20manuals/101/powerplex%20fusion%20system%20protocol.pdf.

[28] S. Verrow, M. Blair, B. Packard, W. Godfrey, Gel-Free Size Selection Using SPRIselect For Next Generation Sequencing, Beckman Coulter, 2013.

[29] B. Brinkmann, A. Sajantila, H.W. Goedde, H. Matsumoto, K. Nishi, P. Wiegand, Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci, Eur. J. Hum. Genet. 4 (1996) 175–182.

[30] M. Heinrich, H. Felske-Zech, B. Brinkmann, C. Hohoff, Characterisation of variant alleles in the STR systems D2S1338, D3S1358 and D19S433, Int. J. Legal Med. 119 (2005) 310–313.

[31] H. Oberacher, F. Pitterl, G. Huber, H. Niederstätter, M. Steinlechner, W. Parson, Increased forensic efficiency of DNA fingerprints through simultaneous resolution of length and nucleotide variability by high-performance mass spectrometry, Hum. Mutat. 29 (2008) 427–432.

[32] C.M. Ruitberg, D.J. Reeder, J.M. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Res. 29 (2001) 320–322.