Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers

Article *in* Forensic Science International: Genetics · May 2014

Impact Factor: 4.6 · DOI: 10.1016/j.fsigen.2014.04.010 · Source: PubMed

4 authors, including:

Odile Loreille

33 PUBLICATIONS **838** CITATIONS

SEE PROFILE

Title: Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers

Author: Melissa Scheible Odile Loreille Rebecca Just Jodi Irwin

PII: S1872-4973(14)00086-6
DOI: http://dx.doi.org/doi:10.1016/j.fsigen.2014.04.010
Reference: FSIGEN 1145

To appear in: *Forensic Science International: Genetics*

Received date: 6-12-2013
Revised date: 12-3-2014
Accepted date: 22-4-2014

Please cite this article as: M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers, *Forensic Science International: Genetics* (2014), http://dx.doi.org/10.1016/j.fsigen.2014.04.010

Short tandem repeat typing on the 454 platform: Strategies and considerations for targeted sequencing of common forensic markers

Melissa Scheible[a,b,*], Odile Loreille[a,b], Rebecca Just[a,b] and Jodi Irwin[a,b,1]

[a]*American Registry of Pathology, Camden, DE*
[b]*Armed Forces DNA Identification Laboratory, Armed Forces Medical Examiner System, Dover AFB, DE*
[*]*Corresponding author: E-mail address: melissa.k.scheible.ctr@mail.mil (M. Scheible)*
[1]*Present address: Federal Bureau of Investigation, 2501 Investigation Parkway, Quantico, VA 22135*

1. Introduction

The past several years have seen a dramatic advance in the methods, chemistries and detection platforms available for DNA sequence data generation. Next generation sequencing (NGS) technologies, which produce large volumes of sequence data at extremely low cost relative to current platforms, are being broadly applied to various questions in medical genetics, evolutionary biology, molecular anthropology, phylogeny, epidemiology and metagenomics. For many of these applications, NGS is being used to produce sequence data covering thousands of loci, or even entire organismal genomes in a single sequencing run. Given this capacity, it is not difficult to envision the potential implications of this technology for criminalistics, missing persons and disaster victim identification purposes. Historically, the recovery of large numbers of markers in a single assay has been restricted by both the technical limitations of current, established capillary based sequencing genotyping technologies, as well as the quality and quantity of DNA originating from the damaged and degraded specimens regularly encountered in forensic casework. These limitations do not apply in quite the same way to NGS, however. As a result, the simultaneous recovery of the standard autosomal DNA, mitochondrial DNA, and X and Y-chromosomal markers regularly assayed in forensic genetics, along with additional markers of interest, may be possible with these new technologies [1].

To date, research into NGS by the forensic community has been relatively limited [2-11]. The high cost of sample processing and the sheer volume of sequence data produced initially made these methods more suitable for much larger-scale applications such as de novo sequencing of organismal genomes and exome sequencing for clinical research/diagnostics. Furthermore, the large volume of data generated with traditional next generation shotgun sequencing approaches generally is not necessary, desirable, nor financially practical for most forensic genetic applications.

More recently, however, the high throughput capacity of next generation sequencing has been harnessed for targeted re-sequencing applications [3-10,12-15]). With these workflows, NGS is used to produce data from numerous samples and often numerous targeted markers in any given reaction. That is, instead of producing genome-wide low coverage sequence data for a single sample in a given run, sequence data for many targeted genomic regions are produced at high depths of coverage for tens or hundreds of individuals. Given that the forensic community is primarily interested in restricted regions of the genome, and given the potential of high throughput sequencing (both in terms of cost-efficiency and judicious use of limited sample material) for recovering genetic information from multiple markers and multiple individuals in a single run, these targeted approaches seem to be the most immediately applicable for forensic genetic applications.

For the most commonly employed forensic markers, short tandem repeats (STRs), limited data exist in any discipline on the use of NGS for STR typing. Although genome-wide data have been used to identify and locate variable tandem repeats in the genome [2,16-21], it has only been recently, and in a handful of studies, that STR sequencing via NGS has been evaluated for the routine genotyping of STRs in forensics [3,4,6,8,9,15].

To gather information on the potential of NGS technologies for short tandem repeat typing of highly degraded forensic specimens in particular, we have evaluated a workflow based on miniSTR amplicons and the Roche 454 platform. We have focused on small (<250bp) fragments because of the sample type generally encountered in our laboratory (highly degraded skeletal remains) and the increased success of short amplicons with fragmented DNA [22-26]. For applications related to degraded DNA in particular, sequencing offers some major advantages over currently employed capillary electrophoresis-based fragment analysis methods, where the limited number of available fluorescent dyes restricts the number of markers with overlapping size ranges that can be multiplexed in a single reaction. Because NGS is not subject to these same chemistry-based limitations, numerous markers with overlapping size ranges can be sequenced simultaneously. Here, we describe our preliminary investigations into the multiplexed-STR typing of highly degraded specimens using next generation sequencing.

2. Materials and methods

2.1 Samples and DNA extraction

The high-quality U.S. population samples used for this work were selected on the basis of the availability of fragment-based genotype information for the majority of the markers targeted in our study. Those population samples were obtained, and DNA extraction was performed, as described in [27]. The non-probative, degraded skeletal remains tested, referred to as "casework" samples from here on, were selected to represent the range of sample quality routinely encountered in our laboratory's missing persons casework. The specimens were of highly variable quality (as assessed by earlier mitochondrial DNA typing success, data not shown) and ranged in age from 40-60 years postmortem. DNA extractions from the casework samples were performed as described in [28].

All work described herein was reviewed and approved by the U.S. Army Medical Research and Materiel Command Institutional Review Board.

2.2 Amplification strategies

Although a number of different target enrichment techniques are available, enrichment of STR markers was performed via the polymerase chain reaction (PCR) for these experiments. Both low and high quality samples were PCR-amplified with three different multiplexing strategies. First, a published multiplex optimized for mass spectrometry was used as described in [29] with the PCR conditions specified for AmpliTaq Gold DNA Polymerase (Life Technologies, Carlsbad, California) and 40 PCR cycles except that Amelogenin was excluded, resulting in a total of thirteen targeted markers ranging from 69-211 base pairs using the primer sequences in [30] (Table S1). The published primer concentrations, as well as primer concentrations altered for this specific application, were tested (Table S2). In total, twenty-four high quality population samples and three degraded specimens (triplicate amplifications of duplicate extractions for a total of six amplifications per degraded sample) were typed using this strategy.

Additionally, a series of twelve multiplexes containing four loci each (for a total of forty-eight markers) was employed. These targeted markers include all Combined DNA Index System

(CODIS) core loci, those recently recommended by the CODIS Core Loci Working Group [31] as well as the entire set of non-CODIS miniSTRs described by the National Institute of Standards and Technology (NIST) in [32]. This series of multiplexes was amplified with the PCR parameters described in [29] with the exception of the annealing temperature, which was adjusted to accommodate the addition of the new primers. The arrangement of markers within each multiplex (grouped by primer melting temperatures), as well as the amplification primer sequences, are shown in Table S1. For high-quality samples, primers were added in equimolar ratios for a final concentration of 5.9 uM per reaction. For degraded samples, primers that yielded few or no sequence read coverage for the high-quality population samples were doubled in each multiplex's master mix, but the final concentration of 5.9 uM total primers was maintained. Amplification products of the twelve multiplexes were pooled for each sample prior to library preparation. A total of fifteen high quality population samples, three controls commonly used in forensic STR typing (Control DNA 9947A [Life Technologies], 9948 Male DNA [Promega Corporation, Madison, Wisconsin], and K562 High Molecular Weight [Promega Corporation]), and thirteen casework extracts (duplicate amplifications per extract) were typed with this strategy. The casework samples typed with this series of twelve multiplexes are not the same samples used for13-plex testing, due to limited extract volume for all casework samples.

Samples were also amplified with the AmpFℓSTR® Identifiler® PCR Amplification Kit (Life Technologies) using the primer mix included in the kit, with standard PCR parameters employed for high-quality samples and a modified amplification strategy for degraded samples [33]. Although previous studies have reported successful 454 sequencing results generated from commercial STR kit amplification product [4,34], our two attempts resulted in an absence of DNA-containing beads following the post-emulsion PCR (emPCR) enrichment. This problem never occurred when using unlabeled primers for sample enrichment; and thus, we believe it was likely related to the fluorescent label on the kit primers. As a result, enrichment by commercial kit amplification was not pursued further in this study and no results are presented in this report.

No samples were quantified during these experiments. Inputs ranging from 5 to 10µl (per 50µl reaction) were used for all population samples. Casework extracts were concentrated to the exact volume required for the planned amplifications in this study to maximize the DNA in each reaction.

2.3 Preparation for GS Junior

The GS Junior System was used for sequencing. At the time of experimental design and execution, this system and the GS FLX were the only massively parallel second-generation sequencing instruments that could produce average read lengths (400-600 base pairs) long enough to span entire miniSTR amplicons and both adaptors [35].

Individual samples were tagged with multiplex identifiers (MIDs) and prepared for sequencing according to manufacturer guidelines [36-39] with a few exceptions. For one, fragmentation by nebulization was unnecessary, since amplified targets were already smaller than the optimal DNA library size range (400-600 base pairs). Second, adaptors and adaptor dimers were removed with Agencourt AMPure XP (Beckman Coulter, Inc., Brea, California**)** using 2µl AMPure XP per 1µl of PCR product and the general protocol recommended by the manufacturer

[40] as opposed to the process described in the GS Junior user manual. The latter selects for fragments larger than 300 base pairs and would have eliminated not only the adaptors and adaptor dimers, but also the miniSTR amplicons. Third, one to 1.8 molecules of sample library per capture bead were targeted for emPCR since initial experiments demonstrated that the recommended two molecules often produced an overabundance of DNA-containing beads. This is generally indicative of multiple templates per bead and, in the end, leads to a drastic reduction in successful reads since those reads representing multiple templates are discarded during the data filtering process. Finally, the quantity of amplification primer was reduced to one quarter the typical volume as suggested for amplicon libraries of short fragments. This modification was intended to decrease signal crosstalk by reducing signal strength and eliminating incomplete extension [39]. Beads were sequenced on the GS Junior system using version 2.5p1 of the sequencing software.

2.4 Data analysis

Sequence data were imported into the CLC Genomics Workbench v5.1 software (CLC bio, Aarhus, Denmark) for post-processing data analysis. Reads representing individual samples were separated based on the unique MID sequence tag and aligned to references designed for this project.

Reference sequences were manually created for each described allele of every locus and spanned both amplification primers in length (Table S3). Exact repeat structure was used when available [30,41], and remaining repeat structures and flanking sequences were deduced by using a combination of sources [32,41-43]. Reference sequence allele variation was limited to those alleles with known repeat structure at the time they were created, and distinctions among variants of the same length used established nomenclature (for example, 17 vs. 17') [41]. The polymorphisms reported in this study (single nucleotide polymorphisms [SNPs] and insertions and deletions [indels], hereafter referred to as "SNPs") include the differences from the allele variants with described repeat structure. That is, the polymorphisms represent additional variation not already reported with a specific existing nomenclature strategy.

The default Genomics Workbench mapping and alignment parameters were adjusted so that only the sequences matching a reference allele by 85% of its length and 85% of its base similarity were captured. Non-specific matches, or reads that could be aligned equally well to multiple reference sequences [44], were ignored (not aligned). These stringent parameters were used to both ensure that STR markers with similar repeats would not align to the incorrect references and eliminate reads ending within the repeat region.

Allele calls were made based on the number of reads that aligned to each reference. In most cases the authentic allele or alleles were obvious, as they were represented by the vast majority of reads for a particular locus. The alignment for each called allele was reviewed to confirm correct alignment and the number of repeats, and any consensus differences from the reference sequence were noted. Alleles that could potentially represent stutter (i.e. alleles with one fewer or one more repeat as compared to the called allele) were not designated as alleles (i.e. they were assumed to be stutter) if the number of reads that aligned was less than approximately 15% of the majority allele [45,46]. Thus, in essence, a 15% stutter filter was applied to the data.

2.5 Comparative CE data

To evaluate the allele calls made from the 454 sequence data, all samples were either typed by standard fragment analysis methods using CE and commercially available kits or compared with known genotypes. In addition, a subset of the alleles was Sanger sequenced to verify concordance between Sanger generated and NGS generated data, and thus confirm the authenticity of NGS detected SNPs. Fifteen alleles representing six different samples were Sanger sequenced for these purposes. For fragment analyses, population samples were typed using the AmpFℓSTR® Identifiler® PCR Amplification Kit (Life Technologies), and known genotypes of the non-CODIS miniSTRs were provided by NIST (C. Hill, personal communication). Casework extracts were typed with the AmpFℓSTR® Identifiler® PCR Amplification Kit (Life Technologies) or the PowerPlex® 16 System (Promega Corporation) using a modified amplification strategy and replicate amplifications [47]. Because the focus of our study was the feasibility of STR sequencing via NGS, rather than the development of an optimized assay, the genotypes generated by CE typing were typically only used to verify alleles called in the sequenced STRs. That is, generally no allele recovery comparisons were performed between the CE-generated and 454 sequencing profiles. The only exception to this was the typing of casework specimens using the NGS 48-marker assay. In the case of these highly compromised samples, which frequently result in limited useful fragment analysis data, such comparisons were performed to assess the potential for additional information recovery with an NGS approach.

3. Results and discussion

-

3.1 Summary of data analysis issues

At the data analysis stage, a number of data features were encountered that we elected to classify into one of two distinct categories: those resulting from the wet laboratory procedures used to produce the data, and those encountered post data production as a result of data analysis and alignment tools not yet optimized for STR markers nor forensic applications of STRs in particular.

Those data characteristics falling into the molecular biology procedure category included 1. stutter - a known PCR artifact that we observed at percentages comparable to previous reports [45,46,48], 2. locus/allele imbalance – reflecting the challenges of multiplex optimization, and 3. unusual sequencing artifacts (described in detail in section 3.2) that couldn't be simply explained by known PCR or NGS/454 issues. In this last case, the unusual features were observed only in the sequence data produced from the 13-plex, and they disappeared completely upon switching to the smaller multiplexes of four markers each. Interestingly, the affected reads were only in one direction, rather than in both forward and reverse reads as one would expect with the nonspecific amplification of a similar repeat elsewhere in the genome.

Issues falling into the second category primarily relate to NGS read alignment difficulties resulting from imperfect data analysis tools and the highly complex repeat structure of a few of the common forensic markers. Specifically, D21S11 and vWA, which exhibit significant variation in repeat structure among alleles of the same length [8], were problematic. Without a customized set of STR analysis tools and unsuccessful efforts with de novo assemblies (further described below), we relied on the set of references we created for each described allele for the purposes of NGS read alignment. However, when not all variation was represented in those references, alignment difficulties were occasionally encountered.

3.2 13-plex

Population sample profiles

Although primer concentrations were slightly adjusted to improve performance, the 13-plex was never fully optimized to produce equal reads among loci. Even so, when considering both the high quality samples and the degraded specimens, the reads from those loci that amplified well reflected repeat motifs that were perfectly consistent with known genotypes. Allele calls for high quality population samples are presented in Table S4 and summarized in Table 1. Nine of the targeted markers exhibited some degree of allele dropout. While four of those markers dropped out in at least half the typed samples, the remaining nine loci in the 13-plex consistently produced sequence data. Of the nine markers that consistently amplified, 99% of all recovered alleles were confirmed with fragment analysis data produced with commercially available kits. The single discordant call, the D21S11 29 allele in sample GC03394, was the result of high (27%) stutter in the multiplex amplification/NGS assay that appeared to be an authentic allele.

In addition to permitting the simultaneous recovery of a larger number of alleles as compared to traditional CE typing, STR sequencing also provides the opportunity to discern sequence variation. A summation of the observed SNPs is listed at the bottom of each sample in Table S4. Between zero and six previously undescribed SNPs were observed per sample across the recovered loci, with an average of 2.3 SNPs per individual. SNPs were observed in 5 of the 9 markers that were consistently amplified with this assay. Fourteen of the sixty-two (23%) observed SNPs were unique within this dataset, and the remainder were shared by two or more individuals. In the marker D8S1179, all six observed polymorphisms were unique in this dataset.

Casework sample profiles

Samples of aged, degraded skeletal elements with "known" autosomal STR profiles were unavailable. Thus, performance of the NGS assay was based on the profile inferred for a particular individual using the data recovered from separate amplifications of two replicate extractions as well as the partial profiles generated with commercial STR kits. Because of the multiple replicates per individual, the total number of different casework evidentiary samples tested was reduced as compared to the population samples.

Table 2 shows the recovered alleles for three non-probative casework specimens using the 13-plex strategy. Each of the six specimens was extracted in duplicate, and each extraction was

amplified in triplicate, resulting in a total of eighteen degraded sample amplifications. The same marker that consistently dropped out with the population samples, D10S1248, also dropped out for all the evidentiary samples. As expected with challenging samples, the number of called alleles for the other twelve markers varied greatly depending on the quality of each particular sample/extraction, with a range of three to nineteen alleles recovered and an average of 11.7 per amplification using the NGS assay. Nevertheless, the amplification and extraction replicates consistently produced profiles concordant with each other as well as the genotypes generated with a commercial STR kit, despite sometimes low coverage.

In addition, between zero and nine SNPs were observed in the NGS profiles with an average of three SNPs per amplification. In amplifications where only a relatively small number of alleles could be recovered, the sequence variants provided additional discrimination potential that would have otherwise been lost with traditional CE typing methods. For example, replicate amplification 3 of Sample 1, extraction B, only produced seven alleles but contains additional discriminatory information in the form of two SNPs that are not detectable with CE typing.

Quality of sequence reads

While the vast majority of NGS reads for any given marker exhibited the sequence data expected based on previously described motifs (i.e. the known repeat structure, along with SNPs and indels in some alleles), a few of the markers showed unusual features that were clearly evident in both pristine and degraded specimens. For instance, Figure S1 shows the alignment of NGS reads from locus D21S11. In addition to the sequences expected based on the known STR profiles and Sanger data, a number of the reads reflected a different sequence. The Sanger data (not shown) exhibited background in the same region, yet the exact sequence observed in the NGS data was not apparent.

D16S539 also showed evidence of a secondary sequence. While the sequences aligned to the proper allele reference, a number of the forward sequences harbored additional guanine residues towards the beginning of the reads (see Figure S2). Although the randomly distributed insertions likely represent the type of homopolymer sequencing errors known to manifest in 454 data, the most abundant G insertions are not as easily explained. In this case, there was no evidence of a secondary sequence, or any other type of background, in the Sanger data originating from single-plex amplification of that marker.

3.3 48-marker assay

Population sample profiles

Because of the aforementioned sequence quality issues with the 13-plex, all remaining experiments were performed with unlabeled primers that covered the thirteen miniSTR markers previously used [29], plus thirty-five more (Table S1). The forty-eight markers were organized into twelve multiplexes of four markers each to simplify the reactions and determine if sequence quality would improve for those markers that exhibited artifacts in the 13-plex. NGS allele calls for high quality population samples, along with those calls confirmed via CE, are listed in Table S5 and summarized in Table 3.

Again, because our goal was simply to assess the feasibility of NGS for STR sequencing, and we did not optimize the multiplexes beyond their initial construction, sequence data were absent or incomplete for a number of markers. Thirty-four of the forty-eight included markers exhibited some degree of allele dropout. Seven of the forty-eight markers never resulted in any 454 sequence data, and thus evaluations are based on the forty-one markers for which any sequence data was produced. For thirty-eight of these markers, the allele calls were confirmed with CE typing methods (three were not confirmed by CE data because they were not included in any available commercial STR kit). Despite the lack of multiplex optimization reflected in the dropout rates with this assay, the alleles recovered successfully were abundant, ranging from thirty-eight to sixty-one alleles called with an average of fifty-two alleles per sample. The 48-marker assay also produced much cleaner alignments as compared to the larger multiplex. As with the 13-plex, the initial NGS assay allele calls that were inconsistent with CE data - two alleles in this dataset, or 0.21%, were due to unusually high stutter (24% and 60% of the authentic allele).

Despite allele dropout, the number of SNPs identified (assessed in comparison to the references sequences to which the data were aligned) for each individual ranged from three to fourteen, with an average of 8.4 SNPs per individual typed. There was a SNP detected in one out of every six alleles recovered, on average. Twenty-seven of forty markers recovered (this number excludes Amelogenin) show the expected repeat structure and have no variation aside from the number of repeats. However, the remaining thirteen markers exhibited sequence variation that cannot be detected with CE typing. The number of total SNPs and unique SNPs (those found in only one sample in our dataset) were tallied for each recovered marker, and are shown in Figure 1. D12S391 had the greatest proportion of unique variants, with twelve unique of the fourteen total observed. An example of variation in D12S391 is shown in Figure S3.

To assess the additional information gleaned by STR sequencing in another way, we compared the number of alleles that would be detected with traditional CE typing methods to the number detected with the NGS assay in this study (Figure 2). While there was no difference in the number of alleles for twenty-eight of the markers, twelve markers (again, excluding Amelogenin) yielded additional alleles when sequence variation was considered. For three of these markers (D5S818, D5S2500, and D12S391) the number of alleles doubled when utilizing sequence information instead of fragment size alone. Overall, forty more alleles were detected with STR sequencing than would have been detected with CE typing, representing an increase of 21%. Given the allele dropout that is known to have occurred at twenty-seven of the recovered markers, these values likely underrepresent the informational increase reaped by STR sequencing. In fact, if only the thirteen markers at which no dropout was observed with the NGS assay are considered (again, excluding Amelogenin), twenty-two more alleles were detected with sequencing than would have been detected via CE typing - an increase of 31%.

As Sanger sequencing of STRs is time-consuming and challenging [49], sequence variation among many of these 48 markers is currently not comprehensively described. The data presented here, some of the first addressing STR sequence variation among randomly sampled individuals, hint at the additional information that is likely to be revealed through sequence characterization of these and other markers. Our data mirror recently published results

describing variation among the markers D21S11 [8] and D12S391 [9], which showed that sequence variation was not only represented by SNPs and indels, but also - for these two complex repeats - by sub-repeat composition differences that are generally undetectable in CE typing. The sub-repeat variation was shown to dramatically increase the diversity and discriminatory power of these markers, highlighting once again the benefits of large scale sequence characterization of these markers.

Casework sample profiles

Because of the primer concentration adjustments made prior to the casework sample amplifications, only five of forty-eight markers completely dropped out in these experiments. However, incomplete multiplex optimization combined with low quantity/quality templates caused low coverage in several other markers. Thus, allele recovery for the skeletal casework samples, as expected for these evidentiary specimens, was highly variable. As shown in Table 4, the alleles recovered ranged from three to fifty-six, with an average of 30.6 alleles called per sample replicate. The sample profiles were consistent between replicate amplifications (barring the expected stochastic effects of low DNA quality/quantity template amplification) and with partial genotypes generated from a commercial STR kit. Although many alleles dropped out of the 48-marker 454 runs due to the non-optimized enrichment amplification, the average allele recovery of 30.6 alleles per replicate assay run exceeded the number of alleles recovered in any single amplification with the commercial STR kit (a maximum of thirty alleles for one sample).

Given the substantial number of sequence variants observed in the pristine samples, we investigated how much additional information in the form of discriminatory SNPs might be recovered in a practical casework degraded sample scenario where a smaller number of alleles would amplify. In total, 186 SNPs were identified in over 795 recovered alleles, averaging one SNP for every four called alleles. The number of SNPs per sample replicate ranged from zero to sixteen, with an average of seven SNPs identified per replicate. For example, sample J (Table 4), which showed the highest allele recovery with the commercial kit amplification (performed in duplicate) had twenty-nine and thirty alleles recovered with no marker dropout. However, even with markers completely dropping out in the 48-marker NGS assay, forty-two and forty-nine alleles were recovered in replicate amplifications and six and nine SNPs, respectively, were identified as well. At the other end of the spectrum, replicate amplifications of a poor quality specimen, sample I, recovered three and four alleles with a commercial STR kit. Yet, our 48-marker STR assay recovered twelve and fifteen alleles with two and three SNPs, respectively. The opportunity to target only small amplicons, and also retrieve discriminatory sequence data will allow for greater information recovery from these challenging casework samples.

Quality of sequence reads

In terms of sequence data quality, we observed a vast improvement in the data produced from the smaller multiplexes. For all samples, the unusual artifacts previously observed in the D21S11 forward sequences were not observed with the 4-plex strategy (Figure S1). Likewise, the extra guanine residues in some D16S539 reads from the 13-plex (Figure S2) disappeared completely

when that marker was amplified in the 4-plex format. We suspect this is due to reduced primer interaction during PCR.

3.4 Data analysis

In terms of secondary data analysis, we approached the NGS data in two different ways early on: via de novo assembly and via reference alignment. Despite multiple attempts to perform de novo assemblies, and regardless of the mapping and alignment parameters used, this strategy never yielded useful contigs for our application. Without an automated method to separate sequences by marker first, de novo assembly attempts resulted in all contigs for all markers in a single large list of generically-named assemblies (i.e. contigs were not identified by marker and/or allele). Further, some contigs contained multiple markers, resulting in a consensus sequence containing concatenated fragments. Additionally, without software customization, de novo assembly consensus sequences (even once properly identified and split into discrete marker sequences) would either need to be aligned to reference sequences or manually assigned to a particular allele/repeat call. It is possible that de novo alignments of multiplexed STRs even with the most stringent alignment settings are simply too difficult with the current software options available. This may be due to repeat motif similarity among the different loci, and the complexities of accurately sorting very similar reads while also allowing for sufficient mismatch in the alignments to accommodate SNPs and sequencing errors. These types of issues have been previously noted in other applications [50].

As a result, reference alignments were used for all aspects of the study. For the majority of markers, alignment to a list of references worked quite well. The authentic reads aligned to the correct references, and stutter, normally one fewer repeat, was generally apparent and excluded from the authentic allele calls. An example mapping summary for one sample is shown in Table S6. Resulting contigs are sorted by marker, and authentic alleles are easily recognizable, possessing the highest total read counts for each marker. Contigs that represented stutter typically presented with read counts less than 10-15% of those for authentic alleles, and aligned to the allele with one fewer repeat. Figure 3 shows an example of a typical reference alignment. Although there are the expected intermittent incorporation errors throughout the reads, the authentic sequence is clear.

Despite the general utility of the reference alignment approach, we did encounter a few problems. For one, reference assemblies were sometimes incorrect for those markers with more complex repeat structures. Because not all possible variation was represented in the reference sequences, sequence reads would occasionally align to an allele with a different number of repeats. In those cases, the consensus of the NGS sequences contained fewer or more repeats than the reference, and thus the consensus repeats had to be counted manually to determine the correct allele call. For example, Figure S4 shows an alignment of D21S11 allele 30 reads to the allele 31 reference. With Sanger data and standard Sanger data analysis software, these types of alignment issues can be corrected by hand. However, available commercial NGS analysis software generally does not allow for this type of manual editing (and ideally it would not be required in an NGS data analysis pipeline) and thus the consensus sequences for novel variation must be reviewed and handled carefully when performing alignments to a reference.

Similar problems of reads not aligning to a reference have been reported in different contexts [51], and this potentially introduces difficulties from the standpoint of establishing that all successfully amplified loci/alleles are also successfully assembled and represented in the final analyzed data. Although our manipulation of the various assembly and alignment parameters was not exhaustive, it is fair to say that STR data analysis and assembly warrants a significant amount of further investigation. Our experience suggests that uniform NGS assembly and alignment parameters may not be appropriate for all STR loci and/or data features. Instead, custom parameters may be required for different markers, and assembly/alignment algorithms will likely require careful development and optimization to accommodate not only the unique features of the various markers and reads, but (assuming reference alignment continues to be pursued moving forward) also the uncharacterized genetic variation that may not be represented in reference sequences.

3.5 Challenges to PCR multiplex optimization for subsequent sequencing

Aside from the issues related to secondary data analysis, the other challenges we encountered were related to multiplex optimization to yield equal representation of all targeted markers. Since our primary goals were to 1. assess the feasibility and utility of STR typing by NGS on authentic casework material, 2. perform a preliminary assessment of the NGS data produced (in terms of quality and ease of analysis) and 3. get an idea of the sequence variation that is present among these markers, we did not pursue multiplex optimization any further than the twelve quadruplexes. Obviously, a highly multiplexed assay with all 48 (or more) miniSTR markers in a single reaction would be a more ideal end-product, and greater multiplexing would indeed be criticial for those forensic cases in which evidence and DNA extract is limited. And, clearly, standard miniSTR fragment analyses could also be performed in quadruplexes. However, it is the potential to multiplex tens if not hundreds of these small amplicons, combined with the additional sequence data recovered that make the NGS approach desirable.

Yet, even with a highly optimized multiplex, or if commercially produced assays are eventually available, the sensitivity of NGS approaches and the granularity of NGS data in terms of displaying/showing "background" signal (that may have otherwise gone undetected with standard CE – fragment analysis or Sanger sequencing) may necessitate a greater understanding of the origin of the kinds of data artifacts we observed in some of our results (see Figures S1 and S2). Identifying their source (which may in some cases be introduced during the NGS lab workflow or data analysis steps) and clearly defining the difference between signal and noise will be helpful in establishing which reads can be ignored as "noise" or "background". In this study, simplification of the multiplexes resulted in a dramatic improvement to the final sequence quality, and while further multiplexing would obviously be desirable for the reasons discussed above, it is also necessary to maintain sequence quality.

It is possible that alternative PCR-based enrichment strategies could minimize the effort required to optimize standard highly multiplexed amplifications, while at the same time reducing the artifacts we observed with traditional multiplex PCR. Enrichment techniques based on PCR via picoliter droplets [52-54] or chip-based PCR (e.g. Fluidigm) are two possible options and, in theory, both of them offer an opportunity to amplify a large number of markers with reduced amplification bias. This would not only facilitate implementation, but would also enhance the

utility of any PCR-based enrichment workflow. These methods of enrichment would seemingly be the most straightforward way to address high quality samples. However, in terms of the most degraded and low template specimens, these approaches may be limited by the quantity of target molecule in the extract. For the picoliter PCR technology, the likelihood of the correct template meeting up with the corresponding microdroplet may be more restricted than it would otherwise be in a standard multiplex reaction, in which all molecules are more or less free to interact with each other. Or, for a chip-based technology, the likelihood of the desired template being represented in the small volume applied to the chip may be low.

From the standpoint of low quality and low DNA template forensic specimens, and based on in-house experience with alternate NGS workflows, target enrichment based on hybridization capture seems to hold promise for enriching multiple targets while also minimizing template sampling problems [55]. When tested in our hands, a hybridization assay for human mitochondrial DNA has yielded relatively uniform coverage across the mtGenome, and others are also having success with this approach (C. Calloway, personal communication). For us, this assay has produced robust entire mitochondrial genome data from skeletal remains that are both extremely low in endogenous DNA and heavily contaminated with microbes (data not shown). We suspect that this approach would potentially prove useful with other sample types encountered in forensics as well. This method (like miniSTR amplification) is particularly effective in recovering small, fragmented templates. However, to the best of our knowledge, probe design for STR capture/enrichment is not something that has been previously described, and probe-based capture for variable-length STR targets is likely to be more complex than for mitochondrial DNA.

4. Conclusion

Regardless of the challenges encountered and the need for additional work both at the assay development and data analysis stages, the results described here demonstrate the potential of NGS for multiplexed sample and STR sequencing from authentic casework material. As previously noted, the ability to simultaneously type and sequence numerous markers with small, overlapping size ranges is likely to be one of the greatest benefits of NGS in forensic applications given the highly compromised sample types regularly encountered in casework. Though we did not attempt to further combine the 48 markers in this study, additional multiplexing that also maintains high data quality is the logical next step. Yet even with small multiplexes and severely compromised samples (which may only yield partial profiles regardless of target amplicon size), large quantities of discriminatory data were recovered in a single NGS run – both in terms of the number of markers typed, and the sequence variation detected. Among the 48 STRs typed on skeletal remains of highly variable quality, the recovered data and resulting aligned reads revealed allele recovery exceeding that of the commercially available STR kits. Furthermore, sequence variation that in some cases doubled the number of represented alleles introduces an opportunity for discrimination not possible with standard CE based STR typing. This potential to recover sequence information in addition to repeat number is likely to be useful in a number of scenarios, but will be especially valuable in those situations where only partial STR profiles are recovered and/or the question involves extended kin. Additionally, our results reveal that particular loci may provide greater potential for additional discriminatory information (Figures 1 and 2). Thus, with further characterization of population data and known

heterozygosity values, typing strategies that first target those markers most likely to yield maximum information could be devised. This type of targeted strategy may be useful in low DNA template quantity or limited evidentiary material situations, or for mixture deconvolution when many alleles may be shared (e.g. mixtures involving close kin).

It is the case, however, that additional SNP information introduces some complications when viewed from the familiar standpoint of repeat-based allele nomenclature. Conversion of the sequence-based information to repeat-based data will undoubtedly be necessary to facilitate comparisons to existing fragment-based STR profiles. However, the added benefits of recovering sequence information would be lost if allele calls were based solely on the number of repeats. Thus, some type of alternative nomenclature system that retains the repeat-based information, yet also captures the SNP variation, will need to be developed.

One potentially viable nomenclature option might parallel the system currently employed in forensic, and other, mitochondrial DNA sequencing applications. MtDNA sequences are aligned to an established reference mtDNA genome - the revised Cambridge Reference Sequence (rCRS, [56,57]), and the consensus sequence is typically reported as position-based differences from that standard reference. Similar methodology could seemingly be applied to STR sequence data: consensus sequences could be compared to a single designated reference for each marker (or perhaps, a set of accepted references representing the described alleles for each locus), and repeat and sequence variation accordingly denoted by position. The difference-based nomenclature could be used to represent and report alleles, while database queries could be performed using string-based sequences of the alleles (again mirroring forensic mtDNA, and most other sequence based applications). See the discussions of Röck et al. [58] for further information regarding the pros and cons of difference-based versus string-based database queries. There are certainly other viable STR nomenclature options as well. We simply present this one in an effort to initiate the discussion. Given the rapid pace at which investigations into NGS for forensic purposes are now proceeding, it would behoove the community to address standardization of STR nomenclature soon.

Generally speaking, our experience demonstrates that one of the greatest challenges to maximizing the information recovered from NGS sequencing of STRs may be the data analysis pipeline – of which profile nomenclature and reporting represent just one aspect. Our multiple attempts to perform de novo assembly of STR sequences using standard software tools but variable assembly parameters produced results that would require either additional data manipulation (subsequent alignment of de novo assembly consensus sequences to a reference) or extensive manual/visual interpretation. Read mapping of multiplexed marker/sample data to our 654 manually-created references (Table S3) worked substantially better in almost all instances; but challenges were encountered with some complex repeat structures and when not all sequence variation was represented in the references. As it will be critical for forensic applications to ensure that sequence read assemblies accurately reflect the amplified alleles for any given sample/marker set, custom analysis tools specific to STR sequencing may need to be developed, or custom parameters may need to be applied on a locus-by-locus basis, regardless of the data assembly method employed.

Additional work is required before STR sequencing using NGS methods can be fully validated and routinely applied in a forensic casework setting, particularly given that new assays, platforms, tools and workflows are continuously emerging. Our results, however, clearly demonstrate the transformative potential of these new technologies for typing the most commonly employed forensic markers, STRs, on highly compromised casework specimens, and highlight some critical issues specific to forensic NGS applications that would benefit from immediate discussion and longer-term research and development efforts.

5. Acknowledgments

The authors would like to thank James Canik, Lanelle Chisolm, Dr. Brion Smith, COL Louis Finelli, Lt Col Laura Regan, Dr. Tim McMahon, James Ross, Shairose Lalani, Marjorie Bland and the American Registry of Pathology for logistical and administrative support; Suzanne Barritt-Ross for permitting the use of non-probative casework samples; Dr. Cassandra Calloway for fruitful discussion; the AFDIL Emerging Technologies Section for continuous support and valuable feedback; Dr. John Butler for permission to use NIST samples for these preliminary experiments; and Becky Hill for providing known genotypes.

The opinions and assertions contained herein are solely those of the authors and are not to be construed as official or as views of the US Department of Defense, its branches, the U.S. Army Medical Research and Materiel Command, the Armed Forces Medical Examiner System, the Federal Bureau of Investigation, or the U.S. Government.

6. References

[1] B.L. Hancock-Hanser, A. Frey, M.S. Leslie, P.H. Dutton, F.I. Archer, P.A. Morin, Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics, Mol.Ecol.Resour. 13 (2013) 254-268.

[2] M. Gymrek, D. Golan, S. Rosset, Y. Erlich, lobSTR: A short tandem repeat profiler for personal genomes, Genome Res. 22 (2012) 1154-1162.

[3] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, 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.

[4] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic.Sci.Int.Genet. 6 (2012) 810-818.

[5] M.M. Holland, M.R. McQuillan, K.A. O'Hanlon, Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy, Croat.Med.J. 52 (2011) 299-313.

[6] 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. 0 (2012) 1-6.

[7] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S. Gomes, L. Souto, et al., Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM), Forensic Science International: Genetics. 7 (2013) 543-549.

[8] E. Rockenbauer, S. Hansen, M. Mikkelsen, C. Borsting, N. Morling, Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing, Forensic.Sci.Int.Genet. 8 (2014) 68-72.

[9] S. Dalsgaard, E. Rockenbauer, A. Buchard, H.S. Mogensen, R. Frank-Hansen, C. Borsting, et al., Non-uniform phenotyping of D12S391 resolved by second generation sequencing, Forensic.Sci.Int.Genet. 8 (2014) 195-199.

[10] O. Loreille, H. Koshinsky, V. Fofanov, J.A. Irwin, Application of next generation sequencing technologies to the identification of highly degraded unknown soldiers' remains, Forensic Science International: Genetics Supplement Series. 3 (2011) e540-e541, doi:10.1016/j.fsigss.2011.10.013.

[11] 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.

[12] T. Maricic, M. Whitten, S. Paabo, Multiplexed DNA sequence capture of mitochondrial genomes using PCR products, PLoS One. 5 (2010) e14004.

[13] E.M. Kenny, P. Cormican, W.P. Gilks, A.S. Gates, C.T. O'Dushlaine, C. Pinto, et al., Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection, DNA Res. 18 (2011) 31-38.

[14] O. Harismendy, P.C. Ng, R.L. Strausberg, X. Wang, T.B. Stockwell, K.Y. Beeson, et al., Evaluation of next generation sequencing platforms for population targeted sequencing studies, Genome Biol. 10 (2009) R32-2009-10-3-r32. Epub 2009 Mar 27.

[15] M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat sequencing on the 454 platform, Forensic Science International: Genetics Supplement Series. 3 (2011) e357-e358.

[16] T.A. Castoe, K.T. Hall, M.L. Guibotsy Mboulas, W. Gu, A.P. de Koning, S.E. Fox, et al., Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing, Genome Biol.Evol. 3 (2011) 641-653.

[17] J. Abdelkrim, B. Robertson, J.A. Stanton, N. Gemmell, Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing, BioTechniques. 46 (2009) 185-192.

[18] M. Allentoft, S.C. Schuster, R. Holdaway, M. Hale, E. McLay, C. Oskam, et al., Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data, BioTechniques. 46 (2009) 195-200.

[19] M.E. Allentoft, C. Oskam, J. Houston, M.L. Hale, M.T. Gilbert, M. Rasmussen, et al., Profiling the dead: generating microsatellite data from fossil bones of extinct megafauna--protocols, problems, and prospects, PLoS One. 6 (2011) e16670.

[20] Q. Santana, M. Coetzee, E. Steenkamp, O. Mlonyeni, G. Hammond, M. Wingfield, et al., Microsatellite discovery by deep sequencing of enriched genomic libraries, BioTechniques. 46 (2009) 217-223.

[21] J. Weber-Lehmann, E. Schilling, G. Gradl, D.C. Richter, J. Wiehler, B. Rolf, Finding the needle in the haystack: Differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing, Forensic.Sci.Int.Genet. 9 (2014) 42-46.

[22] J.M. Butler, Y. Shen, B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA, J.Forensic Sci. 48 (2003) 1054-1064.

[23] M.D. Coble, J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA, J.Forensic Sci. 50 (2005) 43-53.

[24] T.J. Parsons, R. Huel, J. Davoren, C. Katzmarzyk, A. Milos, A. Selmanovic, et al., Application of novel "mini-amplicon" STR multiplexes to high volume casework on degraded skeletal remains, Forensic.Sci.Int.Genet. 1 (2007) 175-179.

[25] P. Grubwieser, R. Muhlmann, B. Berger, H. Niederstatter, M. Pavlic, W. Parson, A new "miniSTR-multiplex" displaying reduced amplicon lengths for the analysis of degraded DNA, Int.J.Legal Med. 120 (2006) 115-120.

[26] M.N. Gabriel, E.F. Huffine, J.H. Ryan, M.M. Holland, T.J. Parsons, Improved MtDNA sequence analysis of forensic remains using a "mini-primer set" amplification strategy, J.Forensic Sci. 46 (2001) 247-253.

[27] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline, Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations, J.Forensic Sci. 48 (2003) 908-911.

[28] O.M. Loreille, T.M. Diegoli, J.A. Irwin, M.D. Coble, T.J. Parsons, High efficiency DNA extraction from bone by total demineralization, Forensic.Sci.Int.Genet. 1 (2007) 191-195.

[29] F. Pitterl, H. Niederstatter, G. Huber, B. Zimmermann, H. Oberacher, W. Parson, The next generation of DNA profiling--STR typing by multiplexed PCR--ion-pair RP LC-ESI time-of-flight MS, Electrophoresis. 29 (2008) 4739-4750.

[30] H. Oberacher, F. Pitterl, G. Huber, H. Niederstatter, 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.

[31] D.R. Hares, Expanding the CODIS core loci in the United States, Forensic.Sci.Int.Genet. 6 (2012) e52-4.

[32] C.R. Hill, M.C. Kline, M.D. Coble, J.M. Butler, Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples, J.Forensic Sci. 53 (2008) 73-80.

[33] J.A. Irwin, R.S. Just, O.M. Loreille, T.J. Parsons, Characterization of a modified amplification approach for improved STR recovery from severely degraded skeletal elements, Forensic.Sci.Int.Genet. 6 (2012) 578-587.

[34] A. Berti, F. Barni, E. Pilli, E. Rizzi, A. Pianese, G. Corti, et al., A new ultradeep LT (low template) DNA profiling approach based on an emulsion-based clonal amplification of an STRs multiplex PCR product followed by massive pyrosequencing. Presented at 24th World Congress of the International Society for Forensic Genetics, Vienna, Austria, 2011.

[35] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature. 437 (2005) 376-380.

[36] Rapid Library Preparation Method Manual: GS Junior Titanium Series, (2010).

[37] emPCR Amplification Method Manual - Lib-L: GS Junior Titanium Series, (2011).

[38] Sequencing Method Manual: GS Junior Titanium Series, (2011).

[39] Amplicon (PCR Product) Sequencing Tips for GS FLX Titanium Reagents, (2010).

[40] Agencourt® AMPure XP®: PCR Purification, (2009).

[41] 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.

[42] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, et al., GenBank, Nucleic Acids Res. 41 (2013) D36-42.

[43] C.R. Hill, M.D. Coble, J.M. Butler, Characterization of 26 New miniSTR Loci. Presented at 17th International Symposium on Human Identification, Nashville, TN, 2006, http://www.cstl.nist.gov/strbase/pub_pres/Promega2006_Hill.pdf.

[44] CLC Genomics Workbench User manual, (2011).

[45] P. Gill, R. Sparkes, C. Kimpton, Development of guidelines to designate alleles using an STR multiplex system, Forensic Sci.Int. 89 (1997) 185-197.

[46] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res. 24 (1996) 2807-2812.

[47] J.A. Irwin, M.D. Leney, O. Loreille, S.M. Barritt, A.F. Christensen, T.D. Holland, et al., Application of low copy number STR typing to the identification of aged, degraded skeletal remains, J.Forensic Sci. 52 (2007) 1322-1327.

[48] G. Levinson, G.A. Gutman, Slipped-strand mispairing: a major mechanism for DNA sequence evolution, Mol.Biol.Evol. 4 (1987) 203-221.

[49] M.C. Kline, C.R. Hill, A.E. Decker, J.M. Butler, STR sequence analysis for characterizing normal, variant, and null alleles, Forensic.Sci.Int.Genet. 5 (2011) 329-332.

[50] T.J. Treangen, S.L. Salzberg, Repetitive DNA and next-generation sequencing: computational challenges and solutions, Nat.Rev.Genet. 13 (2011) 36-46.

[51] D.J. Hedges, T. Guettouche, S. Yang, G. Bademci, A. Diaz, A. Andersen, et al., Comparison of three targeted enrichment strategies on the SOLiD sequencing platform, PLoS One. 6 (2011) e18595.

[52] J.H. Leamon, D.R. Link, M. Egholm, J.M. Rothberg, Overview: methods and applications for droplet compartmentalization of biology, Nat.Methods. 3 (2006) 541-543.

[53] T. Geng, R. Novak, R.A. Mathies, Single-Cell Forensic Short Tandem Repeat Typing within Microfluidic Droplets, Anal.Chem. (2013).

[54] R. Tewhey, J.B. Warner, M. Nakano, B. Libby, M. Medkova, P.H. David, et al., Microdroplet-based PCR enrichment for large-scale targeted sequencing, Nat.Biotechnol. 27 (2009) 1025-1031.

[55] J.E. Templeton, P.M. Brotherton, B. Llamas, J. Soubrier, W. Haak, A. Cooper, et al., DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification, Investig.Genet. 4 (2013) 26-2223-4-26.

[56] S. Anderson, A.T. Bankier, B.G. Barrell, M.H. de Bruijn, A.R. Coulson, J. Drouin, et al., Sequence and organization of the human mitochondrial genome, Nature. 290 (1981) 457-465.

[57] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat.Genet. 23 (1999) 147.

[58] A. Röck, J. Irwin, A. Dur, T. Parsons, W. Parson, SAM: String-based sequence search algorithm for mitochondrial DNA database queries, Forensic.Sci.Int.Genet. 5 (2011) 126-132.

- Study demonstrates feasibility and accuracy of miniSTR typing by NGS
- Large quantities of information recovered from both high and low quality samples
- Discriminatory information (sequence variation) gleaned by NGS exceeds CE typing
- Challenges of NGS data assembly and analysis for STRs are highlighted

Figure 1

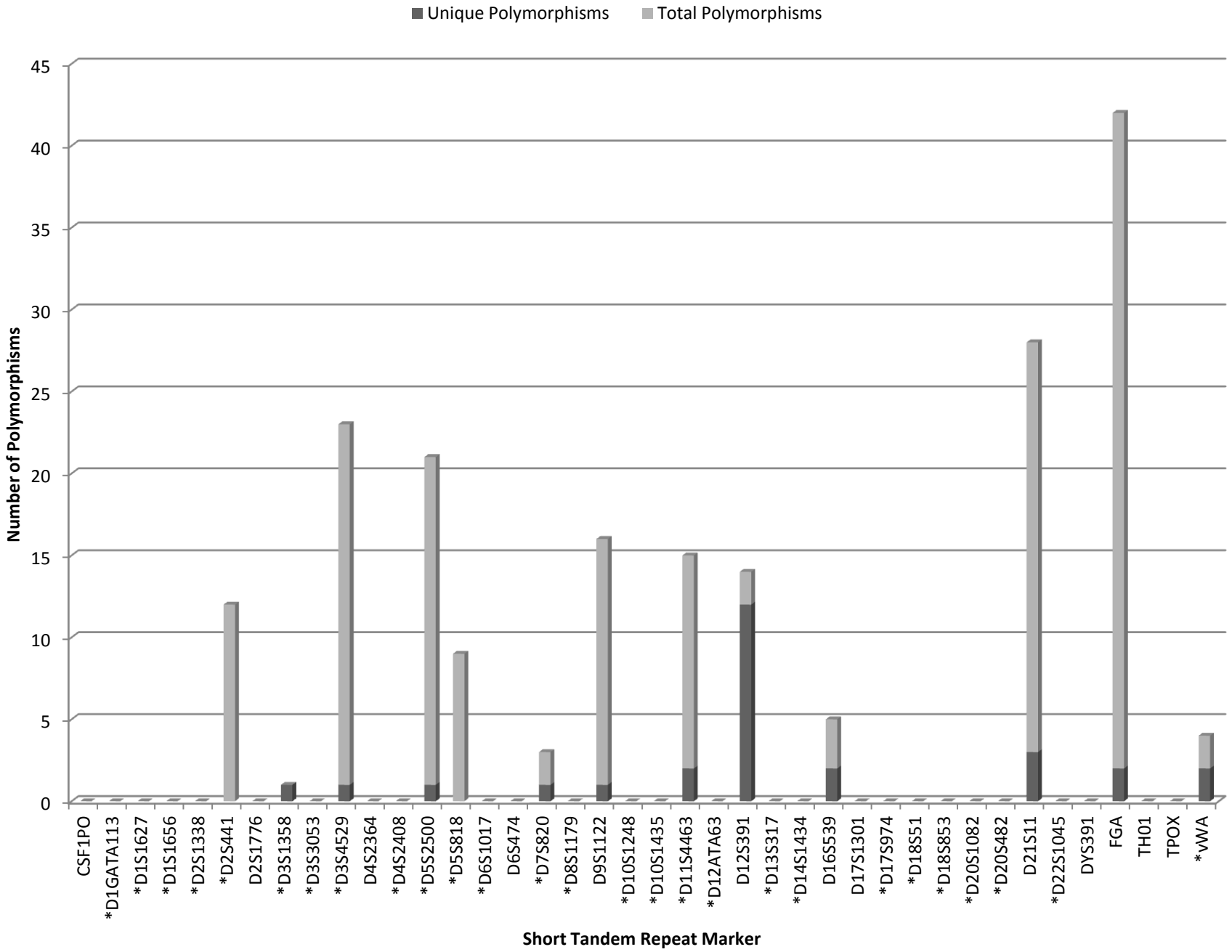

Unique Polymorphisms
Total Polymorphisms
Number of Polymorphisms
0
5
10
15
20
25
30
35
40
45
CSF1PO
*D1GATA113
*D1S1627
*D1S1656
*D2S1338
*D2S441
D2S1776
*D3S1358
*D3S3053
*D3S4529
D4S2364
*D4S2408
*D5S2500
*D5S818
*D6S1017
D6S474
*D7S820
*D8S1179
D9S1122
*D10S1248
*D10S1435
*D11S4463
*D12ATA63
D12S391
*D13S317
*D14S1434
D16S539
D17S1301
*D17S974
*D18S51
*D18S853
*D20S1082
*D20S482
D21S11
*D22S1045
DYS391
FGA
TH01
TPOX
*vWA
Short Tandem Repeat Marker

Figure 2

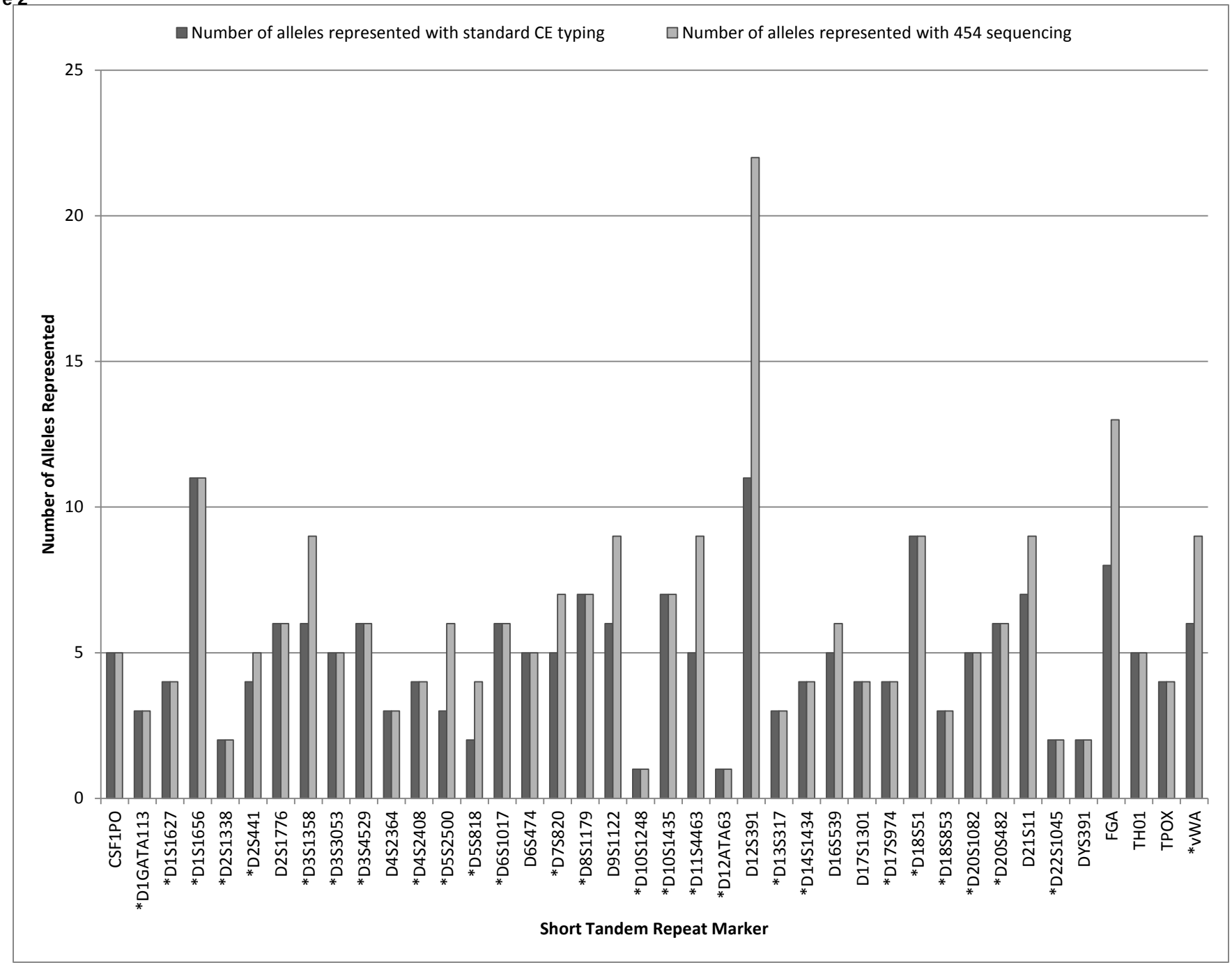
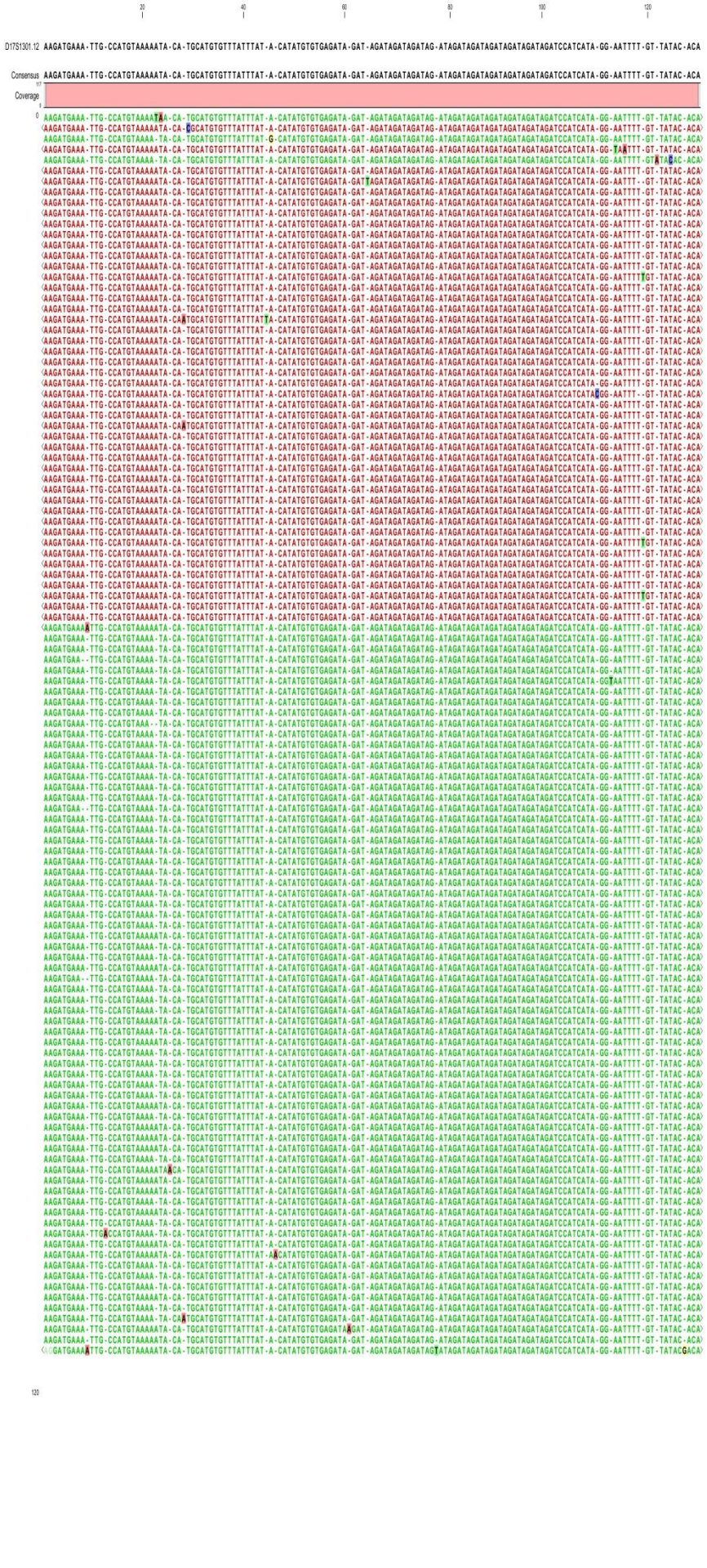

Number of alleles represented with standard CE typing
Number of alleles represented with 454 sequencing
Number of Alleles Represented
0
5
10
15
20
25
CSF1PO
*D1GATA113
*D1S1627
*D1S1656
*D2S1338
*D2S441
D2S1776
*D3S1358
*D3S3053
*D3S4529
D4S2364
*D4S2408
*D5S2500
*D5S818
*D6S1017
D6S474
*D7S820
*D8S1179
D9S1122
*D10S1248
*D10S1435
*D11S4463
*D12ATA63
D12S391
*D13S317
*D14S1434
D16S539
D17S1301
*D17S974
*D18S51
*D18S853
*D20S1082
*D20S482
D21S11
*D22S1045
DYS391
FGA
TH01
TPOX
*vWA
Short Tandem Repeat Marker

Figure 3


D17S1301.12
AAGATGAAA-TTG-CCATGTAAAAATA-CA-TGCATGTGTTTATTTAT-A-CATATGTGTGAGATA-GAT-AGATAGATAGATAG-ATAGATAGATAGATAGATAGATAGATCCATCATA-GG-AATTTT-GT-TATAC-ACA
Consensus
AAGATGAAA-TTG-CCATGTAAAAATA-CA-TGCATGTGTTTATTTAT-A-CATATGTGTGAGATA-GAT-AGATAGATAGATAG-ATAGATAGATAGATAGATAGATAGATCCATCATA-GG-AATTTT-GT-TATAC-ACA
Coverage

Sample	Alleles recovered by NGS	% Alleles verified by CE	SNPs observed via NGS
GT38070	20	100%	0
GT36877	19	100%	5
GT38087	19	100%	2
GT38107	19	100%	2
GC03394	19	95%	1
GT36880	19	79%	4
TT51435	18	100%	3
GT38065	18	100%	3
GT38095	18	100%	3
TT51422	17	100%	2
GT38093	17	100%	4
GT38119	17	100%	3
PT83912	16	100%	1
PT84541	16	100%	4
PT84183	15	100%	1
MT95744	15	100%	2
GT38100	15	100%	2
GT38076	15	100%	1
GT37864	14	100%	3
GT37047	14	100%	2
GT38098	14	100%	1
GT37168	13	100%	6
GT38069	13	100%	1
GT38081	12	100%	0

Sample	1					
Extraction replicate		A			B	
Amplification replicate	1	2	3	1	2	3
Marker						
CSF1PO	10	10	10	10	10	10
D10S1248						
D13S317			12			
D16S539		11		11	11	11
D21S11						30
D22S1045	15					15
D2S441	11	10		10		10,11
D3S1358				18	18	
D5S818						
D7S820	11					
D8S1179						
TPOX	8	8	8	8	8	8
vWA		18				
Alleles called	5	5	3	5	4	7
Total polymorphisms observed	0	1	0	1	0	2

Sample	Alleles recovered by NGS	% Alleles verified by CE	SNPs observed via NGS
GT38100	61	92%	10
GT38107	61	92%	10
GT38081	59	92%	10
GT38093	58	91%	10
GT38076	57	91%	11
9948	54	96%	6
9947A	54	93%	14
GT38069	54	91%	8
GT36880	53	38%	9
GT38095	52	90%	5
GT38087	52	90%	3
GT38070	51	92%	8
TT51422	50	88%	12
GT38098	48	92%	12
GT36877	47	94%	6
K562	43	95%	8
GT38119	43	91%	3
GC03394	38	87%	7

Sample	A		B		C	
Amplification replicate	1	2	1	2	1	2
Marker						
Amelogenin	**X,Y**	**X,Y**	X,Y	Y	Y	X,Y
CSF1PO	11,12	11,12	12	11,12		
D1GATA113	12	12	12			11,12
D1S1627	14	13,14	14	14		
D1S1656	14	14		14		
D1S1677		12				
D2S1338				17		
D2S1776	12	12	12			12
D2S441	11	11,14	11	14		
D3S1358	**15'**	**15'**		15'		
D3S3053		11,12	11,12			
D3S4529		14		14,15		
D4S2364	9	9	9			
D4S2408		10	10,11	10,11		
D5S2500	14,17	14	14	14,17		
D5S818			11			
D6S1017	10	10	10	10		10
D6S474		17	17	17		
D7S820	9	9				
D8S1115						
D8S1179		14	14	14		11
D9S1122	12	12	12	11,12		
D9S2157						
D10S1248						
D10S1435			12,14			
D11S4463	14	16		14,16		
D12ATA63						14
D12S391		17,21'	17			
D13S317						
D14S1434	13	13	13	13		
D16S539	11	11,12		11,12		
D17S1301		11,12	11			
D17S974		9				
D18S51		17		17		
D18S853			14	11,14	11	
D19S433						
D20S1082	11	11	11	11		
D20S482			14,16			14
D21S11	28,**31.2**	28,**31.2**	31.2	31.2		
D22S1045		15,16	15	15,16		16
DYS391	11	11		11		
FGA	23.3,24.2	23.2	23,24			
Penta D						

Figures and Tables

Figure 1. Polymorphism information by locus for eighteen high-quality samples using the 48 marker assay

Total and unique polymorphisms (SNPs or indels) observed across fifteen population samples and three positive controls sequenced using the 48 marker NGS assay. An asterisk (*) designates those markers known to have at least some allele dropout, determined by comparison to CE fragment analysis data. No variation in repeat structure was identified in the alleles recovered for twenty-seven of the markers, while thirteen markers showed SNPs when compared to the reference sequence.

Figure 2. Allele counts by fragment analysis versus sequencing

For the fifteen population samples and three positive controls typed using the 48 marker NGS assay, the data were evaluated to determine the number of alleles that would have been detected by fragment analysis versus the number of alleles revealed when sequence variation was taken into account. An asterisk (*) designates those markers known to have at least some allele dropout (determined by comparison to CE data). For twenty-eight of the loci, the same number of alleles would have been detected by both fragment analysis and sequencing. For the remaining twelve markers, utilizing sequence information would increase the number of alleles detected relative to CE data. Three of the markers (D5S818, D5S2500, and D12S391) contain double the number of alleles when using sequence information instead of fragment size alone.

Figure 3. Example alignment of 454 STR sequence reads to a reference sequence

Although there are expected intermittent base incorporation errors throughout the reads, the authentic sequence is easily identified.

Table 1. Population sample profile summary for 13-plex

For each sample tested, the alleles recovered by NGS, the percentage of those alleles confirmed by CE data, and the number of SNPs observed were tallied. Alleles recovered ranged from 12 to 20, and nearly all (99%) called alleles were verified by commercial STR kit typing and existing CE data. The 5 unconfirmed allele calls were due to 1) incomplete CE genotype information for sample GT36880, and 2) a high stutter allele (drop-in) in sample GC03394.

Table 2. Casework sample profiles for 13-plex

Calls confirmed with traditional CE typing in 2 of 3 replicate amplifications are bolded. For example, no alleles weare duplicated in replicate CE typing for Sample 1, therefore no alleles detected by 454 sequencing are bolded. The number of NGS alleles called and the observed polymorphisms were tallied (at the bottom of the table) for each amplification replicate.

Table 3. Population sample profile summary for 48 marker assay

For the 15 population samples and 3 controls typed using the 48 marker assay, the number of alleles recovered, the alleles verified by CE, and the number of SNPs observed were tallied. Allele recovery for these samples ranged from 38 to 61, and 89% of all called alleles were verified by CE data. Nearly all of the 107 unconfirmed alleles were due to 1) incomplete CE data for one sample, and 2) three loci that are not included in accessible commercial STR kits (D1S1656, D12S391 and DYS391).

Table 4. Casework sample profiles for 48 marker assay

Markers typed and calls confirmed with traditional CE typing in 2 of 2 replicate amplifications are bolded. For example, no alleles were replicated in duplicate CE typing for Sample B, therefore no alleles detected by 454 sequencing are bolded. The number of NGS alleles called and the observed polymorphisms were tallied for each amplification replicate.