Published in final edited form as:

Forensic Sci Int Genet. 2016 March ; 21: 15–21. doi:10.1016/j.fsigen.2015.11.005.

Sequence variation of 22 autosomal STR loci detected by next generation sequencing

Katherine Butler Gettings[1,*], **Kevin M. Kiesler**[1], **Seth A. Faith**[2], **Elizabeth Montano**[3], **Christine H. Baker**[3], **Brian A. Young**[3], **Richard A. Guerrieri**[3], and **Peter M. Vallone**[1]

Katherine Butler Gettings: katherine.gettings@nist.gov; Kevin M. Kiesler: kevin.kiesler@nist.gov; Seth A. Faith: safaith@ncsu.edu; Elizabeth Montano: montanoe@battelle.org; Christine H. Baker: bakerc@battelle.org; Brian A. Young: youngb@battelle.org; Richard A. Guerrieri: guerrierir@battelle.org; Peter M. Vallone: peter.vallone@nist.gov

[1]U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA

[2]North Carolina State University, College of Veterinary Medicine and Forensic Science Institute, 1060 William Moore Drive, Raleigh, NC 27607, USA

[3]Battelle Memorial Institute, 505 King Avenue, Columbus, OH 43201, USA

Abstract

Sequencing short tandem repeat (STR) loci allows for determination of repeat motif variations within the STR (or entire PCR amplicon) which cannot be ascertained by size-based PCR fragment analysis. Sanger sequencing has been used in research laboratories to further characterize STR loci, but is impractical for routine forensic use due to the laborious nature of the procedure in general and additional steps required to separate heterozygous alleles. Recent advances in library preparation methods enable high-throughput next generation sequencing (NGS) and technological improvements in sequencing chemistries now offer sufficient read lengths to encompass STR alleles. Herein, we present sequencing results from 183 DNA samples, including African American, Caucasian, and Hispanic individuals, at 22 autosomal forensic STR loci using an assay designed for NGS. The resulting dataset has been used to perform population genetic analyses of allelic diversity by length compared to sequence, and exemplifies which loci are likely to achieve the greatest gains in discrimination via sequencing. Within this data set, six loci demonstrate greater than double the number of alleles obtained by sequence compared to the number of alleles obtained by length: D12S391, D2S1338, D21S11, D8S1179, vWA, and D3S1358. As expected, repeat region sequences which had not previously been reported in forensic literature were identified.

Introduction

Length variations among individuals in short tandem repeat (STR) loci have been used in forensic applications since the 1990s, due to the ease with which these loci can be multiplexed combined with a high degree of heterogeneity. Over the years, many researchers

[*]Corresponding Author's Contact Information: Katherine Butler Gettings, National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA, Phone: 301-975-6401, Fax: 301-975-8505, katherine.gettings@nist.gov.

have performed Sanger sequencing of forensic STR loci in order to better understand discordances between capillary electrophoresis (CE) kits, microvariant alleles, null alleles, and mutational events [1–7]. However, routine Sanger sequencing of STRs is not practical, as loci cannot be multiplexed and heterozygous alleles must be physically separated prior to sequencing.

Massively parallel sequencing methods (herein referred to as next generation sequencing or NGS) can simultaneously sequence many thousands of genomic regions in a single reaction. Industry competition has led to drastic drops in sequencing costs in recent years, and advances in library preparation methods and read lengths now enable sequencing of forensic STR loci, as demonstrated by several laboratories [8–15]. While these publications highlight the potential gains via NGS, the methods used are low-throughput in samples and/or loci interrogated. A high-throughput approach is needed not only for forensic DNA databasing, but also to reduce the sequencing cost, which is equally important in both casework and databasing applications. In addition, bioinformatics methods for STR sequence data analysis must maintain back compatibility with length-based methods and corresponding existing forensic DNA databases, such as NDIS.

In the work presented here, NGS of 22 autosomal STR loci was performed on 183 population samples, manually in 96-well format. Two bioinformatics pipelines were used to analyze the data, results were compared to CE data, and discrepancies were investigated further. Population genetic analyses including probability of identity (PI) and heterozygosity (Het) were performed on length- and sequence-based alleles. The sequences obtained in this limited data set give an indication of the level of diversity expected in the larger population and provide examples of how isoalleles (alleles of the same length but different sequence) can improve discrimination and mixture deconvolution in forensic casework. Further, demonstrating successful results in a manual 96-well approach indicates the possibility of automated high-throughput sample processing.

Materials and Methods

DNA extracts from NIST population samples (n = 183) were selected to represent individuals of self-identified ancestry from three categories: African American (n = 68), Caucasian (n = 70), and Hispanic (n = 45). These are the three most common population groups in the United States, and published CE data across multiple kits exist for these well-characterized population samples [16].

DNA extracts were quantified with Quantifiler Human DNA Quantification Kit (Life Technologies, Carlsbad CA, USA) on an ABI 7500 Real Time PCR System (Life Technologies). Based on Quantifiler results, samples were normalized to 0.5 ng/μL.

DNA samples were amplified with a prototype version of the PowerSeq Auto System (Promega Corporation, Madison WI, USA), which includes the same loci amplified in PowerPlex Fusion (Promega): 22 autosomal STR loci, one Y-STR locus (DYS391), and Amelogenin. These loci are inclusive of the expanded US CODIS core loci and the 12 core

European Standard Set loci. The PowerSeq Auto System is designed for NGS: it contains non-labeled primers and produces amplicons between 129 and 284 base pairs in size.

The amplification reaction consisted of 5 μL 5× reaction mix, 5 μL 5× primer mix, 14 μL H_20, and 1 μL sample (0.5 ng) for a total reaction volume of 25 μL. Amplification was performed on an Applied Biosystems GeneAmp 9700 thermal cycler with the following parameters: 96 °C hold for 1 min; 30 cycles of 94 °C for 10 s, 59 °C for 1 min, 72 °C for 30 s; 60 °C hold for 10 min; 4 °C soak. Following the instruction of the assay developers, each sample was amplified two times, then both reactions were pooled prior to purification.

Samples were amplified in 96-well plate format containing 94 DNA samples per plate, one negative control consisting of PCR master mix without DNA template (NTC), and one negative control for library preparation consisting of an empty well (LNEG). Within the 188 DNA samples, the five single source components (A, B, C, E, and F) of NIST Standard Reference Material 2391c *PCR-Based DNA Profiling Standard* were run as positive controls; CE and Sanger sequencing data has been published for these samples [17].

PCR products were purified with the QIAquick 96 PCR Purification Kit (Qiagen, Limburg, Germany) according to the manufacturer's instructions. For each plate, a subset (n = 12) of the purified PCR products were quantified with the Qubit dsDNA HS Assay Kit (Life Technologies) according to the manufacturer's instructions. Based on an average value of the Qubit results per plate, approximately 1 μg of each purified PCR product was used as input for sequencing library preparation. Individual samples were not normalized with respect to DNA input for the library preparation procedure.

Sequencing template libraries were prepared with the TruSeq DNA PCR-Free Sample Preparation Kit HS (Illumina, San Diego CA, USA) following the manufacturer's protocol (Illumina part # 15036187 Rev. B). For the first 96-well plate, the size selection option for 550 bp insert libraries was performed (NOTE: manufacturer's protocol recommends using 2 μg purified product as input for this size selection option). Due to low final concentration of the libraries for the first plate, the size selection procedure was changed to the 350 bp library insert option in order to retain smaller library molecules when processing the second 96-well plate. After the ligation and cleanup procedure in the TruSeq protocol, 48 samples from each 96-well plate were quantified using the Kapa Biosystems qPCR Master Mix and Primer Premix (ABI Prism) for Illumina Platforms (Kapa Biosystems, Wilmington MA, USA) on an ABI 7900 quantitative PCR (qPCR) instrument (Life Technologies). Libraries from both plates were sufficient for sequencing.

Libraries were pooled on an equal volume basis and the average concentration from the qPCR quantification was used as the nominal concentration in the procedure "Preparing Libraries for Sequencing on the MiSeq" (Illumina part # 15039740 Rev. D), where both library pools were denatured according to the instructions for a 2 nmol/L library. PhiX control was spiked into the sample library at 5 % to compensate for the possibility of low library diversity. Sequencing was performed on the MiSeq system (Illumina) using the 600 cycle MiSeq Reagent Kit v3 (Illumina) for 2 × 300 paired end sequencing following the

procedure in the “MiSeq Reagent Kit v3 Reagent Preparation Guide” (Illumina part # 15044983 Rev. B) and the “MiSeq System User Guide” (Illumina part # 15027617 Rev. N).

Bioinformatics

Following completion of the sequencing run, FASTQ files were generated by MiSeq Reporter (Illumina). These files were analyzed with two independent bioinformatic pipelines to cross-validate the results.

1. Length-based allele calls, bracketed repeat region sequences, and coverage levels were generated using the proprietary software, ExactID (Battelle Memorial Institute, Columbus OH, USA). True allele and stutter responses were discriminated from noise using an analytical threshold of 160 reads in all cases where the FASTQ files were > 50 MB per sample, and 50 reads for files less than 50 MB per sample.
2. The Perl script STRait Razor version 1.5 [18] was automated for batch sample processing using Bpipe [19]. STRait Razor output files were parsed with a custom Java script which produced length-based allele calls, repeat region sequences, and coverage statistics. Sequences that were in the majority, defined as the highest coverage allele per locus and any additional alleles coverage ≥ 40% of the highest coverage allele with complementary and balanced forward/ reverse reads, were reported as true alleles.

Genotypes from both ExactID and STRait Razor were independently analyzed for concordance to CE-based genotypes generated previously with PowerPlex Fusion (Promega) [16]. The concordance checks were performed using the VLOOKUP function in Excel (Microsoft, Redmond WA, USA). Discordances were evaluated further to determine the true genotype/sequence, and manual corrections were made as needed.

Population Genetics

Sequence variants per locus, per allele, and per population were compiled and counted. The probability of identity (PI) and heterozygosity (Het) were calculated per locus within and across populations, and compared between loci across populations. The genotype for length-based alleles are the two STR repeat numbers (e.g., D2S1338: 17,17) whereas the genotype for sequence-based alleles are the two sequences (e.g., D2S1338: [TGCC]4[TTCC]13, [TGCC]6[TTCC]11). All calculations were performed using Excel.

Results and Discussion

Sequencing results

As this sequencing method was a combination of prototype amplification primers and a library preparation method intended for fragmented DNA, minimal treatment of run performance metrics were assessed. Averaged metrics are followed by +/− standard deviation.

Average library quantitation values for the first and second plates were 0.98 nmol/L (+/− 0.434 nmol/L) and 2.69 nmol/L (+/− 0.759 nmol/L), respectively; final library pool concentration was adjusted to 2.0 nmol/L for the second plate, while the first plate library pool was not diluted for the final library loading procedure. Differences in library yields are attributable to the change in size selection during library construction which targeted 550 bp for the first plate and 350 bp for the second plate. Library construction was not repeated for the first plate as a sufficient quantity of library molecules was present on the MiSeq flowcell to generate satisfactory sequencing coverage.

Cluster densities for the first and second plates were 463 K/mm^2 (+/− 6 K/mm^2) and 1218 K/mm^2 (+/− 28 K/mm^2), respectively. As a result, final sequencing yield was 6.71 Gb for the first plate and 16.01 Gb for the second plate. Sequencing yields correlated to the average library concentration loaded into the MiSeq cartridge.

The total number of paired sequencing reads for the first and second plates were 1.137×10^7 and 2.885×10^7 respectively. Based on these figures, the average number of paired sequencing reads per sample was approximately 9.9×10^4 (+/− 4.0×10^4) and 2.9×10^5 (+/− 9.9×10^4), respectively for the first and second runs. After filtering out non-majority reads using the STRait Razor/Java informatics pipeline, the average per locus coverage (the number of sequence reads per locus) was 2540 × (+/− 430) and 8290 × (+/− 3470) for runs one and two, respectively.

Supplementary Figure 1 contains heatmaps of the coverage per sample and per locus for both plates.

Positive and negative control results

The five single source components (A, B, C, E, and F) of NIST Standard Reference Material 2391c *PCR-Based DNA Profiling Standard* were compared to the published Sanger sequencing data [17]. For these five samples, the NGS sequence data for all STR loci included in the prototype PowerSeq Auto System (22 autosomal STR loci and one Y-STR locus) were concordant with the published Sanger sequence data. Results at the Amelogenin locus were also concordant with published data.

Sequences detected in the negative control samples NTC and LNEG can be categorized as follows: 1) sequences which contain only ambiguous “N” bases, 2) sequences which contain base calls but do not bin to an STR locus in the bioinformatic pipeline, and 3) sequences which contain base calls and do bin to a locus in the bioinformatic pipeline. Sequences in category three are the most likely to interfere with analysis. The highest number of sequences in category three (i.e. coverage of an STR allele) in a negative control sample in this study was 25× coverage. Analysis results from the NTC and LNEG controls from each plate are included in Supplementary Table 1.

CE concordance check results

Concordance was evaluated between the length-based CE genotype and the length-based NGS genotype for all 183 samples and all 24 loci, resulting in evaluation of 4,392 loci. Both bioinformatic pipelines produced numerical genotypes that were over 99% concordant to the

CE genotypes. Instances of discord, described below, were all attributable to bioinformatic configurations. There were no instances of discord attributable to sequencing errors.

Three categories of discord were observed in comparing the two NGS pipeline results to the CE data.

1. Flanking region InDels: Traditional CE analysis accounts for the entire amplified region when determining the STR allele length. Both bioinformatic methods used in this study employ recognition sites adjacent to the repeat region. When an InDel was present outside of the bioinformatic recognition sites but within the CE amplified region, a discordant result was obtained. One example of this was seen at the D13S317 locus, where a four base deletion in the 3′ flanking region resulted in a “9” allele by CE and a “10” allele by NGS. In this case, there were 10 repeat units present; therefore, the NGS result was not in error, but rather was discordant in comparison to CE data.
2. Bioinformatic null allele, Type 1: In this category of discord, deleted bases in the same region as a bioinformatic recognition site caused one allele to go undetected in the bioinformatic pipeline, resulting in the erroneous appearance of a homozygote. This was observed at the Penta D locus, where a 13 base deletion adjacent to the 5′ repeat region caused the sequence to go undetected in one bioinformatic method. This deletion is common in individuals of African descent (resulting in an x.2 allele via length-based CE analysis).
3. Bioinformatic null allele, Type 2: In this case, the bioinformatic configuration file did not contain a bin matching the observed allele. One example of this type of discord was observed at the D12S391 locus, when a “17, 17.1” genotype by length-based CE and one bioinformatic method appeared as a “17” homozygote by the other bioinformatic method. The cause was confirmed by the lack of a 17.1 allele definition in the bioinformatic configuration file.

In all cases of discordant data, it should be noted that the original sequence data was correct. There are several possible approaches to overcoming each category of discord. In the case of flanking region InDels and type 1 bioinformatic null alleles, determining common variants that may be encountered and moving the bioinformatic recognition sites accordingly is one strategy to improving concordance with length-based CE data. Type 2 bioinformatic null alleles can be remedied by accounting for all possible length variants at each locus in a configuration file. For both versions of bioinformatic null alleles, it was also determined that errors such as this could be detected by evaluating expected versus observed locus coverage based on interlocus balance across the sample set. The source of each bioinformatic null allele detected in this study has been addressed in more recent versions of ExactID and STRait Razor [20], as applicable.

The results of this concordance check demonstrate the utility of using multiple bioinformatic analysis methods and comparing results to length-based CE data, particularly at this early

stage in bioinformatic pipeline development. However, this analysis and the resulting categories of discord should not be considered comprehensive for multiple reasons. Analyzing more samples with the methods outlined above may reveal additional sources/types of error. In addition, it is expected that other approaches for obtaining STR genotypes from sequence data will be developed, and each bioinformatic approach may have different limitations. Analyses such as this are an important step prior to forensic implementation of NGS.

Alleles obtained by sequence

At the highest level of analysis, counting the alleles obtained by length and by sequence reveals which loci exhibit the largest increase in alleles via sequencing. Table 1 contains the number of alleles obtained by length and by sequence, sorted by the difference. In this set of 183 individuals, the number of alleles obtained by sequence more than doubles what was obtained by length at six loci: D12S391, D2S1338, D21S11, D8S1179, vWA, and D3S1358. Nine additional loci show moderate gains in number of alleles obtained by sequence (D1S1656, D2S441, FGA, D18S51, Penta E, D19S433, D5S818, CSF1PO, and D10S1248), while seven loci do not gain any additional alleles by sequence (D13S317, D16S539, D22S1045, D7S818, Penta D, TH01, and TPOX).

These findings are largely consistent with the published literature. For example, Scheible et al. [14] studied 19 of the loci herein on n=18 population or control samples, and reported the most gain in alleles by sequence at D12S391, with additional gains by sequence for D2S441, D3S1358, FGA, vWA, and D21S11. In addition, Zeng et al. [13], using the same multiplex as this study on n=24 population samples, detected sequence-based heterozygotes at the D21S11, D2S1338, D3S1358, D8S1179, and vWA loci. Gelardi et al. [8] focused population sample sequencing efforts (n=197 Danish individuals) on three loci confirmed herein as highly discriminating: D3S1358, D12S391, and D21S11. Dalsgaard et al. [15] and Rockenbauer et al. [9] each employ sequence data at one locus to address specific questions of CE allele resolution (D12S391) and mutation in parentage cases (D21S11), respectively.

Scheible et al. [14] reported no gain in alleles by sequence for D2S1338; however, only two alleles total were detected at this locus among the 18 samples sequenced and dropout was confirmed by CE comparison. Lastly, gains were reported for three simple repeat loci (D5S818, D7S820, and D16S539) which are likely owed to flanking region polymorphisms, not investigated herein.

For the gains in alleles to have maximal impact in improving discrimination among individuals for forensic applications, it is important that the sequence variation be well distributed across alleles. Figure 1 shows the number of sequence variants observed per length-based allele, per locus.

For the six loci showing the greatest gain in alleles by sequencing, the correlation coefficient (Pearson’s r) was calculated between the number of sequence-based alleles per length-based allele and the global frequency of that length-based allele, as reported in [16]. At five of these loci (D12S391, D2S1338, D21S11, D8S1179, and D3S1358) the correlation coefficient ranged from 0.56 to 0.83, indicating the gains in alleles by sequence may be

distributed in positive correlation to length-based allele frequencies. At the vWA locus, the correlation coefficient was near zero, inferring no correlation between distribution of gains in alleles by sequence (highest for alleles 14 and 15) to length-based allele frequencies (highest for alleles 16 and 17). Reasons for this finding at the vWA locus derive from sequence variants that appear to be associated with alleles 14 and 15 and the presence of two sequence variants for all other length-based alleles. A full listing of sequence variants obtained at all loci, including notation of alleles not previously reported in forensic literature [21] is included in Supplementary Table 2.

Population genetic analyses

The results of probability of identity (PI) and heterozygosity (Het) analyses can be seen in Table 2, and are separated by population, as well as presented as averages across populations. In addition, the loci were ranked based on their PI and Het averages across populations by length and by sequence, with #1 being the most discriminating locus and #22 being the least. The change in this ranking by going from length to sequence is also reported in Table 2, and allows for comparison between loci. The loci are divided in the same way as Table 1, with the first (left most) group demonstrating the greatest increases in alleles by sequence, the second (middle) group demonstrating moderate increases in alleles by sequence, and the third (right most) group having no additional alleles by sequence.

The increase in Het averages across populations for the first grouping of loci shown in Table 2 range from 4 % to 12 %, and represent the percentage of individuals at each of these loci who were homozygous by length and became heterozygous by sequence. The middle grouping of loci have an increase in Het ranging from zero to 8 %, whereas the last grouping of loci with no additional alleles have no increase in Het. The decrease in PI averages across populations for the first grouping of loci shown in Table 2 range from 1 % to 6 %, and represent the decrease in probability that two individuals will have the same genotype at a locus. The middle grouping of loci have a decrease in PI ranging from zero to 2 %, whereas the last grouping of loci with no additional alleles by sequence have no decrease in PI. The change in locus ranking from length to sequence analysis shows gains or no change in rank for the first grouping of loci, followed by primarily loss or no change in ranking for the middle and last groupings of loci.

For loci such as D2S1338, despite the substantial increase in number of alleles by sequence (+28 as shown in Table 1), the improvements are minimal because this locus is already highly polymorphic by length, with correspondingly low PI and high Het by length (essentially there is little room for improvement in these metrics). For loci such as TPOX, there is no change in Het, PI or rank, because it has an equally low number of alleles by both length and sequence. The locus which demonstrates the greatest overall improvement is D3S1358. Because this locus exhibits only eight alleles by length, it is able to achieve a substantial improvement in these metrics by sequence. Lastly, D2S441 is categorized as a locus that experiences a moderate gain in alleles by sequence in Table 1, but is shown in Table 2 to experience the greatest improvement in PI, Het and associated rankings from the “moderate” group. Moreover, it is the only locus in this “moderate” group to improve in ranking. The reason for this can be found by referring to Figure 1, which shows that the five

alleles gained by sequence at this locus are each variants of a different length-based allele. In addition, the five length-based alleles which contain sequence variants at D2S441 account for over 72% of the total allele frequency at this locus according to global allele frequency data [16].

On a population level, the greatest gains are seen in the African American population, where the six loci with the greatest increase in alleles (D12S391, D2S1338, D21S11, D3S1358, D8S1179, and vWA) have average decrease in PI of −4.4 % and average increase in Het of +11.5 %. The average decrease in PI / increase in Het for the European and Hispanic populations among these six loci were −2.1 % / +5.5 % and −2.6 % / +8.1 %, respectively.

Conclusions

The results of sequencing 183 population samples at 22 commonly used autosomal STR loci are indicative of which loci may routinely benefit from repeat region sequencing. Six loci are included in this category: D12S391, D2S1338, D21S11, D8S1179, vWA, and D3S1358, and these loci are largely consistent with the results of other studies which included fewer samples and/or fewer loci [8, 13, 14]. These benefits will be realized in the form of an increase in alleles, which will increase the statistical power of an inclusion and also decrease the frequency of overlapping alleles. The level of improvement is expected to vary by population. The additional 16 loci may randomly exhibit sequence variation and may provide additional information on a case-by-case or population-specific basis. In addition, sequencing flanking regions is expected to provide further information at some loci and evaluating stutter by sequence may also provide benefits, particularly in the case of mixture analysis. Analyses to characterize flanking region variation and stutter within this dataset are ongoing and will be the subject of future publications. Additional research is needed to qualify which samples or mixture types are expected to benefit from sequencing and quantify the extent of this benefit, so that laboratories can evaluate the cost-benefit of implementing this technology.

These results also show that, at this point in development, the configuration of the bioinformatic pipeline can have a significant impact on the concordance or lack thereof with CE length-based data. While changes have been made to both bioinformatic methods used in this study to address the particular issues detected herein and achieve concordance, sequencing more samples is expected to reveal further discordances. Comparing results across platforms and pipelines is important during developmental validation; more importantly, quality assurance measures to detect lower than expected coverage at a locus should be implemented to guard against “bioinformatic null alleles”.

While the high quality and single source nature of the population samples and the positive controls implemented in this study provide confidence in the repeat region sequences obtained, developmental validation of the commercially available assay will be needed prior to forensic laboratory internal validation. In addition, while this study demonstrates that quality results can be obtained when processing samples in a 96-well format, the manual library preparation used herein would require automation to truly be considered high-throughput.

In addition to validation studies, sequencing a greater number of population samples than were included in this study will be needed to generate allele frequencies. We are expanding our sequencing efforts to include more samples from the populations discussed herein (African American, European, and Hispanic) to address this need in the near future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors express gratitude to Doug Storts and Jaynish Patel at Promega for providing PowerSeq Auto reagents and technical guidance.

NIST Funding Sources and Disclaimers

This work was funded in part by the Federal Bureau of Investigation (FBI) interagency agreement DJF-13-0100-PR-0000080: “DNA as a Biometric”. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the the U.S. Departments of Commerce or Justice. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

References

1. Kline MC, Hill CR, Decker AE, Butler JM. STR sequence analysis for characterizing normal, variant, and null alleles. Forensic Sci Int Genet. 2011; 5:329–332. [PubMed: 20932816]
2. Allor C, Einum DD, Scarpetta M. Identification and characterization of variant alleles at CODIS STR loci. J Forensic Sci. 2005; 50:1128–1133. [PubMed: 16225220]
3. Dauber EM, Kratzer A, Neuhuber F, Parson W, Klintschar M, Bar W, et al. Germline mutations of STR-alleles include multi-step mutations as defined by sequencing of repeat and flanking regions. Forensic Sci Int Genet. 2012; 6:381–386. [PubMed: 21873136]
4. Griffiths RAL, Barber MD, Johnson PE, Gillbard SM, Haywood MD, Smith CD, et al. New reference allelic ladders to improve allelic designation in a multiplex STR system. Int J Legal Med. 1998; 111:267–272. [PubMed: 9728756]
5. Huel RL, Basic L, Madacki-Todorovic K, Smajlovic L, Eminovic I, Berbic I, et al. Variant alleles, triallelic patterns, and point mutations observed in nuclear short tandem repeat typing of populations in Bosnia and Serbia. Croat Med J. 2007; 48:494–502. [PubMed: 17696304]
6. Lins AM, Micka KA, Sprecher CJ, Taylor JA, Bacher JW, Rabbach D, et al. Development and population study of an eight-locus short tandem repeat (STR) multiplex system. J Forensic Sci. 1998; 43:1168–1180. [PubMed: 9846394]
7. Phillips C, Fernandez-Formoso L, Garcia-Magarinos M, Porras L, Tvedebrink T, Amigo J, et al. Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. Forensic Sci Int Genet. 2011; 5:155–169. [PubMed: 20457091]
8. Gelardi C, Rockenbauer E, Dalsgaard S, Borsting C, Morling N. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci Int Genet. 2014; 12:38–41. [PubMed: 24893347]
9. Rockenbauer E, Hansen S, Mikkelsen M, Borsting C, Morling N. Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. Forensic Sci Int Genet. 2014; 8:68–72. [PubMed: 24315591]
10. Fordyce SL, Avila-Arcos MC, Rockenbauer E, Borsting C, Frank-Hansen R, Petersen FT, et al. High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. BioTechniques. 2011; 51:127–133. [PubMed: 21806557]

11. Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, Deforce D. Forensic STR analysis using massive parallel sequencing. Forensic Sci Int Genet. 2012; 6:810–818. [PubMed: 22503403]
12. Bornman D, Hester M, Schuetter J, Kasoji M, Minard-Smith A, Barden C, et al. Short-read, high-throughput sequencing technology for STR genotyping. BioTechniques. 2012
13. Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, et al. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet. 2015; 16:38–47. [PubMed: 25528025]
14. Scheible M, Loreille O, Just R, Irwin J. Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers. Forensic Sci Int Genet. 2014; 12:107–119. [PubMed: 24908576]
15. Dalsgaard S, Rockenbauer E, Buchard A, Mogensen HS, Frank-Hansen R, Borsting C, et al. Nonuniform phenotyping of D12S391 resolved by second generation sequencing. Forensic Sci Int Genet. 2014; 8:195–199. [PubMed: 24315608]
16. Hill CR, Duewer DL, Kline MC, Coble MD, Butler JM. U.S. population data for 29 autosomal STR loci. Forensic Sci Int Genet. 2013; 7:e82–83. [PubMed: 23317915]
17. National Institute of Standards and Technology. Certificate of Analysis, Standard Reference Material 2391c. PCR-Based DNA Profiling Standard. https://www-s.nist.gov/srmors/view_cert.cfm?srm=2391C
18. Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, et al. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci Int Genet. 2013; 7:409–417. [PubMed: 23768312]
19. Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics. 2012; 28:1525–1526. [PubMed: 22500002]
20. Warshauer DH, King JL, Budowle B. STRait Razor v2.0: The improved STR Allele Identification Tool-Razor. Forensic Sci Int Genet. 2014; 14:182–186. [PubMed: 25450790]
21. Gettings KB, Aponte RA, Vallone PM, Butler JM. STR Allele Sequence Variation: Current Knowledge and Future Issues. Forensic Sci Int Genet. 2015; 18:118–130. [PubMed: 26197946]

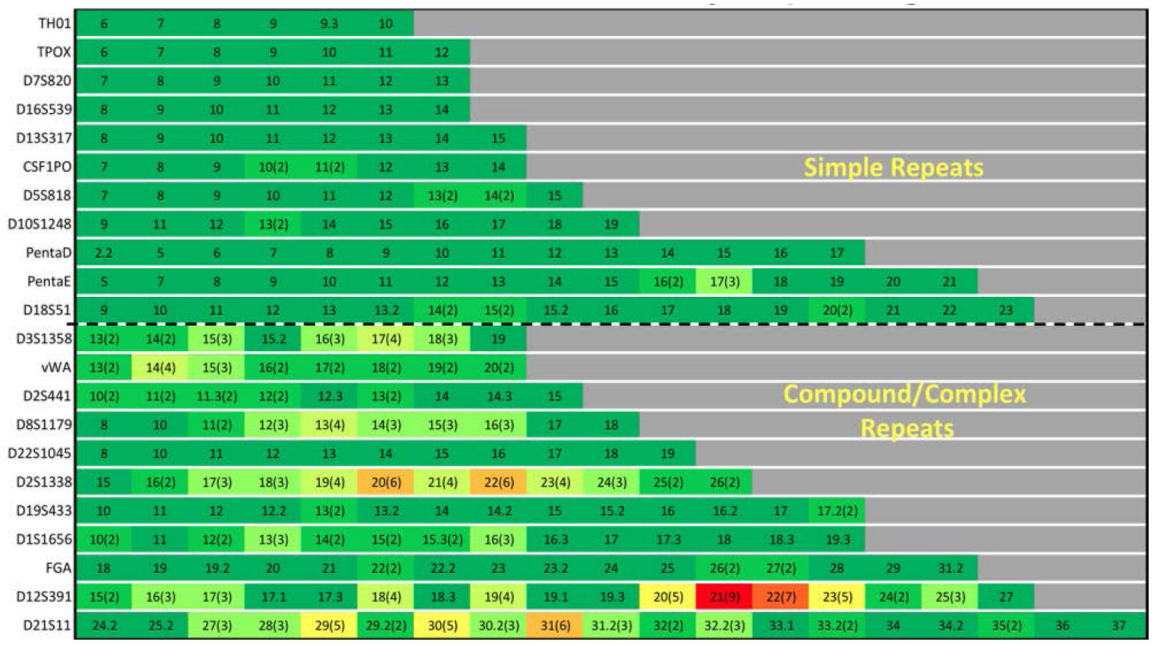


Figure 1.

Within-locus distribution of increase in alleles via sequencing. Numbers not in parentheses reflect the length-based alleles obtained in N=183; whereas numbers in parentheses (when present) represent how many sequence variants were observed in N=183. Loci are arranged by simple repeats ordered by increasing numbers of length-based alleles followed by compound/complex repeats ordered by increasing number of length-based alleles. Color coding ranges from *dark green = least sequence variation* to *red = greatest sequence variation*.

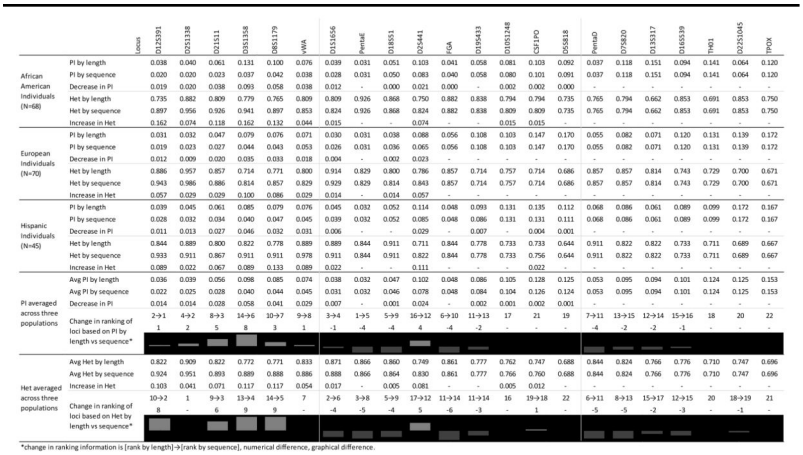
Table 1

Number of unique alleles obtained by length compared to sequence (N=183).

	Alleles obtained by length	Alleles obtained by sequence	Difference
D12S391	17	53	+36
D2S1338	12	40	+28
D21S11	19	46	+27
D8S1179	10	22	+12
D3S1358	8	19	+11
vWA	8	19	+11
D1S1656	14	23	+9
D2S441	9	14	+5
PentaE	16	19	+3
D18S51	18	21	+3
FGA	16	19	+3
D19S433	14	16	+2
CSF1PO	8	10	+2
D5S818	9	11	+2
D10S1248	9	10	+1
PentaD	14	14	-
D22S1045	11	11	-
D13S317	8	8	-
D7S820	7	7	-
D16S539	7	7	-
TPOX	7	7	-
TH01	6	6	-

Table 2

Probability of identity (PI) and heterozygosity (Het) per population by length and sequence, averaged across three populations, locus rankings by sequence, and change in rankings from length to sequence.



		Locus	D12S391	D2S1338	D21S11	D3S1358	D8S1179	vWA	D1S1656	PentaE	D18S51	D22S445	FGA	D19S433	D10S1248	CSF1PO	D5S818
African American Individuals (N=68)	PI by length		0.038	0.040	0.061	0.131	0.100	0.076	0.039	0.031	0.051	0.103	0.041	0.058	0.081	0.103	0.092
	PI by sequence		0.020	0.020	0.023	0.037	0.042	0.038	0.028	0.031	0.050	0.083	0.040	0.058	0.080	0.101	0.091
	Decrease in PI		0.019	0.020	0.038	0.093	0.058	0.038	0.012	-	0.000	0.021	0.000	-	0.002	0.002	0.000
	Het by length		0.735	0.882	0.809	0.779	0.765	0.809	0.809	0.926	0.868	0.750	0.882	0.838	0.794	0.794	0.735
	Het by sequence		0.897	0.956	0.926	0.941	0.897	0.853	0.824	0.926	0.868	0.824	0.882	0.838	0.809	0.809	0.735
	Increase in Het		0.162	0.074	0.118	0.162	0.132	0.044	0.015	-	-	0.074	-	-	0.015	0.015	-
European Individuals (N=70)	PI by length		0.031	0.032	0.047	0.079	0.076	0.071	0.030	0.031	0.038	0.088	0.056	0.108	0.103	0.147	0.170
	PI by sequence		0.019	0.023	0.027	0.044	0.043	0.053	0.026	0.031	0.036	0.065	0.056	0.108	0.103	0.147	0.170
	Decrease in PI		0.012	0.009	0.020	0.035	0.033	0.018	0.004	-	0.002	0.023	-	-	-	-	-
	Het by length		0.886	0.957	0.857	0.714	0.771	0.800	0.914	0.829	0.800	0.786	0.857	0.714	0.757	0.714	0.686
	Het by sequence		0.943	0.986	0.886	0.814	0.857	0.829	0.929	0.829	0.814	0.843	0.857	0.714	0.757	0.714	0.686
	Increase in Het		0.057	0.029	0.029	0.100	0.086	0.029	0.014	-	0.014	0.057	-	-	-	-	-
Hispanic Individuals (N=45)	PI by length		0.039	0.045	0.061	0.085	0.079	0.076	0.045	0.032	0.052	0.114	0.048	0.093	0.131	0.135	0.112
	PI by sequence		0.028	0.032	0.034	0.040	0.047	0.045	0.039	0.032	0.052	0.085	0.048	0.086	0.131	0.131	0.111
	Decrease in PI		0.011	0.013	0.027	0.046	0.032	0.031	0.006	-	-	0.029	-	0.007	-	0.004	0.001
	Het by length		0.844	0.889	0.800	0.822	0.778	0.889	0.889	0.844	0.911	0.711	0.844	0.778	0.733	0.733	0.644
	Het by sequence		0.933	0.911	0.867	0.911	0.911	0.978	0.911	0.844	0.911	0.822	0.844	0.778	0.733	0.756	0.644
	Increase in Het		0.089	0.022	0.067	0.089	0.133	0.089	0.022	-	-	0.111	-	-	-	0.022	-
PI averaged across three populations	Avg PI by length		0.036	0.039	0.056	0.098	0.085	0.074	0.038	0.032	0.047	0.102	0.048	0.086	0.105	0.128	0.125
	Avg PI by sequence		0.022	0.025	0.028	0.040	0.044	0.045	0.031	0.032	0.046	0.078	0.048	0.084	0.104	0.126	0.124
	Decrease in PI		0.014	0.014	0.028	0.058	0.041	0.029	0.007	-	0.001	0.024	-	0.002	0.001	0.002	0.001
	Change in ranking of loci based on PI by length vs sequence*		2→1	4→2	8→3	14→6	10→7	9→8	3→4	1→5	5→9	16→12	6→10	11→13	17	21	19
			1	2	5	8	3	1	-1	-4	-4	4	-4	-2	-	-	-
Het averaged across three populations	Avg Het by length		0.822	0.909	0.822	0.772	0.771	0.833	0.871	0.866	0.860	0.749	0.861	0.777	0.762	0.747	0.688
	Avg Het by sequence		0.924	0.951	0.893	0.889	0.888	0.886	0.888	0.866	0.864	0.830	0.861	0.777	0.766	0.760	0.688
	Increase in Het		0.103	0.041	0.071	0.117	0.117	0.054	0.017	-	0.005	0.081	-	-	0.005	0.012	-
	Change in ranking of loci based on Het by length vs sequence*		10→2	1	9→3	13→4	14→5	7	2→6	3→8	5→9	17→12	11→14	11→14	16	19→18	22
			8	-	6	9	9	-	-4	-5	-4	5	-6	-3	-	1	-

		Locus	PentaD	D7S820	D13S317	D16S539	TH01	D22S1045	TPOX
African American Individuals (N=68)	PI by length		0.037	0.118	0.151	0.094	0.141	0.064	0.120
	PI by sequence		0.037	0.118	0.151	0.094	0.141	0.064	0.120
	Decrease in PI		-	-	-	-	-	-	-
	Het by length		0.765	0.794	0.662	0.853	0.691	0.853	0.750
	Het by sequence		0.765	0.794	0.662	0.853	0.691	0.853	0.750
	Increase in Het		-	-	-	-	-	-	-
European Individuals (N=70)	PI by length		0.055	0.082	0.071	0.120	0.131	0.139	0.172
	PI by sequence		0.055	0.082	0.071	0.120	0.131	0.139	0.172
	Decrease in PI		-	-	-	-	-	-	-
	Het by length		0.857	0.857	0.814	0.743	0.729	0.700	0.671
	Het by sequence		0.857	0.857	0.814	0.743	0.729	0.700	0.671
	Increase in Het		-	-	-	-	-	-	-
Hispanic Individuals (N=45)	PI by length		0.068	0.086	0.061	0.089	0.099	0.172	0.167
	PI by sequence		0.068	0.086	0.061	0.089	0.099	0.172	0.167
	Decrease in PI		-	-	-	-	-	-	-
	Het by length		0.911	0.822	0.822	0.733	0.711	0.689	0.667
	Het by sequence		0.911	0.822	0.822	0.733	0.711	0.689	0.667
	Increase in Het		-	-	-	-	-	-	-
PI averaged across three populations	Avg PI by length		0.053	0.095	0.094	0.101	0.124	0.125	0.153
	Avg PI by sequence		0.053	0.095	0.094	0.101	0.124	0.125	0.153
	Decrease in PI		-	-	-	-	-	-	-
	Change in ranking of loci based on PI by length vs sequence*		7→11	13→15	12→14	15→16	18	20	22
			-4	-2	-2	-1	-	-	-
Het averaged across three populations	Avg Het by length		0.844	0.824	0.766	0.776	0.710	0.747	0.696
	Avg Het by sequence		0.844	0.824	0.766	0.776	0.710	0.747	0.696
	Increase in Het		-	-	-	-	-	-	-
	Change in ranking of loci based on Het by length vs sequence*		6→11	8→13	15→17	12→15	20	18→19	21
			-5	-5	-2	-3	-	-1	-

*change in ranking information is [rank by length]→[rank by sequence], numerical difference, graphical difference.

* change in ranking information is [rank by length]→[rank by sequence], numerical difference, graphical difference.