

University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

U.S. Department of Justice Publications and
Materials

U.S. Department of Justice


2018

Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results

Rebecca S. Just
FBI, rsjust@fbi.gov

Jodi A. Irwin
FBI, jairwin@fbi.gov

Follow this and additional works at: <http://digitalcommons.unl.edu/usjusticematls>

 Part of the [Civil Rights and Discrimination Commons](#), [Constitutional Law Commons](#), [Law and Society Commons](#), [Law Enforcement and Corrections Commons](#), [Other Law Commons](#), [President/Executive Department Commons](#), and the [Public Law and Legal Theory Commons](#)

Just, Rebecca S. and Irwin, Jodi A., "Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results" (2018). *U.S. Department of Justice Publications and Materials*. 46.
<http://digitalcommons.unl.edu/usjusticematls/46>

This Article is brought to you for free and open access by the U.S. Department of Justice at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in U.S. Department of Justice Publications and Materials by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.



Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results

Rebecca S. Just*, Jodi A. Irwin

DNA Support Unit, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, VA, 22135, USA



ARTICLE INFO

Keywords:

Next generation sequencing (NGS)
Short tandem repeat (STR)
Mixture
Probabilistic genotyping
Longest uninterrupted stretch (LUS)
Sequence variation

ABSTRACT

Some of the expected advantages of next generation sequencing (NGS) for short tandem repeat (STR) typing include enhanced mixture detection and genotype resolution via sequence variation among non-homologous alleles of the same length. However, at the same time that NGS methods for forensic DNA typing have advanced in recent years, many caseworking laboratories have implemented or are transitioning to probabilistic genotyping to assist the interpretation of complex autosomal STR typing results. Current probabilistic software programs are designed for length-based data, and were not intended to accommodate sequence strings as the product input. Yet to leverage the benefits of NGS for enhanced genotyping and mixture deconvolution, the sequence variation among same-length products must be utilized in some form.

Here, we propose use of the longest uninterrupted stretch (LUS) in allele designations as a simple method to represent sequence variation within the STR repeat regions and facilitate – in the near term – probabilistic interpretation of NGS-based typing results. An examination of published population data indicated that a reference LUS region is straightforward to define for most autosomal STR loci, and that using repeat unit plus LUS length as the allele designator can represent greater than 80% of the alleles detected by sequencing. A proof of concept study performed using a freely available probabilistic software demonstrated that the LUS length can be used in allele designations when a program does not require alleles to be integers, and that utilizing sequence information improves interpretation of both single-source and mixed contributor STR typing results as compared to using repeat unit information alone. The LUS concept for allele designation maintains the repeat-based allele nomenclature that will permit backward compatibility to extant STR databases, and the LUS lengths themselves will be concordant regardless of the NGS assay or analysis tools employed. Further, these biologically based, easy-to-derive designations uphold clear relationships between parent alleles and their stutter products, enabling analysis in fully continuous probabilistic programs that model stutter while avoiding the algorithmic complexities that come with string based searches. Though using repeat unit plus LUS length as the allele designator does not capture variation that occurs outside of the core repeat regions, this straightforward approach would permit the large majority of known STR sequence variation to be used for mixture deconvolution and, in turn, result in more informative mixture statistics in the near term. Ultimately, the method could bridge the gap from current length-based probabilistic systems to facilitate broader adoption of NGS by forensic DNA testing laboratories.

1. Introduction

Though the potential benefits of next generation sequencing (NGS) technologies for forensic purposes have been understood for some time [1–5], it is only recently that the forensic genetics community has gained a better handle on their practical utility for DNA casework. Dozens to hundreds of markers with small, overlapping amplicon sizes can be multiplexed and simultaneously sequenced, offering a considerable improvement over current methods in terms of data recovery

and discriminatory power. These advantages have been shown to be particularly useful when sample material is limited and/or only partial profiles are likely to be recovered [5–9], as well as in complex kinship scenarios where the sheer volume of data from multiple marker systems can substantially improve resolution [7].

In addition to the benefits that come from large multiplexes of small amplicons are those that come from the sequence data themselves – particularly for short tandem repeat (STR) loci. For more than two decades, intra-allelic variation has been known to exist, having been

* Corresponding author.

E-mail addresses: rsjust@fbi.gov (R.S. Just), jairwin@fbi.gov (J.A. Irwin).

identified and at least preliminarily characterized via older sequencing technologies, mass spectrometry and other methods [10–18]. However, there has simply not been a straightforward means of accessing STR sequence data in DNA testing laboratories built around capillary electrophoresis (CE) and STR repeat unit based chemistries. Additionally, since CE-based STR typing results are generally sufficiently discriminating, there has historically not been widespread interest in STR sequence variation for forensic casework purposes. With the likely future adoption of NGS in forensic laboratories, however, sequence data will be easily accessible and the practical benefits of those data may be exploited. For example, in cases for which only partial data are recovered, the rarity of sequence-based alleles versus repeat unit based alleles can improve the discrimination power for the limited number of loci. Similarly, in kinship scenarios, more informative statistics can be developed when the inheritance of identical repeat unit based alleles can be clarified with sequence information. Yet, one of the most anticipated benefits of NGS-derived STR data is the potential of sequence level information to improve the interpretation of mixtures.

Many studies have now demonstrated that the number of distinct alleles at any given marker increases when sequence variation is considered ([5,14,19–28], for example). At some loci, these gains are modest, with an average increase in alleles of only 6% and a maximum of 23% for the commonly employed simple repeats [23]. For other markers, the compound repeats in particular, the gains can be dramatic. On average, the increase in alleles by sequence versus length in these cases is 100%, with some markers showing gains greater than 200%. The fact that length-based alleles can be further resolved by sequence intuitively suggests advantages for the interpretation of mixed DNA profiles. In theory, sequence-based alleles would permit 1) resolution of repeat unit homozygotes into sequence-based heterozygotes (isometric heterozygotes), 2) distinction of contributors sharing repeat unit based alleles, and 3) differentiation of stutter products and true alleles. Indeed, in the few studies to date that have evaluated mixtures with NGS, these very advantages have borne out in practice [9,29,30]. The same studies also indicate that commercially available NGS-based STR assays may offer greater sensitivity to, and therefore detection of, low-level contributors. In at least two cases minor components could be detected in 99:1 mixtures [9,26], and all or nearly all loci exhibited mixture with 19:1 contributor ratios [9,26,30]. Though it will take time to develop a comprehensive understanding of the routine practical value that these features will impart to mixture interpretation, there seems to be little doubt that the sensitivity of currently available assays and, to an even greater extent, the granularity of the NGS (sequence) data will help in the detection of low-level contributors, in the determination of number of contributors and in the deconvolution process itself [9,29,30].

Despite the fact that the value of STR sequence information for mixture deconvolution is becoming better understood, the use of NGS technology, and STR sequence data in particular, is still some ways off from being widely adopted in forensic laboratories. Further characterization and understanding of STR sequence variation is required, sufficient reference population data for sequence-based alleles are needed, and globally accepted standards for STR sequence nomenclature still must be established [31,32]. As a result, nearly all operational forensic laboratories currently continue to employ CE-based technologies. CE technologies and repeat unit based chemistries sufficiently address the primary needs of the DNA testing community; and thus near-term laboratory improvements have not centered on NGS. Instead, a topic garnering substantial attention lately has been the interpretation of CE-based mixture data via probabilistic genotyping (for example, [33–36]).

Though probabilistic genotyping can be applied to any STR typing results, the approach was developed specifically to aid in the interpretation of complex data that are typical of evidentiary specimens, such as partial profiles, results developed from low template samples, and mixtures of DNA from multiple contributors [33,37–47]. Probabilistic methods address uncertainty in the DNA data through use of allelic drop-out and/or drop-in models, which are employed to

determine the probability of the observed STR typing results according to all possible contributor genotype combinations. As probabilistic methods do not rely on the application of a stochastic threshold [48–51], they leverage more of the available genotyping data than a binary approach and have proven to produce more informative statistics with respect to contributors and non-contributors to the DNA evidence [52]. Some systems additionally employ probability models for interpretation of STR typing features such as stutter and degradation, which assists in reducing the subjectivity and/or inconsistency that can result from manual designation of products as either allelic or stutter [53]. These latter programs are often termed “fully continuous” in reference to their combined use of quantitative profile information (e.g. peak heights) and models of peak behavior [54]. Fully continuous programs result in probabilities that reflect that some of the possible genotype sets are more likely to have produced the observed data than others.

As a result of the advantages of using probability models for DNA evidence interpretation, as well as the availability of a number of developmentally validated software programs (both freely and commercially; for example, Refs [40,43–45,55]), more and more laboratories are planning to implement probabilistic genotyping. Surveys of laboratories in the United States, for example, indicate that while only 20% currently employ probabilistic mixture interpretation, an additional 60% plan to validate and implement in the next year (D. Hares, personal communication). With this widespread adoption of probabilistic interpretation, laboratories interested in pursuing NGS for sequence data benefits will find themselves restricted by the availability of probabilistic genotyping software that can exploit sequence information. To facilitate the adoption of STR sequencing technology in forensic casework laboratories that currently use, or will soon be implementing, probabilistic genotyping for CE data, we have been considering steps towards probabilistic genotyping of NGS-derived STR typing results that can leverage at least some sequence information for mixture deconvolution in the near term.

The longest uninterrupted stretch (LUS) refers to the greatest number of contiguous identical sequence repeats within an STR [12,56]. Several studies have demonstrated that for most current forensic autosomal STR (aSTR) loci the LUS is the region of the amplicon that typically results in stutter products (via the loss or gain of identical repeat units), and that the length of the LUS is often correlated with stutter ratios [12,26,56–59]. For these reasons, the LUS identified by sequencing of some alleles has been used to model expected stutter for probabilistic genotyping of CE-based data [56,58,59]. The LUS has also been employed as part of an NGS data analysis tool to help filter stutter products and other noise from sequence-based STR data [60]. As has been noted [26,58,61], sequence variants within the repeat region may result in different lengths of the LUS for alleles of the same size/represented by the same repeat unit designation. Thus, it is clear that the LUS length itself could serve to distinguish many same-length alleles that differ by sequence.

Here, we propose use of both the repeat unit and the LUS length in allele designations (in the format RepeatUnit.PartialUnit_LUSlength) as a means to represent sequence alleles for probabilistic genotyping of NGS-based STR results. We analyzed published population data to assess the degree to which a single reference LUS region could be determined for each of 27 aSTRs, and the extent to which use of both the repeat unit and the LUS length in allele designations can represent known sequence variation. Additionally, we conducted a proof of concept study in an open-source probabilistic genotyping program (LRmix Studio [55,62]) to demonstrate the feasibility of the LUS method for allele designation, and to preliminarily investigate the practical utility of sequence information for the interpretation of mixed and low-level DNA typing results. Finally, we examined the stutter products that occur at the compound D12S391 locus to explore how the LUS concept for allele designation could be applied in probabilistic programs that model stutter.

2. Materials and methods

2.1. Analyses of population data

Published sequence data for 777 individuals from four U.S. population groups [25] and 400 individuals from two British population groups [28], both developed using the ForenSeq DNA Signature Prep Kit [63] and MiSeq FGx instrument (Illumina, Inc.) for sample typing, were used to investigate potential LUS length reference regions. For each of the 27 aSTR loci included in the ForenSeq assay, the sequence alleles were examined to determine the frequency with which different motifs in the core repeat region produced the LUS. From these analyses, the repeat region most commonly producing the LUS (from here forward termed the “LUS reference region”) for each locus was identified.

The same two data sets were used to determine the percentage of sequence alleles per locus that would be uniquely represented using the RepeatUnit.PartialUnit_LUSlength format. To account for the fact that the data reported in Ref [25] contains additional flanking region sequence that is considered by STRait Razor [64] but not reported by the ForenSeq Universal Analysis Software (UAS; [65]), the Ref [25] data were also assessed considering only the portions of each STR that are included in the ForenSeq UAS output (Table S1). That is, sequence-based alleles were considered in two ways for this data set: with the UAS reported regions only, and then again with the entire sequence string that included flanking regions. For all three data sets (alleles from Ref [25] Table S1, a modified version of the table containing only the ForenSeq UAS regions for each allele; and the alleles from Ref [28] Table 1), and using the LUS reference regions previously identified, the LUS length was determined for each sequence and appended to the repeat unit allele. Counts of the distinct LUS-based alleles were subsequently compared to both the full sequence allele and repeat unit allele counts for each locus.

2.2. Proof of concept study

The proof of concept study used aSTR data developed via the MiSeq

FGx Forensic Genomic System, with DNA Primer Mix B and UAS version 1.2.1 [63,65]. All manufacturer protocols were followed, except where described below. A schematic of the samples and tests for the proof of concept study is displayed in Fig. S1.

For the single-source portion of the study, two samples were used. Each sample was amplified in duplicate using the following DNA inputs for PCR: 125 pg, 62.5 pg and 31.25 pg. These 12 ForenSeq libraries were subsequently sequenced in the same pool containing 40 total libraries. For the mixture portion of the study, a total of 14 two-person mixtures were amplified and sequenced. Mixture Set 1 consisted of seven mixtures of Male 1 and Male 2, while Mixture Set 2 consisted of seven Female:Male 1 mixtures. For both mixture sets, the contributors were combined in the ratios 10:1, 5:1, 2:1, 1:1, 1:2, 1:5 and 1:10. The two mixture sets were processed in separate MiSeq FGx runs in which 32 total libraries were pooled.

For the UAS analyses of both the single-source and mixture data, a custom set of aSTR stutter thresholds, determined from ForenSeq data [66,67], were applied. Sample Detail Reports were exported from the UAS, and the data were further reviewed to apply an analytical threshold of 30 reads (regardless of total locus coverage) and to remove any instances of stutter that exceeded the filters. The four single-source 31.25 pg typing results were also reinterpreted using an analytical threshold of 100 reads to simulate lower DNA input typing results. Poor typing of the Penta D locus was observed across all mixed samples, and thus the results for this marker were removed from the mixture data.

The resulting data (16 total single-source results and 14 total two-person mixture results) as well as the five reference genotypes were then transformed into both 1) repeat unit based profiles, and 2) LUS-based profiles. The LUS profiles were generated by translating each sequence allele into a bracketed format (using in-house lookup tables) from which the length of the LUS (in the LUS reference region; Table 1) was then derived. The LUS length was appended to each repeat unit in the format RepeatUnit.PartialUnit_LUS length. For example, the D12S391 allele 15 sequence (AGAT)8 (AGAC)6 AGAT was designated as “15_8” in the LUS profile. The resulting repeat unit and LUS-based “evidence” profiles and five reference genotypes were subsequently

Table 1
LUS length reference regions for the STR loci typed by the ForenSeq assay.

Locus	Example Alleles Showing LUS Length Reference Regions ^a	CE/Repeat Unit Designation	LUS Concept Allele Designation
CSF1PO	<u>(ATCT)7 ACCT (ATCT)3</u>	11	11_7
D10S1248	<u>(GGAA)6 GTAA (GGAA)7</u>	14	14_7
D12S391	<u>(AGAT)8</u> (AGAC)6 AGAT	15	15_8
D13S317	<u>(TATC)9</u> (AATC)2 (ATCT)3 TTCT GTCT GTC	9	9_9
D16S539	<u>(GATA)2 CATA (GATA)9</u>	12	12_9
D17S1301	<u>(AGAT)10 AGGT AGAT</u>	12	12_10
D18S51	<u>(GAA)4 AGGA (AGAA)10</u> AAAG AGAG AG	15	15_10
D19S433	CT CTCT TTCT TCCT CTCT <u>(CCTT)9 CCTA CCTT CTTT CCTT</u>	11	11_9
D1S1656	CA (CACA)2 <u>CCTA (TCTA)9 TGTA (TCTA)4</u>	15	15_9
D20S482	<u>(AGAT)11</u>	11	11_11
D21S11	(TCTA)4 (TCTG)6 (TCTA)3 TA (TCTA)3 TCA (TCTA)2 TCCA TA <u>(TCTA)10</u>	28	28_10
D22S1045	<u>(ATT)13</u> ACT (ATT)2	16	16_13
D2S1338	<u>(GGAA)2 GGAC (GGAA)9</u> (GGCA)6	18	18_9
D2S441	<u>(TCTA)8 TCTG TCTA</u>	10	10_8
D3S1358	TCTA (TCTG)2 <u>(TCTA)12</u>	15	15_12
D4S2408	<u>ATCT GTCT (ATCT)7</u>	9	9_7
D5S818	<u>CTCT (ATCT)3 ATGT (ATCT)10</u>	14	14_10
D6S1043	<u>(ATCT)5 ATGT (ATCT)9</u>	15	15_9
D7S820	AAAC TATC AATC TGTC <u>(TATC)7 TACC (TATC)3</u>	11	11_7
D8S1179	<u>TCTA TCTG (TCTA)10</u>	12	12_10
D9S1122	<u>TAGA TCGA (TAGA)9</u>	11	11_9
FGA	(GAAA)2 GGAG <u>AAAG AAAC (AAAG)13</u> AGAA AAAA (GAAA)3	23	23_13
Penta D	AAAAG <u>AAAGA GAAGA (AAAGA)9</u>	11	11_9
Penta E	<u>(TCTTT)13 TCCTT (TCTTT)3</u>	17	17_13
TH01	<u>(AATG)9</u>	9	9_9
TPOX	<u>(AATG)9</u>	9	9_9
vWA	TAGA TGGA <u>(TAGA)10</u> (CAGA)4 TAGA	15	15_10

LUS length reference regions are bolded and underlined.

^a Alleles are displayed using the ForenSeq UAS data range, but in accordance with ISFG recommendations as to strand [31].

formatted for interpretation in the LRmix Studio program [55,62].

U.S. Caucasian allele frequency tables for repeat unit alleles and LUS-based alleles were developed using the allele counts in the ForenSeq UAS-modified version of Table S1 from Ref [25], as described above. The allele frequency files, as well as repeat unit and LUS-formatted files for two mixtures (Mixture Set 1, contributor ratios 10:1 and 5:1) and their relevant contributors (Mixture Male 1 and Mixture Male 2) are provided as part of the Supplementary material (Folder S1) as example data.

LRmix Studio version 2.1.3 Community Edition was used for all probabilistic interpretations. For the single-source profiles, the contributor number was set to 1, Hp proposed the true donor as the person of interest, and Hd proposed an unknown contributor. For the two-person mixtures, the contributor number was set to 2, Hp proposed the true donor as the person of interest plus one unknown contributor, and Hd proposed two unknown contributors.

In addition to interpretations performed considering all 27 aSTR loci typed by the ForenSeq assay, the single-source profiles were also interpreted using only the 20 core CODIS loci by de-selecting the seven non-CODIS markers in the Profile Summary Table. As a result, a total of 64 total Hp true single-source tests (32 with repeat unit alleles and 32 with LUS-based alleles) and 56 total Hp true mixture tests (14 each with repeat unit and LUS-based alleles, for each known contributor) were performed in the software (see Fig. S1).

All interpretations used the default theta value (0.01) and drop-in probability (0.05). For each interpretation the dropout probability for the hypothesized person of interest and all unknowns was set to reflect the observed dropout for the person of interest (see examples in Folder S1), rather than determined by using the Sensitivity Analysis and Dropout Estimation functions available in the software. This approach was taken to minimize the variability among interpretations, since the primary goal was to assess the difference in likelihood ratio (LR) values between the repeat unit and LUS-based interpretations. For the two-person mixtures, only the obligate (unshared) alleles for a donor were used to determine the observed dropout probability. For all interpretations that used a dropout probability greater than 0, non-contributor tests (Hd true) were run in the software for 500 iterations. From each of these runs, the 99% Log₁₀ LR value was recorded.

LRmix Studio interpretations used the default minimum allele frequency of 0.001. Among the five distinct samples used for the proof of concept study (Table S2), one repeat unit allele (allele 22 at locus D18S51) and two LUS-based alleles (allele 22_22 at locus D18S51, and allele 22_13 at locus D2S1338) were not observed in the U.S. Caucasian population data used for the allele frequencies. These alleles were thus assigned the minimum allele frequency in interpretation.

2.3. Examination of D12S391 stutter products

The D12S391 locus is a compound STR with two repeat regions, AGAT and AGAC, from which stutter may occur. To consider how the LUS concept for allele designation could be applied in practice when multiple stutter products by sequence are encountered, we examined the relative frequency and magnitude of stutter produced from the AGAT and AGAC portions of the repeat using D12S391 alleles typed from single source ForenSeq sample libraries. The ForenSeq data were developed as described in Ref [27], using DNA Primer Mix B for typing and the ForenSeq UAS for analysis. The sample libraries came from sequencing runs 1, 3, 4 and 5 performed for that study.

The aSTR data were exported from the UAS to Excel via the Sample Detail Report function. All D12S391 sequences reported by the UAS, to include any clusters of identical sequences represented by greater than 10 reads (the fixed detection threshold in the UAS), were examined and manually classified as allelic, stutter or noise. To accurately assess stutter ratios, alleles that could not be distinguished from potential -1 repeat unit stutter were removed from the data set. Owing to the overall low incidence of -2 repeat unit stutter, as well as low stutter ratios

when it was detected, alleles that could not be distinguished by sequence from potential -2 stutter were not removed from the dataset. In total, 196 D12S391 alleles from 103 sample libraries were retained for further analysis.

For each D12S391 allele, the LUS length for each portion of the compound repeat region (AGAT and AGAC) was determined. Sequences differing from the allele by one repeat unit via the loss of an AGAT repeat or an AGAC repeat were classified as -1 AGAT stutter and -1 AGAC stutter, respectively. Sequences that differed from the allele sequence by two repeat units via the loss of two AGAT repeats were classified as -2 AGAT stutter. No -2 AGAC stutter, and no +1 stutter from either portion of the repeat region were detected in the sequence data. Any sequence that could not be identified as either allelic or a logical stutter product (-1 of the core repeat units, or -2 repeat units from the same portion of the repeat) was classified as noise. Where stutter products were detected, the stutter ratio was calculated by dividing the stutter sequence read count by the allele sequence read count.

3. Results

3.1. Investigation of LUS length reference regions

The 746 aSTR sequence alleles detected in 777 individuals in Ref [25] and 503 aSTR sequence alleles identified among 400 individuals in Ref [28] were reviewed to determine whether the LUS was consistently identified within the same region of each STR. For 25 of the 27 loci in the Ref [25] dataset, the same region of the STR produced the LUS for every sequence variant detected. For the remaining two loci, D12S391 and vWA, the same region of the STR produced the LUS in all but four sequences (Table S1). Of the 79 distinct alleles reported for D12S391, the AGAT portion of the repeat produced the LUS in 77 of the sequences, whereas the remaining two sequences had equal AGAT and AGAC repeat lengths. For 33 of the 35 distinct vWA sequence alleles, the TAGA portion of the repeat produced the LUS, whereas the CAGA repeat produced the LUS with the remaining two vWA alleles. Among the 503 alleles reported in Ref [28], the same two exceptions were observed with vWA, and one additional exception was identified in one D21S11 sequence (in which the first and last TCTA repeats were of equal length). As the same region of each aSTR produced the LUS in nearly all instances (99.5% of all alleles reported in Ref [25], and 99.4% of all alleles reported in Ref [28]), the repeat regions designated in Table 1 were used to determine the LUS length for all further work discussed herein.

3.2. Percentage of sequence alleles captured by use of both repeat unit and LUS

To assess the extent to which use of the LUS can capture sequence variation, we assigned a repeat unit plus LUS allele designation (in the format RepeatUnit.PartialUnit_LUSlength) to each of the aSTR sequence alleles in Table S1 of Ref [25], and to the sequence alleles in the modified version of the table that contained only the regions of each locus that are reported by the ForenSeq UAS (Table S1, this paper). Overall, when considering only the portions of each aSTR reported by the UAS, 88.3% of the allele variation by sequence, and 75.7% of the increase in the number of distinct alleles by sequence (as compared to by repeat unit alone) was captured by using both the repeat unit and the LUS length in the allele designation (Fig. 1A). Among the three loci that are most variable by sequence (D21S11, D12S391 and D2S1338) the increase in distinct alleles captured via the LUS length was 61.6%. For the remaining 24 aSTR loci, > 90% of the increase was captured. If considering the full sequence information available from the STRait Razor analyses, use of the RepeatUnit.PartialUnit_LUSlength format represented 82.2% of the allele variation by sequence, which translates to 65.7% of the increase in the number of distinct alleles by sequence as

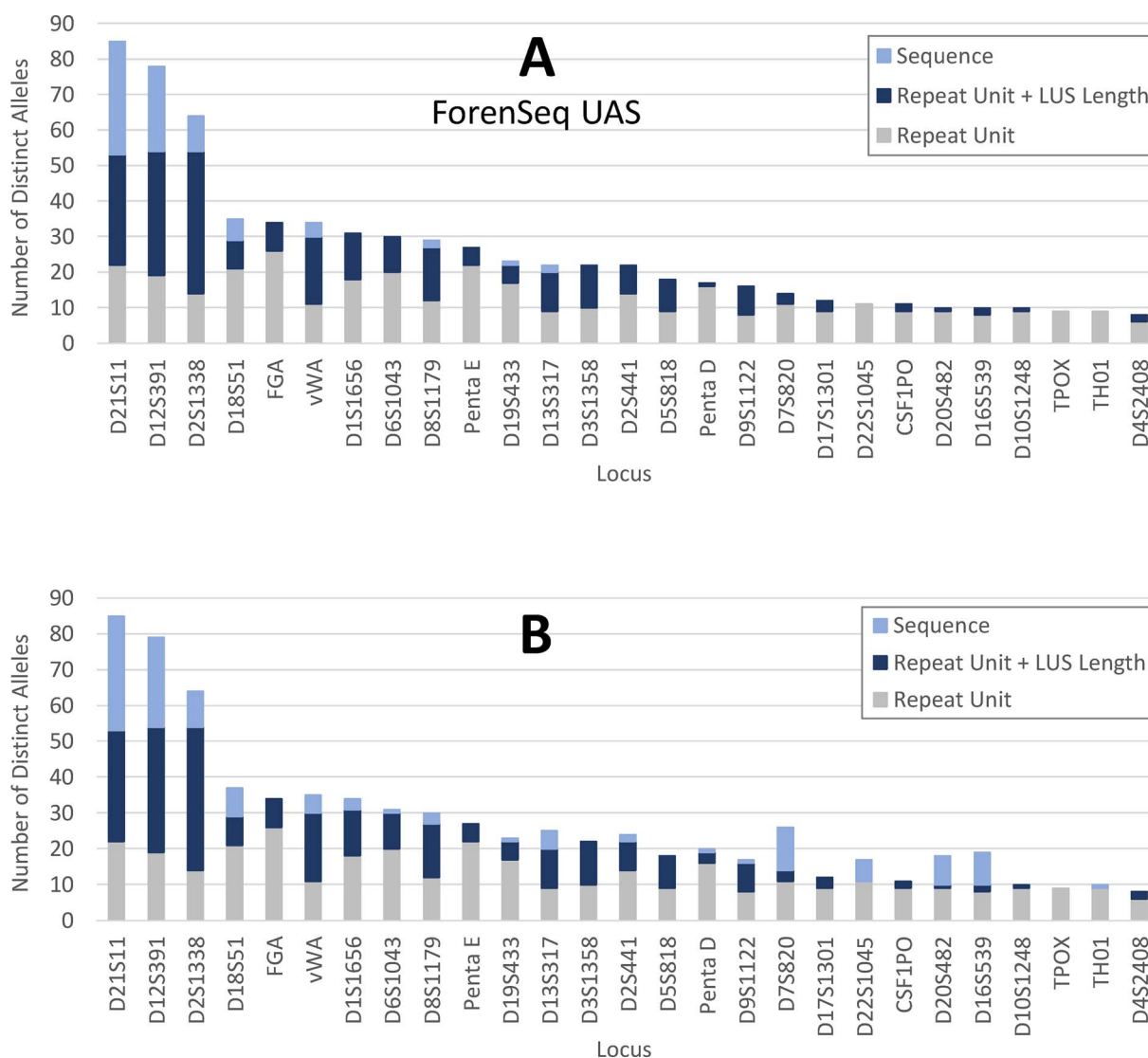


Fig. 1. Extent of sequence variation captured by use of the LUS.

Using published ForenSeq data for 777 individuals from four population groups [25], the number of distinct alleles per locus that would be captured by use of the LUS length in allele designations (dark blue bars) was compared to the alleles detected by sequencing (light blue bars) and the number of alleles represented by repeat unit alone (gray bars) for A) the regions of each locus reported by the ForenSeq UAS, and B) the complete sequence data reported by Ref [25].

compared to repeat unit alone (Fig. 1B). The lower percentages when considering the data accessed by STRait Razor are primarily due to additional flanking region sequence variation for four loci (D7S820, D16S539, D20S482 and D22S1045) that is not reported by the UAS and does not impact the LUS length. An examination of the ForenSeq alleles from two British populations reported by Devesse et al. [28] revealed a highly similar picture: overall, 86.9% of the allele variation by sequence was captured by use of the LUS allele designation concept (data not shown).

3.3. Tests of the LUS concept

For the five distinct samples used for the proof of concept study (Table S2), all sequence alleles were uniquely represented by use of the repeat unit plus LUS length designation. For the two samples used for the single-source testing, one had two loci with isoalleles (alleles identical by length but different by sequence) whereas the other had no isoalleles. For the first mixture set, neither Male 1 nor Male 2 had isoalleles individually, but three alleles shared by repeat unit were not shared by LUS concept designation. For Mixture Set 2, the Female sample had two loci with isoalleles, and two loci with alleles shared by

repeat unit but not by LUS concept designation with Male 1.

LR values obtained by interpretation of the single-source and mixed contributor profiles in LRmix Studio demonstrated two general trends. First, the LR values obtained for true contributors to the LUS-based profiles were nearly always higher than those for the repeat unit profiles (Fig. 2). In two of the 42 comparisons, however, the LUS-based interpretations produced lower LRs. Both instances occurred with interpretation of Mixture Set 1 when the minor contributor was proposed as the donor to the 1:10 and 10:1 mixtures (Fig. 2B, “ > 75%” obligate allele dropout category). In both cases, alleles at three loci (D12S391, D2S441 and D9S1122) were shared by the major and minor contributors by repeat unit, but were not minor contributor alleles by sequence/LUS allele designation. That is, the typing results exhibited dropout of alleles specific to the minor contributor, but this dropout was only apparent when the allele sequences were considered. A review of the per-locus LR values in the LRmix Studio reports revealed that at these three loci the minor contributor LR value was higher for the repeat unit interpretation than the LUS interpretation, whereas the LR value was the same or higher for the LUS interpretation at all other loci.

The second general trend observed in the data was that as allelic dropout decreased for a contributor, the magnitude of the difference

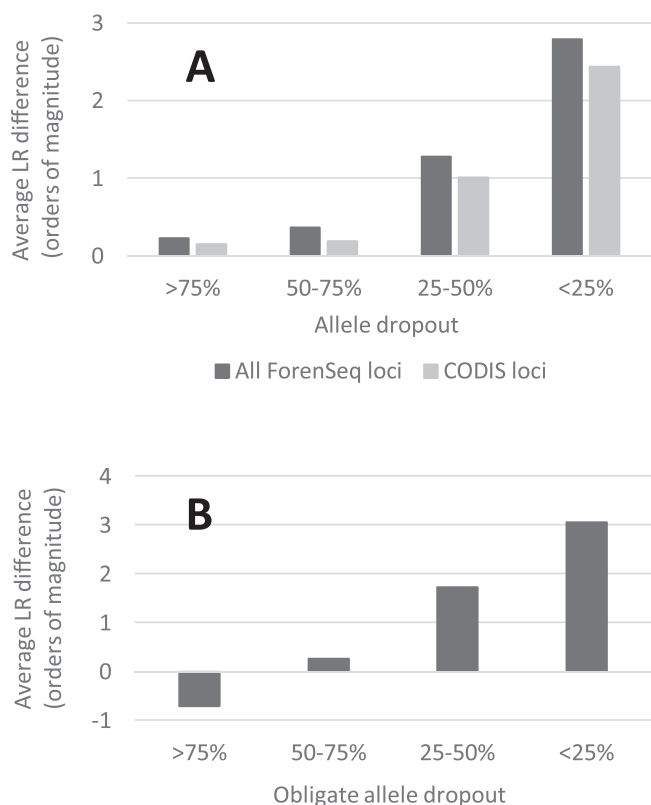


Fig. 2. LR differences for true contributors using LUS-based versus repeat unit only allele designations.

The LRs resulting from LUS-based versus repeat unit based allele designations interpreted in LRmix Studio were compared for true contributors to A) single-source typing results, and B) two-person mixtures at various levels of contributor allelic dropout. Allelic dropout was determined using the percentage of a contributor's unshared alleles not detected in the interpreted profile. The data are displayed as the orders of magnitude by which the LUS-based interpretations increased the LRs compared to the repeat unit interpretations of the same typing results and hypotheses.

between the repeat unit and LUS-based LRs increased (Fig. 2). Considering both the single-source and two-person mixture results (and regardless of the number of loci interpreted), use of the LUS in allele designations on average doubled the LR values when contributor dropout ranged from 50 to 75%. However, when greater than 50% of a contributor's alleles were recovered, the LUS interpretation LRs were at least an order of magnitude greater than the repeat unit LRs in 39 of 44 instances, and on average the LR increased by 2.48 orders of magnitude (s.d. 1.04).

The same two general trends were observed when considering non-contributors to the DNA evidence. Overall, 82% of the LUS interpretations produced 99% Log_{10}LR values lower (indicating greater exclusionary support) than those obtained with the repeat unit based data. The 99% Log_{10}LR values were 0.86 orders of magnitude lower on average when contributor dropout ranged from 50 to 75%, compared to an average of 3.44 orders of magnitude lower when contributor dropout was less than 50% (data not shown).

At all levels of dropout, interpretations of the single-source profiles that used only the 20 core CODIS loci resulted in smaller gains in the LR going from repeat unit to LUS-based alleles as compared to interpretations that used the full 27 ForenSeq loci (Fig. 2A). This was the expected result. However, we note that the magnitude of the difference observed when considering different numbers of loci will depend on multiple factors, including the specific assays, loci, and contributor genotypes involved.

3.4. D12S391 stutter products by sequence

To consider how the LUS concept could be applied in fully continuous programs that model stutter, we assessed the incidence of AGAT and AGAC repeat stutter products produced when sequencing the compound D12S391 locus. As Table S3 demonstrates, 92.8% of the alleles exhibited -1 repeat stutter from the AGAT portion of the STR, and the -1 AGAT stutter was detected in nearly every instance (89 of 90) when the associated allele was represented by greater than 200 reads. Minus 1 repeat stutter from the AGAC portion of the repeat was observed less frequently, and was never detected when allele read coverage was less than 200. Overall, -1 AGAC stutter occurred with 17.3% of the D12S391 alleles. When only those alleles with coverage greater than 200 reads were considered, -1 AGAC stutter was detected 37.8% of the time.

The AGAT and AGAC -1 stutter ratios for all alleles with read coverage greater than 200 were plotted as a function of their repeat motif (AGAT or AGAC) LUS lengths to assess 1) the ratios for AGAT versus AGAC -1 stutter at the same LUS length, and 2) the general fit of a linear LUS stutter model [58] for D12S391 when the two repeat motifs are considered separately (Fig. S2). While the -1 AGAC stutter data was limited due to its lower incidence, the plot indicates that at the same LUS length, the AGAT portion of D12S391 produces more stutter than the AGAC portion of the repeat. Given its application to the AGAT and AGAC portions of the repeat separately, the LUS stutter model fit to the data (assessed by the R^2 value) is similar to what has been previously reported for some other STR loci [59,61].

Flanking region sequence variation has recently been reported to impact stutter ratios for some STR loci with simple repeats [68]. While D12S391 data were not delineated by flanking region sequence in this study, we note that the presence or absence of a final AGAT repeat following the AGAC repeat region of D12S391 may affect stutter ratios and the resulting fit of a linear LUS stutter model.

4. Discussion

To take full advantage of the allele resolution achieved via sequencing for mixture interpretation, the ultimate aim would be a probabilistic genotyping program that can utilize sequence strings. A number of string-based search and alignment algorithms are available, and some sequence alignment tools have been developed for secondary and tertiary analyses of NGS-based STR products (e.g. [60,69,70]). However, the implementation of string-based comparison tools for STRs within a probabilistic interpretation software is unlikely to be as simple as the straight adoption of any one of the currently available alignment algorithms. The ideal algorithm within a probabilistic framework (particularly for fully continuous programs) must be sophisticated enough to assess different alignments, evaluate sequence information in the context of previously characterized and categorized data (i.e. alleles), and incorporate biologically relevant information. If the development of consistent and forensically appropriate string alignments (and nomenclature) for mitochondrial DNA [71–73] are any indication, more sophisticated algorithms than those currently available will be required, and will likely take some time to realize for practical use in probabilistic programs.

In the interim, one of the options for probabilistic genotyping of NGS data is STR typing by sequencing, but interpretation of the results based on repeat unit alleles alone. This approach would permit use of any of the presently available probabilistic programs for NGS data interpretation, but would eliminate one of the primary advantages of NGS for mixtures: the potential to improve interpretation via the sequence variation among alleles of the same length.

A second feasible near-term option is representation of sequence alleles in a format that can be accommodated by current programs. To capture all sequence variation, distinct alleles could be artificially numbered, or otherwise given a unique alphanumeric identifier to

designate both the repeat unit and specific sequence variant (as has been suggested by EUROFORGEN to standardize sequence allele nomenclature [74]). Yet, such a system poses challenges. As the identifiers would not be based on the underlying biology of the STRs, regular use would require common agreement for the designations, maintenance and curation of a database of named sequence variants, and processes for handling previously unseen alleles encountered in typing results. More importantly, from the perspective of probabilistic interpretation, implementation in programs that model stutter would require identifier consistency up and down each allelic ladder to associate potential stutter products with parent alleles.

While the LUS method we propose here does not represent 100% of observed sequence variation, its use resolves the challenges of the other near-term options. As the biology-based LUS length is derived directly from the sequence itself, allele designations that combine the repeat unit and LUS length do not require a reference database for lookup or special processes for new variants. Further, with the representation of sequence variants in the RepeatUnit.PartialUnit_LUSlength format, the allele/stutter relationship is straightforward to maintain: generally, the primary stutter product (–1 repeat unit) will be –1_1 compared to the parent allele since stutter typically occurs from the LUS.

A complication in this allele/stutter relationship arises only when complex and compound STRs are considered. These loci have more than one sequence motif within the repeat region, and stutter could conceivably occur from any of them. For example, the D12S391 locus has the general tetranucleotide repeat structure [AGAT]₈₋₂₁[AGAC]₅₋₁₁[AGAT]₀₋₁ (see Ref. [13]). In CE-developed STR data, the stutter associated with a D12S391 allele will be –1 repeat unit regardless of whether the stutter occurs from the AGAT or AGAC portion of the repeat region. However, as the data in Table S3 demonstrate, we may detect two distinct –1 repeat unit stutter products in sequence data for D12S391: the stutter product that occurs from the AGAT portion of the repeat region, and that which occurs from the AGAC portion. This potential for multiple stutter products can be accommodated, however, in a clear-cut and consistent manner using the LUS concept for allele designation. By common definition of one of the repeat motifs in a complex or compound STR as the reference region for the LUS length determination (Table 1), all products for the locus, including stutter, would be designated by reference to that region. For D12S391, the two potential –1 repeat unit stutter products would thus be represented by relation to the parent allele as –1_1 (any stutter from the AGAT region) and –1_0 (any stutter NOT from the AGAT region). An illustration of this using D12S391 typing results for a single-source sample is depicted in Fig. 3. For application in probabilistic programs that model stutter, the isoforms could be associated with stutter ratios developed from empirical data, or the secondary product could simply be allowed some probability of occurrence.

The results of the proof of concept study reported here demonstrate both that 1) allele designations using repeat unit and LUS length can be accommodated by programs that do not require products to be integers, and 2) the use of sequence information for probabilistic interpretation generally increases LR for true contributors to the DNA typing results (Fig. 2) and produces more support for exclusion of non-contributors. While LR gains for true donors were modest when less than 50% of a contributor's alleles were recovered, it is likely that interpretation in a fully continuous program (rather than the semi-continuous program used here) would produce larger differences between repeat unit and LUS-based LR values for mixtures when limited information is recovered in the evidence profile. This trend has been previously observed when comparing semi-continuous versus fully continuous programs [75,76]. Additionally, the use of stutter modeling would likely also further improve contributor resolution due to additional information that could be used for interpretation. For instance, if –1 repeat stutter products were present and considered in the interpretation of the two-person mixtures examined here, the LUS-based designations would add additional information as compared to repeat unit alone at

Sample	Locus	Repeat Unit	LUS Concept Designation	Allele Coverage	Sequence
136	D12S391	19	19_12	57	(AGAT)12(AGAC)6 AGAT
136	D12S391	20	20_13	395	(AGAT)13(AGAC)6 AGAT
136	D12S391	22	22_14	25	(AGAT)14(AGAC)7 AGAT
136	D12S391	22	22_13	74	(AGAT)13(AGAC)8 AGAT
136	D12S391	23	23_14	330	(AGAT)14(AGAC)8 AGAT

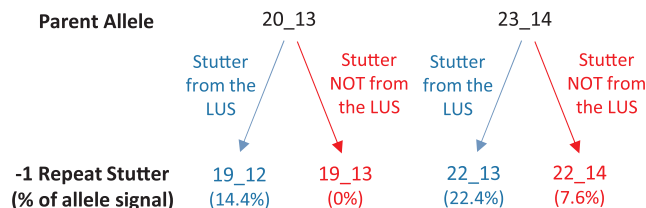


Fig. 3. Illustration of LUS concept product designations for stutter isoforms.

Actual D12S391 ForenSeq typing results for a single-source sample were used to depict how multiple –1 repeat stutter products would be designated using the LUS concept. Stutter occurring from the LUS reference region would be represented as –1_1 as compared to the parent allele, whereas stutter that occurs from any other region of the STR would be represented as –1_0 as compared to the parent allele.

four loci in each mixture (Table S2). We also note that the total sample size in this proof of concept study was very small (N = 5 individuals) and only mixtures of up to two individuals were examined. Thus, in many instances the LR differences observed between the repeat unit and LUS-based interpretations resulted exclusively from the increased rarity of the LUS-based alleles, as no isoalleles were present among the recovered loci. Further testing of more genotypes and a greater range of contributor ratios, DNA inputs and numbers of contributors will be needed to better understand the true practical value of sequence information for the interpretation of low level typing results and mixtures.

The use of the LUS concept for fully continuous probabilistic genotyping would seem to require only two primary changes. First, the programs would need to accommodate non-integer (i.e. string) products for interpretation. Second, if stutter modeling was to be used, the programs would need to allow for some probability of a secondary –1 repeat product. We note that the latter is true regardless of whether the LUS concept or some other sequence product designation scheme is applied – and that in the case of an artificial numbering system (for example), some alleles at complex or compound loci could have more than two –1 repeat stutter products by sequence [66]. While the consideration of two typing products (rather than one) as potential –1 repeat stutter is more complex than current probabilistic genotyping algorithms permit, the approach would still seem to represent a reasonable intermediate step between current software options for fully continuous interpretation of CE-based STR data and those that would need to be developed to accommodate full sequence strings. An alternative to allowing for more than one stutter product could be the application of a threshold that eliminates secondary products prior to probabilistic interpretation. With this approach, fully continuous programs would only need to be modified to handle the RepeatUnit.PartialUnit_LUSlength format and recognize –1_1 products as potential stutter. As with any use of a stutter ratio threshold, though, this approach could exclude otherwise useful information and/or risk inclusion of stutter that exceeds the threshold.

In addition to facilitating the interpretation of some sequence variation in a probabilistic framework, the LUS concept for allele designation has several additional advantages. The LUS length is easily derived from the sequence data, and the simple RepeatUnit.PartialUnit_LUSlength format is highly similar to current repeat unit STR nomenclature. Maintaining the repeat unit in the allele designation enables backward compatibility to existing STR databases, and also permits straightforward visual comparisons (e.g., for exclusionary purposes) between NGS-based

and CE-based profiles without the need to further transform either typing result. Additionally, as the LUS length relates only to the repeat region of the STR, allele designations using LUS concept will be concordant regardless of the NGS assay or analysis software used for typing (which may differ in the portions of the flanking region targeted or reported). This consistency would enable known specimen profiles to be compared to single source and mixed contributor typing results from evidence items even if the NGS typing was performed using different assays, sequencing platforms, data analysis programs or by different laboratories. Similarly, it would permit use of allele frequency data developed via any NGS-based STR typing system.

5. Conclusions

NGS has been shown to have substantial advantages over CE for aSTR typing with respect to 1) the greater number of markers that can be simultaneously typed, 2) the minimization of STR amplicon size, and 3) the greater information content of the sequence data themselves. A combination of these factors, together with potential increased sensitivity to low-level DNA contributors, is likely to improve both the detection of mixtures and the resolution of distinct contributors. Unfortunately, the tools necessary to exploit sequence information for the interpretation of mixed DNA profiles in a probabilistic framework are lacking, and it may be some time before a probabilistic solution that makes full use of the complete sequence information is developed. As a result, we propose an intermediate solution based on the LUS that would enable near-term use of greater than 80% of the currently known STR sequence variation in mixture analyses. Though this allele designation concept is largely suggested as a practical solution to a short-term problem, it is an approach that has a number of advantages. Because the concept is based on the core repeat region(s), the complexities related to the range of the locus sequenced (which may vary based on assay, NGS data analysis software, and STR sequence reference databases) that would otherwise require consideration and accommodation can be avoided. In addition, as the LUS length is derived directly from the sequence, allele designations using the concept do not require a curated list of sequence variants and their associated identifiers. Further, because the LUS concept format maintains the numeric repeat unit allele designations used in extant STR databases and familiar to forensic practitioners, it may ease the transition from CE-based genotyping to sequence-based genotyping. Already, currently available NGS data analysis packages transform and represent abstract digital data in familiar “electropherogram” form [64,65] for, presumably, similar reasons. Ultimately, by facilitating probabilistic interpretation of sequence-based STR data in software programs that currently exist (as we have shown here) or are highly similar to those that already exist, and at the same time leveraging the vast majority of currently characterized STR sequence variation, use of the LUS length in allele designations could result in more informative mixture statistics in the very near term.

Acknowledgements

The authors thank Lilliana Moreno, Michelle Galusha, Anthony Onorato, Nicholas Vlachos, Thomas Callaghan, Tamyra Moretti, Lara Adams and Jade Gray of the FBI Laboratory for their contributions to this work. This research was supported in part through the FBI's Visiting Scientist Program, an educational opportunity administered by the Oak Ridge Institute for Science and Education (ORISE). Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government. This is FBI Laboratory publication #17-20. Conflicts of interest: none.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.fsigen.2018.02.016>.

References

- [1] 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.
- [2] J. Irwin, R. Just, M. Scheible, O. Loreille, Assessing the potential of next generation sequencing technologies for missing persons identification efforts, *Forensic Sci. Int. Genet. Suppl. Ser. 3* (2011) e447–e448.
- [3] M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat sequencing on the 454 platform, *Forensic Sci. Int. Genet. Suppl. Ser. 3* (2011) e357–e358.
- [4] 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).
- [5] 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 Sci. Int. Genet.* 12 (2014) 107–119.
- [6] A.D. Ambers, J.D. Churchill, J.L. King, M. Stoljarova, H. Gill-King, M. Assidi, et al., More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing, *BMC Genomics* 17 (2016) 750.
- [7] F. Calafell, R. Anglada, N. Bonet, M. Gonzalez-Ruiz, G. Prats-Munoz, R. Rasal, et al., An assessment of a massively parallel sequencing approach for the identification of individuals from mass graves of the Spanish Civil War (1936–1939), *Electrophoresis* 37 (2016) 2841–2847.
- [8] P. Fattorini, C. Previdere, I. Carboni, G. Marrubini, S. Sorcaburu-Cigliero, P. Grignani, et al., Performance of the ForenSeq™ DNA Signature Prep kit on highly degraded samples, *Electrophoresis* 38 (April (8)) (2017) 1163–1174, <http://dx.doi.org/10.1002/elps.201600290> Epub 2017 Feb 9.
- [9] C. Xavier, W. Parson, Evaluation of the illumina ForenSeq DNA signature prep kit – MPS forensic application for the MiSeq FGx benchtop sequencer, *Forensic Sci. Int. Genet.* 28 (2017) 188–194.
- [10] A. Urquhart, C.P. Kimpton, T.J. Downes, P. Gill, Variation in short tandem repeat sequences—a survey of twelve microsatellite loci for use as forensic identification markers, *Int. J. Legal Med.* 107 (1994) 13–20.
- [11] B. Brinkmann, A. Sajantila, H.W. Goedde, H. Matsumoto, K. Nishi, P. Wiegand, Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci, *Eur. J. Hum. Genet.* 4 (1996) 175–182.
- [12] 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.
- [13] M.V. Lareu, C. Pestoni, M. Schurenkamp, S. Rand, B. Brinkmann, A. Carracedo, A highly variable STR at the D12S391 locus, *Int. J. Legal Med.* 109 (1996) 134–138.
- [14] M.V. Lareu, M.C. Pestoni, F. Barros, A. Salas, A. Carracedo, Sequence variation of a hypervariable short tandem repeat at the D12S391 locus, *Gene* 182 (1996) 151–153.
- [15] H. Oberacher, W. Parson, R. Muhlmann, C.G. Huber, Analysis of polymerase chain reaction products by on-line liquid chromatography-mass spectrometry for genotyping of polymorphic short tandem repeat loci, *Anal. Chem.* 73 (2001) 5109–5115.
- [16] 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.
- [17] J.V. Planz, K.A. Sannes-Lowery, D.D. Duncan, S. Manalili, B. Budowle, R. Chakraborty, et al., Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry, *Forensic Sci. Int. Genet.* 6 (2012) 594–606.
- [18] 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.
- [19] B. Glock, E.M. Dauber, D.W. Schwartz, W.R. Mayr, Additional variability at the D12S391 STR locus in an Austrian population sample: sequencing data and allele distribution, *Forensic Sci. Int.* 90 (1997) 197–203.
- [20] Y. Shigeta, Y. Yamamoto, Y. Doi, S. Miyaishi, H. Ishizu, Polymorphism of the D12S391 microsatellite in a Japanese population sample, *Forensic Sci. Int.* 102 (1999) 61–66.
- [21] S. Hering, E. Muller, New alleles and mutational events in D12S391 and D8S1132: sequence data from an eastern German population, *Forensic Sci. Int.* 124 (2001) 187–191.
- [22] C. Phillips, L. Fernandez-Formoso, M. Garcia-Magarinos, L. Porras, T. Tvedebrink, J. Amigo, 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.* 5 (2011) 155–169.
- [23] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, *Forensic Sci. Int. Genet.* 18 (2015) 118–130.
- [24] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, et al., Sequence variation of 22 autosomal STR loci detected by next generation sequencing, *Forensic Sci. Int. Genet.* 21 (2016) 15–21.
- [25] N.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of

- genetic sequence variation of 58 STR loci in four major population groups, *Forensic Sci. Int. Genet.* 25 (2016) 214–226.
- [26] K.J. van der Gaag, R.H. de Leeuw, J. Hoogenboom, J. Patel, D.R. Storts, J.F. Laros, et al., Massively parallel sequencing of short tandem repeats-population data and mixture analysis results for the PowerSeq system, *Forensic Sci. Int. Genet.* 24 (2016) 86–96.
- [27] R.S. Just, L.I. Moreno, J.B. Smerick, J.A. Irwin, Performance and concordance of the ForenSeq system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens, *Forensic Sci. Int. Genet.* 28 (2017) 1–9, <http://dx.doi.org/10.1016/j.fsigen.2017.01.001>.
- [28] L. Devesse, D. Ballard, L. Davenport, I. Riethorst, G. Mason-Buck, D. Syndercombe Court, Concordance of the ForenSeq system and characterisation of sequence-specific autosomal STR alleles across two major population groups, *Forensic Sci. Int. Genet.* 34 (May) (2018) 57–61, <http://dx.doi.org/10.1016/j.fsigen.2017.10.012>.
- [29] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the Illumina (RR) beta version ForenSeq DNA signature prep kit for use in genetic profiling, *Forensic Sci. Int. Genet.* 20 (2016) 20–29.
- [30] A.C. Jager, M.L. Alvarez, C.P. Davis, E. Guzman, Y. Han, L. Way, et al., Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories, *Forensic Sci. Int. Genet.* 28 (2017) 52–70.
- [31] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, et al., Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, *Forensic Sci. Int. Genet.* 22 (2016) 54–63.
- [32] A. Alonso, P. Muller, L. Roewer, B. Budowle, W. Parson, European survey on forensic applications of massively parallel sequencing, *Forensic Sci. Int. Genet.* 29 (July) (2017) e23–e25, <http://dx.doi.org/10.1016/j.fsigen.2017.04.017> Epub 2017 Apr 26.
- [33] P. Gill, L. Gusmao, H. Haned, W.R. Mayr, N. Morling, W. Parson, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, *Forensic Sci. Int. Genet.* 6 (2012) 679–688.
- [34] Scientific Working Group on DNA Analysis Methods (SWGDM), Guidelines for the Validation of Probabilistic Genotyping Systems, (2015) https://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf.
- [35] F.R. Bieber, J.S. Buckleton, B. Budowle, J.M. Butler, M.D. Coble, Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion, *BMC Genet.* 17 (2016) 125-016–0429-7.
- [36] T.R. Moretti, R.S. Just, S.C. Kehl, L.E. Willis, J.S. Buckleton, J.A. Bright, et al., Internal validation of STRmix for the interpretation of single source and mixed DNA profiles, *Forensic Sci. Int. Genet.* 29 (2017) 126–144.
- [37] I.W. Evett, P.D. Gill, J.A. Lambert, Taking account of peak areas when interpreting mixed DNA profiles, *J. Forensic Sci.* 43 (1998) 62–69.
- [38] R.G. Cowell, S.L. Lauritzen, J. Mortera, A gamma model for DNA mixture analyses, *Bayesian Anal.* 2 (2007) 333–348.
- [39] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, *Forensic Sci. Int. Genet.* 4 (2009) 1–10.
- [40] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele(R) DNA mixture interpretation, *J. Forensic Sci.* 56 (2011) 1430–1447.
- [41] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, *Forensic Sci. Int. Genet.* 5 (2011) 265–268.
- [42] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, *Forensic Sci. Int. Genet.* 6 (2012) 749–761.
- [43] K.E. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, *J. Forensic Sci.* 58 (Suppl. 1) (2013) S243–9.
- [44] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights allowing for multiple donors, allelic dropout and stutters, *Forensic Sci. Int. Genet.* 7 (2013) 555–563.
- [45] D. Taylor, J.A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, *Forensic Sci. Int. Genet.* 7 (2013) 516–528.
- [46] R. Puch-Solis, T. Clayton, Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRA software, *Forensic Sci. Int. Genet.* 11 (2014) 220–228.
- [47] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dorum, T. Egeland, Genotyping and interpretation of STR-DNA: low-template, mixtures and database matches—twenty years of research and development, *Forensic Sci. Int. Genet.* 18 (2015) 100–117.
- [48] T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, J.B. Smerick, B. Budowle, Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples, *J. Forensic Sci.* 46 (2001) 647–660.
- [49] T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, A.L. Brown, B. Budowle, Validation of STR typing by capillary electrophoresis, *J. Forensic Sci.* 46 (2001) 661–676.
- [50] P. Gill, R. Puch-Solis, J. Curran, The low-template-DNA (stochastic) threshold—its determination relative to risk analysis for national DNA databases, *Forensic Sci. Int. Genet.* 3 (2009) 104–111.
- [51] Scientific Working Group on DNA Analysis Methods (SWGDM), Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, (2010) swgdam.org.
- [52] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, *Forensic Sci. Int. Genet.* 11 (2014) 144–153.
- [53] S. Cooper, C. McGovern, J.A. Bright, D. Taylor, J. Buckleton, Investigating a common approach to DNA profile interpretation using probabilistic software, *Forensic Sci. Int. Genet.* 16 (2015) 121–131.
- [54] H. Kelly, J.A. Bright, J.S. Buckleton, J.M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, *Sci. Justice* 54 (2014) 66–70.
- [55] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, *Forensic Sci. Int. Genet.* 6 (2012) 762–774.
- [56] C. Brookes, J.A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, *Forensic Sci. Int. Genet.* 6 (2012) 58–63.
- [57] M. Klintschar, P. Wiegand, Polymerase slippage in relation to the uniformity of tetrameric repeat stretches, *Forensic Sci. Int.* 135 (2003) 163–166.
- [58] J.A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, *Forensic Sci. Int. Genet.* 7 (2013) 296–304.
- [59] J.A. Bright, K.E. Stevenson, M.D. Coble, C.R. Hill, J.M. Curran, J.S. Buckleton, Characterising the STR locus D6S1043 and examination of its effect on stutter rates, *Forensic Sci. Int. Genet.* 8 (2014) 20–23.
- [60] J. Hoogenboom, K.J. van der Gaag, R.H. de Leeuw, T. Sijen, P. de Knijff, J.F. Laros, FDSTools a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, *Forensic Sci. Int. Genet.* 27 (2017) 27–40.
- [61] A.A. Westen, L.J. Grol, J. Hartevelde, A.S. Matai, P. de Knijff, T. Sijen, Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM, *Forensic Sci. Int. Genet.* 6 (2012) 708–715.
- [62] H. Haned, J. de Jong, LRmix Studio 2.1 User Manual, (2016) <http://lrmixstudio.org/download/manual.pdf>.
- [63] Illumina ForenSeq DNA Signature Prep Reference Guide, Document #15049528 v01, (2015) (September 28).
- [64] D.H. Warshauer, J.L. King, B. Budowle, STRait razor v2.0: the improved STR allele identification tool—razor, *Forensic Sci. Int. Genet.* 14 (2015) 182–186.
- [65] Illumina, ForenSeq Universal Analysis Software Guide Part # 15053876 v01, (2016).
- [66] M.B. Galusha, R.S. Just, L.I. Moreno, Internal validation of the ForenSeq DNA Signature Prep Kit for reference samples: successes and lessons learned, Oral Presentation at the Green Mountain DNA Conference, Burlington, VT, 2017 [vfl.vermont.gov/sites/pslab/files/pdfs/conference/Moreno.pdf](http://vermont.gov/sites/pslab/files/pdfs/conference/Moreno.pdf).
- [67] L.I. Moreno, M.B. Galusha, R.S. Just, Tuning out the noise: A closer look at ForenSeq autosomal and Y-STR data to develop protocols for routine reference sample typing, submitted.
- [68] A.E. Woerner, J.L. King, B. Budowle, Flanking variation influences rates of stutter in simple repeats, *Genes (Basel)* 8 (2017), <http://dx.doi.org/10.3390/genes8110329>.
- [69] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, et al., STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, *Forensic Sci. Int. Genet.* 7 (2013) 409–417.
- [70] S.L. Friis, A. Buchard, E. Rockenbauer, C. Borsting, N. Morling, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, *Forensic Sci. Int. Genet.* 21 (2016) 68–75.
- [71] M.R. Wilson, M.W. Allard, K. Monson, K.W. Miller, B. Budowle, Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region, *Forensic Sci. Int.* 129 (2002) 35–42.
- [72] H.J. Bandelt, W. Parson, Consistent treatment of length variants in the human mtDNA control region: a reappraisal, *Int. J. Legal Med.* 122 (2008) 11–21.
- [73] A. Rock, 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.
- [74] S. Willuweit, NOMAUT ? NGS STR nomenclature for forensic genetics, Oral Presentation at the 27th Congress of the International Society for Forensic Genetics, Seoul, South Korea, 2017.
- [75] O. Bleka, C.C.G. Benschop, G. Storvik, P. Gill, A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles, *Forensic Sci. Int. Genet.* 25 (2016) 85–96.
- [76] O. Bleka, M. Eduardoff, C. Santos, C. Phillips, W. Parson, P. Gill, Open source software EuroForMix can be used to analyse complex SNP mixtures, *Forensic Sci. Int. Genet.* 31 (2017) 105–110.