Article:

Bright, J. A., Stevenson, K. E., Coble, M. D., Hill, C. R., Curran, J. M., & Buckleton, J. S. (2014). **Characterising the STR locus D6S1043 and examination of its effect on stutter rates**. *Forensic Science International: Genetics*, *8*(1), 20–23.

INSTITUTE OF ENVIRONMENTAL SCIENCE AND RESEARCH LIMITED
Kenepuru Science Centre: 34 Kenepuru Drive, Kenepuru, Porirua 5022 | PO Box 50348, Porirua 5240, New Zealand
T: +64 4 914 0700 **F:** +64 4 914 0770
www.esr.cri.nz

Characterising the STR locus D6S1043 and examination of its effect on stutter rates

Jo-Anne Bright[1,2*], Kate E. Stevenson[1], Michael D. Coble[3], Carolyn R. Hill[3], James M. Curran[2], John S. Buckleton[1]

[1] *ESR Ltd, Private Bag 92021, Auckland 1025, New Zealand*

[2] *Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1025, New Zealand*

[3] *U.S. National Institute of Standards and Technology, Gaithersburg, MD, USA*

* Corresponding author at: Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142, New Zealand. Email address: Jo.bright@esr.cri.nz .

The forensic analysis of DNA is most often undertaken by the amplification of short tandem repeats (STR) using the polymerase chain reaction (PCR). DNA amplification can result in production of the target allele amplicon and a by-product called stutter. Stutter is the result of the miscopy of the target allele and is typically one repeat smaller. Stutter is traditionally described as a ratio of stutter and allele height; stutter ratio (*SR*). The challenge to DNA profile interpretation is most serious whenever stutter products are of a similar height to the minor allelic peaks in a mixed DNA profile. An accurate assignment of peaks and the prediction of their height is important when objectively interpreting forensic DNA profiles. The longest uninterrupted stretch (*LUS*) of tandem repeats within the allele has previously been shown to be a good predictor of stutter ratio. *LUS* is determined by sequencing a range of observed alleles at a locus. The locus D6S1043 is a relatively new locus to appear in commercial forensic DNA testing kits. To date however, there has been no comprehensive report of sequencing of this locus. In this work, we sequence a sample of D6S1043 alleles to determine *LUS* values and investigate allele repeat number and *LUS* as explanatory variables for *SR*.

Keywords: Forensic DNA; sequencing; stutter; interpretation; D6S1043

Introduction

The analysis of forensic DNA is predominantly undertaken by the amplification of short, tandemly repeated lengths of DNA (STRs) using the polymerase chain reaction (PCR). During the PCR process, primers flanking the region of interest are attached to the template DNA during the annealing phase and then the sequence is transcribed during the extension phase, resulting in a theoretical doubling of DNA after each cycle [1]. The primers flanking different STRs can be multiplexed into one reaction. Commercial multiplexes are available that can target up to 26 loci simultaneously.

An unwanted product of the PCR reaction is stutter. A proposed mechanism for stuttering is slipped strand mispairing (SSM) during PCR where the template strand "loops out" resulting in the new strand being one repeat unit shorter than the template strand [2-4]. SSM occurs during DNA replication *in vivo* and also results during PCR *in vitro*. The DNA polymerase enzyme stalls, dissociates from the DNA, and a loop of one or more repeat units may form in either the nascent or the template strand. This causes the insertion or deletion respectively, of one or more STR repeats. Generally the template strand loops out resulting in the new strand being one repeat unit shorter than the template strand (sometimes referred to as back stutter or *N*-1 stutter) [3].

Rates of stutter vary across a DNA profile depending on the locus [5]. The challenge to interpretation is most significant whenever stutter products are of a similar height to minor allelic peaks within a mixed DNA profile or when a large stutter peak raises suggestions of an additional trace contributor.

Back stutter is typically quantified as a stutter ratio (*SR*):

$$SR = \frac{O_{a-1}}{O_a}$$

where O_{a-1} refers to the observed height of the stutter peak, and O_a the parent peak. Stutter can be interpreted either on the basis of a given rule applied per locus, or per multiplex (for example 15%) [6], or directly by modelling the probability of all peaks in the profile given a proposed genotype [7, 8]. It has previously been shown that the longest uninterrupted stretch of repeats (LUS) at an allele is a good predictor of stutter [3, 9] at that allele. LUS is the longest stretch of basic repeat motifs within an allele. It has been shown that alleles with large LUS values stutter more than alleles with small LUS values and amplify less. Values for LUS are determined by sequencing STRs [9].

The locus D6S1043 is a relatively new locus in commercial forensic DNA testing kits. It is available in Applied Biosystems' AmpFlSTR Sinofiler™ multiplex (Life Technologies, Carlsbad, CA) and Promega's PowerPlex® 21 multiplex (Madison, WI). DS61043 has been reported to be highly discriminatory in Asian populations [10-12]. D6S1043 is described as a compound repeat with a core sequence of AGAT and a less common sequence AGAC [13, 14].

In this paper, allele repeat number and LUS are investigated as explanatory variables for stutter ratio at locus D6S1043. An accurate assignment of stutter peaks and the prediction of their height are important factors when interpreting profiles with low level mixtures. Stutter rates were determined using a large dataset of single source DNA profiles from eight laboratories prepared as part of the implementation of an expert software. Sanger sequencing of D6S1043 alleles from 40 volunteers was undertaken to determine values for LUS.

Methods

Stutter ratio variability

Single source PowerPlex® 21 (Promega Corporation, Madison WI) DNA profiles were submitted for analysis from eight laboratories either as previously analysed outputs or as raw, unanalysed .fsa files. All raw data was analysed using Applied Biosystems' GeneMapper™ ID v 3.2.1 with an analysis threshold of 30 rfu. Previously analysed data sets provided by the laboratories were analysed with a maximum analysis threshold of 30 rfu, with some at thresholds below this. All profiles were obtained from samples amplified with 30 cycles. Amplified products from laboratories two and seven were separated using Applied Biosystem's 3500 capillary electrophoresis instruments with the remaining laboratories using 3130 instruments. Samples with large parent allele peak heights that are likely to be affected by saturation effects were removed from the analysis. At these loci the relationship between amount of DNA and allele height is no longer linear therefore resulting in higher estimates of stutter.

Laboratory methods

Laboratory methods followed closely those described in Kline et al. [15]. Within the methods, manufacturers' recommendations were followed except where noted.

DNA extraction

Buccal swabs collected from 40 volunteers were extracted using Promega's DNA IQ™ magnetic bead extraction chemistry (Madison, WI) and quantified using Applied Biosystems Quantifiler™ real time PCR quantitation kit (Life Technologies, Carlsbad, CA). All homozygote samples and heterozygote samples where the repeat numbers differed by greater than one (and therefore were not affected by stutter) were selected for sequencing. A total of 42 alleles were sequenced.

DNA amplification

All samples were amplified using Promega's PowerPlex® 21 multiplex kit (Madison, WI) to determine the expected repeat number.

Homozygotes

The following unlabelled primer sequences were used in the DNA amplification and sequencing reactions:

Forward 5'– TTCGGTATTCTCCACATGGTT – 3'

Reverse 5'– TTCTCTGCCCTTTGTACTCCA – 3'

Primers were synthesized by Eurofins MWG Operon (Huntsville, AL, USA).

A target of 5 ng of DNA was amplified using the unlabelled primers (1.0 μM each primer) with AmpliTaq Gold mastermix (Life Technologies, Carlsbad CA) in a 25 μL reaction.

Samples were denatured on an Applied Biosystems 9700 thermal cycler with a silver block for 10 minutes at 95 °C, followed by 35 cycles of PCR of 1 minute at 94 °C, 1 minute at 59 °C, and 1 minute at 72 °C. Final extension was for 45 min at 60 °C.

Heterozygotes

The amplified products were separated on an E-Gel EX 4% agarose gel (Invitrogen) run for 40 minutes. Separated heterozygote alleles were excised from the gel using a scalpel. Care was taken to avoid stutter bands. DNA was purified from the gel using the Zymoclean™ gel DNA recovery kit (Zymo Research, Orange CA). A 5 µL aliquot of extracted DNA from the agarose gel was re-amplified as described above.

DNA Sequencing

The amplification product (amplified homozygote and re-amplified individual heterozygote alleles) were purified by using 2 µL of illustra ExoStar 1-step (GE Healthcare) per 5 µL of sample on a 9700 thermal cycler.

Sequencing was then undertaken using the Applied Biosystems BigDye Terminator 3.1 cycle sequencing kit. The sequencing mixture consisted of a 10 µL total reaction volume: 4 µL of PCR product, 1 µL of 2.5x Ready reaction mix, 1 µL BigDye 5x Sequencing Buffer, 3 µL H_2O, and 1/10 dilution 1 µL of the forward or reverse primer (at a concentration of 1.0 µM).

Samples were denatured for 5 min at 94°C followed by 25 cycles of 10 s at 94 °C, 5 s at 50 °C, and 2 min at 60 °C, and a final extension for 5 min at 60 °C on a 9700 thermal cycler.

The cycle sequencing products were purified using the CleanSEQ - Dye Terminator Removal technology (Agencourt) prior to electrophoresis on an Applied Biosystems 3130xl Genetic Analyzer. 10 µL of sample was run (injection parameters 22 seconds, 1 kV) in POP-4 on a 36 cm capillary array with the appropriate dye set. Sequences were analysed using Sequence Analysis v3.7 (Applied Biosystems) and aligned using Geneious version 6.0.3 (Biomatters, http://www.geneious.com/).

Data analysis

All data interpretation was undertaken in the statistics program R [16]. Further information on some of the statistical analyses used in this report can be found in Curran [17].

Results

The plot *SR* versus allele repeat number for the locus D6S1043 is presented in Figure 1. Inspection of Figure 1 shows a poor fit of the data (R^2 = 0.30). There appears to be two trends within the data: one for repeats 10 to 14 and a second with repeats from 17 to 20. A small number of samples with allele repeat numbers 15 and 16 appear to split evenly between the two sets.

A summary of the sequenced allele repeats and their *LUS* values is in Table 1. The plot of *SR* versus *LUS* is given in Figure 2. The plot of *SR* versus *LUS* demonstrates a better fit to the data than allele repeat number (R^2 = 0.60).

Figure 1: A plot of *SR* versus allele repeat number

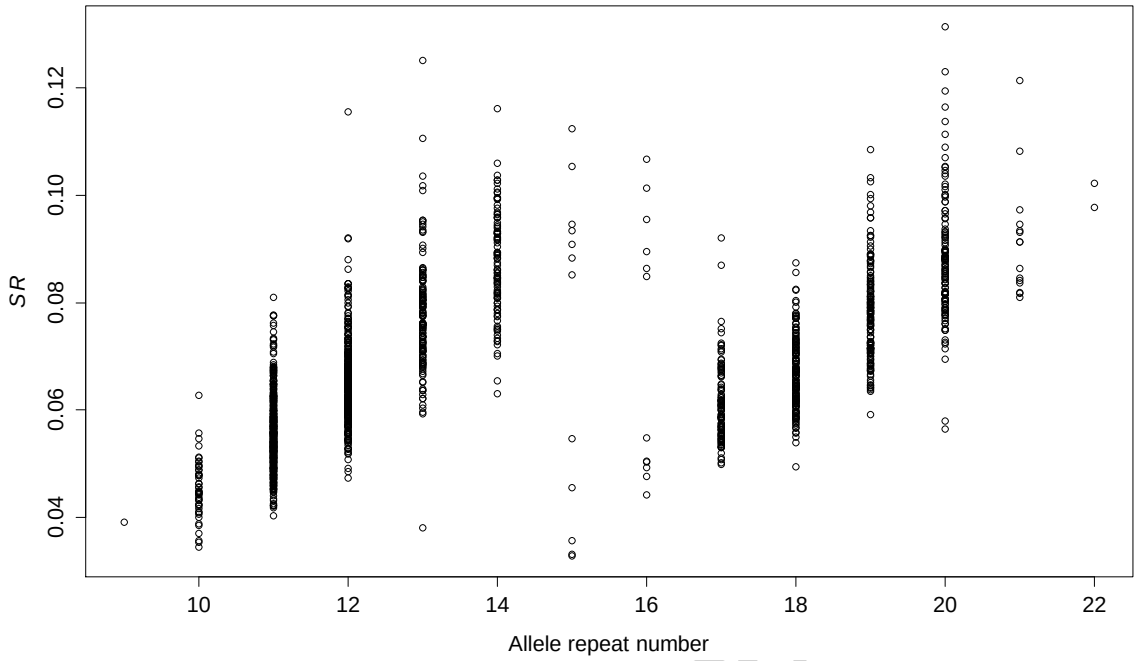
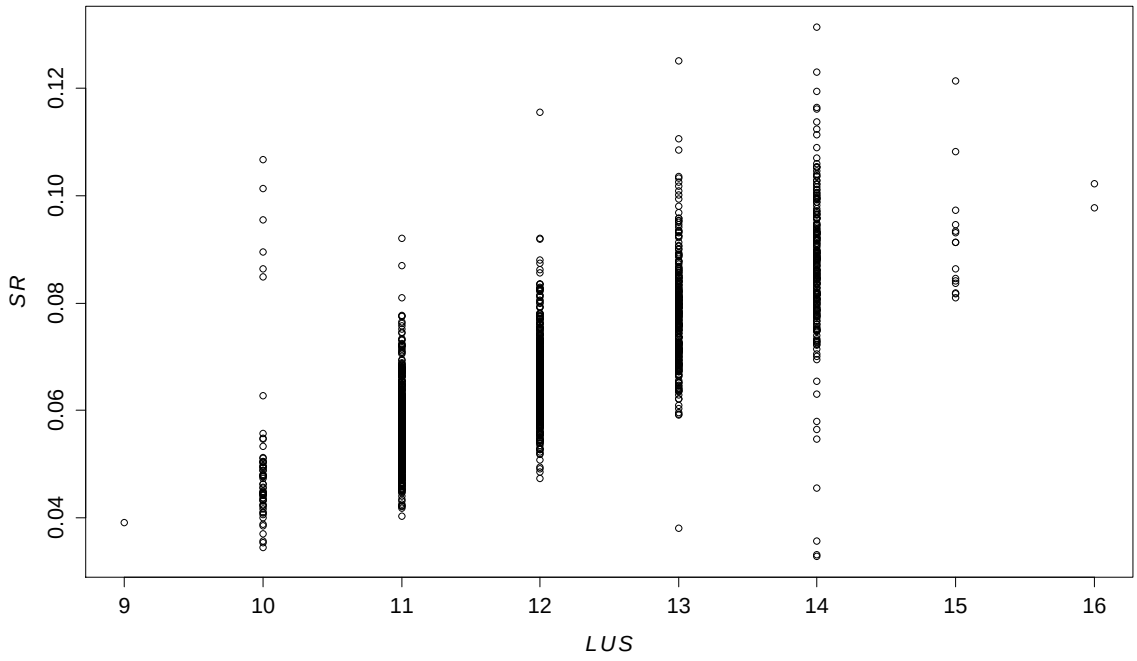


Table 1: Allele repeat number, DNA sequence and *LUS* value

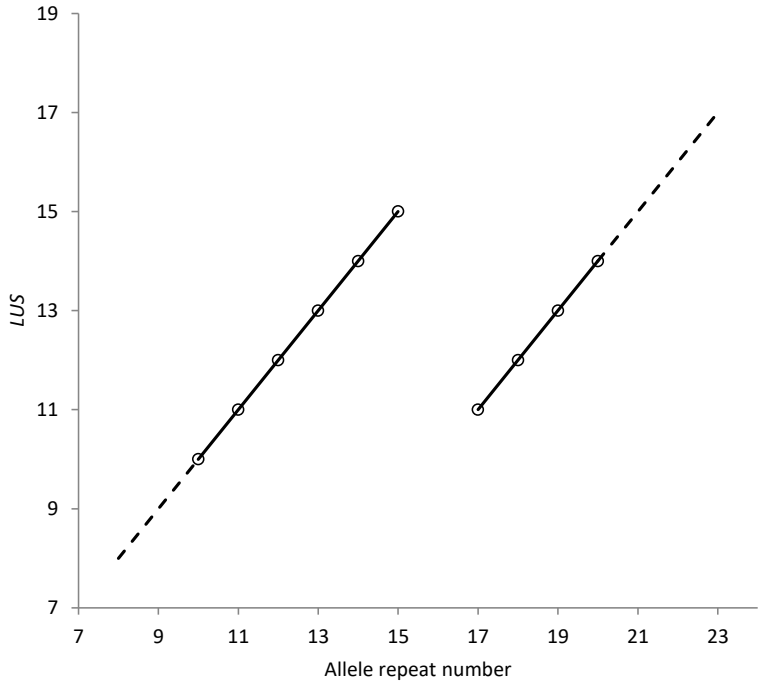
Repeat number	Sequence	*LUS*	Number sequenced
10	$[AGAT]_{10}$	10	2
11	$[AGAT]_{11}$	11	10
12	$[AGAT]_{12}$	12	6
13	$[AGAT]_{13}$	13	3
14	$[AGAT]_{14}$	14	2
15	$[AGAT]_{15}$	15	1
17	$[AGAT]_{11}ACAT[AGAT]_5$	11	4
18	$[AGAT]_{12}ACAT[AGAT]_5$	12	4
19	$[AGAT]_{13}ACAT[AGAT]_5$	13	8
20	$[AGAT]_{14}ACAT[AGAT]_5$	14	2

Figure 2: A plot of *SR* versus *LUS*



For alleles that were not sequenced, the *LUS* value was obtained by extrapolation from Table 1 as in Figure 3. Allele 9 was assigned a *LUS* value of 9 and alleles 21 and 22 LUS values of 15 and 16, respectively. This action mimics the application of the *LUS* model in casework where the allele will be described simply by its allele designation and the true sequence will be unknown. In cases of ambiguous sequences or for alleles not previously sequenced (for example rare alleles) extrapolation from a source of previously sequenced and published profiles must suffice.

Figure 3: LUS value versus allele repeat number



Inspection of Figure 1 shows two populations of data for *SR* for samples with an allele repeat number of 15 and 16. It is likely that there are two corresponding DNA sequences; one with low rates of stutter and the second with higher rates. Only one 15 allele was sequenced in this small study leading to the assignment of a *LUS* value of 15. The second (unobserved) sequence is likely to contain an interrupt in the same sequence as the 17 to 20 alleles (following the same trend) thus resulting in smaller rates of stutter. No 16 alleles were sequenced in this study. The *LUS* value for allele 16 was taken to be 10 to generate Figure 2. Sequencing of more 15 and some 16 alleles is warranted.

Discussion

The introduction of continuous (probabilistic) methods for the interpretation of forensic DNA profiles has the potential to increase the efficiency of laboratories, and improve the consistency and transparency of results. Stutter is an unwanted by-product of the PCR process and can interfere with the assignment of peaks to individuals, especially in the presence of minor contributors. This paper investigates allele repeat number and the longest uninterrupted stretch of short tandem repeats at an allele as a predictor of stutter rate at the locus D6S1043. DNA sequences of alleles at this locus had not been previously reported. LUS was shown to be a better predictor of stutter ratio than allele repeat number as previously reported for other forensic loci [8]. Sequencing of more 15- and 16-repeat alleles (not sequenced in this study) is warranted as the variation in stutter ratio indicates that both simple and complex variations of these alleles exist. When using a continuous model of interpretation, a weighted average of the different motifs would assist with interpretation of ambiguous sequences, such as alleles 15 and 16. The population proportions of the different sequences would be required. This would require a significant population study of the different sequences.

The accurate prediction of the height of stutter peaks is important when stutter products are of a similar height to minor allelic peaks within a mixed DNA profile. Models for the prediction of stutter can be used in expert systems and remove the requirement for the manual assignment of peaks as allelic or stutter within evidence profiles. This work demonstrates the effectiveness of the LUS model for predicting the height of stutter peaks at locus D6S1043.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. We acknowledge the useful advice of Julia Allwood and Joanne Simons (ESR) and the comments of Johanna Veth and Joanne Simons which have greatly improved this paper. The authors also gratefully acknowledge the source of the stutter ratio data that lies behind Figures 1 and 2. These data were provided by the state and territory government forensic biology laboratories of Australia.

References

[1] J. M. Butler. Advanced Topics in Forensic DNA Typing: Methodology: Academic Press; 2011.

[2] X. Y. Hauge, M. Litt, A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR, Human Molecular Genetics. 2(4) (1993) 411-415.

[3] 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.

[4] G. Levinson, G. Gutman, Slipped Strand Mispriming : a major mechanism for DNA sequence evolution, Molecular Biology and Evolution. 4(3) (1987) 203-221.

[5] J. S. Buckleton, C. M. Triggs, S. J. Walsh. Forensic DNA Evidence Interpretation. Boca Raton, Florida: CRC Press; 2004.

[6] P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Science International. 160(2–3) (2006) 90-101.

[7] M. W. Perlin, M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose, et al., Validating TrueAllele® DNA mixture interpretation. , Journal of Forensic Sciences. 56 (2011) 1430-1447.

[8] 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 Science International: Genetics. 7(2) (2013) 296-304.

[9] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Science International: Genetics. 6(1) (2012) 58-63.

[10] S. Huang, Y. Zhu, X. Shen, X. Le, H. Yan, Genetic variation analysis of 15 autosomal STR loci of AmpFℓSTR® Sinofiler™ PCR Amplification Kit in Henan (central China) Han population, Legal Medicine. 12(3) (2010) 160-161.

[11] D. J. Lu, Q. L. Liu, H. Zhao, Genetic data of nine non-CODIS STRs in Chinese Han population from Guangdong Province, Southern China, International Journal of Legal Medicine. 125(1) (2011) 133-137.

[12] D. Tong, Y. Chen, X. Ou, W. Chen, S. Liu, Y. Zhang, et al., Polymorphism analysis and evaluation of 19 STR loci in the Han population of Southern China, Annals of Human Biology. 40(2) (2013) 191-196.

[13] J. M. Butler, C. R. Hill, D. L. Duewer, M. C. Kline, K. O'Connell. Characteristics of 24 Commonly Used Autosomal STR Loci and U.S. Population Data with the Recently Announced Expanded CODIS Core Loci. International Symposium on Human Identification; National Harbor, MD2011.

[14] Applied Biosystems, AmpFlSTR Sinofiler PCR Amplification Kit User Guide, 2012.

[15] M. C. Kline, C. R. Hill, A. E. Decker, J. M. Butler, STR sequence analysis for characterizing normal, variant, and null alleles, Forensic Science International: Genetics. 5(4) (2011) 329-332.

[16] R Development Core team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2004.

[17] J. M. Curran. Introduction to data analysis with R for forensic scientists. Boca Raton, FL: CRC Press; 2010.