E/S/R
Science for Communities

Article:

Bright, J., Buckleton, J. S., Taylor, D., Fernando, M. A. and Curran, J. M. (2014), ***Modeling forward stutter: Toward increased objectivity in forensic DNA interpretation***. ELECTROPHORESIS, 35: 3152-3157.

This is the **Accepted Manuscript** (final version of the article which included reviewers' comments) of the above article published by **Wiley** at
https://doi.org/10.1002/elps.201400044

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

Modelling forward stutter: towards increased objectivity in forensic DNA interpretation

Jo-Anne Bright[1,2*], John S. Buckleton[1], Duncan Taylor[3], M. A. C. S. S. Fernando[2], James M. Curran[2]

[1] *ESR Ltd*

[2] *University of Auckland, Department of Statistics*

[3] *Forensic Science South Australia*

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

Forward stutter, or over stutter, one repeat unit length larger than the parent allele (*N*+1 stutter), is a relatively rare product of the PCR amplification of short tandem repeats used in forensic DNA analysis. We have investigated possible explanatory variables for the occurrence and size of forward stutter for four different autosomal multiplexes. In addition, we have investigated models used to predict the expected heights of forward stutter. For all tetra and penta-nucleotide repeats we can find no correlation between allelic peak height, marker or longest uninterrupted sequence in the allele. The data fit a gamma distribution with no explanatory variables. For the single tri-nucleotide repeat present in two of the four multiplexes (D22S1045) forward stutter is much more common and the best explanatory variable appears to be back stutter height. This suggests some fundamental co-causation of high backward and forward stutter for this locus.

Keywords: Forensic DNA, interpretation, forward stutter, continuous interpretation models, Identifiler™, NGM SElect™, PowerPlex® 21, GlobalFiler™

Accepted Manuscript

Introduction

Analysis of short tandem repeat (STR) DNA sequences using PCR is the dominant forensic DNA analysis technique. The loci were selected for their short length, variability between individuals and their suitability for amplification via multiplexed PCR reactions. STR loci used in modern day multiplexes are typically tetra nucleotide repeats. Tri- and pentanucleotide repeats are also used.

The analysis of STRs is complicated by the occurrence of stutter, a by-product of the PCR process. Stutter differs in size from the allele by multiples of the repeat unit length [1]. The challenge to DNA profile interpretation comes when stutter products are of a similar height to the minor allelic peaks from an additional contributor in a mixed DNA profile or when a large peak in a stutter position raises the suggestion of an additional, trace contributor. The amount of stutter has been observed to be inversely proportional to the repeat unit length. Di- and trinucleotide repeats stutter more than tetranucleotide repeats which in turn stutter more than pentanucleotide repeats [2]. The most common type of stutter is one repeat unit length smaller than the parent allele. Traditionally, this is referred to as 'stutter' but is also known as back stutter or *N*-1 stutter. In this paper, when we use the term 'stutter' we refer to back stutter and the term 'forward stutter' will be used when we discuss stutter in the *N*+1 position. There have been many reported investigations into the occurrence and cause of back stutter [1, 3, 4]. Forward stutter, or over stutter, which is one repeat unit length larger than the allele, is less common, with few reports of this phenomenon [5, 6].

Stutter ratio (*SR*) is described in terms of the ratio of stutter height to allele height. It is quantified as:

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

where O_{a-1} refers to the height of the stutter peak and O_a, the height of the parent allele peak.

The variability in *SR*, or more often the maximum observed, is routinely measured by individual laboratories as part of an internal validation of a new multiplex or analysis platform. Previous work has investigated the longest uninterrupted sequence of core repeats within an allele (*LUS*) as a predictor of *SR* [1, 4]. *SR* has been shown to be linearly related to *LUS* for tetranucleotide repeat loci [4, 7]. It explains approximately 61% of the variation in back stutter ratio in NGM SElect™ [7]. There is a small, but significant, difference in this relationship between different loci (a locus effect).

Initial thinking emphasised the importance of characterising stutter to facilitate decision thresholds to differentiate between true alleles and artefacts. Such decisions are trivial for single source samples but may be quite problematic when analysing mixed DNA profiles where the DNA from the minor contributor is of a similar height to the stutter peaks. As forensic DNA typing technologies advance, more low level and mixed DNA profiles are obtained [8].

In recent times there is a move away from heuristic, threshold-based interpretation strategies, towards continuous interpretation strategies. The introduction of continuous (probabilistic) methods has the potential to reduce error rates in decision making, to increase the efficiency of forensic laboratories, and improve the consistency and transparency of the reported results [9-12]. They remove the requirement for some or all thresholds and increase objectivity of interpretation. Continuous based methods are software-based solutions because of their complexity, and there is a need for the development of models that underpin the theory behind these methods. In this paper we investigate the occurrence of forward stutter in four different commercial STR multiplexes.

Accepted Manuscript

Materials and Method

Single source DNA profiles were analysed using four commercial STR profiling kits. The detail of the extraction technology, profiling kit and number of profiles analysed are provided in Table 1.

Table 1: Number of profiles analysed for four commercial STR profiling kits

Profiling kit	*N*	Sample type	Number of PCR cycles
NGM SElect™ (Life Technologies, Carlsbad, CA)	290	Saliva on FTA® Elute card	29
Identifiler™ (Life Technologies)	330	Saliva on FTA® Elute card	28
PowerPlex® 21 (Promega Corp, Madison, WI)	177	Blood and saliva on FTA® classic card	30
GlobalFiler™ (Life Technologies)	344	Saliva on FTA® classic card	29

DNA was recovered off the FTA® Elute card (Whatman, Maidstone, England) for samples amplified using NGM SElect™ and Identifiler™ using a simple automated elute method [13]. The saliva samples amplified using PowerPlex® 21 were extracted using Promega's DNA IQ™ extraction (Promega Corp, Madison, WI) and blood samples using ChargeSwitch® Forensic DNA purification kit (Invitrogen, Carlsbad, CA). The samples amplified using GlobalFiler™ were extracted using DNA IQ™.

Prior to amplification all samples were quantified using Applied Biosystems' Quantifiler™ (Life Technologies, Carlsbad, CA) according to the manufacturer's instructions. A target of 1 ng was amplified using NGM SElect™ and Identifiler™, 0.5 ng for PowerPlex® 21 and 0.4 ng for GlobalFiler™ following manufacturer's instructions in a 9700 silver block thermal cycler. Amplified products were separated on an Applied Biosystems' 3130*xl* Genetic Analyser (Life Technologies, Carlsbad, CA) and data was analysed using Applied Biosystems' GeneMapper™ *ID* version 3.2.1 (Life Technologies, Carlsbad, CA) using a 30 rfu analytical threshold. GlobalFiler™ data was analysed using GeneMapper™ *ID-X* version 1.4 also using an analytical threshold of 30 rfu.

The analytical threshold of 30 rfu used for data analysis is lower than that used normally for casework to lower the bias towards alleles more likely to stutter by the omission of small, but real, stutter peaks. It is still sufficiently clear of baseline noise to avoid false positives. In addition, only samples with parent allele heights greater than or equal to 500 rfu were selected. All data where the parent alleles were greater than 7000 rfu were removed from the data set to avoid saturation effects. When the camera on a capillary electrophoresis instrument reaches saturation (typically 7000 rfu for a 3130) the true quantity of DNA is no longer accurately represented and peak heights are lower than expected. Any alleles where the forward stutter peak fell into a back stutter position of the heterozygote allele at that locus were also removed from the dataset.

The forward stutter ratio, *FS*, was calculated as

$$FS = \frac{O_{a+1}}{O_a}$$

where O_{a+1} refers to the observed height of the forward stutter peak and O_a the parent allele peak height as before. Exploratory data analysis was undertaken to explore any possible relationship between O_{a+1} and *FS* and a number of potential explanatory variables including parent allele

Accepted Manuscript

height (O_a), back stutter allele height (O_{a-1}), back stutter ratio (*SR*), locus and *LUS*. *LUS* is defined as the longest stretch of basic repeat motifs within the allele. *LUS* values were obtained from the short tandem repeat DNA internet database (STRBase www.cstl.nist.gov/biotech/strbase) [14]. The average *LUS* value across the reported variants observed was taken where multiple values for *LUS* were available.

Results

The number of forward stutter observations and the average *FS* for each locus for each of the four multiplexes are given in Table 2. NGM SElect™ exhibited the most forward stutter >30 rfu, 10.3% of the total number of alleles analysed, followed by GlobalFiler (9.5%), PowerPlex® 21 (8.8%) and Identifiler™ (3.3%). A total of 1007 of the combined 1646 observed forward stutter peaks were less than 50 rfu in height (61%). 50 rfu is a common analytical threshold for forensic laboratories and hence these peaks would not be observed in normal casework. The one marker D22S1045 accounted for 22.7% of all the forward stutter peaks observed in the NGM SElect™ dataset and 34.2% in the GlobalFiler™ dataset. D22S1045 is a trinucleotide repeat and therefore is known to stutter more often [15]. As expected the pentanucleotide repeats within the PowerPlex® 21 multiplex, Penta D and Penta E, did not forward stutter significantly.

Exploratory data analysis suggested little or no effect of parent allele height (Oa), back stutter allele height (Oa-1), back stutter ratio (SR), locus and LUS for all loci in the four multiplexes except the single trinucleotide repeat locus, D22S1045. A series of exploratory plots for the NGM SElect™ dataset (excluding D22S1045) for forward stutter peaks above the analytical threshold (347 from 4190 possible values) are provided in Figure 1. As there is no observed effect of parent allele height, the heights of forward stutter peaks (Oa+1) are plotted and not forward stutter ratio (FS). Data where no forward or back stutter peaks were observed above the analytical threshold have been added to the dataset at 15 rfu (half the analytical threshold). Within the exploratory plots the missing data were jittered vertically between 0 and 15 rfu to give an indication of the number of missing data.

Accepted Manuscript

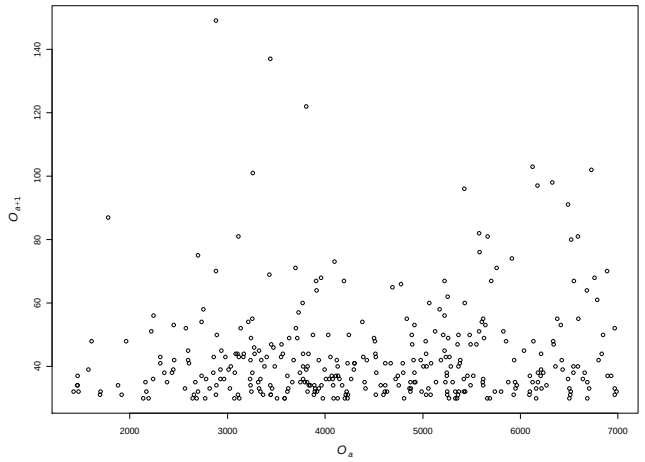
Table 2: The number of forward stutter observations and the average *FS* for each locus per kit

Locus	Repeat length	NGM SElect™		Identifiler™		PowerPlex® 21		GlobalFiler™	
		# Obs	Average *FS*	# Obs	Average *FS*	# Obs	Average *FS*	# Obs	Average *FS*
CSF1PO	4	-	-	39	1.4%	3	1.1%	12	1.1%
D10S1248	4	23	1.2%	-	-	-	-	21	1.1%
D12S391	4	6	1.6%	9	1.0%	-	-	17	1.9%
D13S317	4	-	-	32	1.0%	24	0.9%	25	1.3%
D16S539	4	28	0.8%	38	1.1%	13	1.3%	30	1.0%
D18S51	4	54	1.3%	24	1.1%	13	2.5%	34	1.3%
D19S433	4	8	1.5%	-	-	3	1.0%	3	3.2%
D1S1656	4	35	1.2%	27	1.2%	-	-	51	1.2%
D21S11	4	30	0.8%	43	1.1%	14	1.5%	58	1.1%
D2S1338	4	2	0.8%	-	-	1	0.9%	10	1.5%
D2S441	4	44	1.1%	-	-	-	-	30	1.9%
D3S1358	4	25	1.0%	9	1.0%	7	1.7%	29	1.0%
D5S818	4	-	-	33	1.1%	8	0.9%	32	1.1%
D6S1043	4	-	-	0	0.0%	-	-	-	-
D7S820	4	-	-	26	1.0%	0	0.0%	3	0.7%
D8S1179	4	60	1.0%	-	-	23	0.9%	39	1.3%
DYS391	4	-	-	-	-	-	-	10	1.3%
FGA	4	4	1.5%	16	1.1%	11	1.9%	7	1.0%
SE33	4	19	1.3%	-	-	-	-	41	1.8%
TH01	4	4	0.9%	-	-	8	1.1%	0	0.0%
TPOX	4	-	-	-	-	10	1.2%	0	0.0%
vWA	4	5	1.2%	16	1.3%	26	1.2%	22	1.8%
D22S1045	3	102	3.3%	-	1.4%	-	1.1%	247	3.2%
PentaD	5	-	-	0	0	-	-	-	-
PentaE	5	-	-	0	0	-	-	-	-
Total		271		281		124		331	

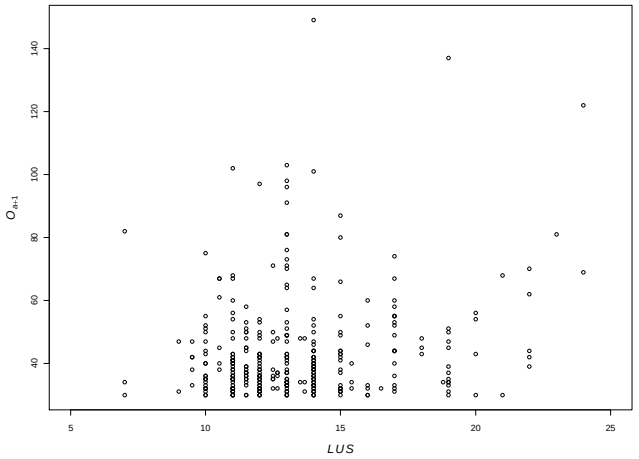
Accepted Manuscript

Figure 1: Exploratory data analysis for the NGM SElect™ (excluding D22S1045) dataset showing forward stutter height (O_{a+1}) versus parent height (O_a, pane A), *LUS* (pane B), marker (pane C) and back stutter height (O_{a-1}, pane D).

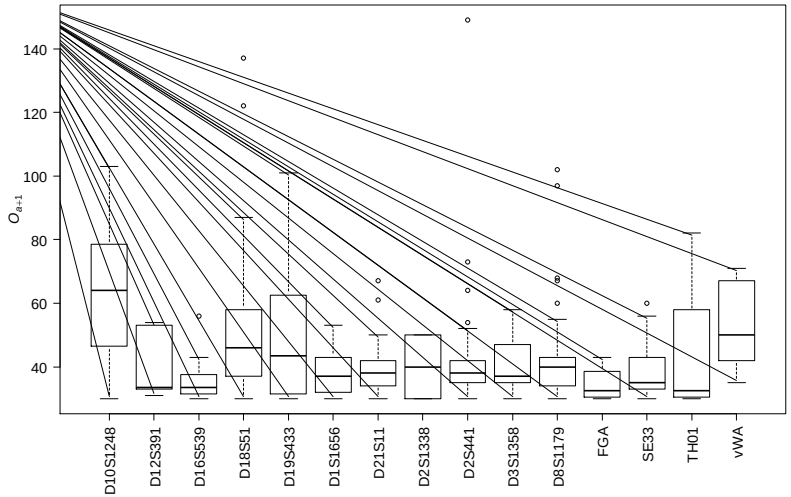
Pane A



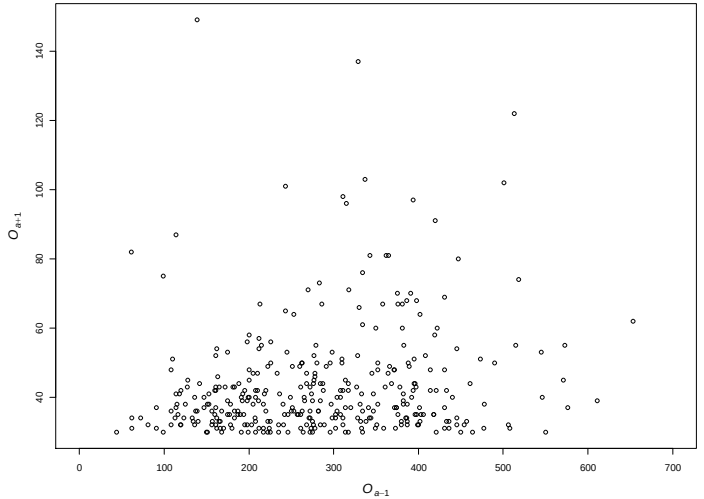
Pane B



Pane C



Pane D

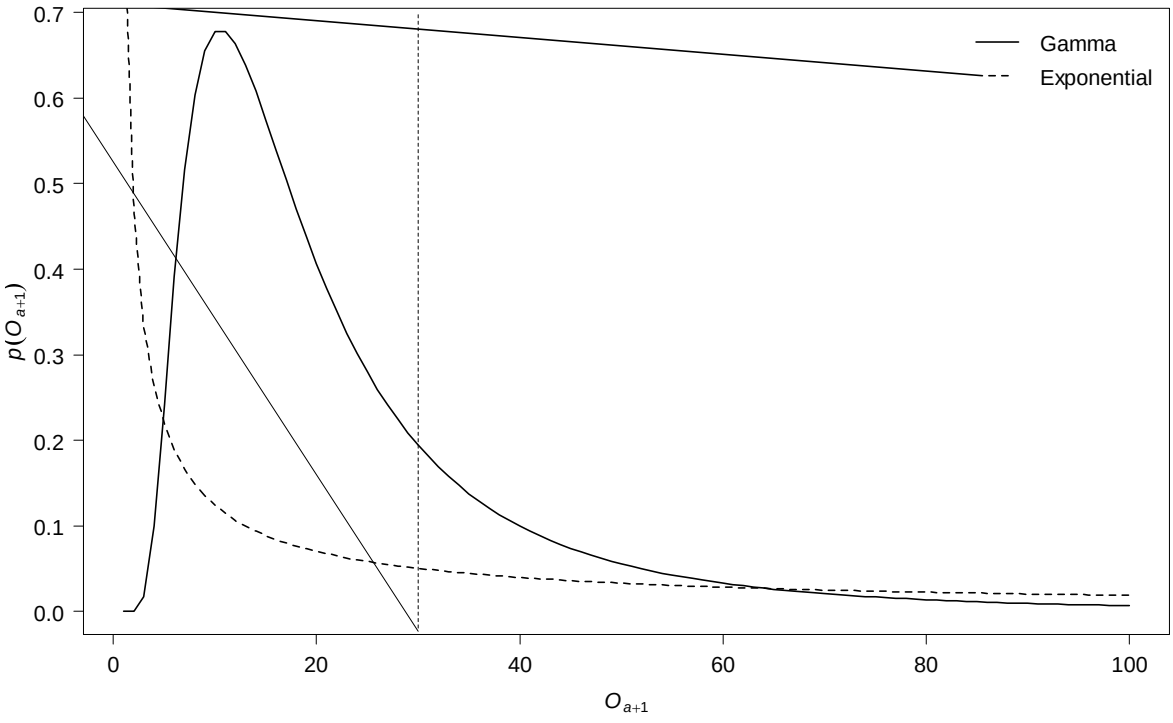


Given the absence of any significant explanatory variables for all tetra and pentanucleotide repeats we fitted a model with no explanatory variables. The probability density of $\log(O_{a+1})$ was modelled as both a gamma distribution described by two parameters; shape, α, and rate, κ, where $\alpha, \kappa > 0$, and as an exponential distribution with the one parameter lambda (λ). The exponential distribution is a special case of the gamma distribution where $\exp(\lambda) = \gamma(\alpha = 1, \kappa = \lambda)$ and was therefore tried for simplicity.

The package `rjags` was used to fit the models for each multiplex in R [16]. JAGS is variant of BUGS [17], a statistical package that allows the user to fit Bayesian models using Markov Chain Monte Carlo (MCMC) techniques. The `dinterval` in JAGS was used to cope with the left censored data, that is forward stutter peaks below the analytical threshold.

A plot of the fitted gamma and exponential probability density functions for the NGM SElect™ tetranucleotide markers is given in Figure 2. This shows the probability density of a forward stutter peak being certain height given that the parent allele, *a*, is present for both distributions. The gamma distribution puts most of its density below the detection threshold of 30 rfu, with a mode around 12 rfu. Under the gamma model 95% of forward stutters would be between approximately 4 rfu and 45 rfu. By contrast, the exponential distribution has its mode at zero, and would place 95% of forward stutters between 1 rfu and 7 rfu. Perhaps most telling, however, is the fact that the gamma model would have approximately 7.6% of forward stutters above the detection threshold of rfu, compared to 0.2% under the exponential model. This highlights the utility of the gamma over the exponential model given that our NGM SElect™ had 10.3% of its observations showing detectable forward stutter.

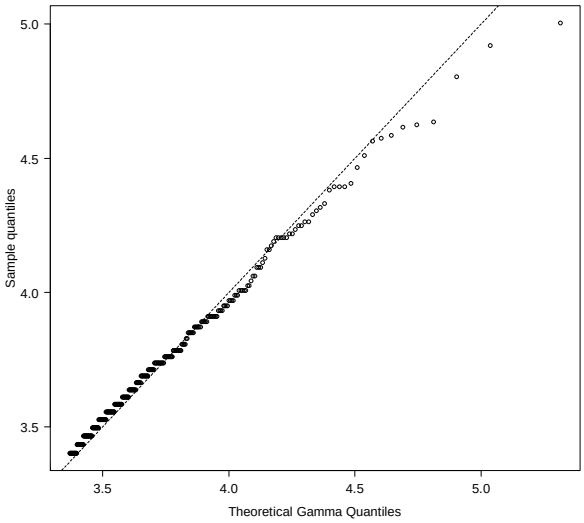
Figure 2: Fitted gamma and exponential probability density functions for the NGM SElect™ tetranucleotide markers



d Manuscript

Figure 3: A quantile-quantile plot of the gamma model (pane A) and exponential model (pane B) for the tetranucleotide repeat markers from the NGM SElect™ dataset

Pane A



Pane B

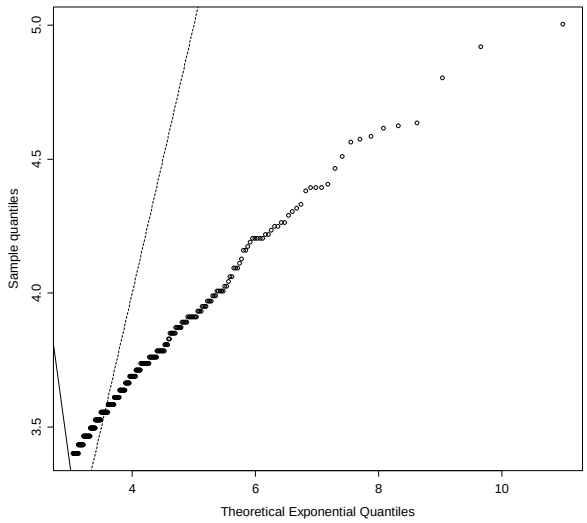


Table 3: Summary of estimates for lambda for each of the four datasets, excluding D22S1045

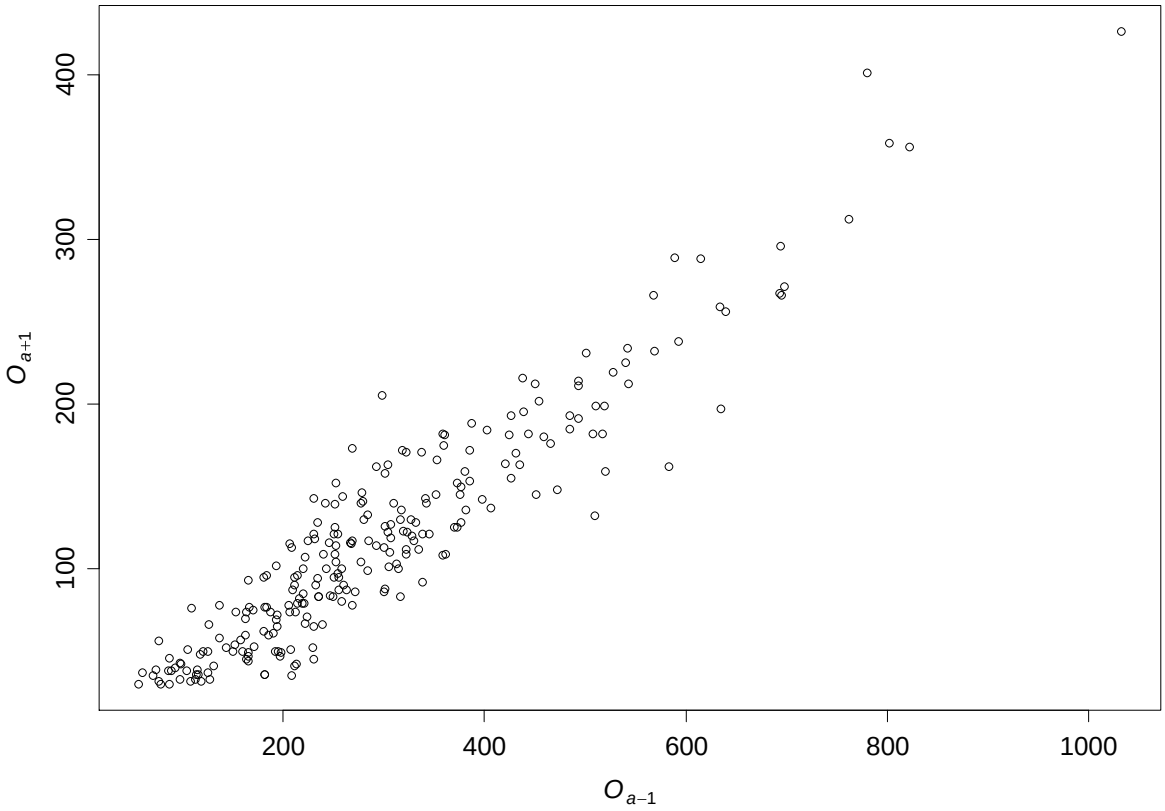
Dataset	Shape, α	95% credible interval	Rate, κ	95% credible interval
NGM SElect™	17.49	(14.30, 20.95)	6.99	(5.89, 8.18)
Identifiler™	11.43	(7.83, 13.01)	5.38	(4.03, 5.97)
PowerPlex® 21	18.46	(15.46, 22.11)	7.20	(6.13, 8.42)
GlobalFiler™	9.54	(8.24, 11.36)	4.37	(3.89, 5.04)

A quantile-quantile (Q-Q) plot of observed forward stutter heights versus theoretical quantiles for each distribution is presented in Figure 3 for the NGM SElect™ tetranucleotide dataset. The Q-Q plots are drawn so that only the quantiles above the detection threshold are used. The Q-Q plot suggests that assumption of a gamma distribution is acceptable but not the assumption of an exponential. A summary of the gamma estimates for shape and rate for each dataset is provided in Table 3.

The trinucleotide repeat locus D22S1045 behaves differently to the tetranucleotide repeats. Exploratory data analysis indicated that the explanatory variable back stutter height (O_{a-1}) was the best for predicting forward stutter peak height (O_{a+1}). The R^2 value for the relationship $O_{a+1} \sim O_{a-1}$ for the NGM SElect™ dataset was 0.78 and for the GlobalFiler™ dataset 0.41. Contrary to expectations, this was superior to any of the other explanatory variables including *LUS*. *LUS* would have been our *a priori* candidate. R^2 for *LUS* against O_{a+1} was 0.38 for the NGM SElect™ and 0.17 for the GlobalFiler™ dataset. A plot of O_{a+1} versus O_{a-1} for the D22S1045 data from the NGM SElect™ dataset is provided in Figure 4.

Accepted Manu

Figure 4: A plot of O_{a+1} versus O_{a-1} for the D22S1045 data from the NGM SElect™ dataset and relationship



The probability of $\log(O_{a+1})$ for the locus D22S1045 was modelled as a normal distribution. The model specified that the logarithm of O_{a+1} is normally distributed for a given mean and variance:

$$\log(O_{a+1\_i}) \sim N(\mu_i, \sigma_i^2)$$

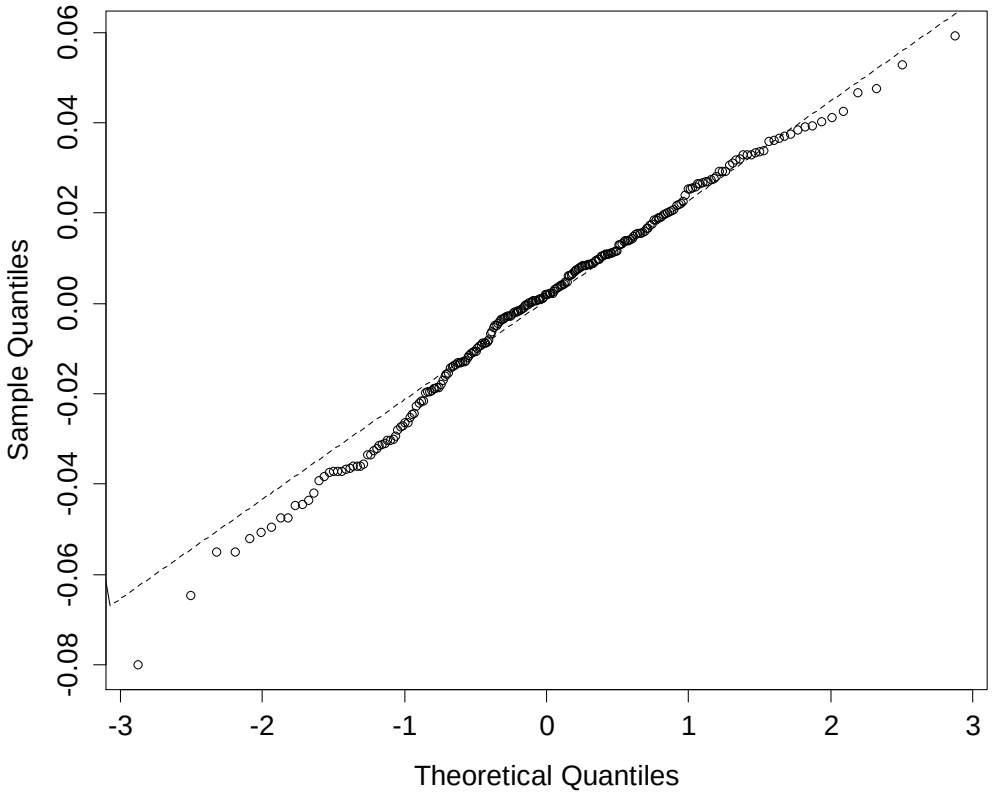
The parameters μ_i and σ_i^2 were modelled by:

$$\mu_i = \beta_0 + \beta_1 O_{a-1} \text{ and } \sigma_i^2 = \sigma^2$$

The package `rjags` was used to fit the model for D22S1045 from the NGM SElect ™ and GlobalFiler™ datasets. JAGS is a variant of BUGS [17], a statistical package that allows the user to fit Bayesian models using Markov Chain Monte Carlo (MCMC) techniques. A normal quantile-quantile (Q-Q) plot of the residuals from the model versus theoretical quantiles from a normal distribution is presented in Figure 5. The Q-Q plot suggests that assumption of normality is acceptable.

Accepted

Figure 5: A quantile-quantile plot of the log normal model for O_{a+1} at D22S1045



Discussion

Forward stutter is a relatively rare event. The majority of forward stutter observations in the datasets from the four different multiplexes were below the commonly applied laboratory analytical threshold of 50 rfu. Over one third of all forward stutters observed in the GlobalFiler™ dataset and one fifth in the NGM SElect™ dataset were from the one trinucleotide repeat marker, D22S1045.

Exploratory data analysis suggested explanatory variables previously used to predict back stutter height such as parent allele height (O_a), locus and the longest uninterrupted sequence (*LUS*) were not suitable for predicting the height of forward stutter peaks for all tetra and pentanucleotide repeats. The data fit a gamma distribution with no explanatory variables and inspection of the Q-Q plot indicated that the assumption of a gamma distribution was sustainable.

Back stutter allele height (O_{a-1}) was shown to be the best predictor of forward stutter height for the trinucleotide repeat D22S1045. A lognormal normal model was fitted to the data and a Q-Q plot demonstrated a good fit to the model. These models should not be extrapolated to single and dinucleotide repeat markers.

Accepted Manuscript

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We gratefully acknowledge the source of the PowerPlex® 21 stutter ratio data. These data were provided by the laboratories of Forensic Science South Australia and Australian Federal Police. We also thank Catherine McGovern and Stuart Cooper and three anonymous reviewers for their helpful comments that improved this manuscript.

The authors have declared no conflict of interest.

References

[1] Walsh, P. S., Fildes, N. J., Reynolds, R., *Nucleic Acids Res.* 1996, *24*, 2807-2812.

[2] Bacher, J., Schumm, J. W., *Profiles in DNA* 1998, *2*, 3-6.

[3] Hill, C. R., Duewer, D. L., Kline, M. C., Sprecher, C. J., McLaren, R. S., Rabbach, D. R., Krenke, B. E., Ensenberger, M. G., Fulmer, P. M., Storts, D. R., Butler, J. M., *Forensic Science International: Genetics* 2011, *5*, 269-275.

[4] Brookes, C., Bright, J.-A., Harbison, S., Buckleton, J., *Forensic Science International: Genetics* 2012, *6*, 58-63.

[5] Bright, J.-A., Huizing, E., Melia, L., Buckleton, J., *Forensic Science International: Genetics* 2011, *5*, 381-385.

[6] Gibb, A. J., Huell, A.-L., Simmons, M. C., Brown, R. M., *Science and Justice.* 2009, *49*, 24-31.

[7] Bright, J.-A., Taylor, D., Curran, J. M., Buckleton, J. S., *Forensic Science International: Genetics* 2013, *7*, 296-304.

[8] Carracedo, A., Schneider, P. M., Butler, J., Prinz, M., *Forensic Science International: Genetics* 2012, *6*, 677-678.

[9] Cowell, R. G., Lauritzen, S. L., Mortera, J., *Forensic Science International: Genetics Supplement Series* 2008, *1*, 640-642.

[10] Perlin, M. W., Legler, M. M., Spencer, C. E., Smith, J. L., Allan, W. P., Belrose, J. L., Duceman, B. W., *Journal of Forensic Sciences* 2011, *56*, 1430-1447.

[11] Taylor, D., Bright, J.-A., Buckleton, J. S., *Forensic Science International: Genetics* 2013, *7*, 516-528.

[12] Puch-Solis, R., Rodgers, L., Mazumder, A., Pope, S., Evett, I., Curran, J., Balding, D., *Forensic Science International: Genetics* 2013, *7*, 555-563.

[13] Parsons, L., Bright, J.-A., *Australian Journal of Forensic Sciences* 2012, *44*, 392-402.

[14] Ruitberg, C. M., Reeder, D. J., Butler, J. M., *Nucleic Acids Research* 2001, *29*, 320 - 322.

[15] Butler, J. M., in: Butler, J. M. (Ed.), *Advanced Topics in Forensic DNA Typing*, Academic Press, San Diego 2012, pp. 99-139.

[16] Plummer, M., 2012.

[17] Lunn, D. J., Spiegelhalter, D., Thomas, A., Best, N., *Statistics in Medicine* 2009, *28*, 3049-3082.

Accepted Manuscript