Estimating Stutter Rates for Y-STR Alleles

Mikkel Meyer Andersen[a], Jill Olofsson[b], Poul Svante Eriksen[c], Helle Smidt Mogensen[b], Niels Morling[b]

[a]Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark, mikl@math.aau.dk
[b]The Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark
[c]Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark

AALBORG UNIVERSITY

Abstract

Stutter peaks are artefacts that arise during PCR amplification of short tandem repeats. Stutter peaks are especially important in forensic case work when DNA mixtures are accounted. To analyse mixtures properly, good estimates of stutter rates must be available.

Aim of study

The aim of the study was (1) to estimate the stutter rates of the AmpFlSTR Yfiler kit, (2) to investigate whether stutter rates differ at the allelic level, and (3) to investigate the influence of the parental peak height on the stutter rate.

Data

Two 1.2 mm punches of FTA® cards with buccal samples from each of 360 persons were amplified in 10 µl reaction volume with AmpF*l*STR® Yfiler® kit with 27 cycles. PCR products were separated on an AB3130xl Genetic Analyzer and fragments analysed using GeneScan 3.7 and Genotyper 3.7 with a 5 RFU threshold. For each sample, the highest peak at each locus was taken as the parental peak if the height was between 50 and $7,000$ RFU. The heights of the parental and -1 repeat stutter peaks were further analysed.

Model

We estimated the stutter rates using weighted linear regression with intercept to allow for stutter rates varying with the parent peak height. The inverse peak heights were used as weights to incorporate that the variance increases with the signal strength. We estimated the stutter rate at the allelic level. Thus, for each locus and allele, the structure of the model is

$$s_i = \alpha p_i + \beta, \tag{1}$$

where s_i is the height of the i'th stutter peak and p_i the height of the corresponding parental peak. This gives regression coefficients, α and β, for each locus and allele. These have impact on the stutter rate, $\frac{s_i}{p_i}$.

Multiple regression model

In order to make an overall model per locus, i.e. taking all alleles into account, a weighted multiple linear regression model with intercept was made. This model has the form

$$s_i = \beta_0 + \beta_1 a_i + \beta_2 p_i + \beta_3 a_i p_i \tag{2}$$

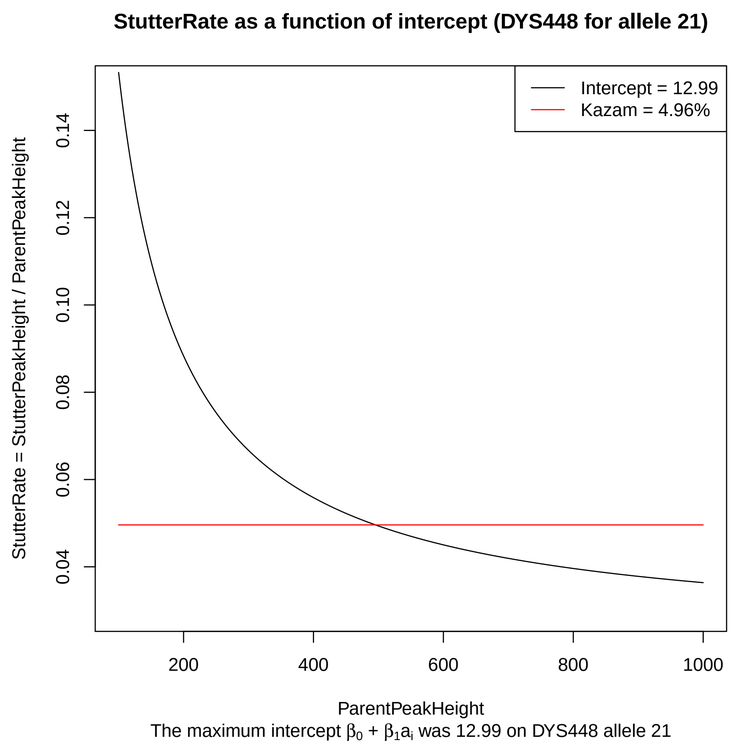
with the same notation as earlier with the addition that a_i is the allele. This gives regression coefficients, $\beta_0, \beta_1, \beta_2, \beta_3$, for each locus and allele, which both have impact on the stutter rate, $\frac{s_i}{p_i}$. Note, that the stutter rate has the form

$$\frac{s_i}{p_i} = \frac{\beta_0}{p_i} + \beta_1 \frac{a_i}{p_i} + \beta_2 + \beta_3 a_i \tag{3}$$

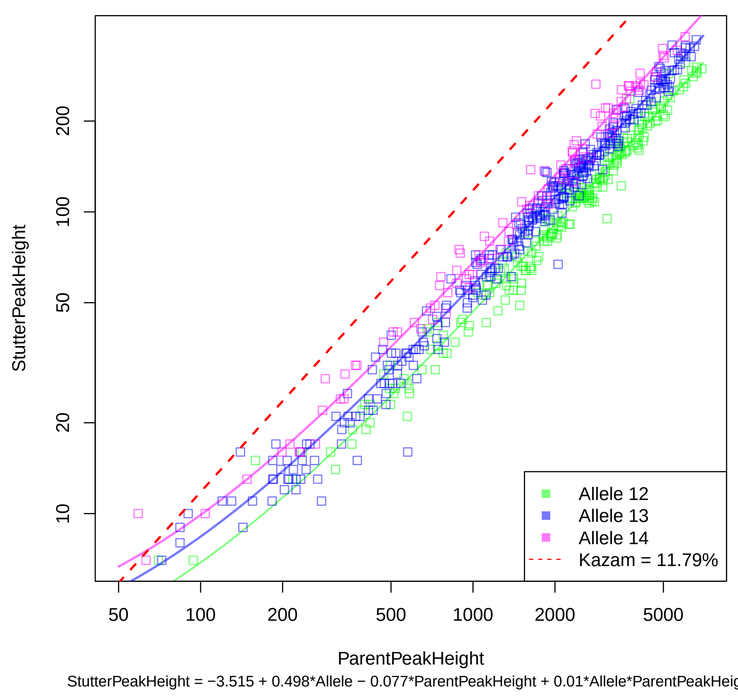
$$= p_i^{-1}(\beta_0 + \beta_1 a_i) + \beta_2 + \beta_3 a_i. \tag{4}$$

Intercept interpretation

From (1), it follows that $\frac{s_i}{p_i} = \alpha + \frac{\beta}{p_i}$ that results in the following interpretation when assuming $\alpha > 0$ and $p_i > 1$: For a positive intercept, the stutter rate decreases when p_i increases.



Multiple regression plot



Multiple regression of DYS389I

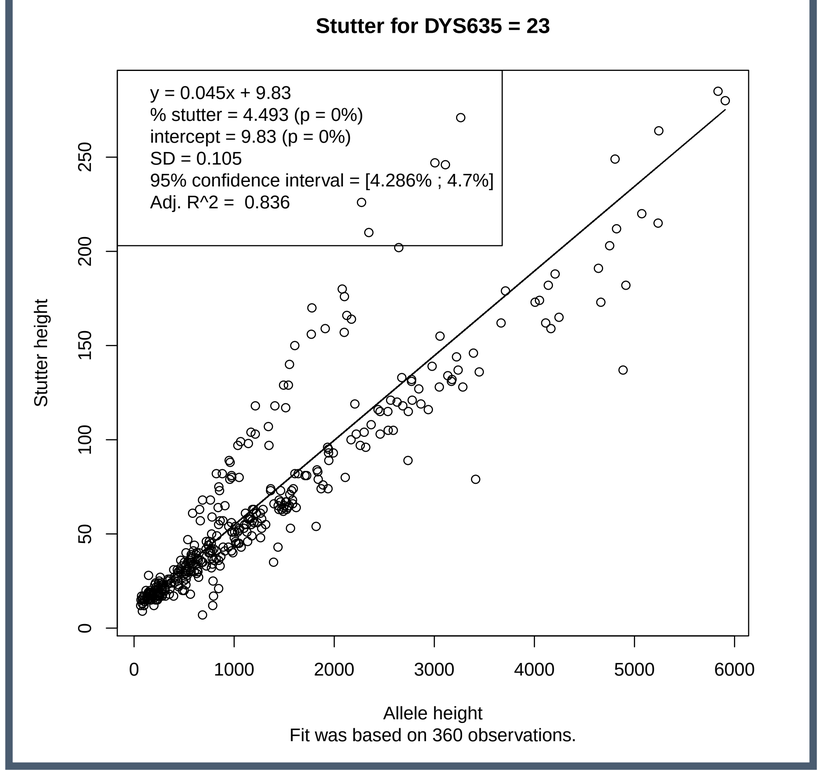
Stutter height and rate predictions for DYS389I by allele and parental heights using weighted multiple linear regression. Kazam refers to Applied Biosystems' recommended stutter filter.

	Parental peak height		
Allele	50	500	2000
12	2.8	23.6	93.2
	5.5 %	4.7 %	4.7 %
13	4.5	29.7	113.8
	9.0 %	5.9 %	5.7 %
14	6.2	35.8	134.3
	12.5 %	7.2 %	6.7 %
Kazam	5.9	59.0	235.8
	11.79 %	11.79 %	11.79 %

Stutter groupings

DYS635 yielded a poor fit, especially for allele 23, where two groups of stutter rates were identified. Sequencing of 14 samples using BigDye Termination v1.1 Cycle Sequencing Kit showed that 9 samples had sequences with longest uninterrupted stretch (LUS) of 9 repetitive units and 5 samples had sequences with LUS equal to 13 repetitive units. The sequence variants were in accordance with the previously published sequences of DYS635 [1]. All samples with LUS 13 were in the group with high stutter rates and all samples with LUS 9 were in the group with lower stutter rates.

Stutter groupings plot



Results

1. Stutter rates differ at the allelic level, hence one stutter rate per locus is not optimal (stutter rates seem to increase with the number of repeats)
2. Applied Biosystems' recommended stutter filter rates, in general, seem to be rather high, which can cause problems when analysing DNA mixtures
3. Intercepts need to be included in the model
4. A weighted multiple linear regression model allowed us to predict stutter heights at almost all loci using allele and parental peak heights as explanatory variables
5. Intra-allelic differences in stutter rates exist, especially among DYS635 alleles that have complex structures with several repetitive sequences of varying lengths together with intervening sequences

References

[1] Gusmão L, Gonzalez-Neira A, Alves C, *et. al*. Chimpanzee homologous of human Y specific STRs. A comparative study and a proposal for nomenclature. Forensic Sci Int. 2002; 126: 129-136.