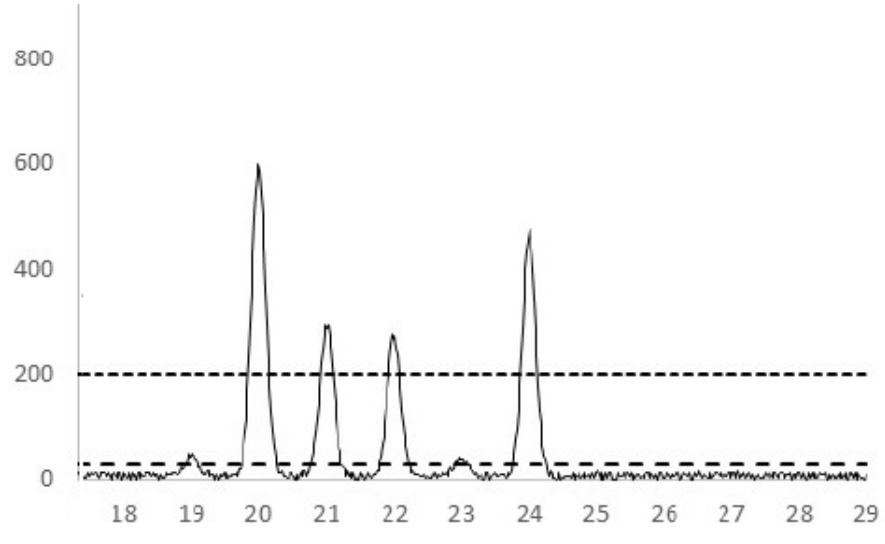
Peak Height Modeling

Continuous Model

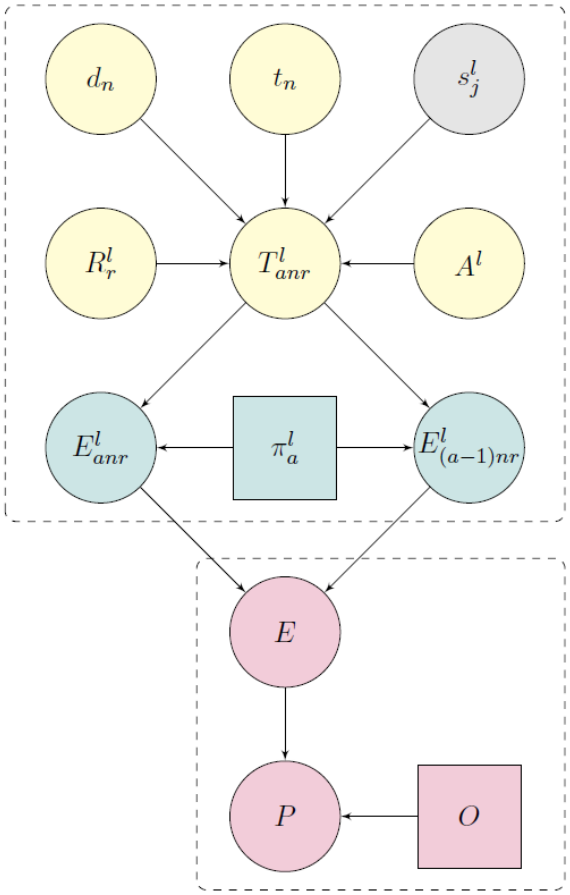
The key point of a fully continuous model is that it considers peak heights as a continuous variable.



Donor 1	Donor 2	Weights (Qualitative)	Weights (Quantitative)
20, 21	22, 24	1	0.05
20, 22	21, 24	1	0.05
20, 24	21, 22	1	0.75
21, 22	20, 24	1	0.05
21, 24	20, 22	1	0.05
22, 24	20, 21	1	0.05

Continuous Model Network

The continuous model we are going to discuss consists of several elements:



Adapted from: The interpretation of single source and mixed DNA profiles (Taylor et al., 2013).

Peak Height Modeling

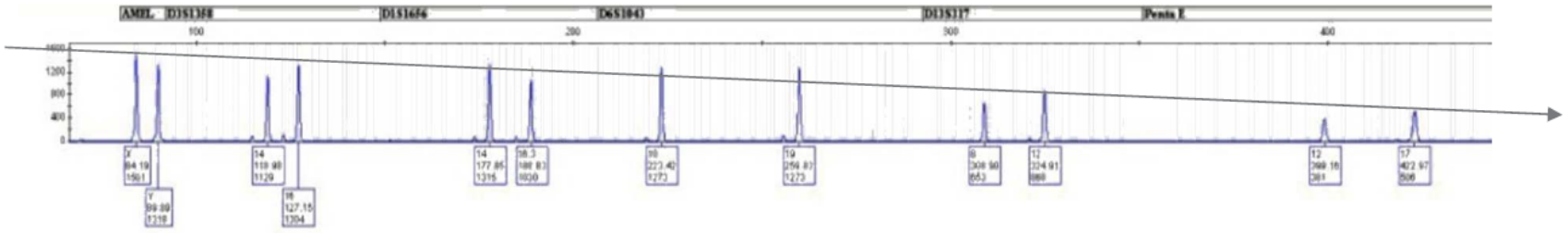
Peak heights can be modeled by defining the *total allelic product* (TAP), which will be a function of

- the template amount t_n;
- a measure of degradation d_n;
- a locus-specific amplification efficiency A^l;
- a replicate multiplier R_r;
- and allele dosage X^l_{an}.

T^l_{arn} then describes the TAP of allele a at locus l, for replicate r from contributor n.

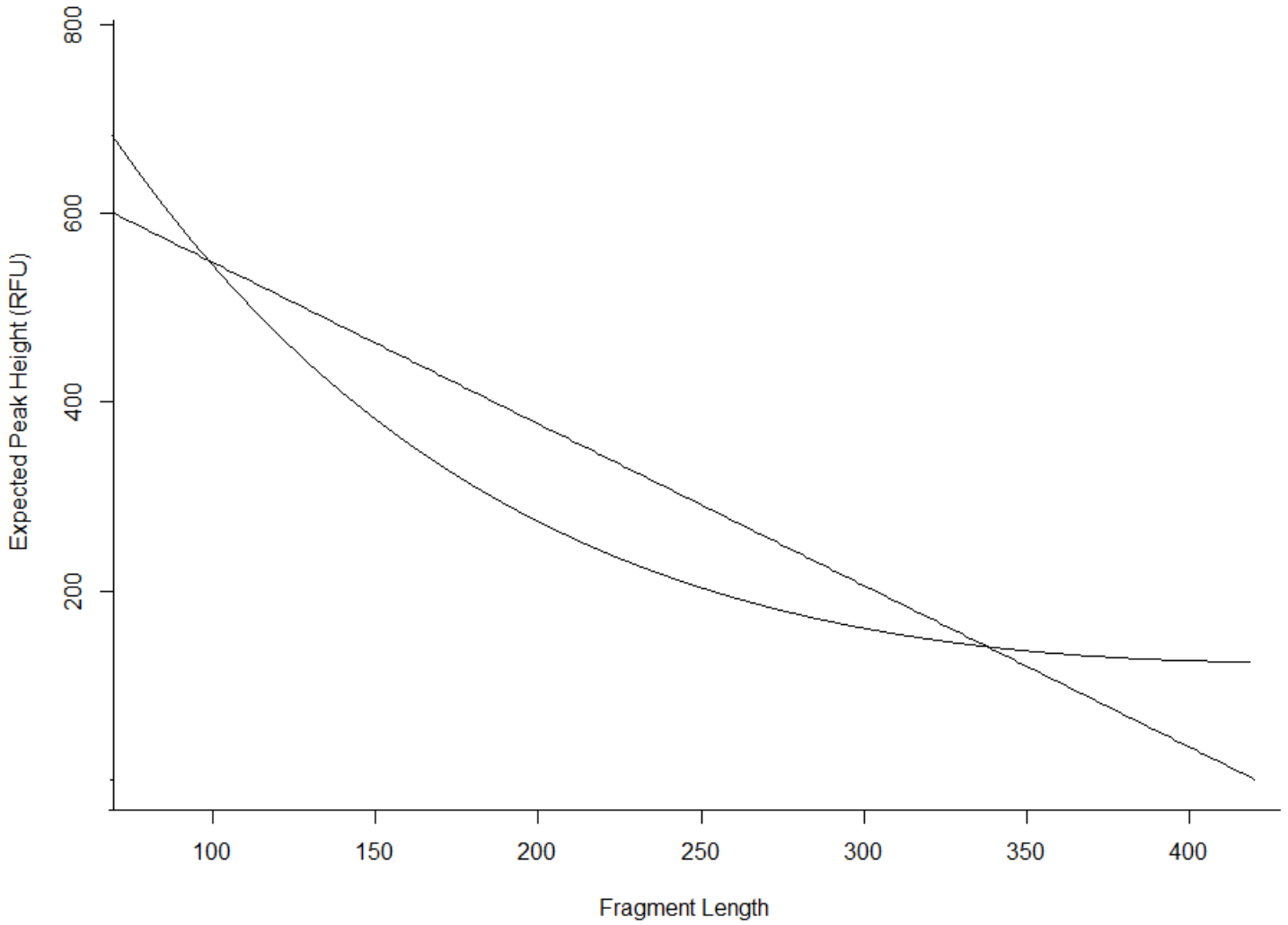
Modeling Degradation

A simple model for degradation would be a linear model, i.e. peak heights decline constantly with respect to molecular weight.

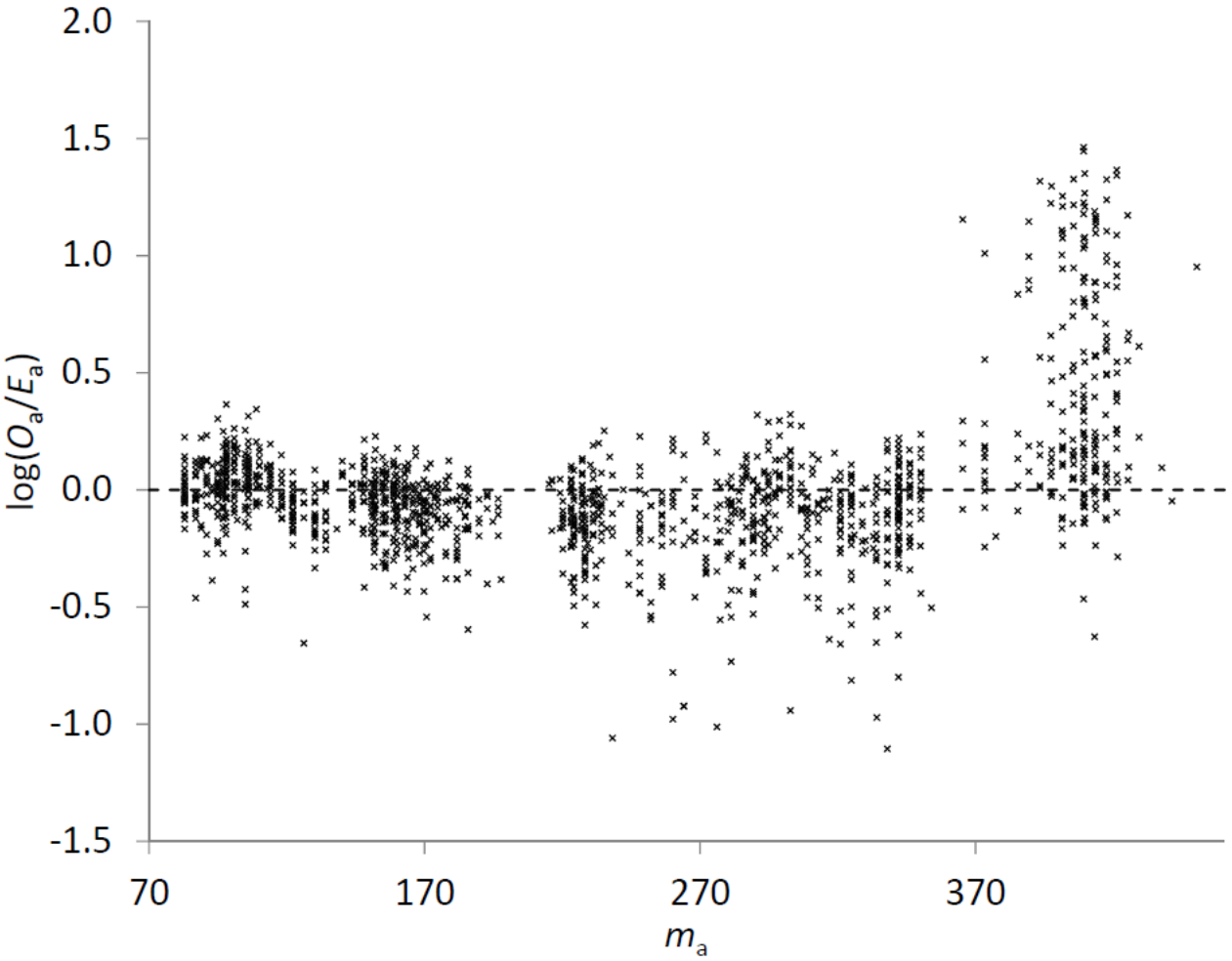


If we assume that the breakdown of a DNA strand is random with respect to location, an exponential model seems more reasonable.

Modeling Degradation

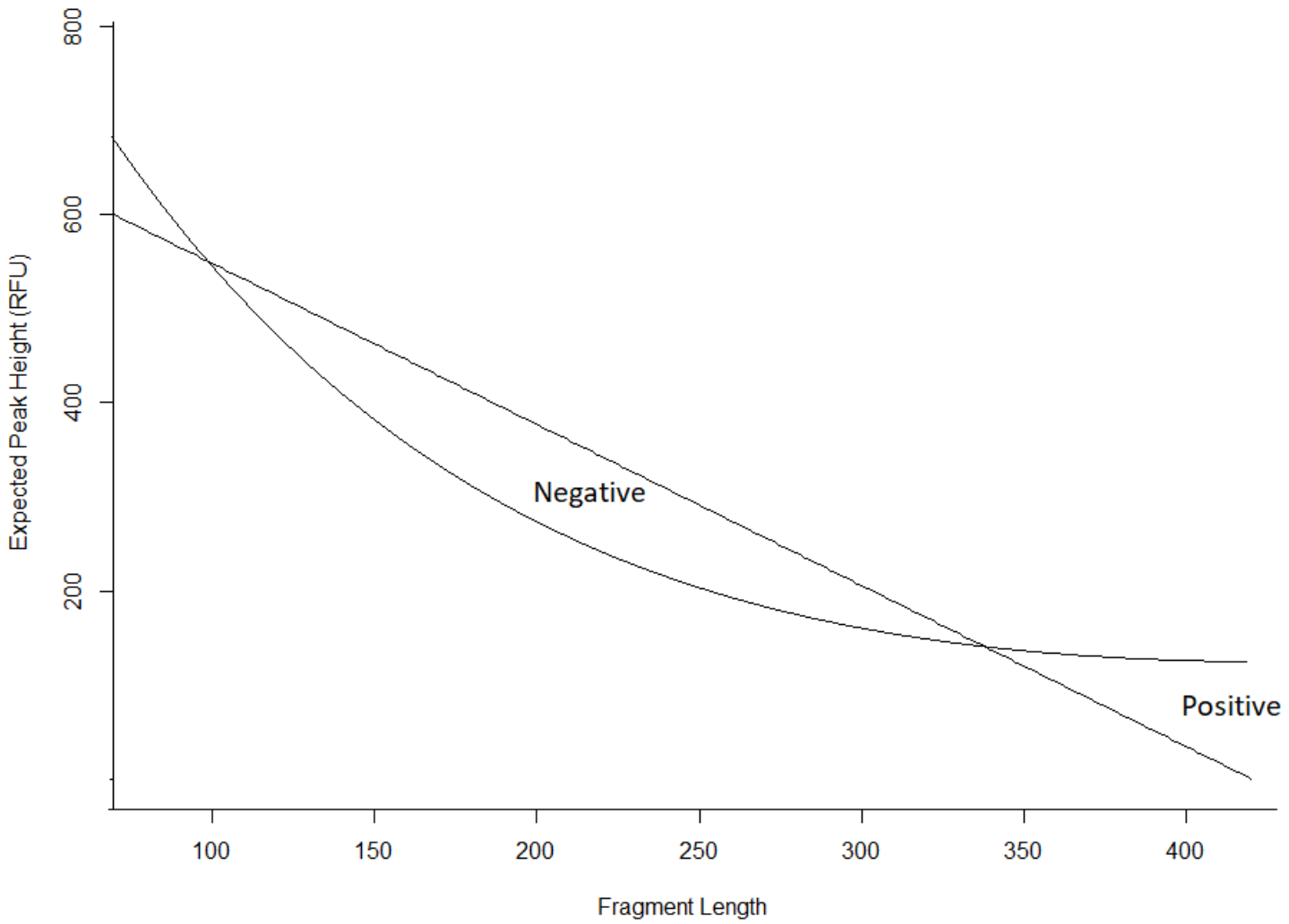


Modeling Degradation

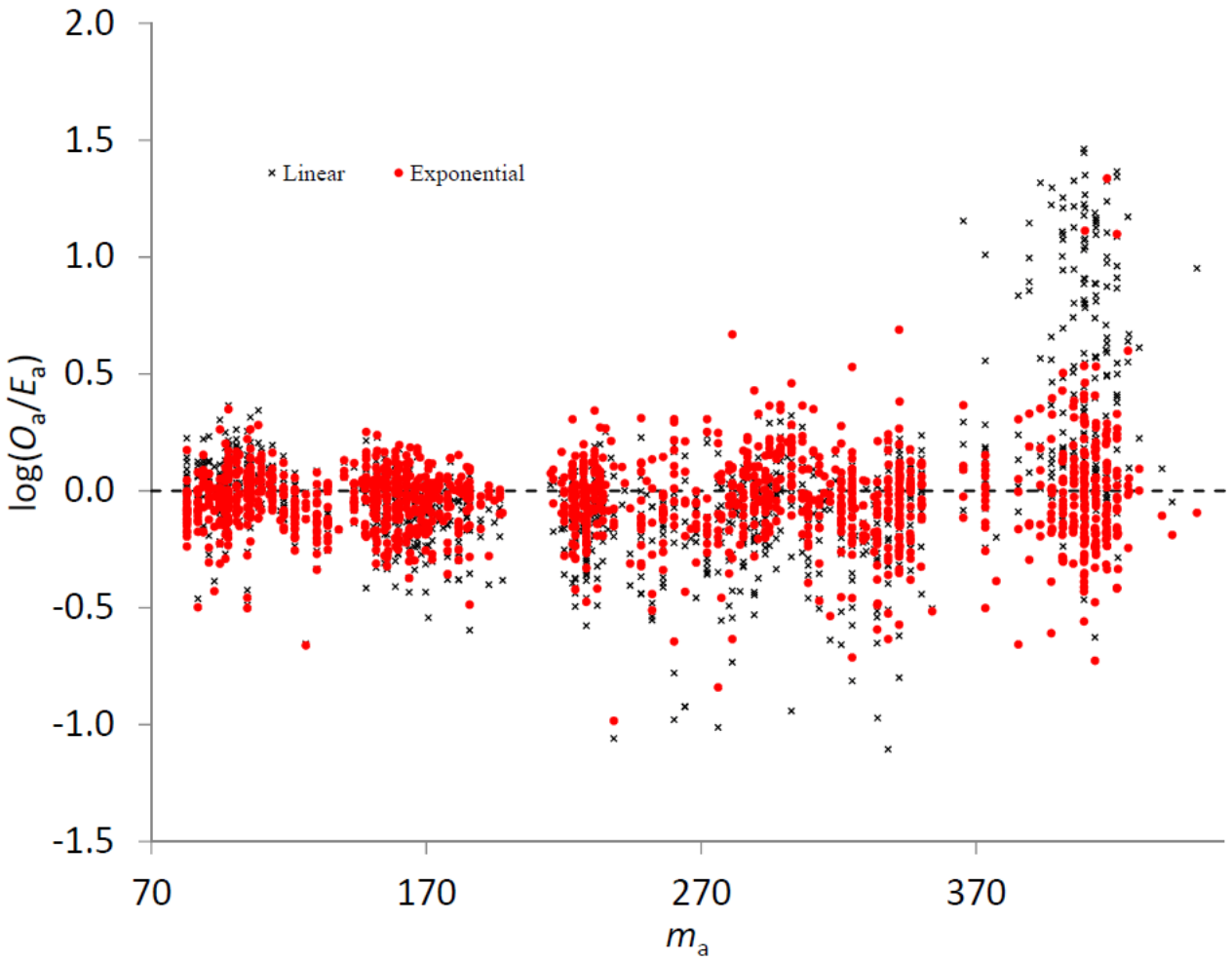


Source: Degradation of Forensic DNA Profiles (Bright et al., 2013).

Modeling Degradation



Modeling Degradation



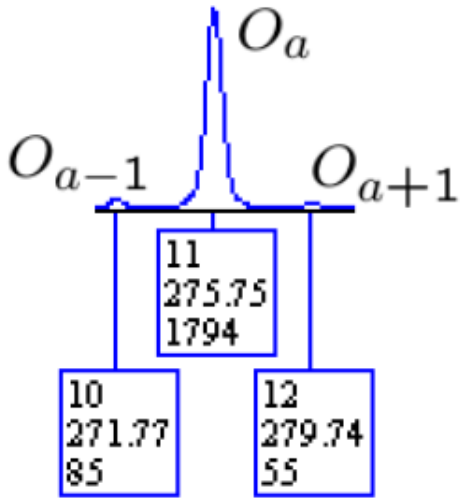
Source: Degradation of Forensic DNA Profiles (Bright et al., 2013).

TAP Modeling

Theoretically, the TAP models the peak heights, but in practice, we will observe slightly different values. This is because we haven't incorporated the concept of stutter yet.

If we allow for back stutter and forward stutter, we can write:

$$T_a = O_{a-1} + O_a + O_{a+1}.$$



Stutter Modeling

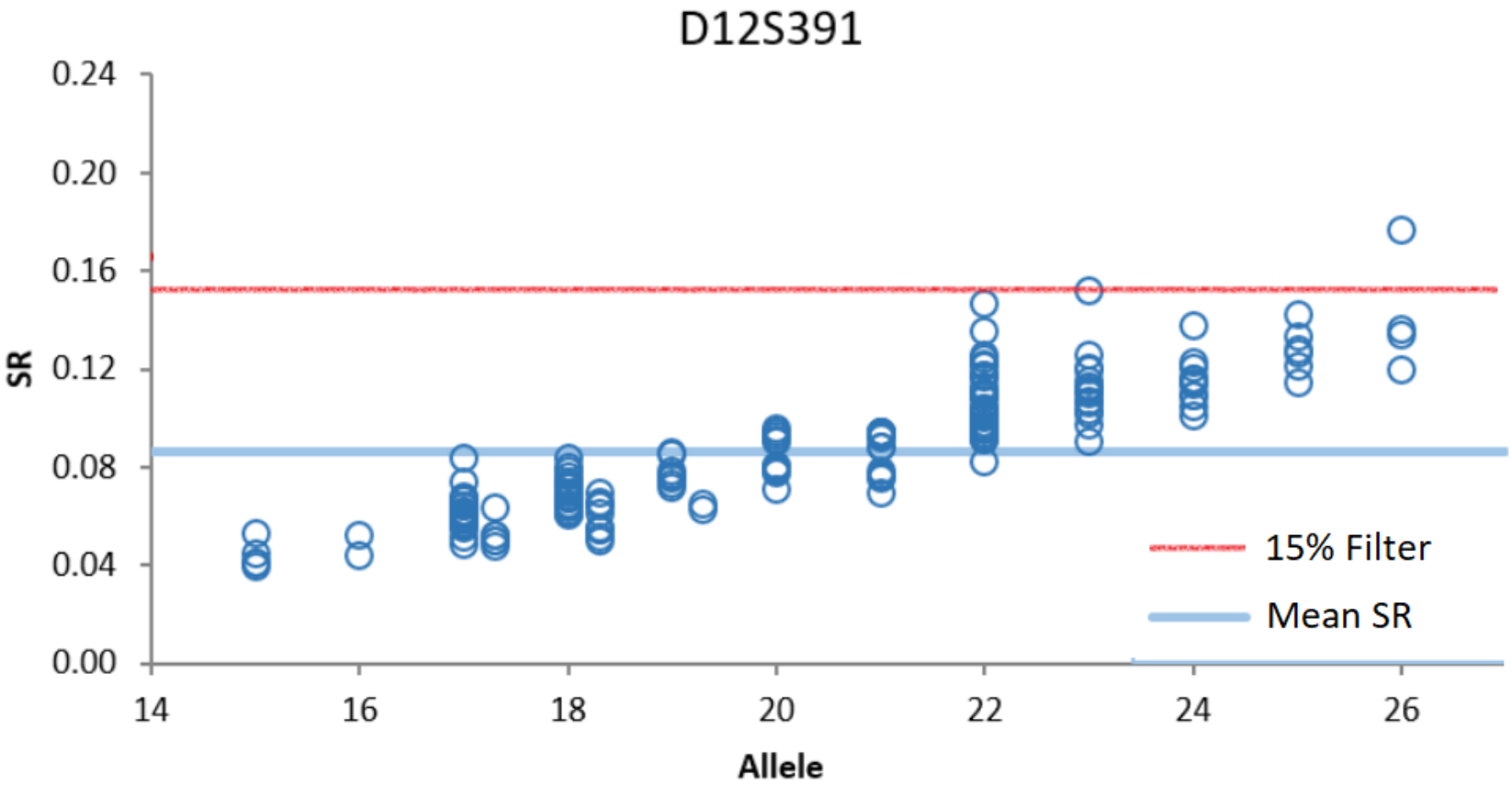
Stutter modeling becomes especially important in case of mixtures, when a true (minor) contributor's alleles are approximately the same height as stutter products from the major contributor.

Stutter is typically modeled by a stutter ratio (SR):

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

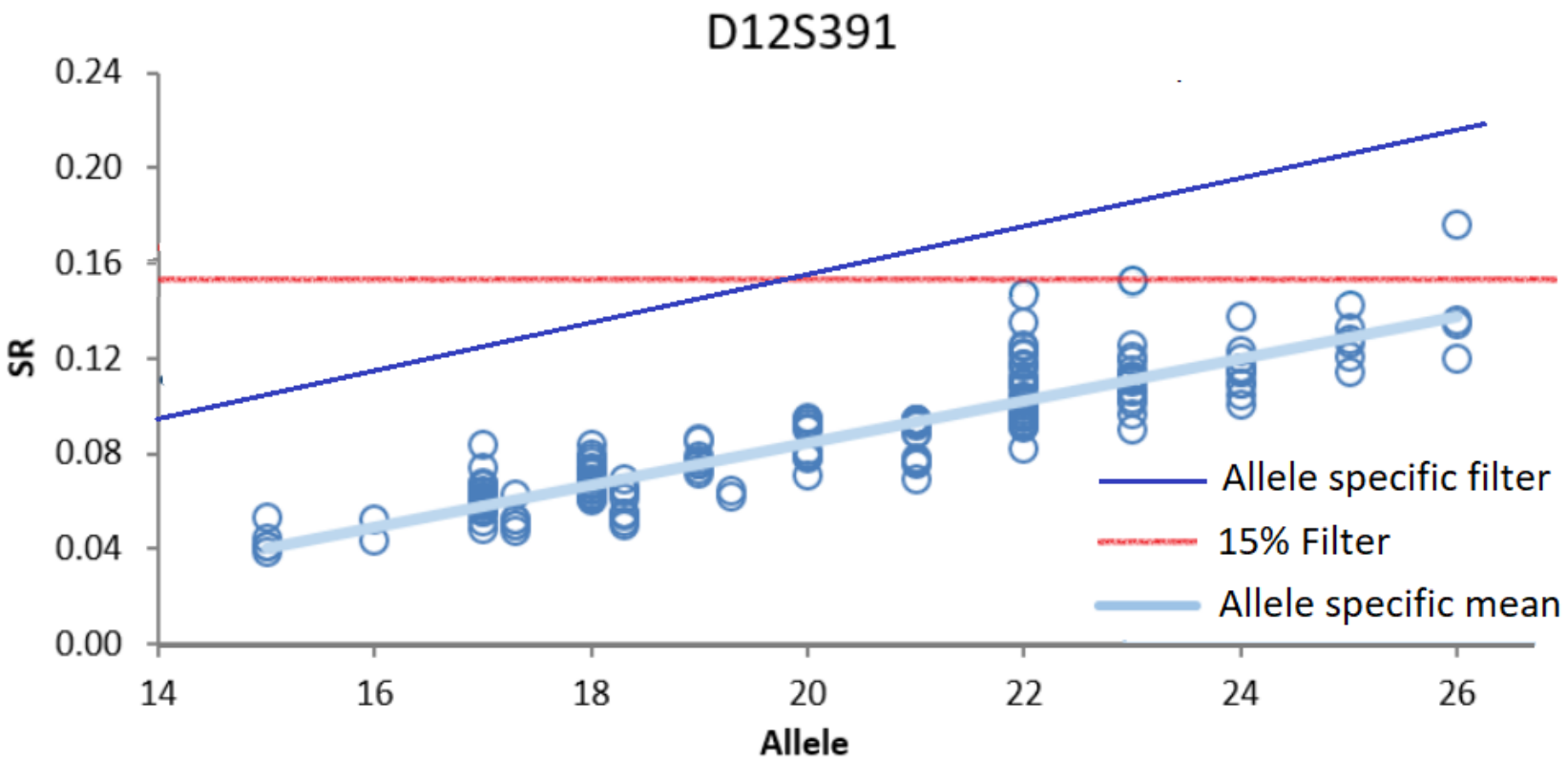
where O_{a-1} refers to the observed peak height of the back stutter of parent peak O_a.

Stutter Modeling - Locus Specific Thresholds



Source: Implementation and validation of an improved allele specific stutter filtering method for epg interpretation (Buckleton et al., 2017).

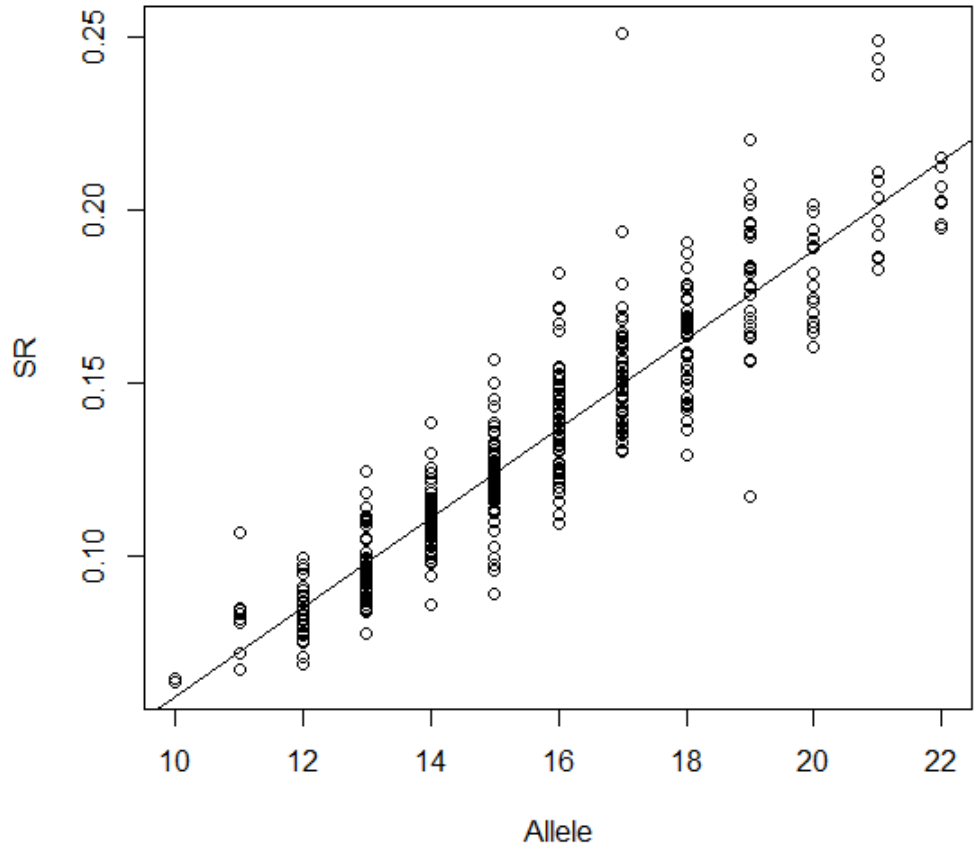
Stutter Modeling - Allele Specific Thresholds



Source: Implementation and validation of an improved allele specific stutter filtering method for epg interpretation (Buckleton et al., 2017).

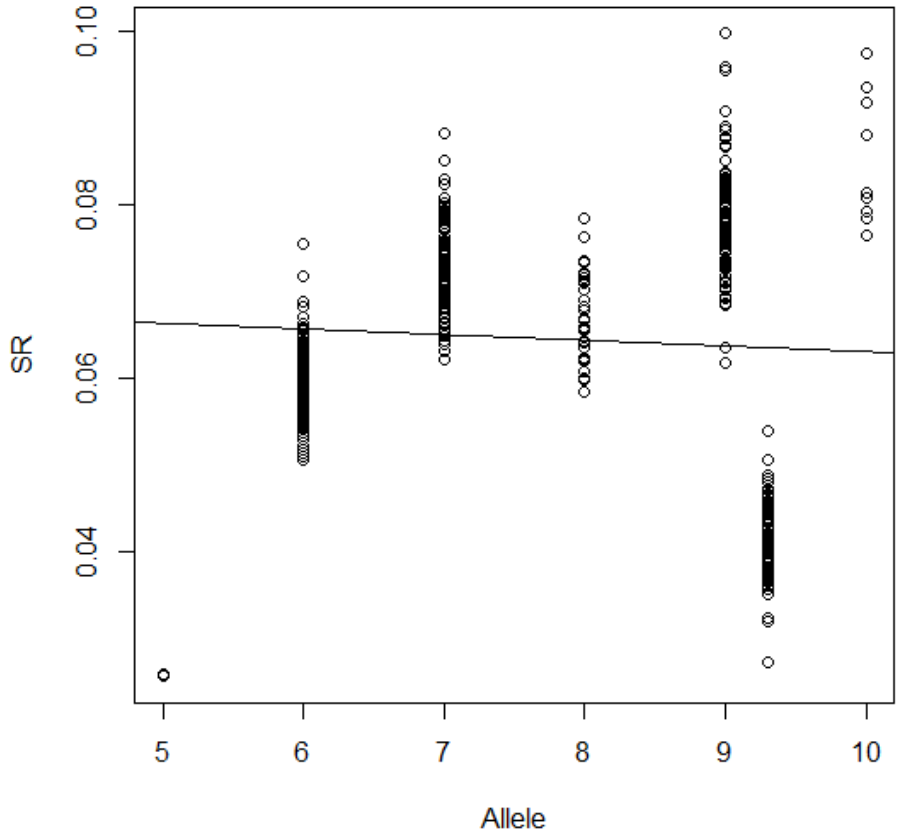
Stutter Modeling - Allele Model

The following figure shows locus D18S51 with a fitted model of $SR = 0.013a - 0.07$ ($R^2 = 85\%$).



Stutter Modeling - Allele Model

But this does not seem to work for all loci:



Locus TH01

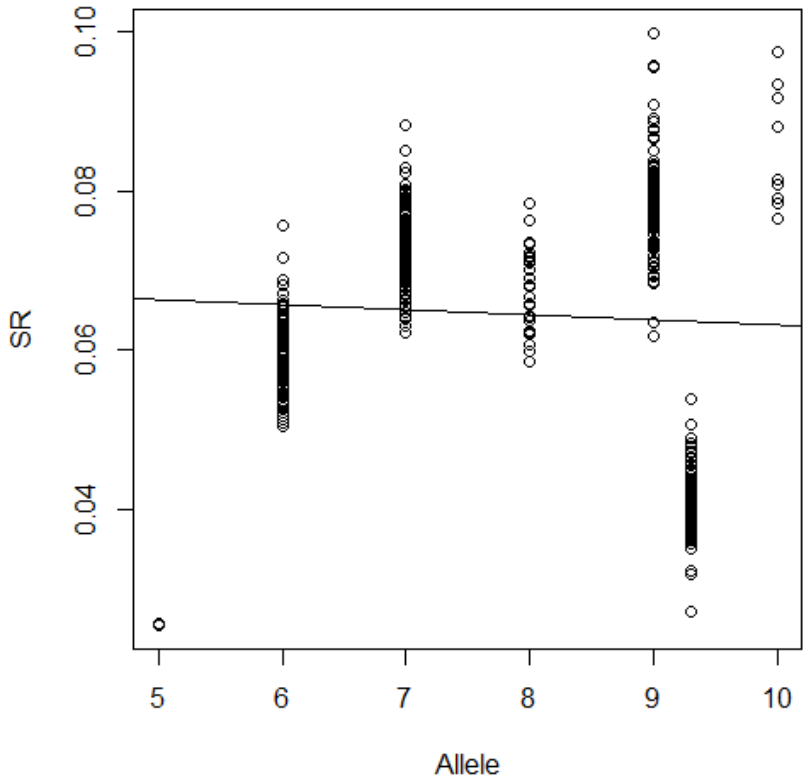
Stutter Modeling - LUS

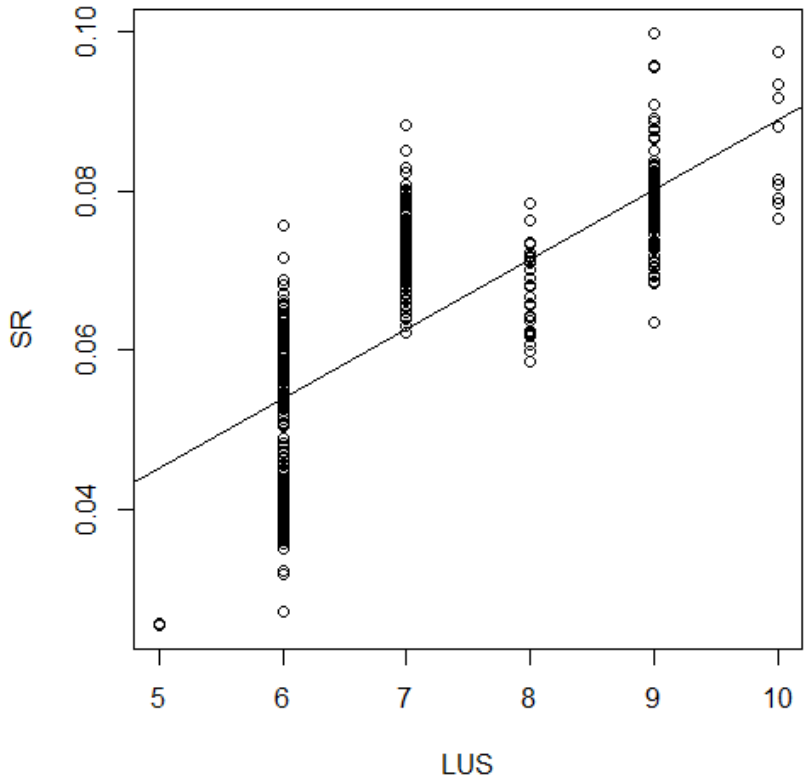
Such observations suggested that there exists a linear relationship between stutter ratio and the *longest uninterrupted stretch* (LUS).

Repeat motif	**Allele**	**LUS**
$[AATG]_6$	6	6
$[AATG]_7$	7	7
$[AATG]_8$	8	8
$[AATG]_9$	9	9
$[AATG]_6ATG[AATG]_3$	9.3	6

Common TH01 allele sequences.

Stutter Modeling - LUS Model





Locus TH01 allele vs. LUS

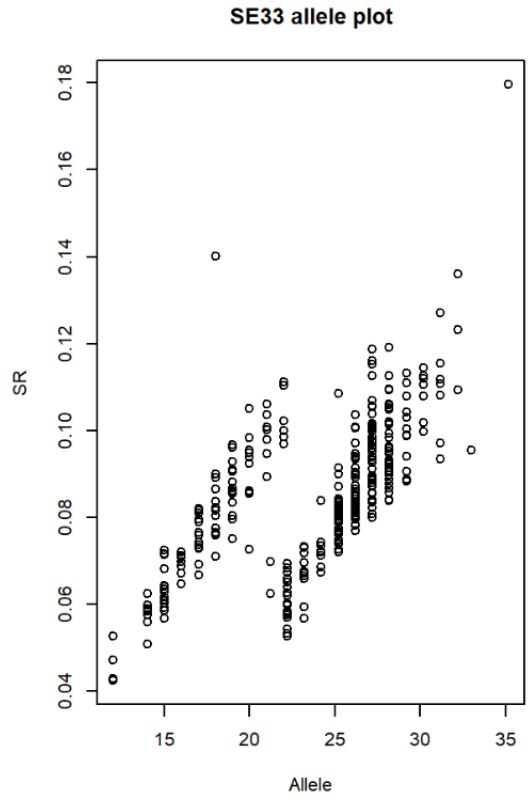
Stutter Modeling - AUS

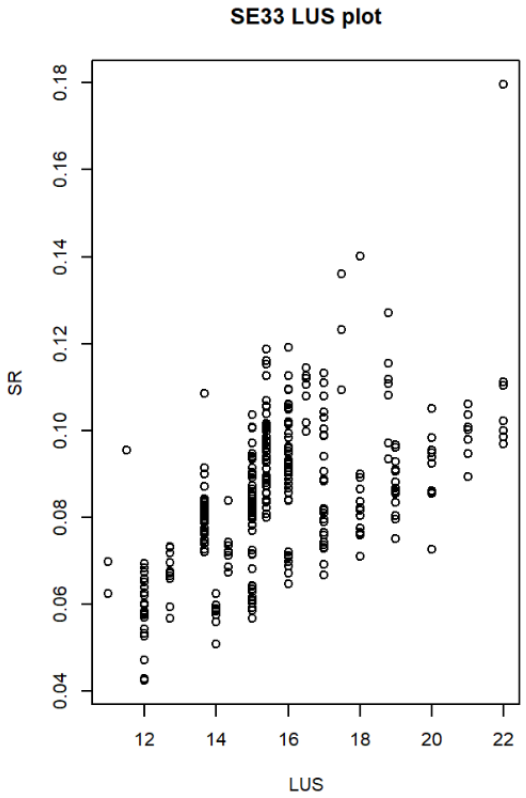
It seems like the LUS still leaves some of the stutter variation unexplained. A multi-sequence model takes into account all uninterrupted stretches (AUS) as potentially contributing to stuttering.

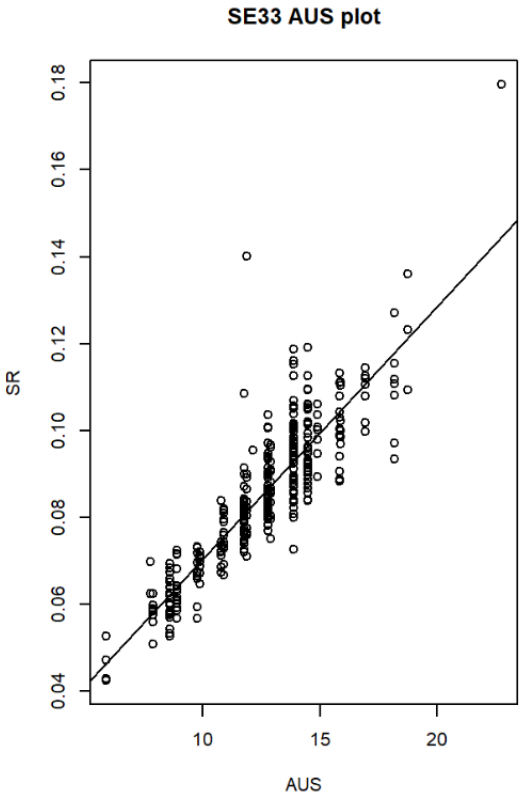
Allele	Repeat motif
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{11}G\ AAGG[AAAG]_2AG$
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{11}AA\ AAAG[AAAG]_9G\ AAGG[AAAG]_2AG$
22	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{22}G[AAAG]_3AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_7AA\ AAAG[AAAG]_{14}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_8[AG]_5[AAAG]_{12}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{12}GAAGG[AAAG]_2AG$

Examples of locus SE33 sequences.

Stutter Modeling - AUS Model







Stutter Modeling - AUS Model

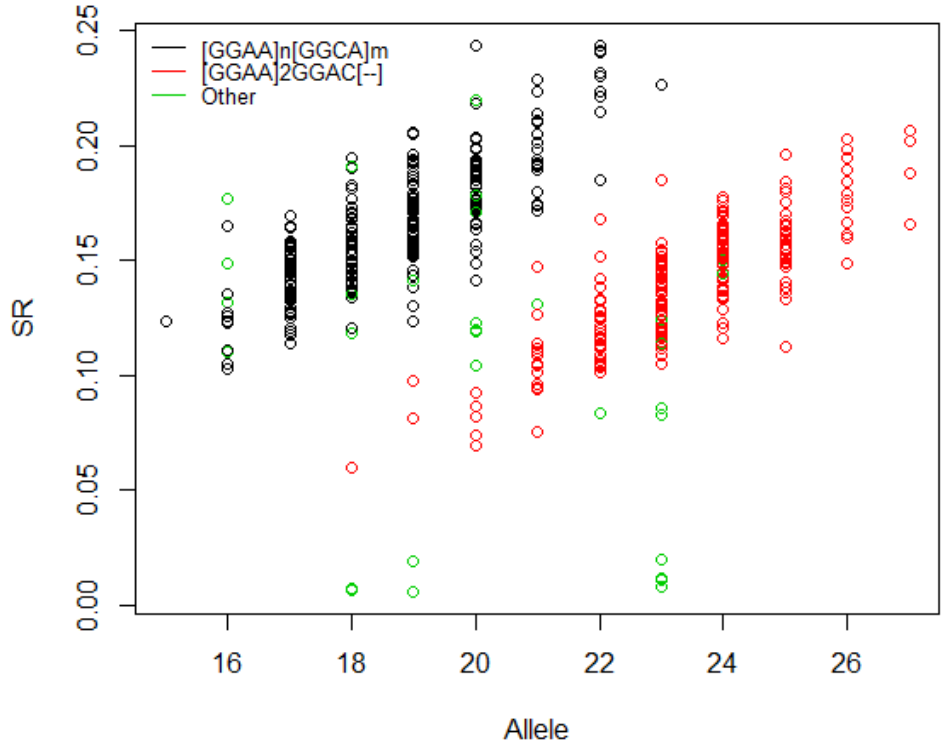
How to determine the length of the stretches for CE data?

Allele	Repeat motif
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{11}G\ AAGG[AAAG]_2AG$
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{11}AA\ AAAG[AAAG]_9G\ AAGG[AAAG]_2AG$
22	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{22}G[AAAG]_3AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_7AA\ AAAG[AAAG]_{14}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_8[AG]_5[AAAG]_{12}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{12}GAAGG[AAAG]_2AG$

Examples of locus SE33 sequences.

Stutter Modeling - AUS Model

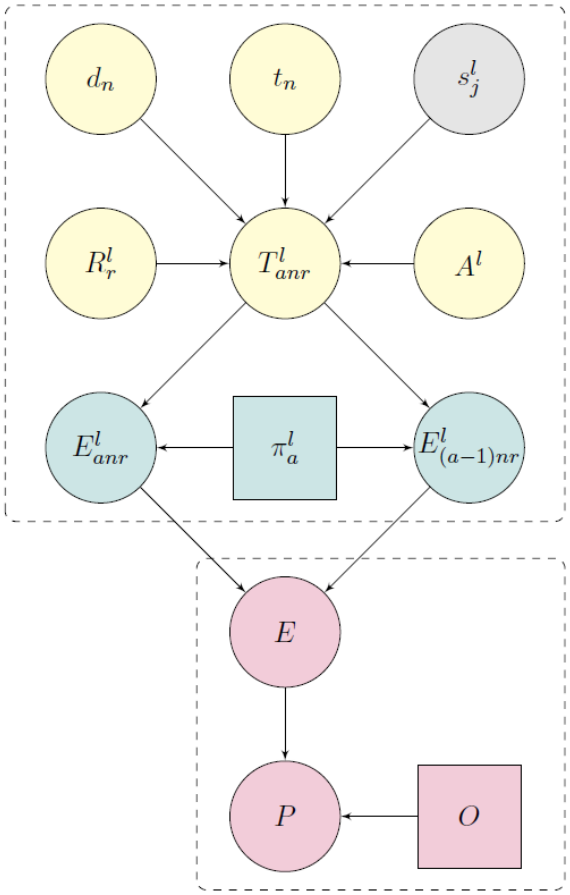
What about variation that is suggested to be attributable to sequence motif? Models fitted based on AUS still left some variability unexplained for some loci.



Stutter ratios for locus D2S1338.

Continuous Model Network

The continuous model we are going to discuss consists of several elements:



Adapted from: The interpretation of single source and mixed DNA profiles (Taylor et al., 2013).