Section 4: DNA Interpretation and Modeling

DNA Interpretation and Modeling

- Thresholds

- Weight of evidence

- LR calculations

- LR modeling

Thresholds

The most straightforward way to interpret an STR profile is with the use of thresholds.

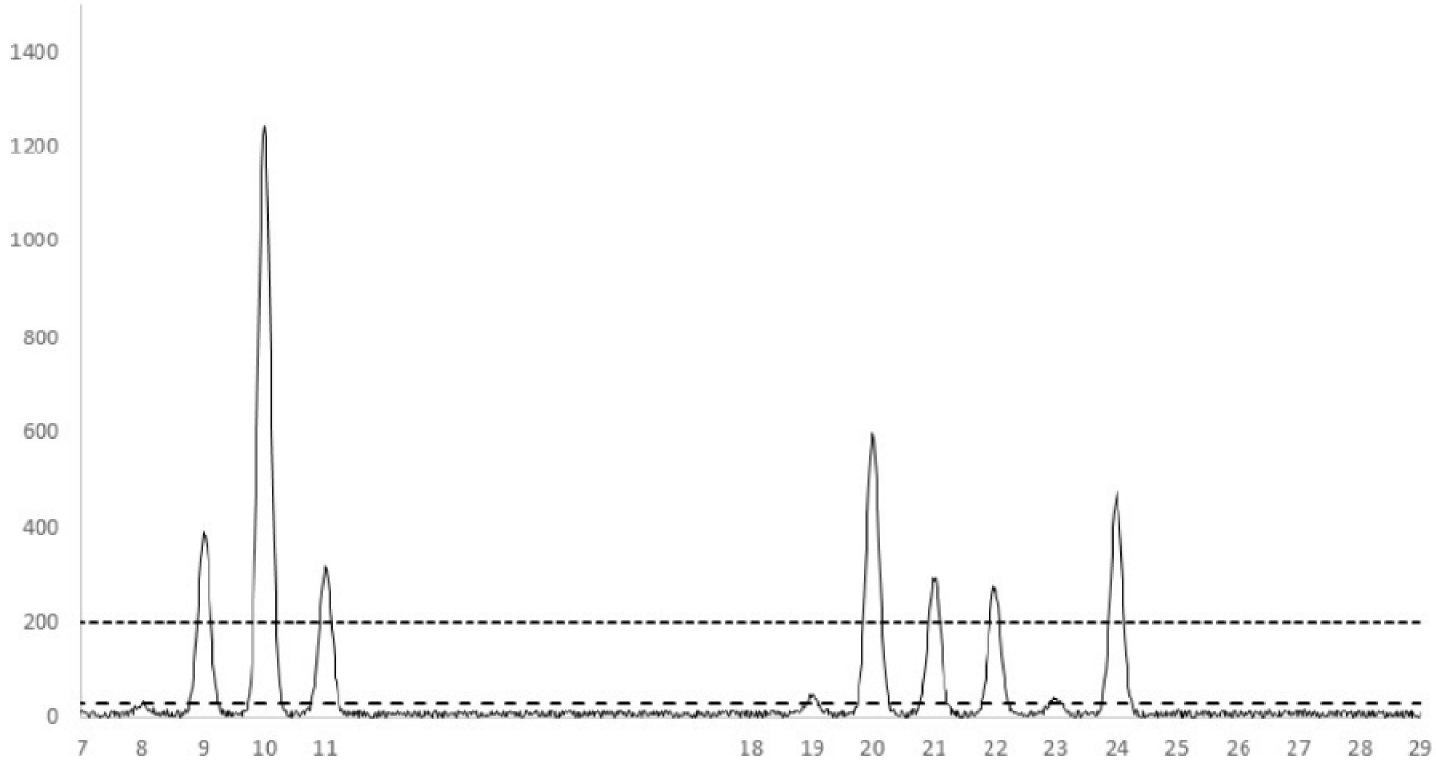
- **High thresholds**: will reduce the number of artifacts and remove a lot of background noise. However, it may potentially lead to a number of drop-outs.

- **Low thresholds**: will detect more authentic alleles, but have a higher probability of showing drop-ins.

Thresholds

An *analytical threshold* (AT) is usually set as a limit above which method response is interpreted as an authentic allele.

Additional stutter thresholds can help improve mixture profile interpretation (e.g. $5 - 15\%$ of the main allele).



Weight of Evidence

An STR profile obtained from a crime scene sample can be compared to a person of interest, and it may be found that this person cannot be excluded. An ‘inclusion’ may be reported, but is practically worthless without some expression on the strength of this evidence.

The Island Problem

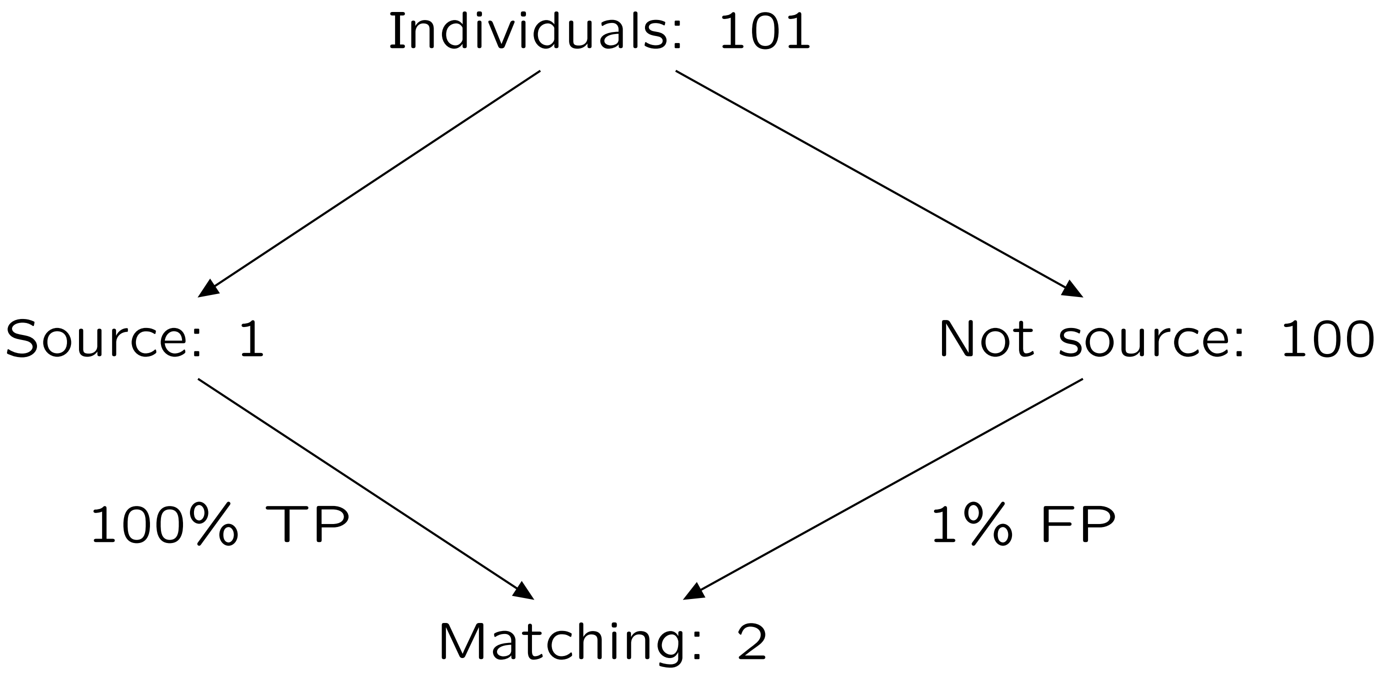
Suppose there is a crime committed on a remote island with a population of size 101. A suspect Q is found to match the crime scene profile. What is the probability that Q is the source of the profile, assuming that:

- All individuals are equally likely to be the source.
- The DNA profiles of all the other individuals are unknown.
- We expect 1 person in 100 to possess this observed profile.

Source: Weight-of-Evidence for Forensic DNA Profiles (Balding & Steele, 2015)

The Island Problem - Solution

In addition to Q, we expect one other individual on the island to match. So, even though the profile is rare, there is only a 50% chance that Q is the source.



Individuals: 101

Source: 1

Not source: 100

100% TP

1% FP

Matching: 2

The Island Problem - Odds Version

Recalling the odds form of Bayes' theorem:

$$\frac{\text{Pr}(H_p|E)}{\text{Pr}(H_d|E)} = \frac{\text{Pr}(E|H_p)}{\text{Pr}(E|H_d)} \times \frac{\text{Pr}(H_p)}{\text{Pr}(H_d)},$$

with

$$\text{Pr}(H_p) = \frac{1}{101} \qquad \text{Pr}(E|H_p) = 100\%$$

$$\text{Pr}(H_d) = \frac{100}{101} \qquad \text{Pr}(E|H_d) = 1\%,$$

yielding prior odds of $\frac{1}{100}$ and a likelihood ratio of 100. Combining this gives posterior odd of 1, or equivalently, a 50%/50% chance.

The Island Problem - Odds Version

A more general formula can be derived by writing:

$$\text{Pr}(H_p|E) = \frac{\text{Pr}(E|H_p)\,\text{Pr}(H_p)}{\text{Pr}(E|H_p)\,\text{Pr}(H_p) + \text{Pr}(E|H_d)\,\text{Pr}(H_d)}$$

$$= \frac{1}{1 + \frac{\text{Pr}(E|H_d)}{\text{Pr}(E|H_p)}\frac{\text{Pr}(H_d)}{\text{Pr}(H_p)}}$$

Note that it is assumed that H_p and H_d are mutually exclusive and collectively exhaustive.

The Island Problem - Odds Version

When N denotes the number of individuals on the island other than the suspect, and p is the profile probability of the observed DNA sample:

$$\Pr(H_p|E) = \frac{1}{1 + Np}$$

Extreme oversimplification of assessing the weight of evidence:

- Uncertainty about N and p
- Effect of searches, typing errors, other evidence
- Population structure and relatives

The Island Problem - Searches

Now suppose Q was identified through a search, with the suspect being the only one among 21 tested individuals who matches the crime scene profile.

- How does this knowledge affect the probability of being the source?
- What is the general expression for the probability of being the source, using k for the number of individuals who have been excluded?

The Island Problem - Searches

In this case we can exclude individuals from our pool of possible donors, such that our prior odds will slightly increase.

Out of the $N-k=80$ individuals, we expect another 0.8 matches, yielding a probability of being the source of $1/1.8 \approx 56\%$. Or, in formula:

$$\Pr(H_p|E) = \frac{1}{1+(N-k)p},$$

where setting $k=0$ gives the original expression and $k=N$ gives $\Pr(H_p|E)=1$.

Likelihood Ratio

As seen previously, the forensic scientist is concerned with assigning the likelihood ratio

$$\text{LR} = \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)},$$

which is equivalent to the reciprocal of the *profile probability* for the island problem:

$$\text{LR} = \frac{1}{\Pr(G_C|H_d, I)} = \frac{1}{p},$$

although we observed that the *match probability* is a more relevant quantity:

$$\text{LR} = \frac{1}{\Pr(G_C|G_S, H_d, I)}.$$

Match Probabilities

Recall the match probabilities for homozygotes:

$$\Pr(AA|AA) = \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

$$= p_A^2 \qquad \text{(if } \theta = 0\text{)},$$

and for heterozygotes:

$$\Pr(AB|AB) = \frac{2[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}$$

$$= 2p_Ap_B \qquad \text{(if } \theta = 0\text{)}.$$

LR for a Single Locus

Consider a simple two-person mixture profile (e.g. contributors are unrelated, ignoring population structure, no drop-outs/drop-ins), where $G_C = ABCD$. What is the appropriate single-locus LR (assuming HWE and p_A, p_B, p_C and p_D are known) when:

- $G_S = AB$ and $G_K = CD$, with

  $$H_p : \text{K + POI (S)} \quad \text{and} \quad H_d : \text{K + Unknown (U)}$$

- $G_S = AA$ and $G_K = CD$, with:

  $$H_p : \text{K + S} \quad \text{and} \quad H_d : \text{K + U}$$

- $G_S = AB$ and the second contributor is unknown

  $$H_p : \text{S + U} \quad \text{and} \quad H_d : \text{2U}$$

LR for a Single Locus

Consider a simple two-person mixture profile (e.g. contributors are unrelated, ignoring population structure, no drop-outs/drop-ins), where $G_C = ABCD$. What is the appropriate single-locus LR (assuming HWE and p_A, p_B, p_C and p_D are known) when:

- $\text{LR} = \frac{\Pr(ABCD|AB,CD,H_p)}{\Pr(ABCD|CD,H_d)} = \frac{1}{2p_A p_B}$;

- $\text{LR} = \frac{\Pr(ABCD|AA,CD,H_p)}{\Pr(ABCD|CD,H_d)} = 0$;

- $\text{LR} = \frac{\Pr(ABCD|AB,H_p)}{\Pr(ABCD|H_d)} = \frac{2p_C p_D}{6 \cdot 4 p_A p_B p_C p_D} = \frac{1}{12 p_A p_B}$.

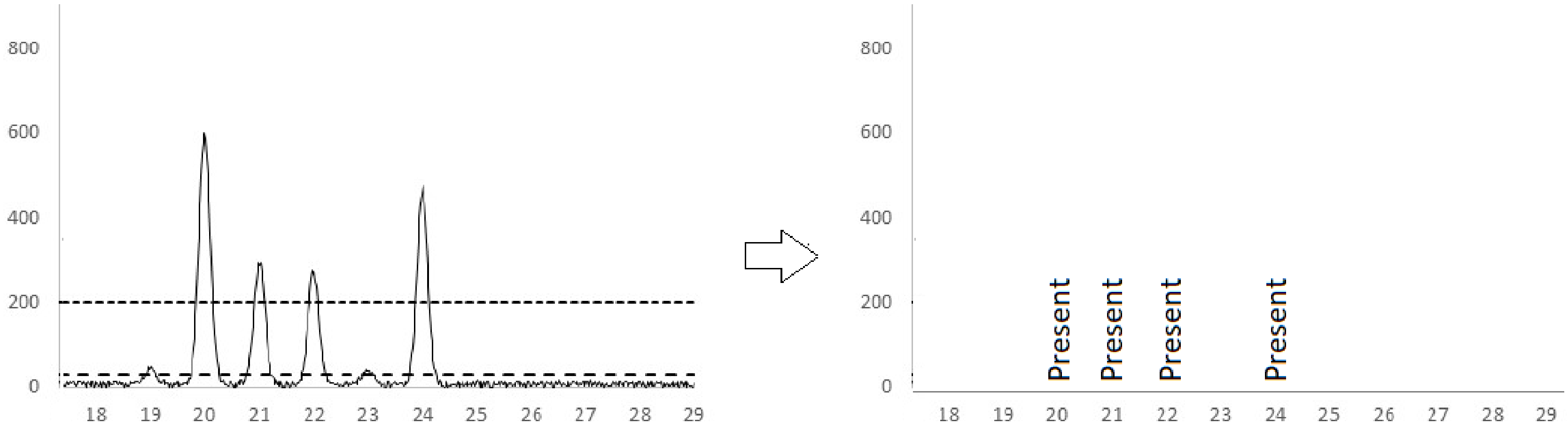
LR Modeling

Different approaches can be used to assess the likelihood ratio:

- Binary model
- Semi-continuous model
- Continuous model

Binary Model

A binary model limits interpretation of DNA profiles to qualitative allele callings only, without any attempt to infer the underlying genotypes (i.e. each are regarded as equally likely).



Just as in our previous example, single-locus LRs can be calculated and combined across loci via multiplication.

Semi-continuous Model

A semi-continuous model retains the simplicity of binary methods, but combines this with probabilistic modeling of known phenomena such as drop-ins and drop-outs.

Since these models still suffer from a significant loss of information, a more quantitative approach might be preferred.

Ideally, a statistical framework utilizes as much available quantitative information as possible, while maintaining comprehensibility.

Semi-continuous Model

The semi-continuous model will be of value when quantitative data is not available (e.g. old cases may only consist of allelic profiles).

Until now we have assumed that the definition of a match is clear. In practice, however, Pr$(E|H_p) \neq 1$.

Alleles carried by (hypothesized) contributors may not be detected in the evidence or vice versa. Drop-out and drop-in probabilities allow us to consider such situations.

Semi-continuous Model - Drop-out

For simplicity, consider a single-source profile evaluated while allowing for drop-out only in the crime scene profile G_C, as it will commonly be the stain that is of limited quantity or quality.

Two drop-out probabilities are usually considered: the probability D that an allele of a heterozygote drops out and the probability D_2 that both alleles of a homozygote drop out, with $D_2 < D^2$.

Assuming that drop-out is independent over alleles and markers, for $G_C = A$ and $G_S = AB$ the LR becomes:

$$\text{LR} = \frac{\Pr(G_C|G_S, H_p)}{\Pr(G_C|G_S, H_d)} = \frac{D(1-D)}{(1-D_2)P_{AA} + D(1-D)\sum_{Q\neq A} P_{AQ}}$$

Semi-continuous Model - Drop-out

Other LRs can be constructed in a similar fashion:

	$\Pr(G_C \mid G_S, H_p)$		
G_C	$G_S = AB$	$G_S = AA$	$\Pr(G_C \mid G_S, H_d)$
A	$D(1-D)$	$1-D_2$	$(1-D_2)P_{AA} + D(1-D)\sum_{Q \neq A} P_{AQ}$
AB	$(1-D)^2$	0	$(1-D)^2 P_{AB}$
$\emptyset$	D^2	D_2	$D^2 \sum_Q P_{QQ} + D^2 \sum_{QQ'} P_{QQ'}$

Omitting loci where no data has been observed in the crime scene profile would only be acceptable if LR ≥ 1, which is not true in general. Ignoring such loci may raise concern that those potentially fail to exclude non-contributors.

Semi-continuous Model - Drop-in

Let C denote the probability that a single allele has dropped in at a particular locus. If drop-ins at different loci are mutually independent and furthermore also independent of any drop-outs:

	$\Pr(G_S \rightarrow G_C)$	
G_C	$G_S = AB$	$G_S = AA$
A	$D(1-D)(1-C)$	$1 - D_2(1-C)$
AB	$(1-D)^2(1-C)$	$(1-D_2)Cp^*_B$
AQ	$D(1-D)Cp^*_Q$	$(1-D^2)Cp^*_Q$
ABQ	$(1-D)^2Cp^*_Q$	0
Q	$D^2Cp^*_Q$	$D_2Cp^*_Q$
$\emptyset$	$D^2(1-C)$	$D_2(1-C)$

Literature usually interprets p^*_Q as the allele frequency of allele Q, estimated as the sample frequency or a variation while allowing for sampling uncertainty.

Semi-continuous Model - Drop-in

LRs can now be constructed for different scenarios. For $G_C = A$ and $G_S = AB$:

- Under H_p, $\Pr(G_C|G_S, H_p) = \Pr(AB \rightarrow A)$

- Under H_d, G_S can be AA, AQ, QQ or QQ' with $Q, Q' \neq A$ such that $\Pr(G_C|G_S, H_d) = \Pr(AA \rightarrow A)P_{AA} + \sum_{Q \neq A}[\Pr(AQ \rightarrow A)P_{AQ} + \Pr(QQ \rightarrow A)P_{QQ}] + \sum_{Q,Q' \neq A} \Pr(QQ' \rightarrow A)P_{QQ'}$

Multiple drop-ins in a profile may be better interpreted as an additional (unknown) contributor.

Estimating Drop-in and Drop-out Probabilities

Drop-in and drop-out probabilities may be assigned by the forensic laboratory.

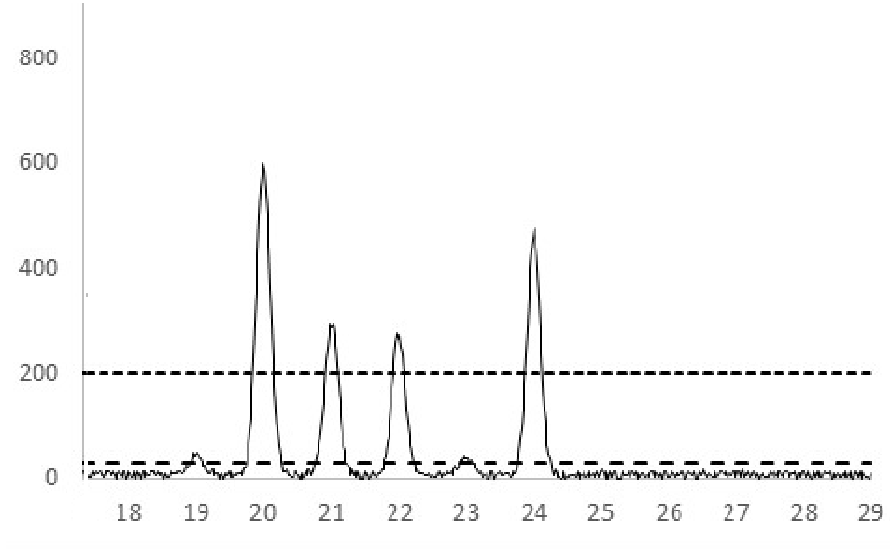
- Several models have been proposed for modeling drop-out probabilities, such as a multidose drop-out model and degradation model. Laboratory trials can be used to choose α when modeling $D_2 = \alpha D^2$, with $0 < \alpha \leq 1$. Instead of assigning probabilities to the drop-out rate they can be integrated out over a range of values[1].

- In case of independence, only a single drop-in probability C is needed, which may be calculated based on observations from negative controls: $C = \frac{x}{NL}$, where x is the number of observed drop-ins in N profiles over L loci.

[1] Accurate assessment of the weight of evidence for DNA mixtures by integrating the likelihood ratio (Slooten, 2017).

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

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}.

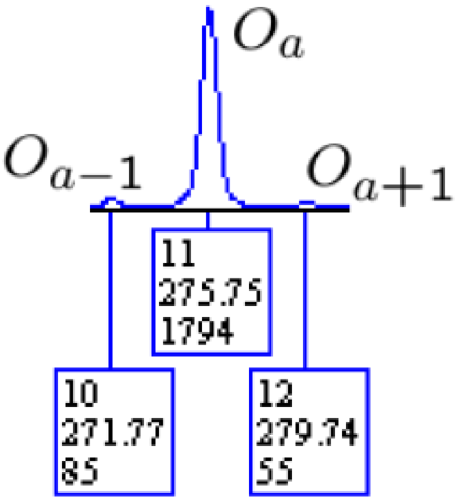
T^l_{arn} then describes the TAP of allele a at locus l, for replicate r from contributor n.

TAP Modeling

Theoretically, the previous slide 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:

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



Stacking

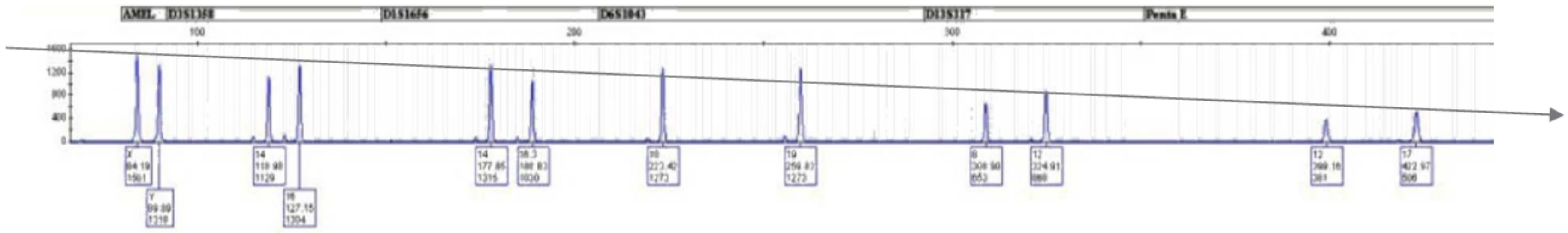
Note that we assume that expected peak heights are additive, i.e. if there are multiple sources of a single allele, the height of that allele will equal the sum of the individual expected heights from each source.

This assumption of additivity is called *stacking*.

Recent talks (Rudin, AAFS 2017) emphasize that this assumption has not been validated. To determine if this practice is scientifically supportable, it would be good to obtain a large set of mixtures from known profiles to look at the expected combined versus observed combined peak heights.

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.

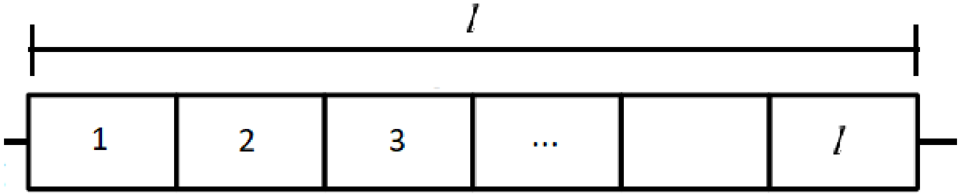
Modeling Degradation

- Consider a DNA fragment of length l.

- Let p be the probability of a break at any of the locations $1, \ldots, l$.

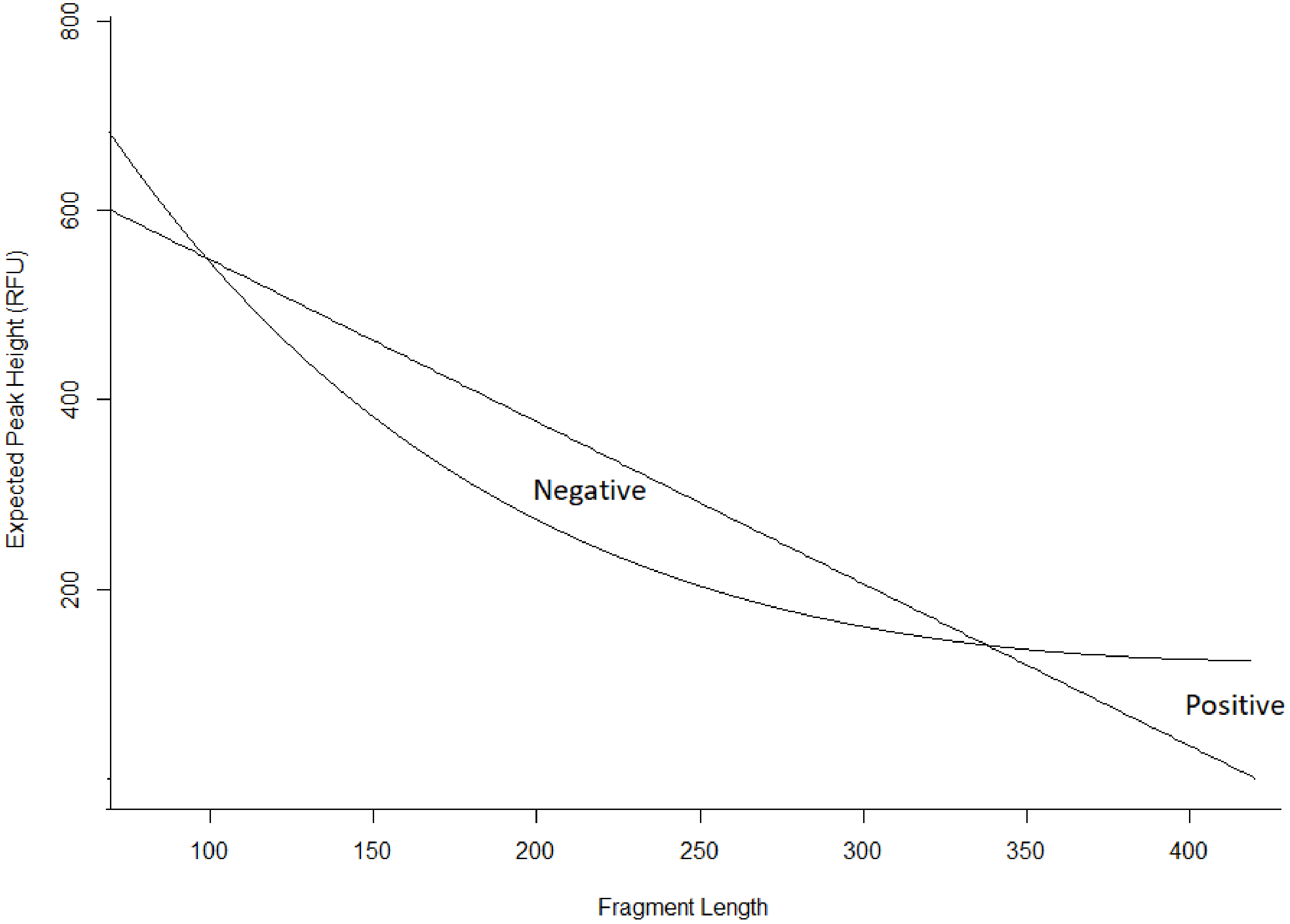
- The chance of the full fragment being amplified is $(1-p)^l$.

- This describes an exponential decline in peak heights.

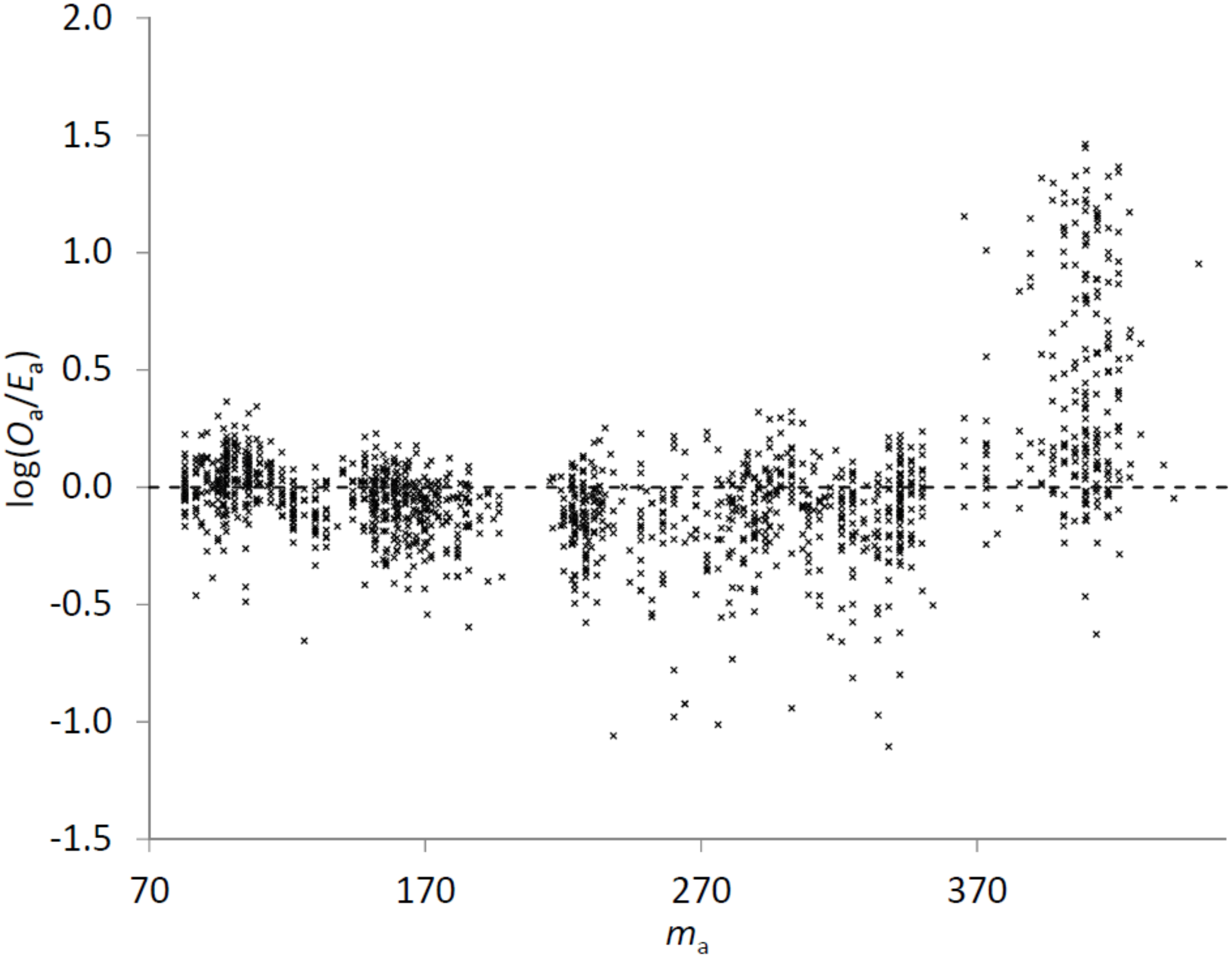


Source: Forensic DNA Evidence Interpretation (Buckleton et al., 2016).

Modeling Degradation

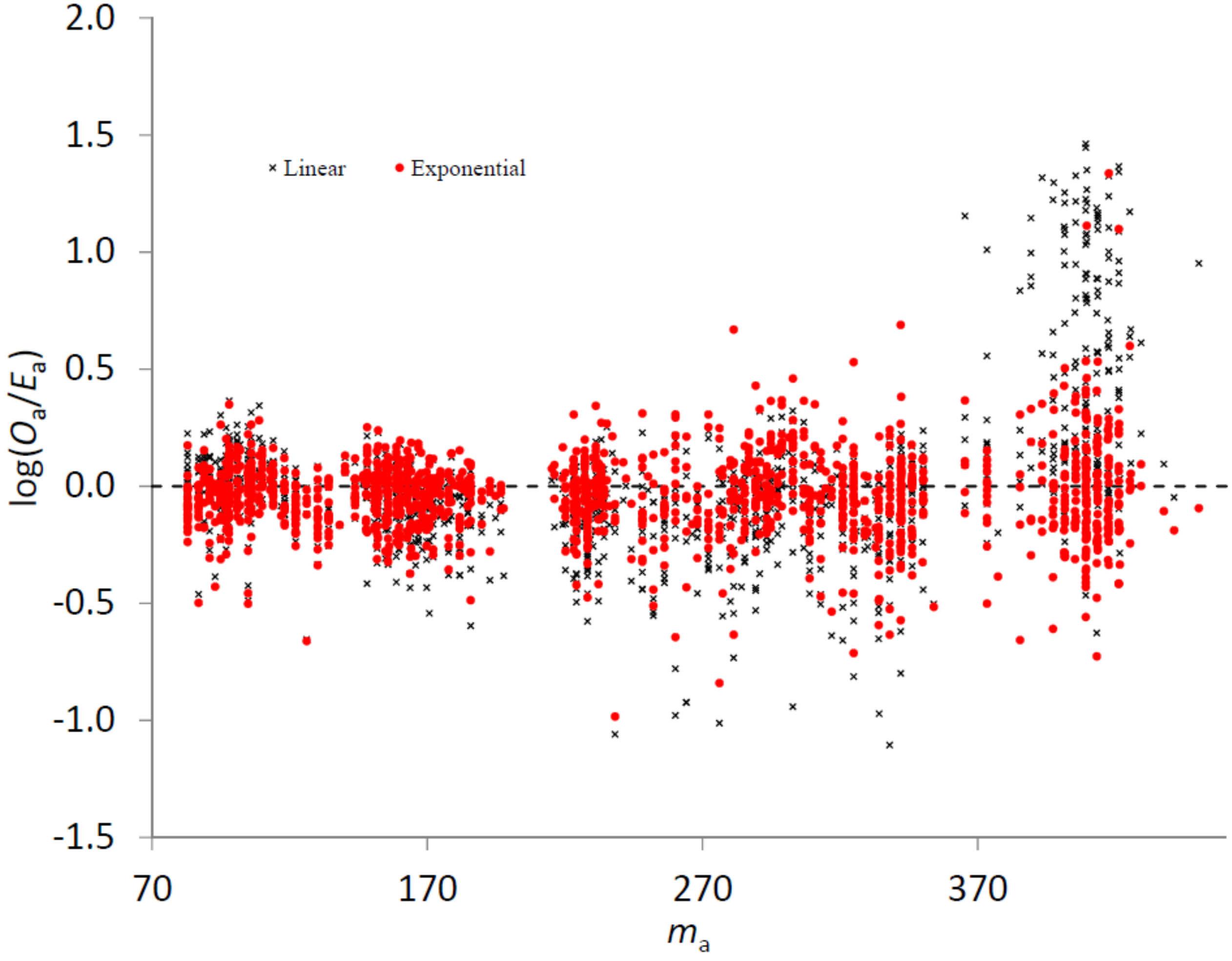


Modeling Degradation



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

Modeling Degradation



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

Modeling Heterozygote Balance

The heterozygote balance (Hb) is usually expressed as a peak height ratio, i.e. the ratio of two heterozygote peaks at a locus.

There are two common definitions:

$$\text{Hb}_1 = \frac{O_{\text{HMW}}}{O_{\text{LMW}}}, \qquad \text{and} \qquad \text{Hb}_2 = \frac{O_{\text{smaller}}}{O_{\text{larger}}},$$

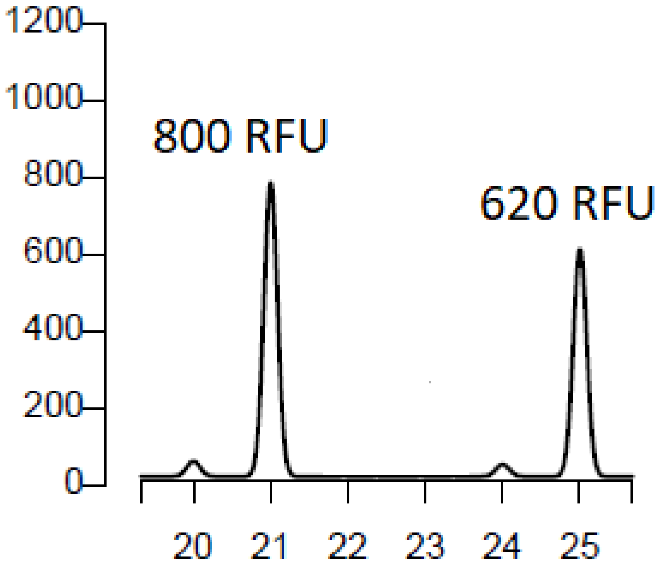
where O is the observed peak height; *smaller* and *larger* refer to the height of the alleles, and HMW and LMW refer to the higher and lower molecular weight allele, respectively.

Modeling Hb

$$\mathrm{Hb}_1 = \frac{O_{\mathrm{HMW}}}{O_{\mathrm{LMW}}}$$

$$= \frac{620}{800} = 0.775$$

$$= \mathrm{Hb}_2$$

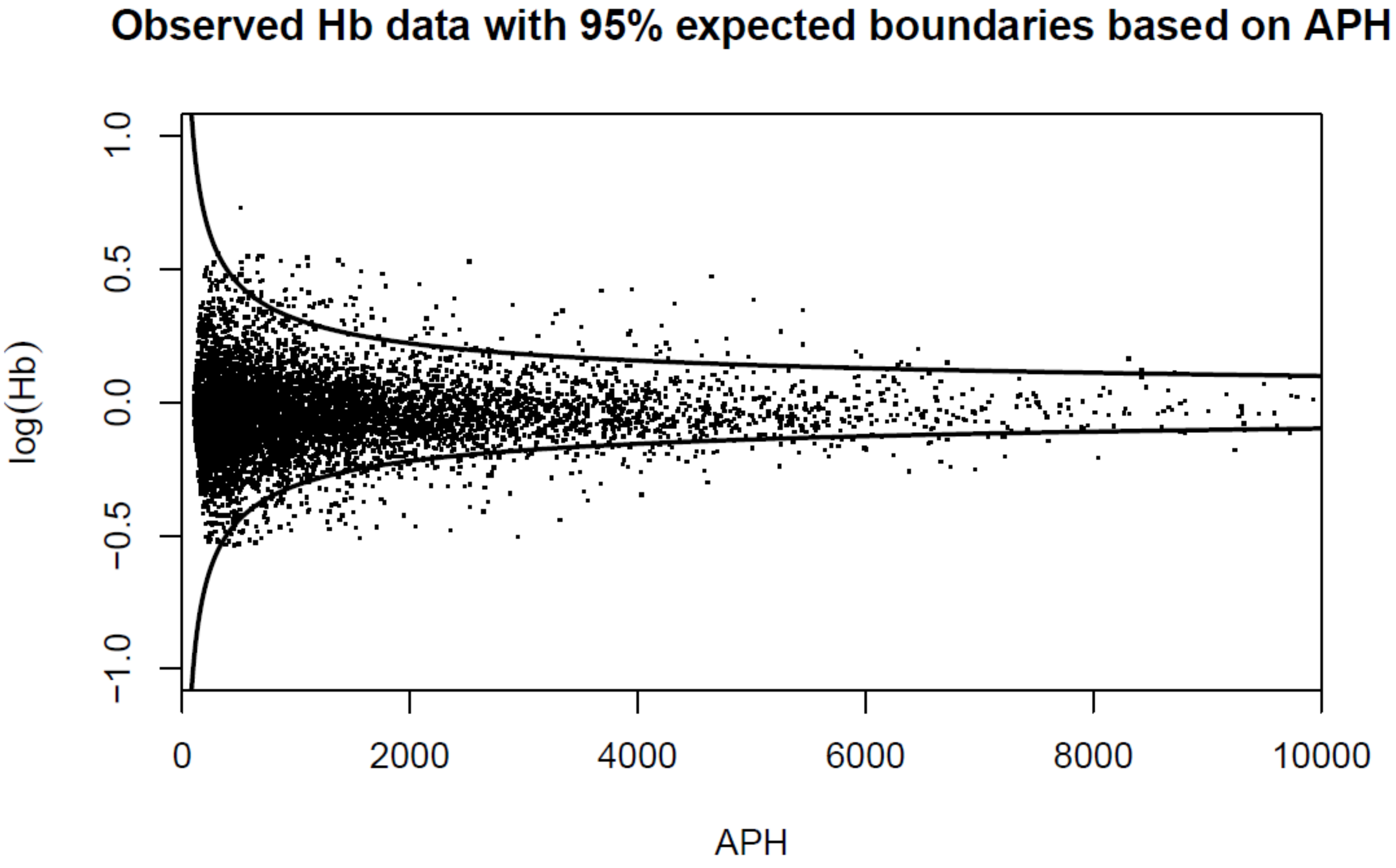


- Hb_1 has the highest information content, because it maintains peak order.

- Hb_2 may be obtained from Hb_1, but not vice versa.

Modeling Hb

The following figure shows Hb rates versus the *average peak height* (APH), which is simply the average of two observed heterozygote alleles at a locus.



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.

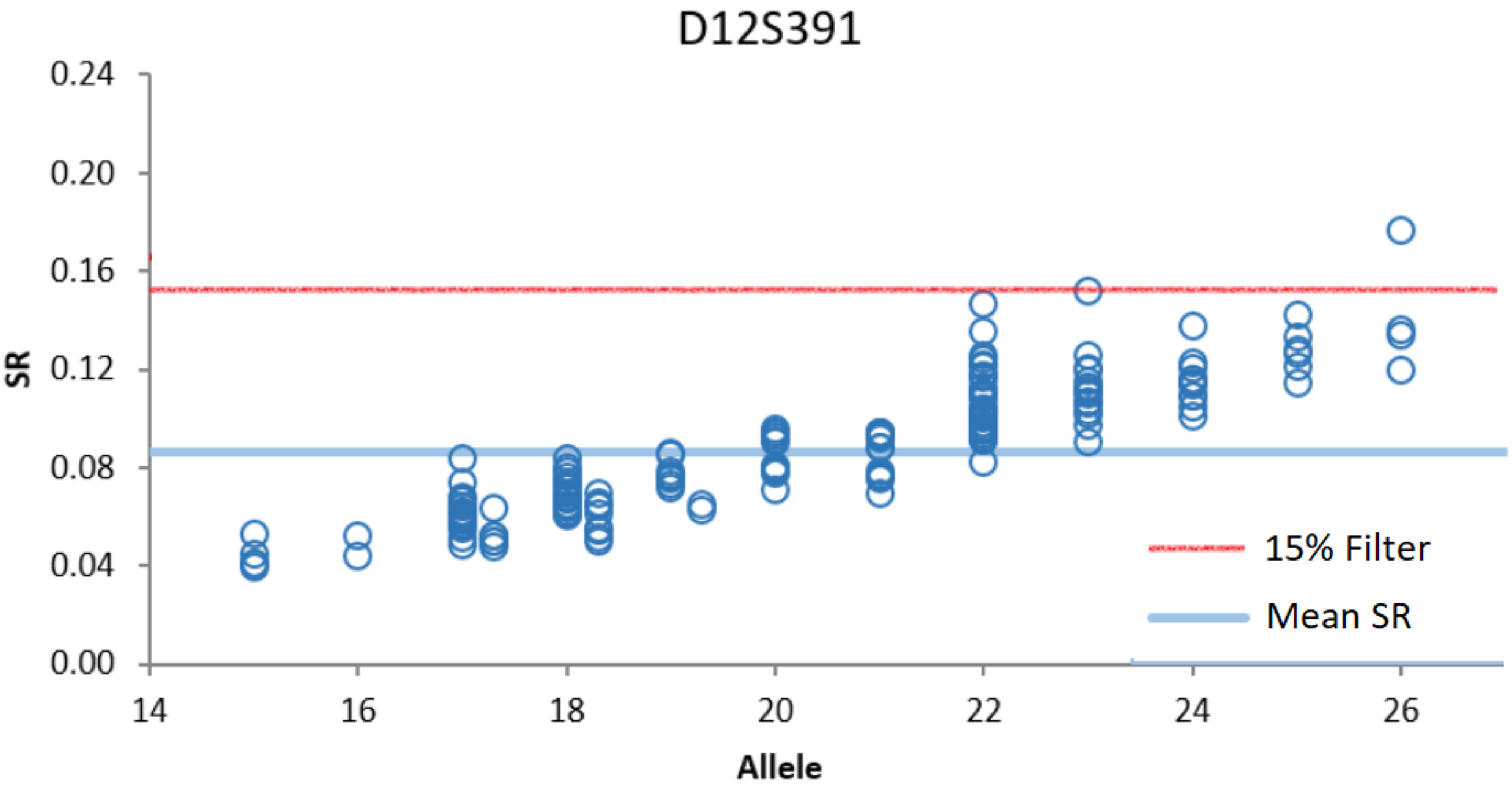
Stutter Modeling

As we've seen earlier, stutter thresholds can be set to help interpret a mixture profile. Locus-specific thresholds account for the variability observed between loci. Traditionally, fixed rates of around 15% are used to remove stutter.

Locus	**Stutter Filter (%)**
TH01	5
D2S441	9
vWA	11
FGA	11.5
SE33	15
D22S1045	17

However, fixed stutter thresholds have the disadvantage that they do not incorporate the well-known stutter characteristics (such as the correlation with the number of repeats).

Stutter Modeling - Locus Specific Thresholds



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

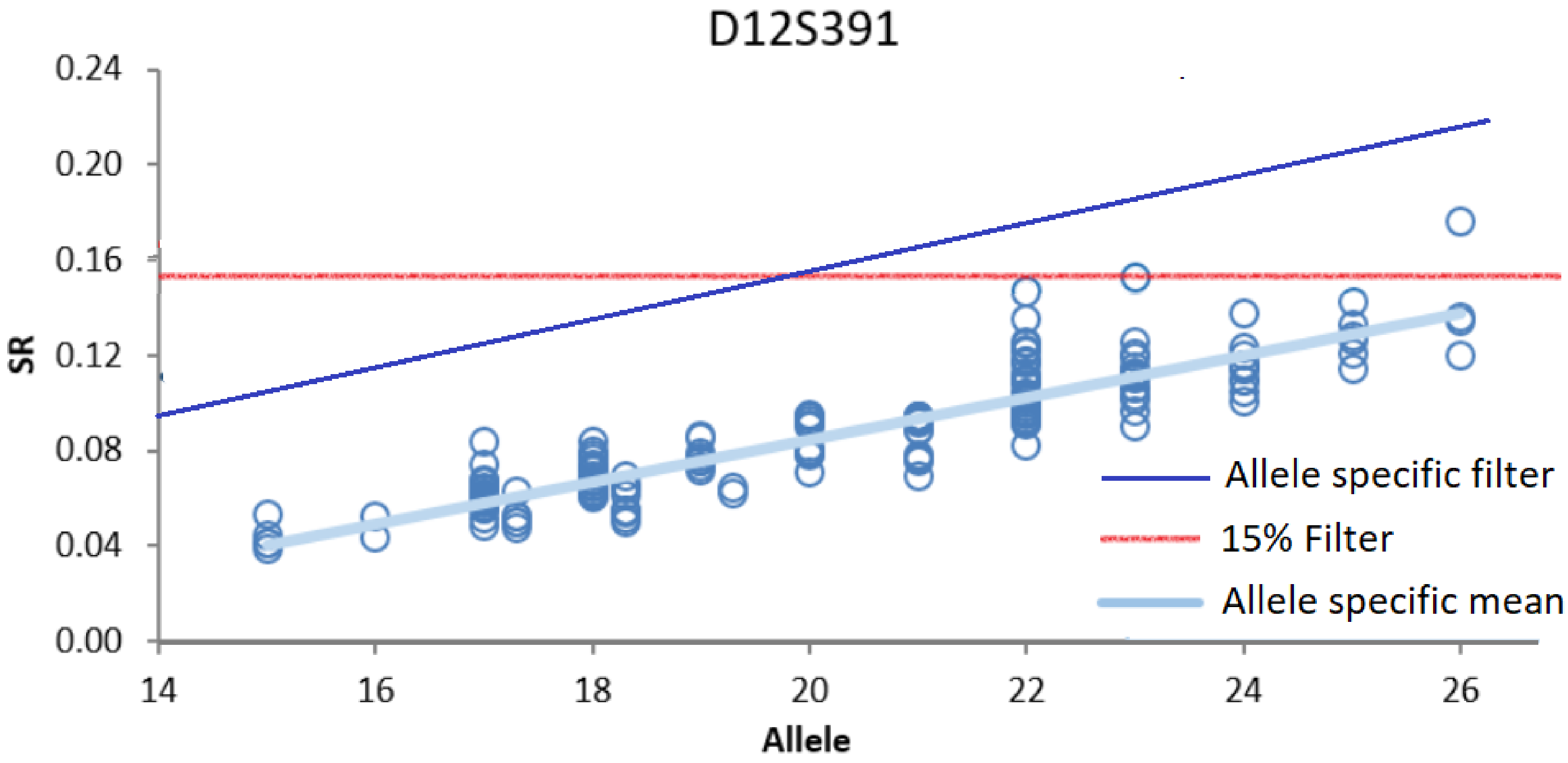
Stutter Modeling - Locus Specific Thresholds

Fixed stutter thresholds lead to over filtering and under filtering:

- **Over filtering**: leads to potential data loss and difficulties in interpretation when true allelic peaks of a minor contributor get filtered.

- **Under filtering**: leads to the possibility that stutter peaks are treated as allelic, and difficulties in determining genotypes for a minor contributor and the number of contributors.

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 - Thresholds

These observations suggest that stutter thresholds should not only be locus-based, but at a minimum also allele-based. Moreover:

- Thresholds do not account for more complex situations such as composite stutter;
- And still result in a binary decision (i.e. the peak is either ignored or labeled as allelic).

Fully continuous models have the potential to overcome such problems, since there is no need for thresholds within a probabilistic approach.

Stutter Modeling - Allele Model

A simple linear, allele specific, model can be fitted for each locus:

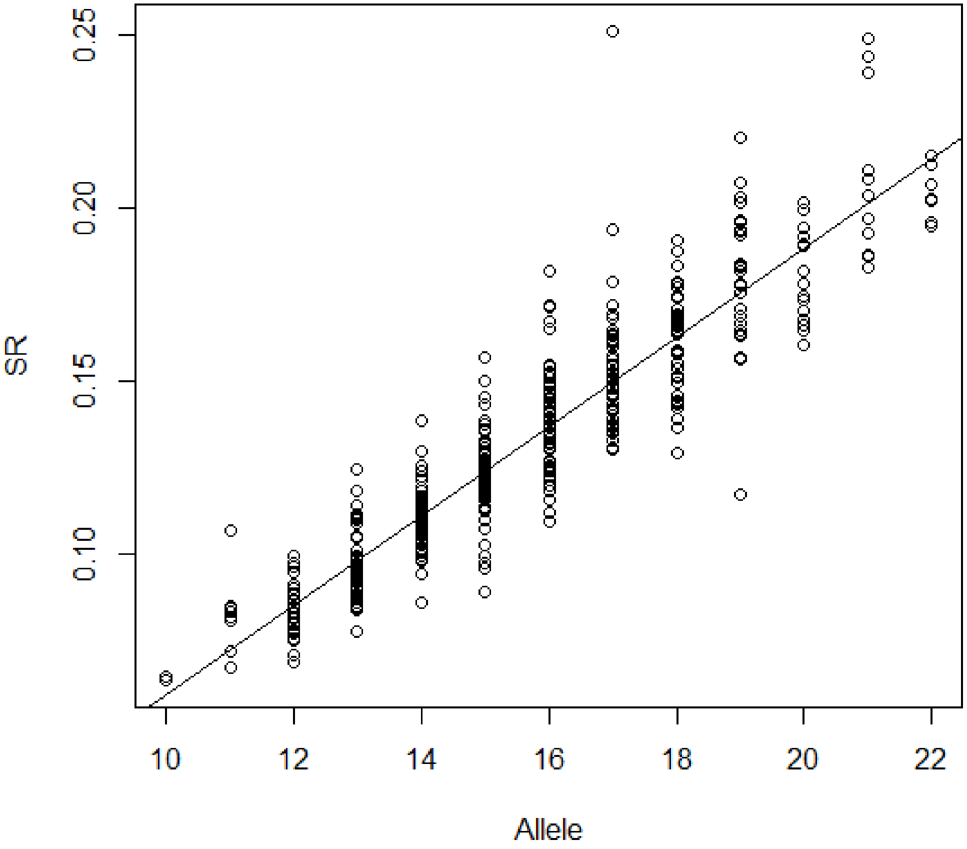
$$SR \sim \text{Allele number} \qquad \Rightarrow \qquad SR = ma + c,$$

with a the allele number, and m and c are constants that can be fitted to the data.

An R-squared measure (R^2) can be used to measure how close the data are fitted to the regression line.

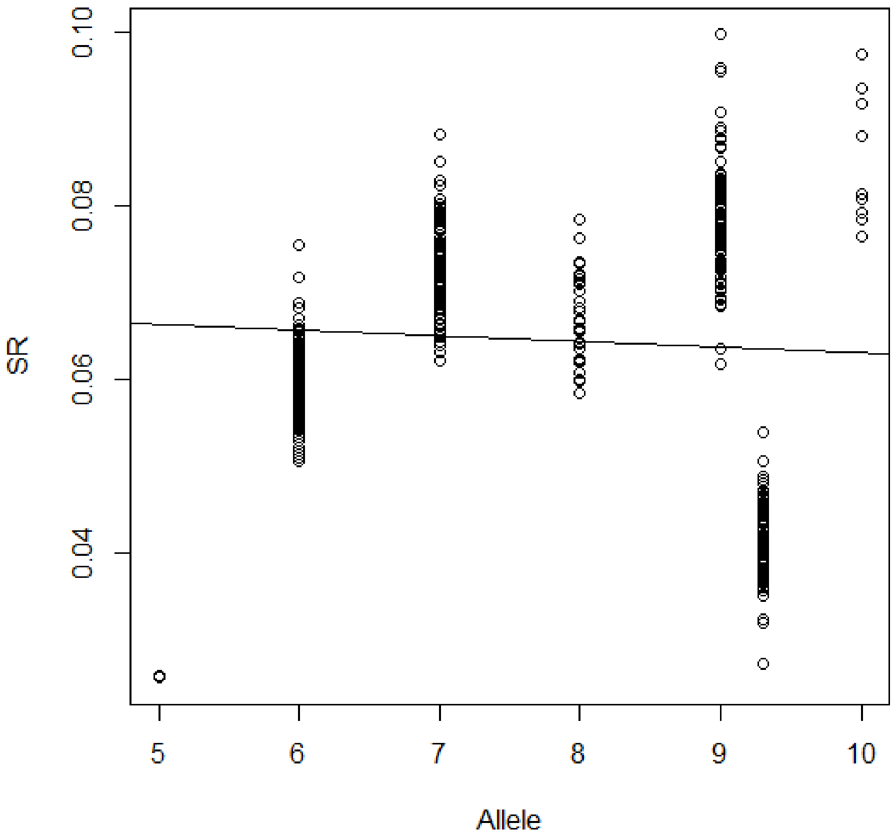
Stutter Modeling - Allele Model

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.

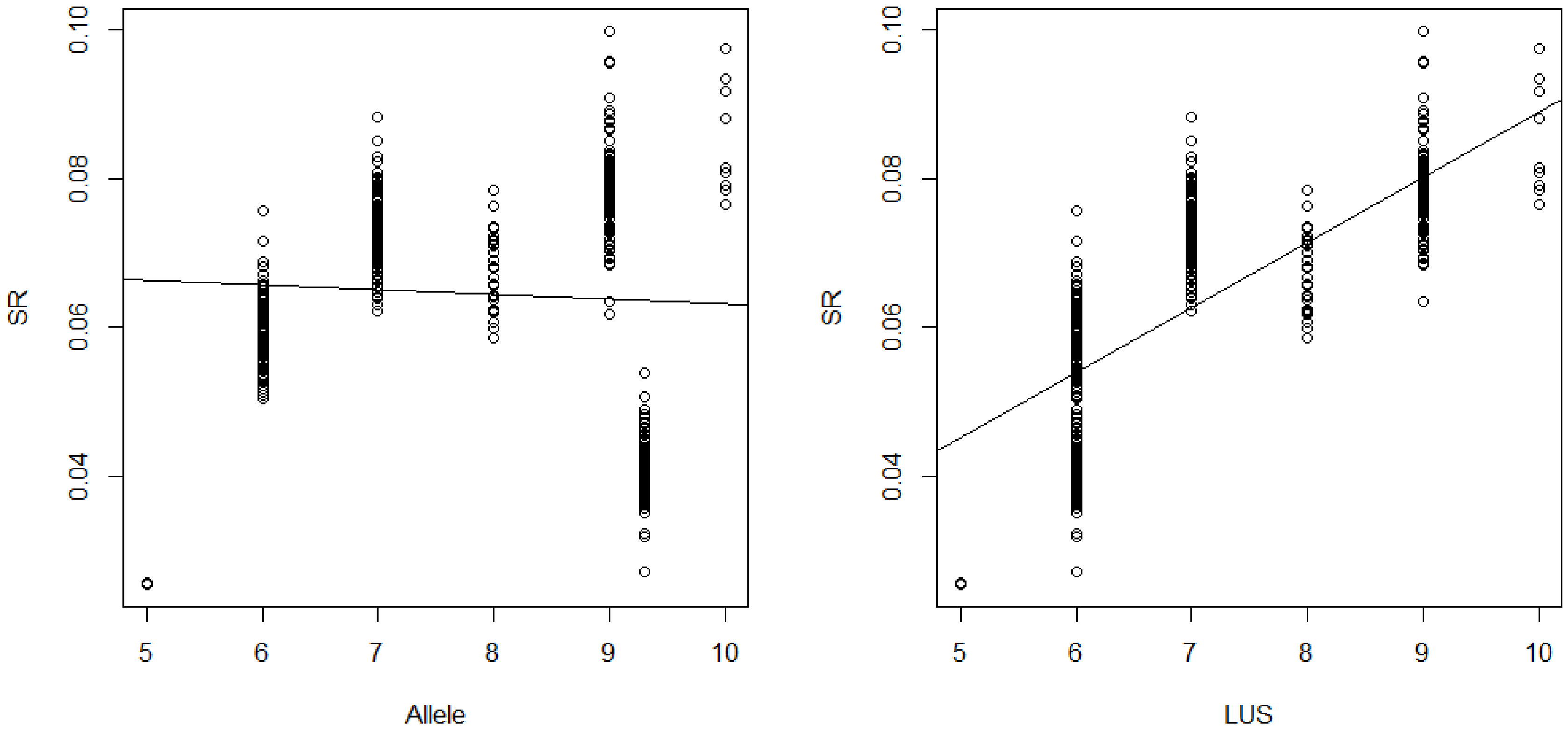
Stutter Modeling - LUS Model

A model based on the LUS can be fitted as follows:

$$SR \sim \text{LUS} \qquad \Rightarrow \qquad SR = ml + c,$$

with l the LUS, and m and c are constants that can be fitted to the data.

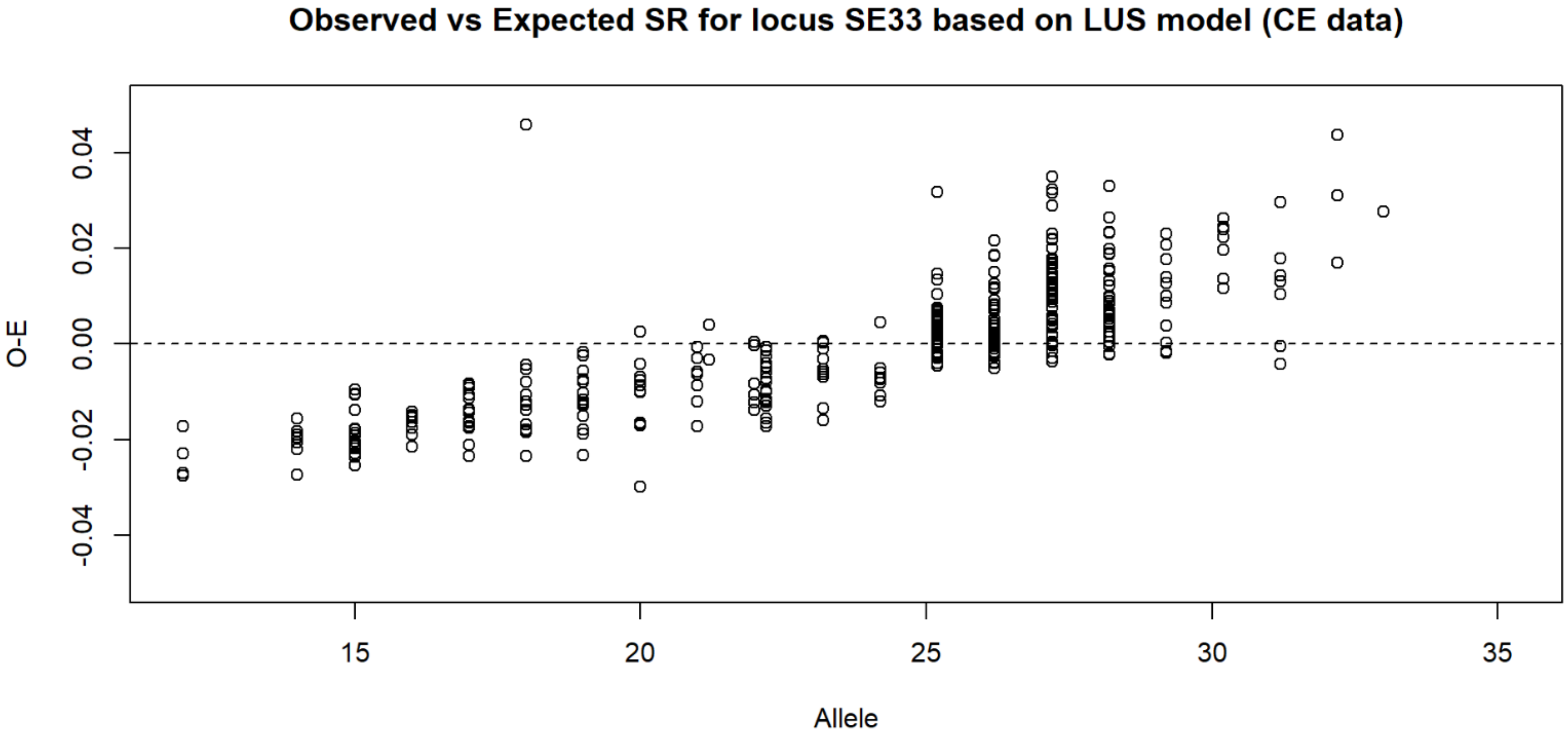
Stutter Modeling - LUS Model



Locus TH01 allele vs. LUS

Stutter Modeling - LUS Model

What about more complex loci?



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.

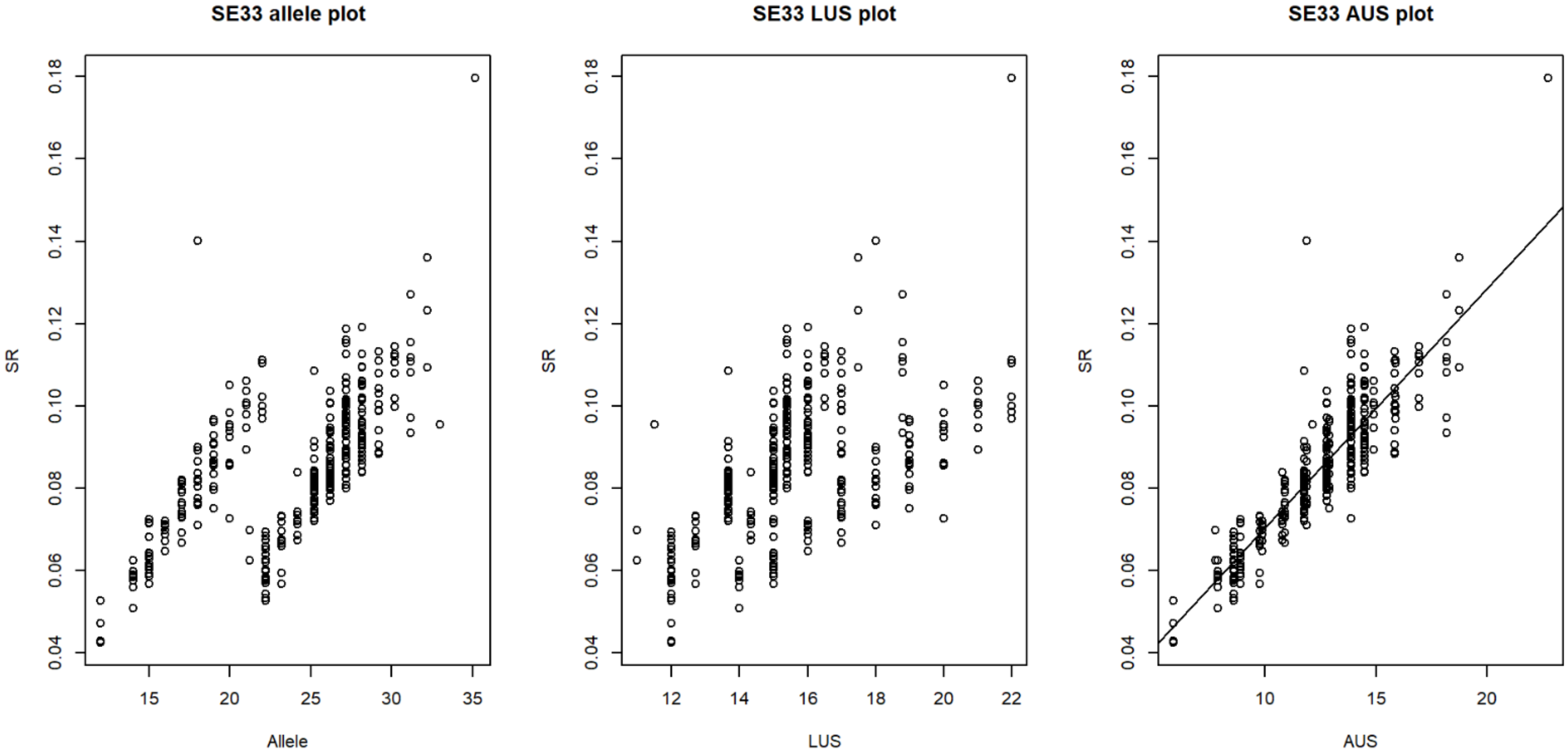
Allele	**Repeat motif**
21.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{9}$AA AAAG[AAAG]$_{11}$G AAGG[AAAG]$_{2}$AG
21.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{11}$AA AAAG[AAAG]$_{9}$G AAGG[AAAG]$_{2}$AG
22	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{22}$G[AAAG]$_{3}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{7}$AA AAAG[AAAG]$_{14}$GAAGG[AAAG]$_{2}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{8}$[AG]$_{5}$[AAAG]$_{12}$GAAGG[AAAG]$_{2}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{9}$AA AAAG[AAAG]$_{12}$GAAGG[AAAG]$_{2}$AG

Examples of locus SE33 sequences.

$$SR \sim \text{AUS} \qquad \Rightarrow \qquad SR = m \sum_{i} \max\left(l_i - x, 0\right) + c,$$

where l_i is the length of sequence i, and m, c and x are constants. The term x is called the lag, and can be interpreted as the number of repeats before stuttering begins.

Stutter Modeling - AUS Model



$$SR = m \sum_i \max(l_i - 6.11, 0) + c$$

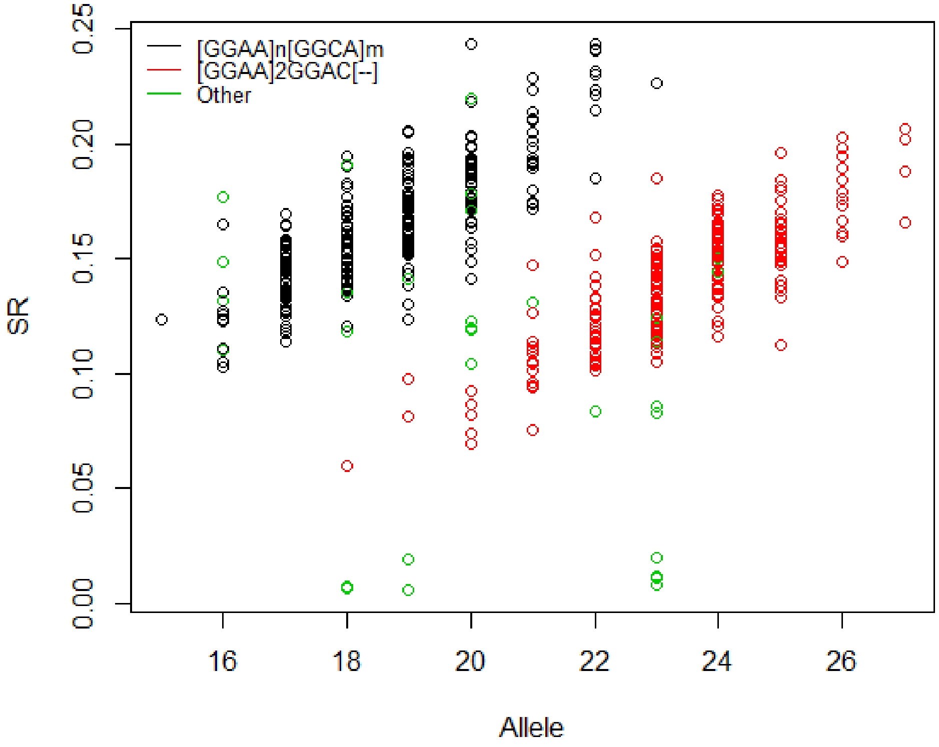
Stutter Modeling - AUS Model

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

Allele	**Repeat motif**
21.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{9}$AA AAAG[AAAG]$_{11}$G AAGG[AAAG]$_{2}$AG
21.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{11}$AA AAAG[AAAG]$_{9}$G AAGG[AAAG]$_{2}$AG
22	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{22}$G[AAAG]$_{3}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{7}$AA AAAG[AAAG]$_{14}$GAAGG[AAAG]$_{2}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{8}$[AG]$_{5}$[AAAG]$_{12}$GAAGG[AAAG]$_{2}$AG
22.2	[AAAG]$_{2}$AG[AAAG]$_{3}$AG[AAAG]$_{9}$AA AAAG[AAAG]$_{12}$GAAGG[AAAG]$_{2}$AG

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.

Stutter Modeling

- Note that for simple repeats there is no difference between the three approaches:

$$[\mathrm{AATG}]_8 \quad \Rightarrow \quad \text{Allele nr} = \text{LUS} = \text{AUS} = 8$$

- What about other stutter products?

We can model forward stutter as well, and can now use these expectations to decompose peak heights (e.g. for composite stutter or stutter affected heterozygotes).

However, the occurrence of artifacts such as double back and 2bp stutter is likely to be so rare that modeling them statistically can hardly be justified.

Forward Stutter Modeling

Forward stutter can be quantified by a stutter ratio as well (FSR):

$$FSR = \frac{O_{a+1}}{O_a},$$

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

Forward stutter is observed less often than back stutter, and peaks are more likely to fall below the limit of detection:

Locus	**Stutter Filter (%)**
TH01	0.06
vWA	0.33
FGA	0.30
D2S441	0.55
SE33	0.59
D10S1248	1.28

LR Modeling

The LR can now be assessed by writing the ratio in the form:

$$\text{LR} = \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)}$$

$$= \frac{\sum_j \Pr(G_C|S_j) \Pr(S_j|H_p)}{\sum_{j'} \Pr(G_C|S_{j'}) \Pr(S_{j'}|H_d)}$$

$$= \frac{\sum_j w_j \Pr(S_j|H_p)}{\sum_{j'} w_{j'} \Pr(S_{j'}|H_d)}.$$

The two propositions each define sets of genotypes S, and the weights w describe how well these sets fit our observed data G_C. Under H_p all the genotype sets S_j usually include G_S.

LR Modeling

The full profile weight can be obtained as a product of the weights at each locus:

$$w_j = \prod_l w_j^l.$$

In case of the binary model, the weights are set either as 1 or 0, depending on whether or not the crime scene profile can be explained based on the genotype set under consideration.

		Weights	**Weights**
Donor 1	**Donor 2**	**(Binary)**	**(Continuous)**
$20, 21$	$22, 24$	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

Modeling Strategies

Now that a model has been developed, we require information about the input parameters.

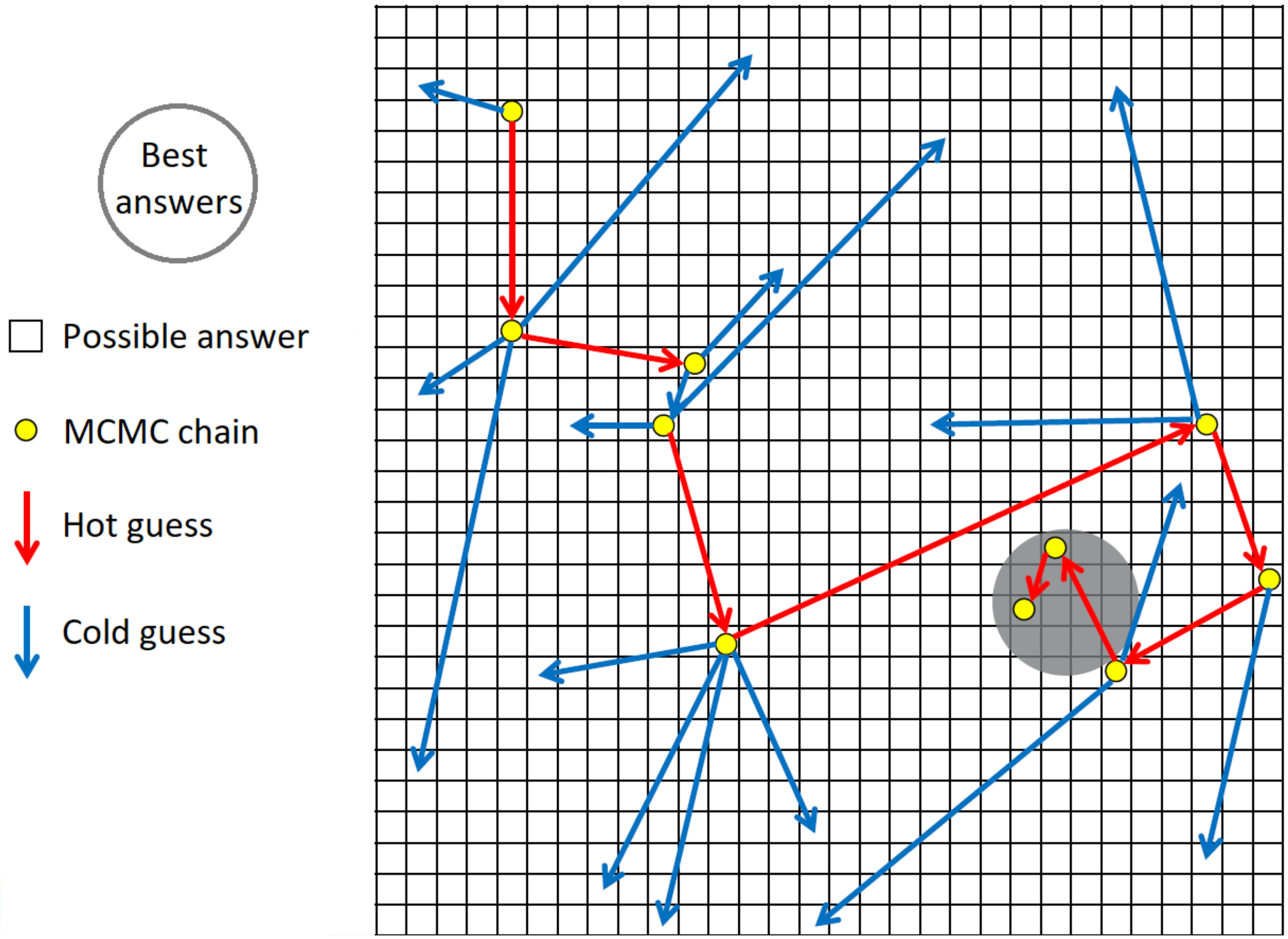
- **Maximization**: Parameters can be chosen that maximize the likelihood of the observations under each hypothesis.

- **Integration**: Rather than knowing the true values of the parameters, we need to know the effect they have on the probability of the observed data.

- **Markov chain Monte Carlo**: Instead of testing every possible combination of parameters, only a small distribution of parameter values and genotype sets will accurately describe the data.

Markov chain Monte Carlo (MCMC)

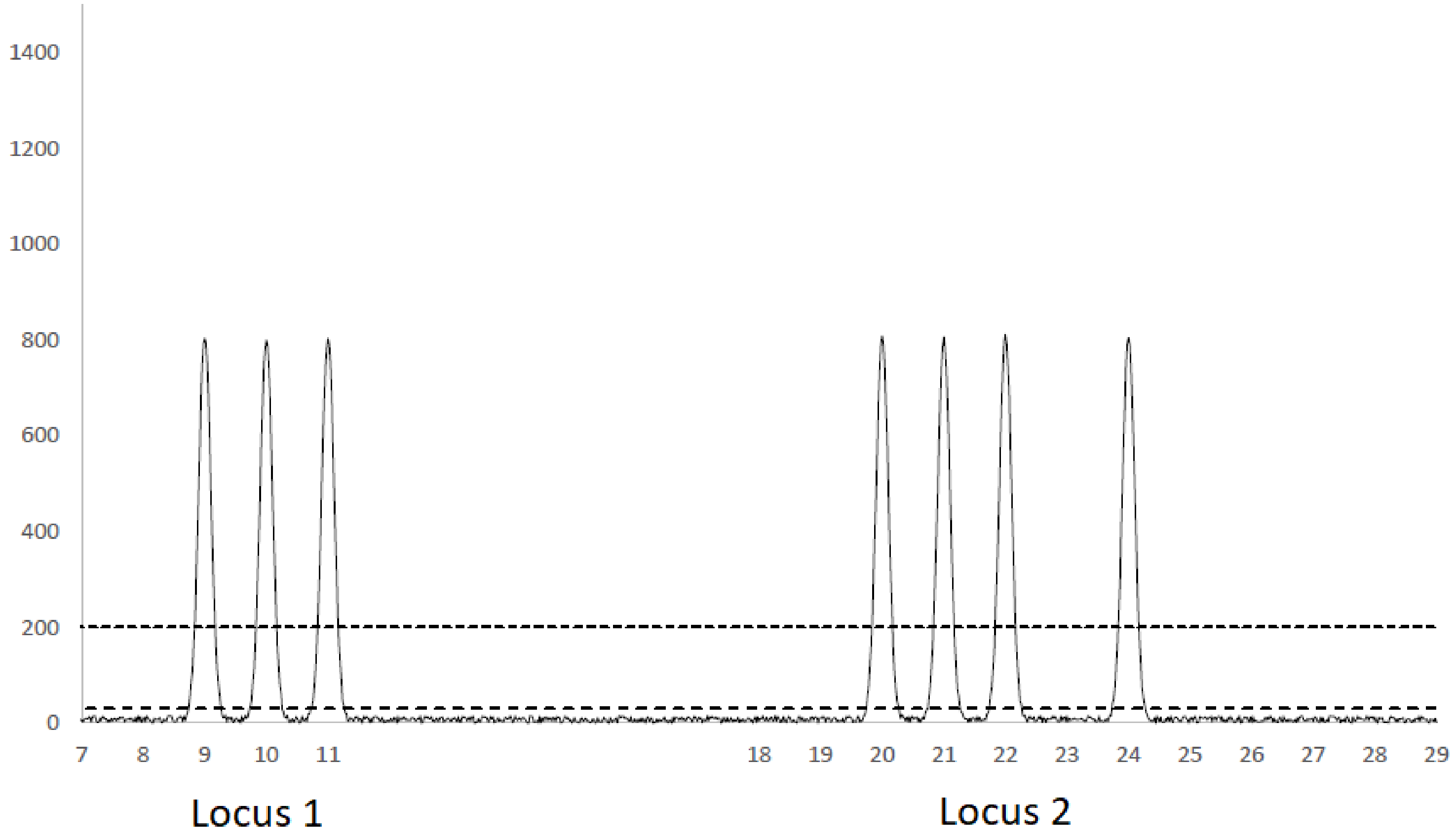
MCMC will start by choosing parameter values at random, eventually leading to more sensible options, until it has reached an equilibrium state.



Expected Peak Heights

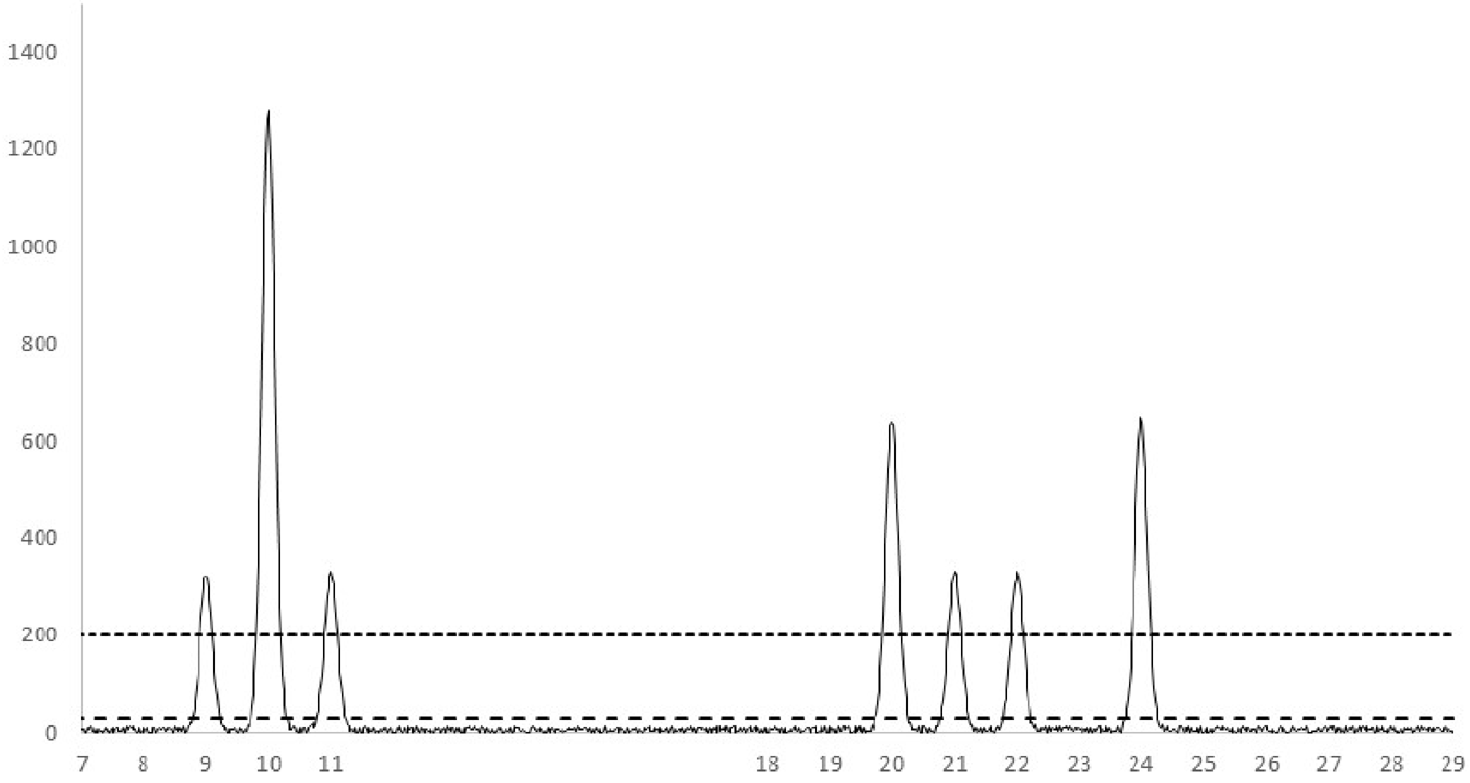
Based on a set of input parameters, an expected profile can be generated.

Step 1: Genotypes are chosen.



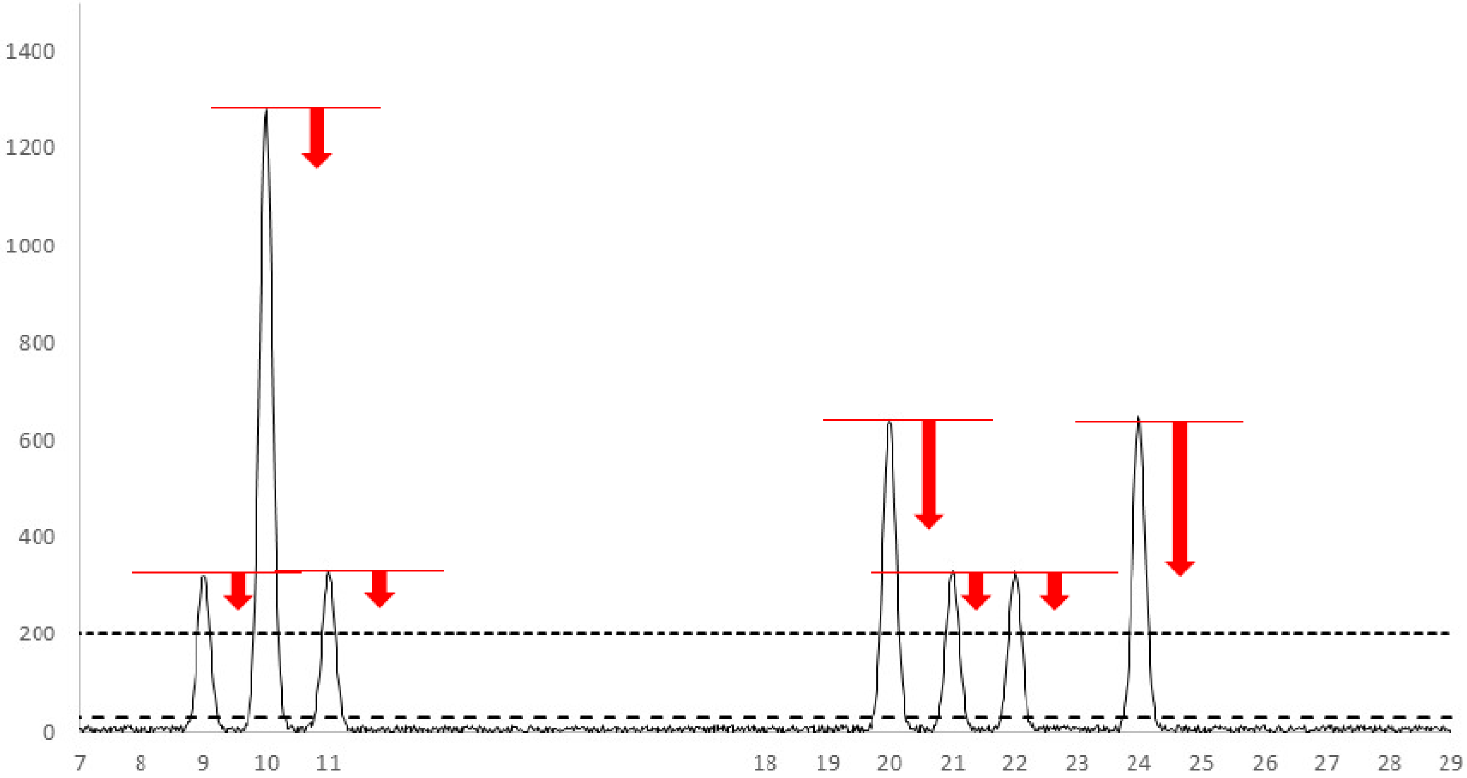
Expected Peak Heights

Step 2: Template amounts per contributor are incorporated.



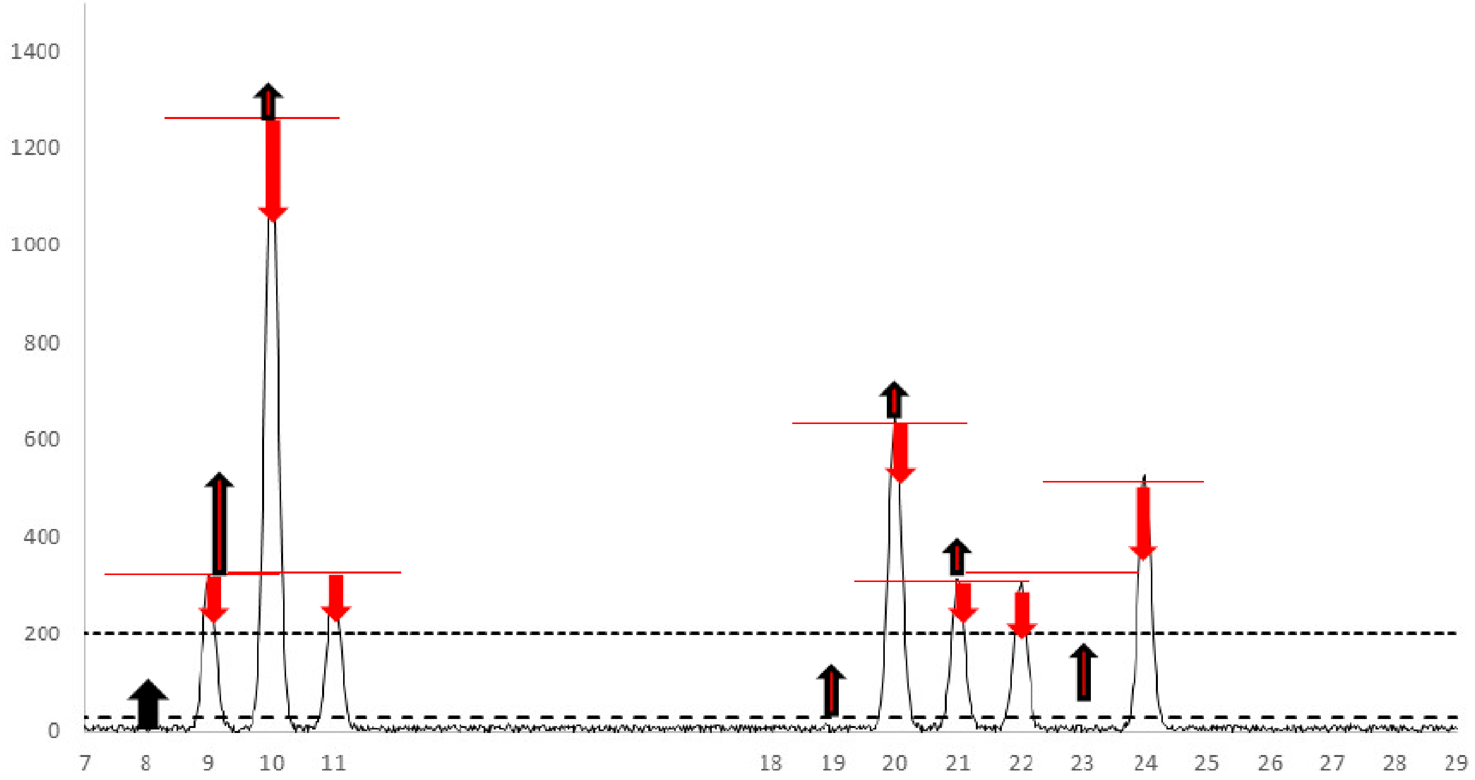
Expected Peak Heights

Step 3: Degradation is taken into account.



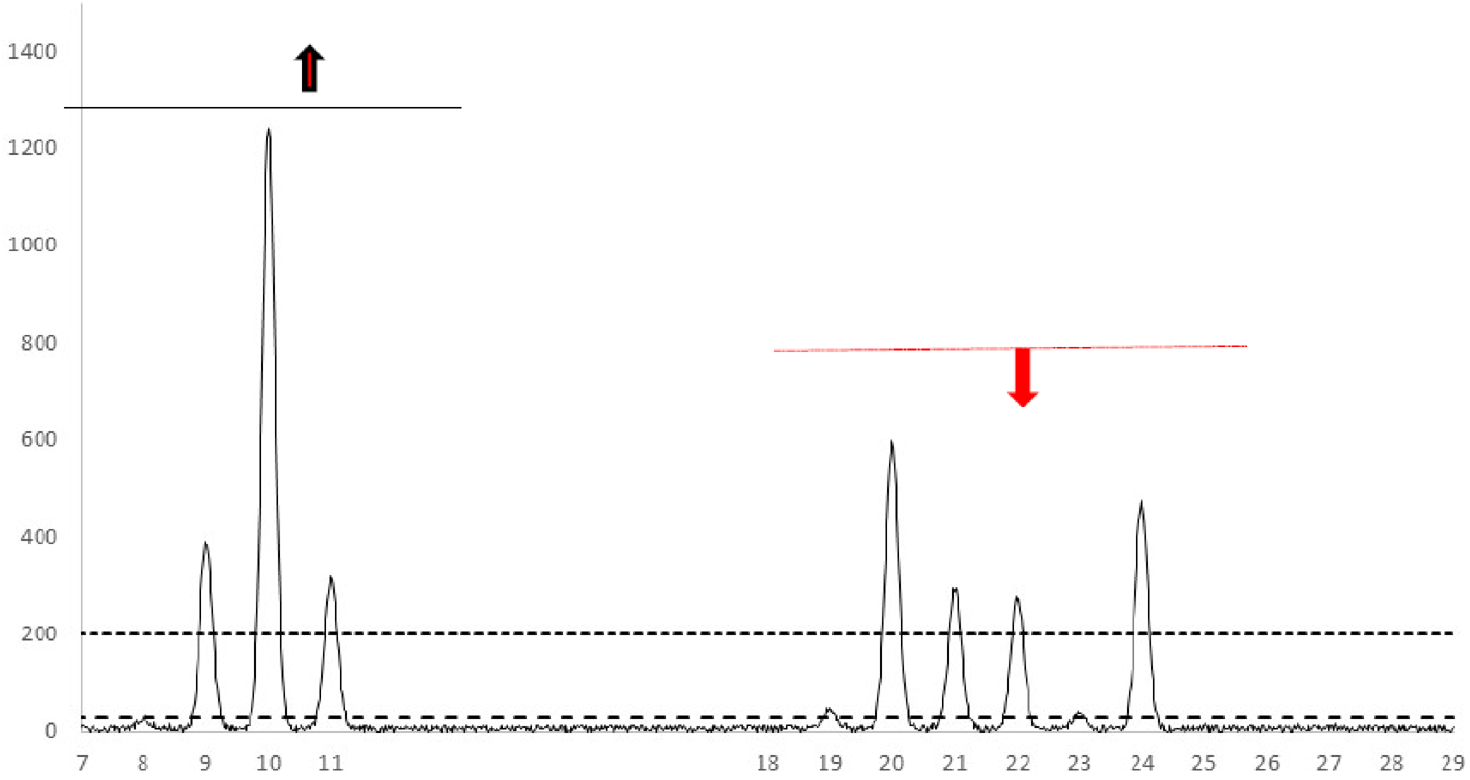
Expected Peak Heights

Step 4: Stutter is taken into account.



Expected Peak Heights

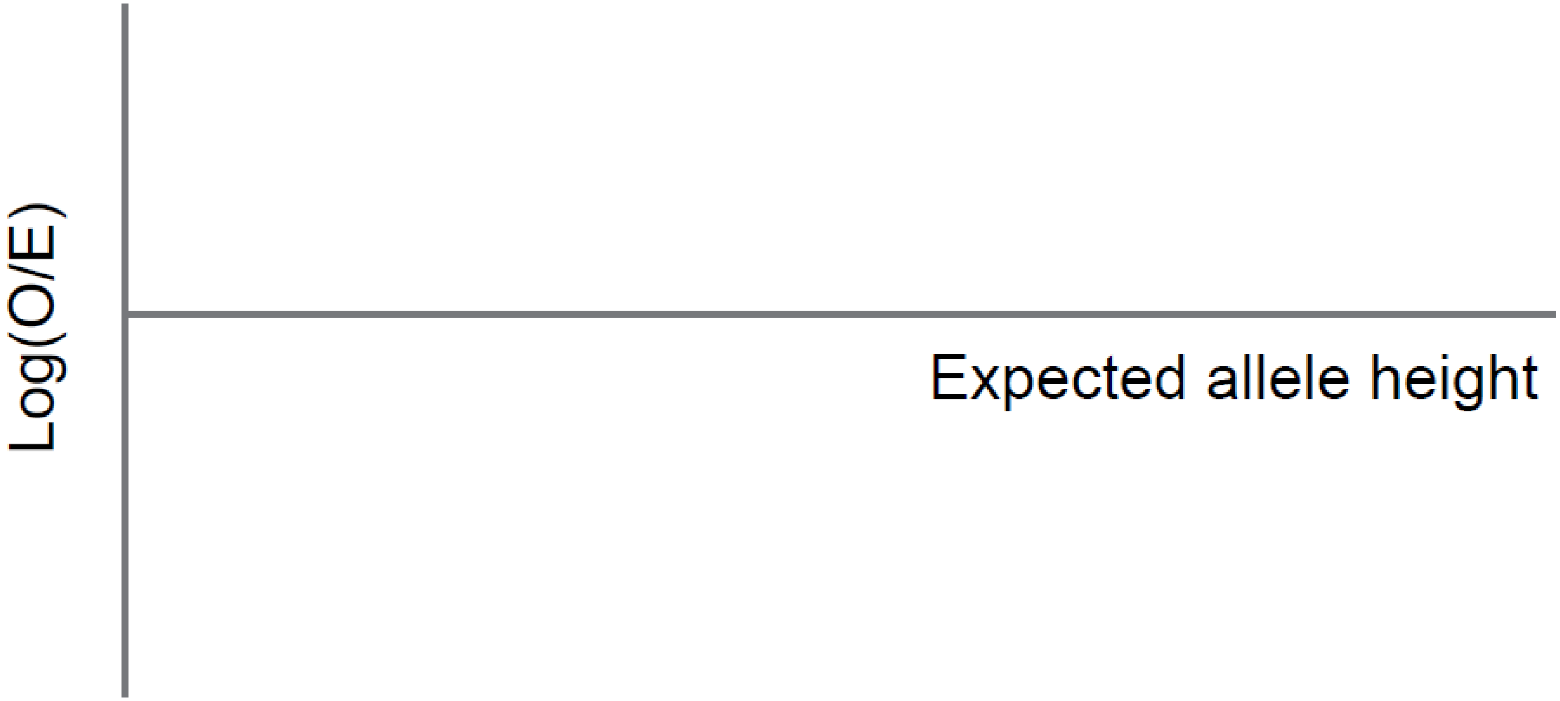
Step 5: Locus specific amplification efficiencies are introduced.



The Perfect Model

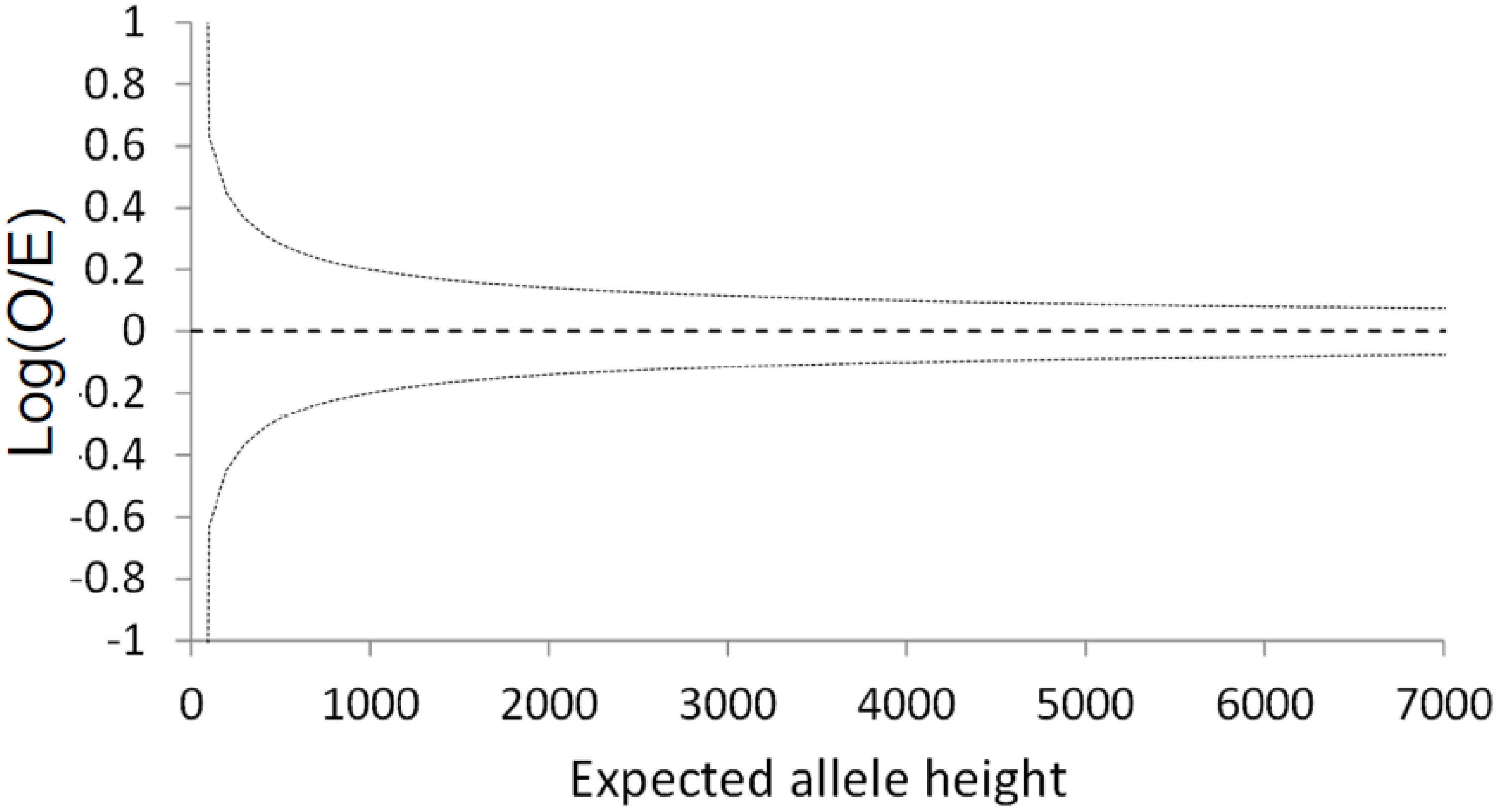
We can now compare our expected profile with the observed STR profile.

What would a perfect model look like?



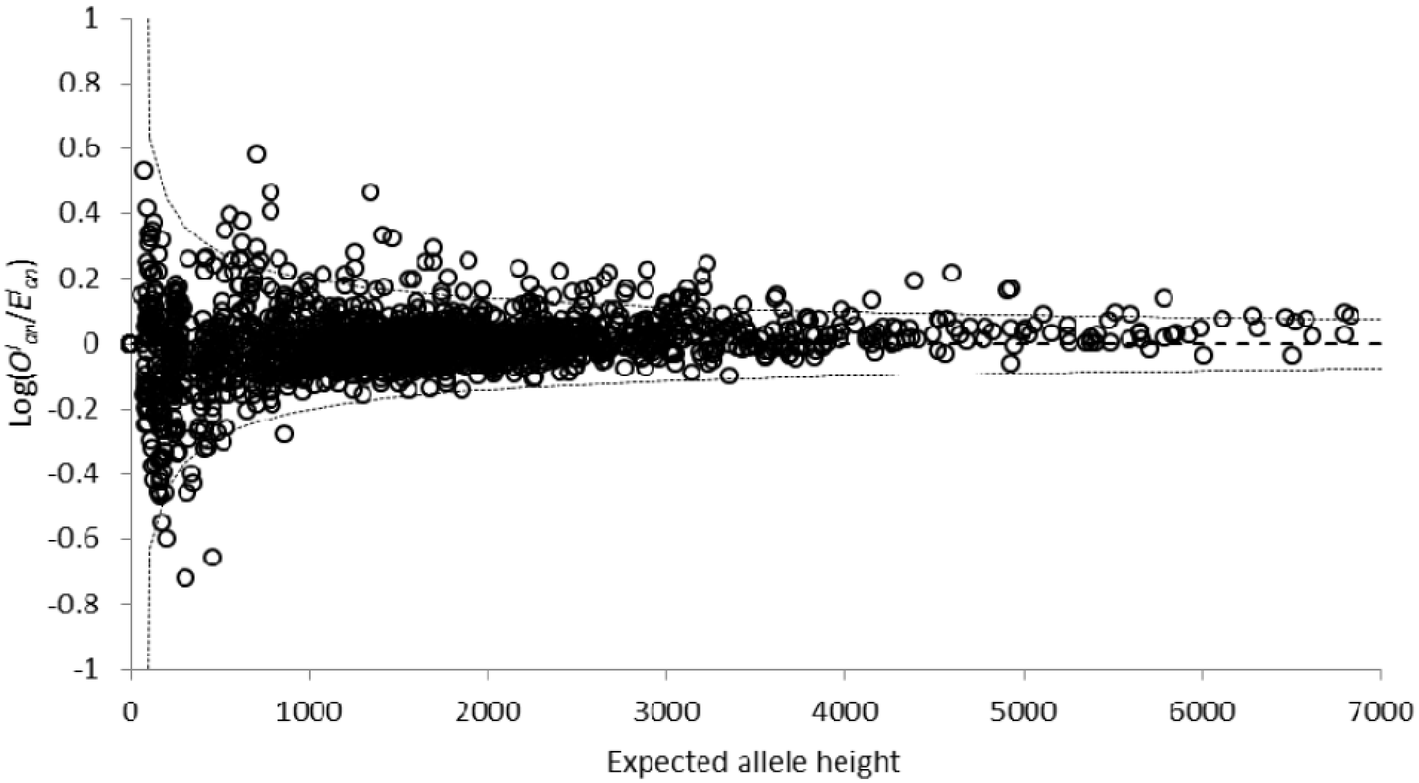
The Perfect Model

Observations show that the relative variance of small peaks is large and the relative variance of large peaks is small. This suggests that the variance is inversely proportional to the expected peak height.



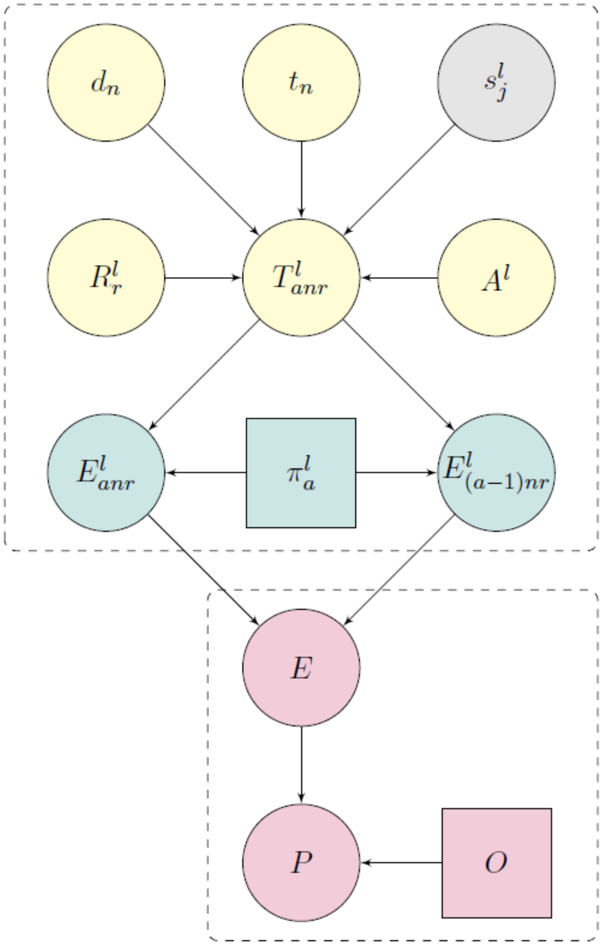
Generating Weights

The weights can now be calculated by considering the ratio of the observed and expected peak heights, assuming the log of this ratio has mean 0 and variance proportional to $1/E$.



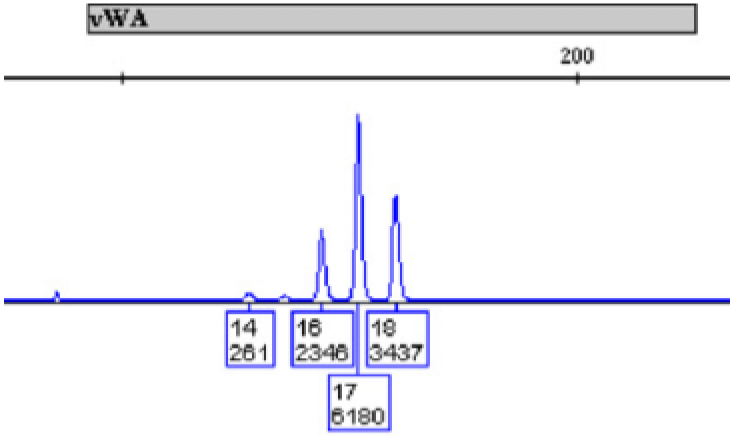
Continuous Model Network

Combining all elements leads to an overall continuous model network:



Worked Example for the Continuous Model

The epg for a 3-person mixture at locus vWA is as follows:



We would like to assess the LR under the hypothesis that:

H_p : $G_S = 17, 18$ and 2U are the source of the sample.

H_d : 3U are the source of the sample.

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

Worked Example for the Continuous Model

The LR can now be assessed by writing the ratio in the form:

$$\mathsf{LR} = \frac{\mathsf{Pr}(G_C|G_S, H_p, I)}{\mathsf{Pr}(G_C|G_S, H_d, I)}$$

$$= \frac{\sum_j \mathsf{Pr}(G_C|S_j)\,\mathsf{Pr}(S_j|H_p)}{\sum_{j'} \mathsf{Pr}(G_C|S_{j'})\,\mathsf{Pr}(S_{j'}|H_d)}$$

$$= \frac{\sum_j w_j\,\mathsf{Pr}(S_j|H_p)}{\sum_{j'} w_{j'}\,\mathsf{Pr}(S_{j'}|H_d)}.$$

The two propositions each define sets of genotypes S, and the weights w describe how well these sets fit our observed data G_C. Under H_p all the genotype sets S_j usually include G_S.

Worked Example for the Continuous Model

Suppose the following weights have been established for locus vWA:

Genotype Set	Donor 1	Donor 2	Donor 3	Weight
S_1	$16, 18$	$17, 17$	$14, 14$	0.00045
S_2	$16, 18$	$17, 17$	$14, 15$	0.00017
S_3	$16, 16$	$17, 17$	$14, 16$	0.00008
S_4	$16, 18$	$17, 17$	$14, 17$	0.00002
S_5	$16, 18$	$17, 17$	$14, 18$	0.00054
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_{15}	$16, 17$	$17, 18$	$14, 15$	0.15800
S_{16}	$16, 17$	$17, 18$	$14, 16$	0.28700
S_{17}	$16, 17$	$17, 18$	$14, 17$	0.21000
S_{18}	$16, 17$	$17, 18$	$14, 18$	0.11400
S_{19}	$17, 17$	$17, 18$	$14, 16$	0.00016

The actual reference profiles of the three known contributors are:

Locus	Donor 1	Donor 2	Donor 3
vWA	$16, 17$	$17, 18$	$14, 16$

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

Worked Example for the Continuous Model

Under H_p only the genotype sets containing G_S are relevant:

Set	Donor 1	Donor 2	Donor 3	Weight	$\Pr(S_j \mid H_p)$
S_1	$16, 18$	$17, 17$	$14, 14$	0.00045	0
S_2	$16, 18$	$17, 17$	$14, 15$	0.00017	0
S_3	$16, 16$	$17, 17$	$14, 16$	0.00008	0
S_4	$16, 18$	$17, 17$	$14, 17$	0.00002	0
S_5	$16, 18$	$17, 17$	$14, 18$	0.00054	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_{15}	$16, 17$	$17, 18$	$14, 15$	0.15800	$2p_{16}p_{17} \cdot 2p_{14}p_{15}$
S_{16}	$16, 17$	$17, 18$	$14, 16$	0.28700	$2p_{16}p_{17} \cdot 2p_{14}p_{16}$
S_{17}	$16, 17$	$17, 18$	$14, 17$	0.21000	$2p_{16}p_{17} \cdot 2p_{14}p_{17}$
S_{18}	$16, 17$	$17, 18$	$14, 18$	0.11400	$2p_{16}p_{17} \cdot 2p_{14}p_{18}$
S_{19}	$17, 17$	$17, 18$	$14, 16$	0.00016	$p_{17}^2 \cdot 2p_{14}p_{16}$

Note that these calculations can be modified to allow for population substructure.

Multiplication of the weights with the probabilities, and summing over them, results in the numerator of the LR $\Pr(E \mid H_p)$.

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

Worked Example for the Continuous Model

Using allele frequencies (in this case from an Australian Caucasian sub-population):

Allele	Frequency
14	0.1146
15	0.1071
16	0.2044
17	0.2726
18	0.2090

yields: $\text{Pr}(E|H_p) = 4.4 \times 10^{-3}$. Similarly, we can calculate the probabilities under H_d, now considering all genotype sets and corresponding donors, we get: $\text{Pr}(E|H_d) = 5.0 \times 10^{-4}$.

Combining this gives us the LR for this specific locus:

$$\text{LR} = \frac{\text{Pr}(E|H_p)}{\text{Pr}(E|H_d)} = \frac{4.4 \times 10^{-3}}{5.0 \times 10^{-4}} = 8.8$$

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

Worked Example for the Continuous Model

The overall LR is a combination of all loci (here compared with a binary model):

Locus	**LR**$_B$	**LR**$_C$
D10S1248	0.97	4.69
vWA	1.24	8.21
D16S539	0.45	5.32
D2S1338	2.27	31.22
D8S1179	0.51	7.79
D21S11	0.94	9.98
D18S51	3.85	52.08
D22S1045	4.32	59.18
D19S433	0.92	7.17
TH01	0.97	13.31
FGA	1.39	21.14
D2S441	0.65	4.84
D3S1358	0.93	13.22
D1S1656	5.55	106.14
D12S391	1.42	21.34
SE33	6.23	69.53
Overall LR	356	3.13×10^{19}

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

Available Software

The Scientific Working Group on DNA Analysis Methods (SWGDAM) defines probabilistic genotyping as

> *". . . the use of biological modeling, statistical theory, computer algorithms, and probability distributions to calculate likelihood ratios (LRs) and/or infer genotypes for the DNA typing results of forensic samples ("forensic DNA typing results")".*

Over the years, several probabilistic genotyping programs have been developed across the globe, ranging from commercial packages to open-source platforms, with the main goal to interpret complex DNA mixtures for CE data.

Available Software

Not all models as published in literature have been translated into software. A non-exhaustive list:

Software	Class	Availability	Optimization
LRmix Studio	semi-continuous	open-source	ML
Lab Retriever	semi-continuous	open-source	ML
MixKin	semi-continuous	in-house	Integration
DNA LiRA	(semi-)continuous	open-source	Bayes
likeLTD	(semi-)continuous	open-source	ML
STRmix	continuous	commercial	Bayes
TrueAllele	continuous	commercial	Bayes
DNA·VIEW	continuous	commercial	ML
DNAmixtures	continuous	open-source*	ML
EuroForMix	continuous	open-source	ML or Bayes
DNAStatistX	continuous	in development	ML

See also: Probabilistic Genotyping Software: An Overview (Coble & Bright, 2019).

Available Software - Discussion

There are no ground truths for probabilistic genotyping calculations. Moreover, the 2016 PCAST (President's Council of Advisors on Science and Technology) report stated:

> *"[w]hile likelihood ratios are a mathematically sound concept, their application requires making a set of assumptions about DNA profiles that require empirical testing. Errors in the assumptions can lead to errors in the results".*

- Under what circumstances have the methods been validated? What are their limitations?

- Commercial software has received criticism regarding their black-box nature. Should source code be made accessible (to the defense)?

Available Software - Discussion

What about the consistency between software programs when they examine the same evidence?

Method	Sample A	Sample B	Sample C
LRmix Studio	1.29	1.85×10^{14}	0.0212
Lab Retriever	1.20	1.89×10^{14}	0.0241
DNA·VIEW	1.09×10^{-14}	4.66×10^{11}	2.24×10^{8}
Combined	Inconclusive	Support to H_p	Inconclusive

Another example can be found in the *People v. Hillary* (NY) case: TrueAllele reported no statistical support for a match (LR < 0), whereas STRmix inculpated the defendant with a likelihood ratio of 360 000. The evidence consisted of an LTDNA sample with an extreme mixture ratio.

Source: An alternative application of the consensus method to DNA typing interpretation for Low Template-DNA mixtures (Garofano et al., 2015).