Forensic Genetics

Module 19 – Topic 4

Schedule – Day 2

Topic 3	Allelic Independence	8:00-8:50
Topic 4	DNA Interpretation and modeling	9:05-9:55
Review	Topics 3 and 4 exercises	10:10-11:00
Topic 5	Population structure and relatedness	11:30-12:20
Review	Topic 5 exercises	12:35-1:25
Topic 6	Reporting likelihood ratios	1:40-2:30

Topic 4 - DNA Interpretation and Modeling

- Thresholds and Modeling Types
  - Binary model
  - Semi-continuous model
  - Continuous model

- Peak Height Modeling
  - Total Allelic Peak Height
  - *Degradation*
  - Stutter
  - *Heterozygote Balance*

- Likelihood Ratio Modeling
  - *Markov Chain Monte Carlo*
  - Probabilistic Genotyping Software

Thresholds

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

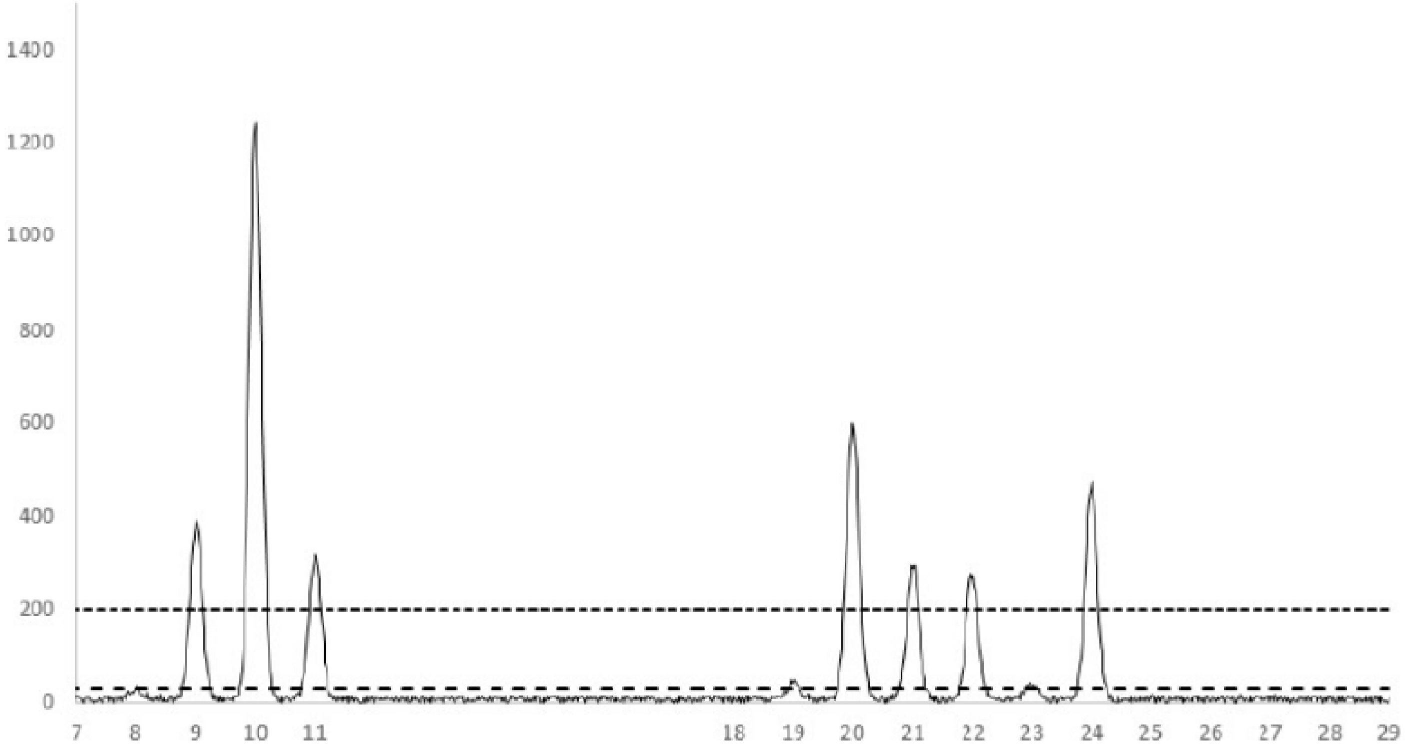
- **High thresholds**: will reduce the number of artifacts and remove a lot of background noise. However, it may potentially lead to a number of drop-outs.

- **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.

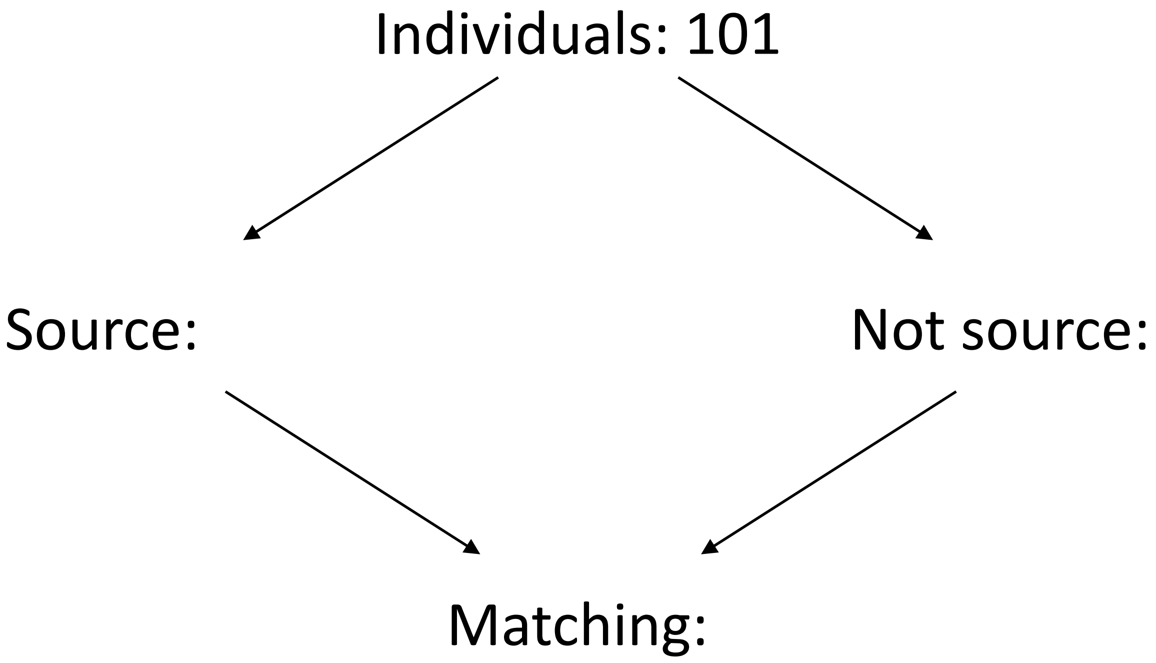
- The DNA profiles of all the other individuals are unknown.

- 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

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?



Individuals: 101

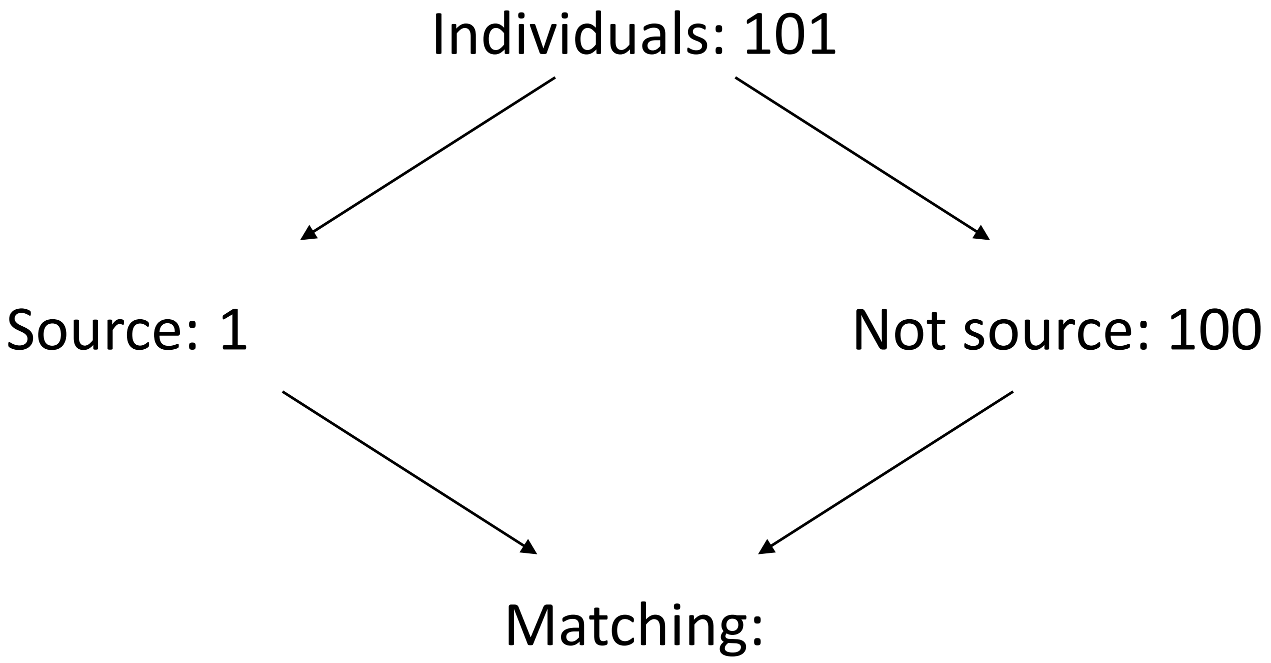
Source:

Not source:

Matching:

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?



Individuals: 101

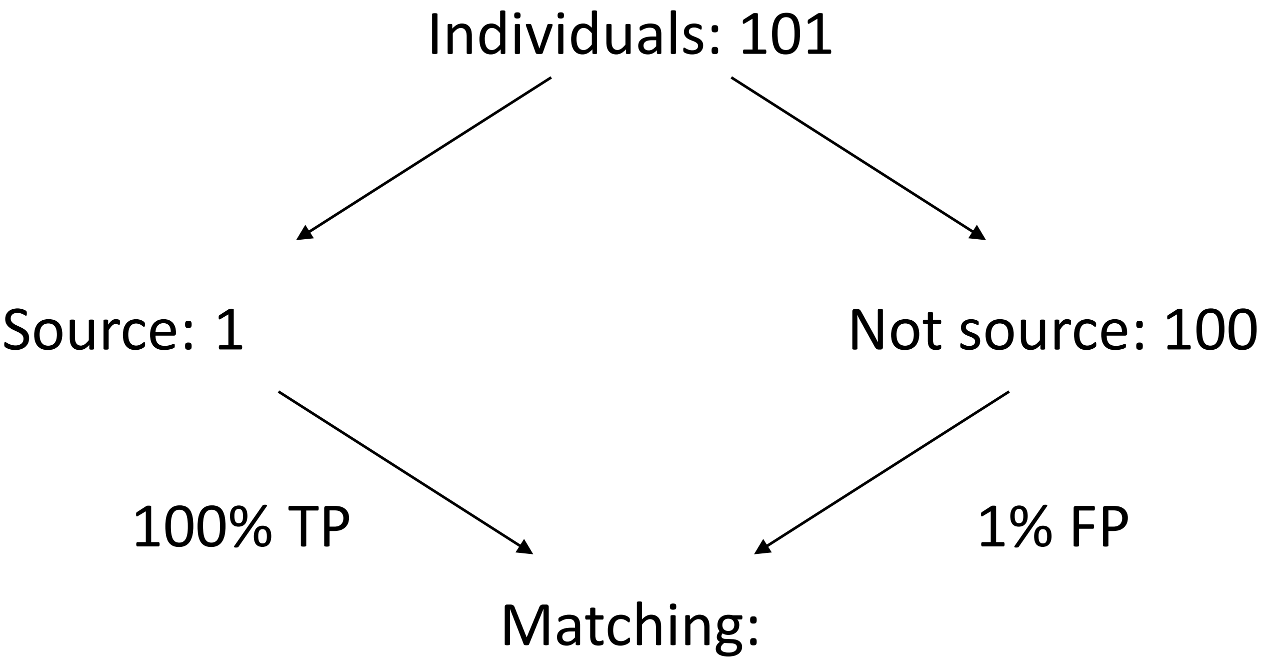
Source: 1

Not source: 100

Matching:

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?


Individuals: 101

Source: 1

Not source: 100

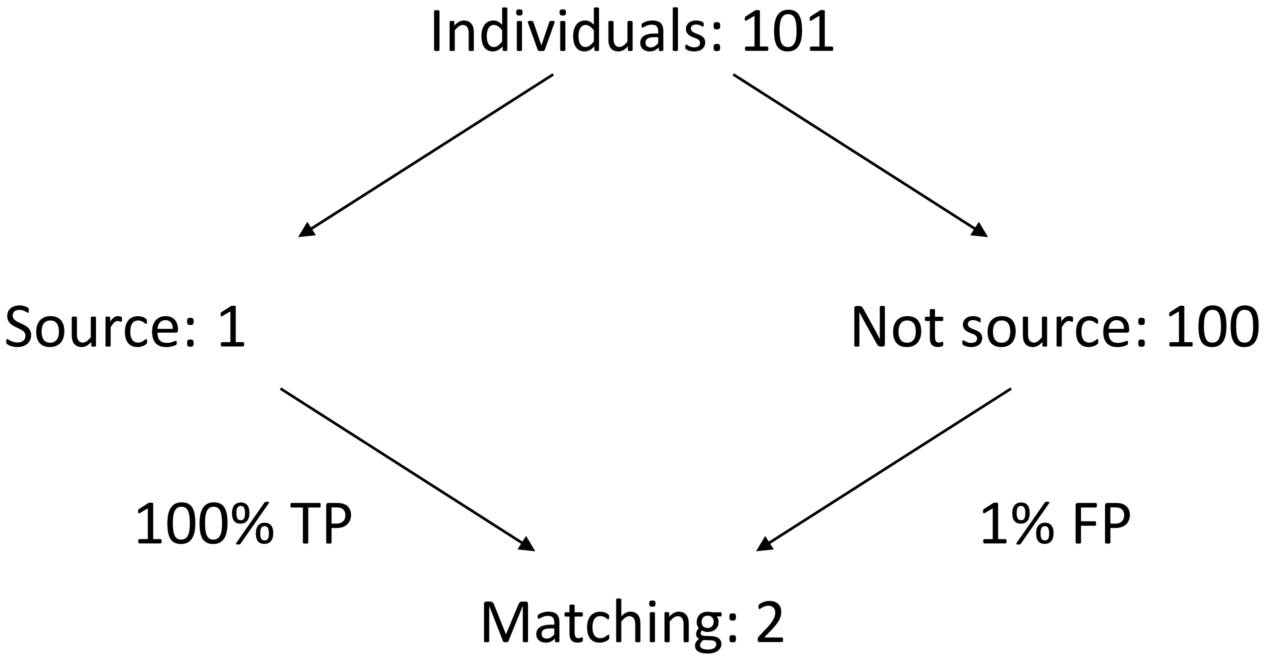
100% TP

1% FP

Matching:

The Island Problem

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

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?

Zoom Poll

How does knowledge about Q being identified through a search affect the probability of being the source?

- It decreases
- It increases
- It stays the same

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:

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

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

and for heterozygotes:

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

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

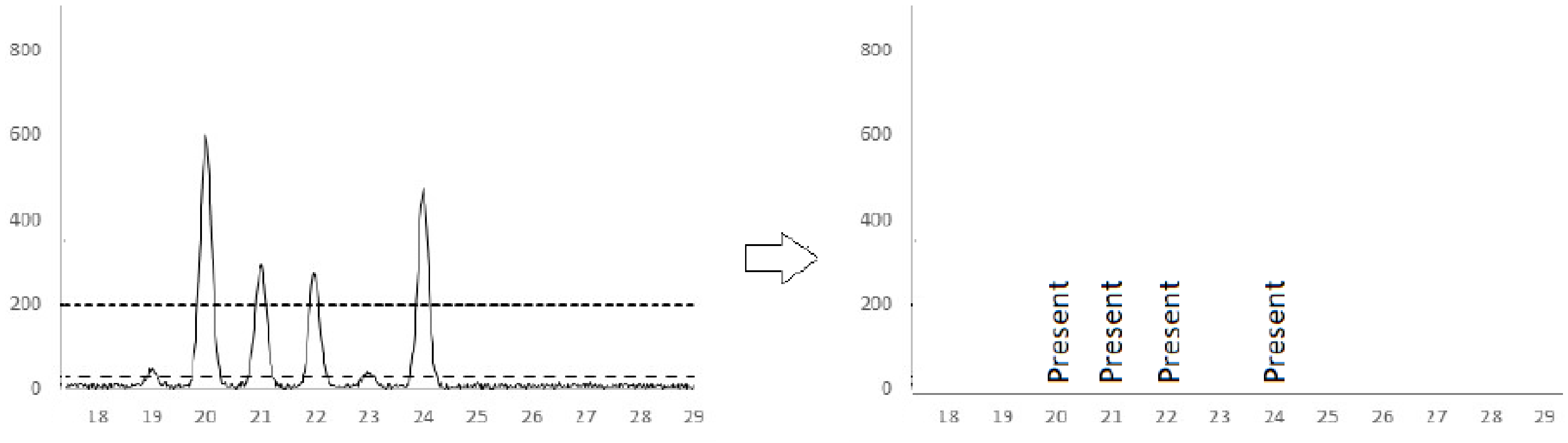
LR Modeling

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).



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.

These models will be of value when quantitative data is not available (e.g. old cases may only consist of allelic profiles).

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:

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

Estimating Drop-in and Drop-out

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

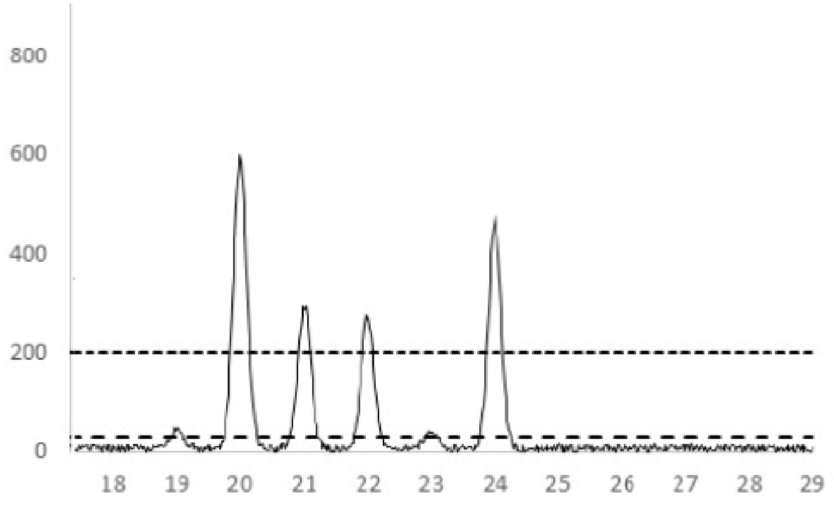
- Several models have been proposed for modeling drop-out probabilities, such as a multidose drop-out model and degradation model. Laboratory trials can be used to choose α when modeling $D_2 = \alpha D^2$, with $0 < \alpha \leq 1$. Instead of assigning probabilities to the drop-out rate they can be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

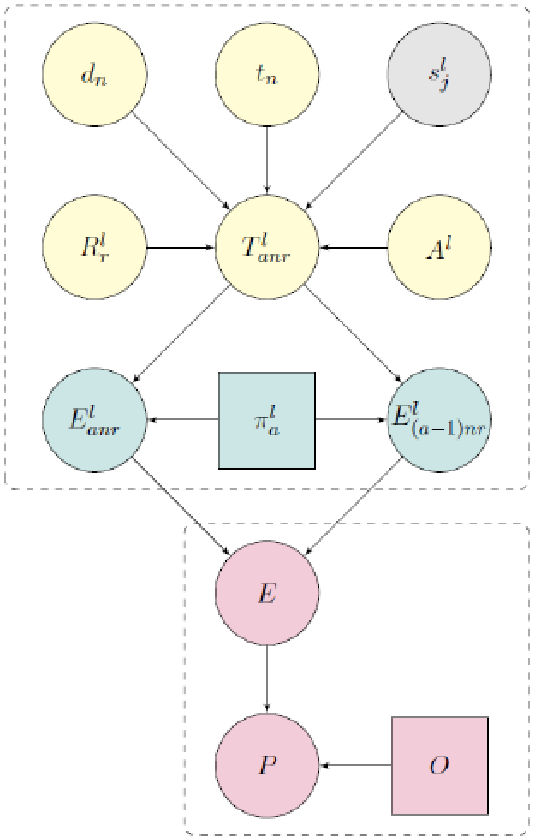
The key point of a fully continuous model is that it considers peak heights as a continuous variable.



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

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

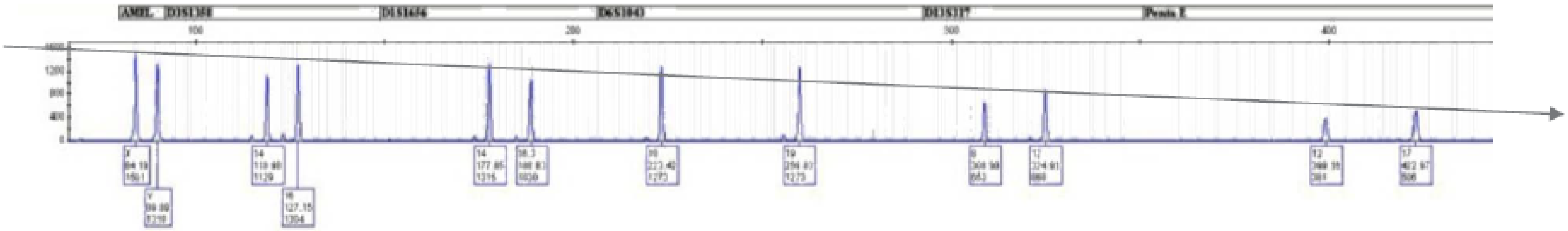
Peak heights can be modeled by defining the *total allelic product* (TAP), which will be a function of

- 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.

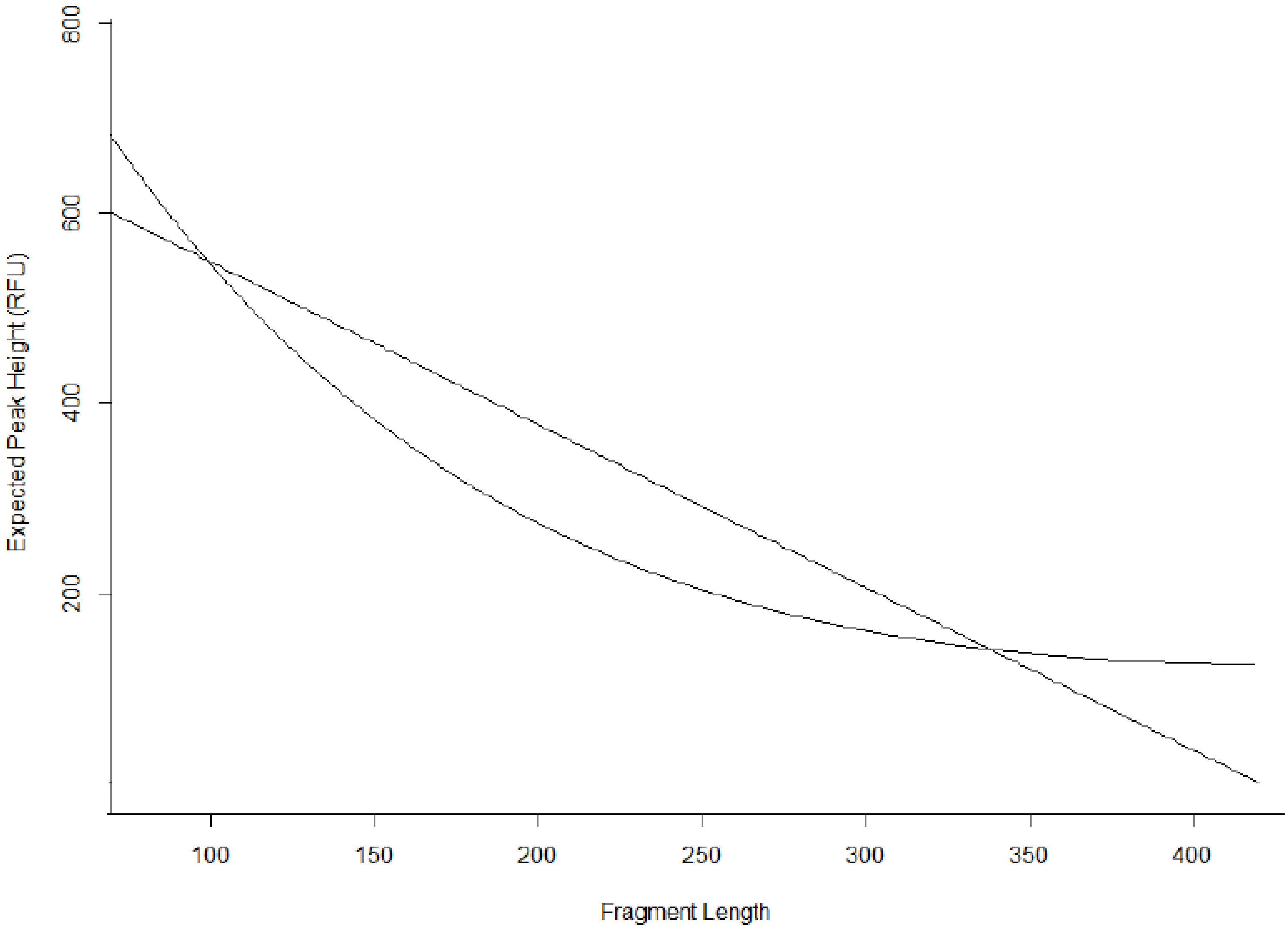
Modeling Degradation

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

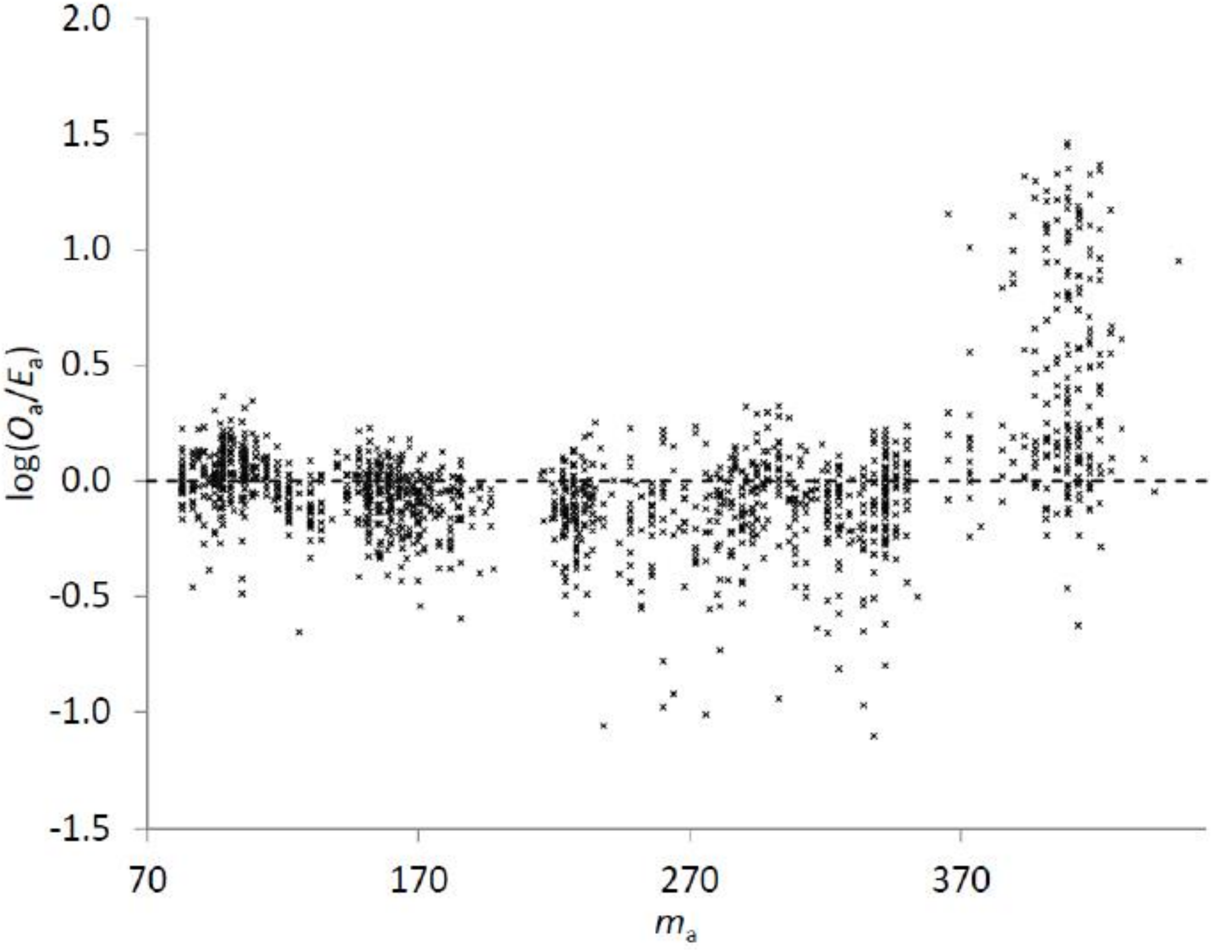


If we assume that the breakdown of a DNA strand is random with respect to location, an exponential model seems more reasonable.

Modeling Degradation

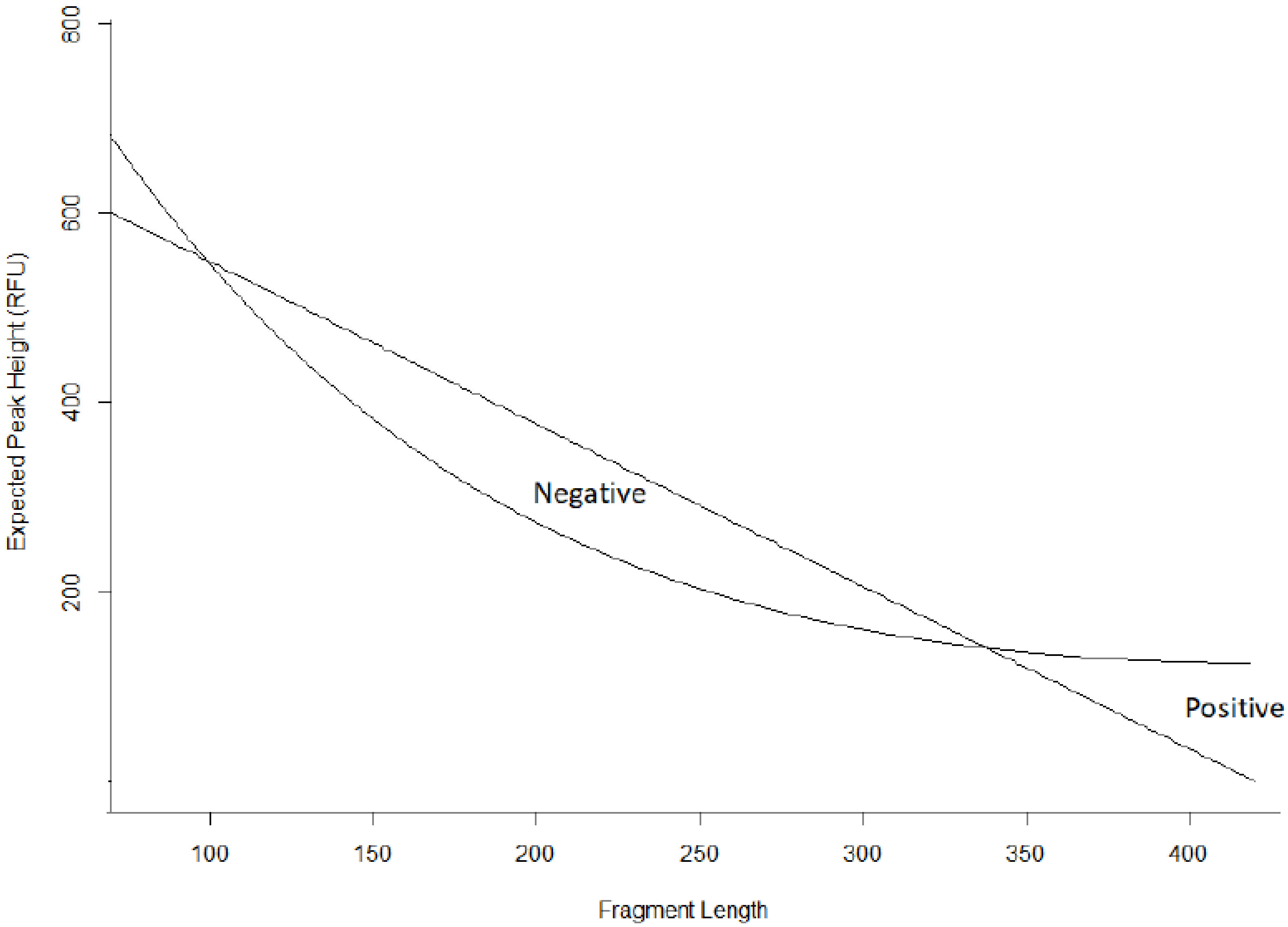


Modeling Degradation

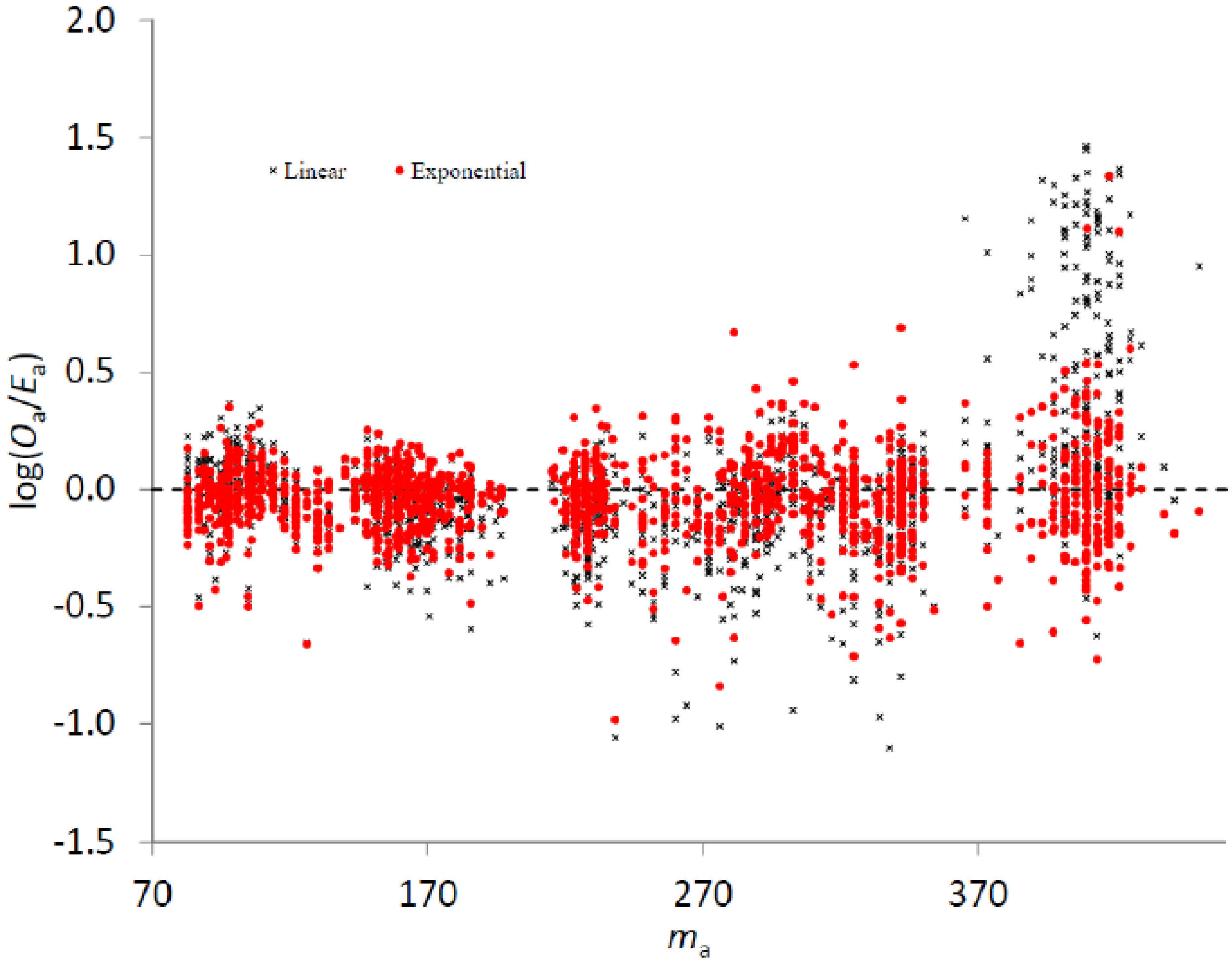


Source: Degradation of Forensic DNA Profiles (Bright et al., 2013).

Modeling Degradation



Modeling Degradation



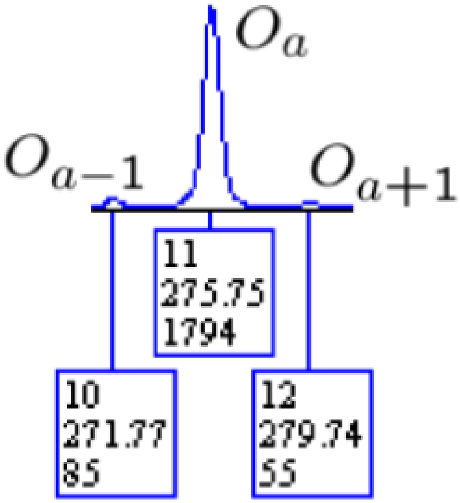
Source: Degradation of Forensic DNA Profiles (Bright et al., 2013).

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.

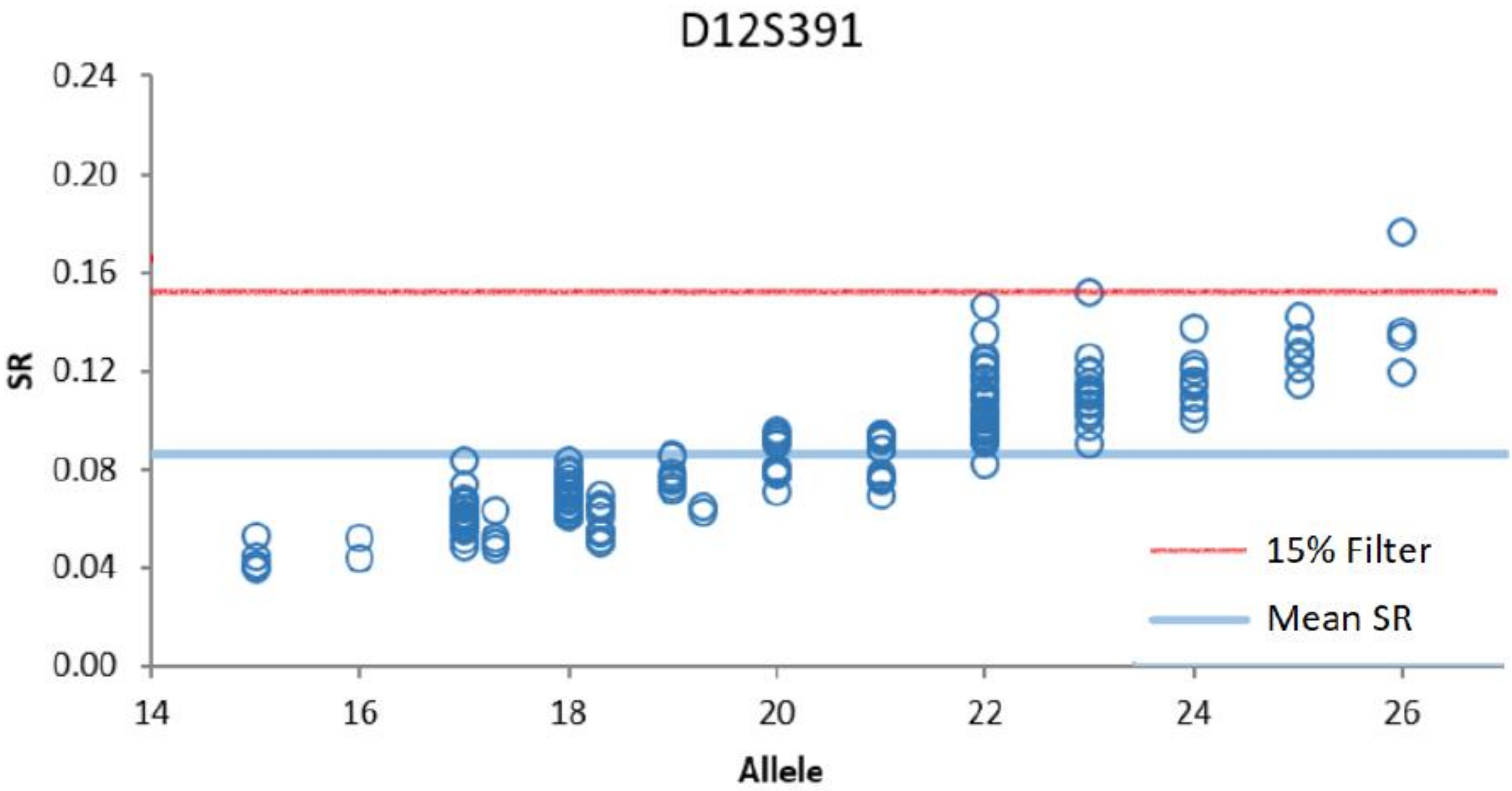
Stutter Modeling

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).

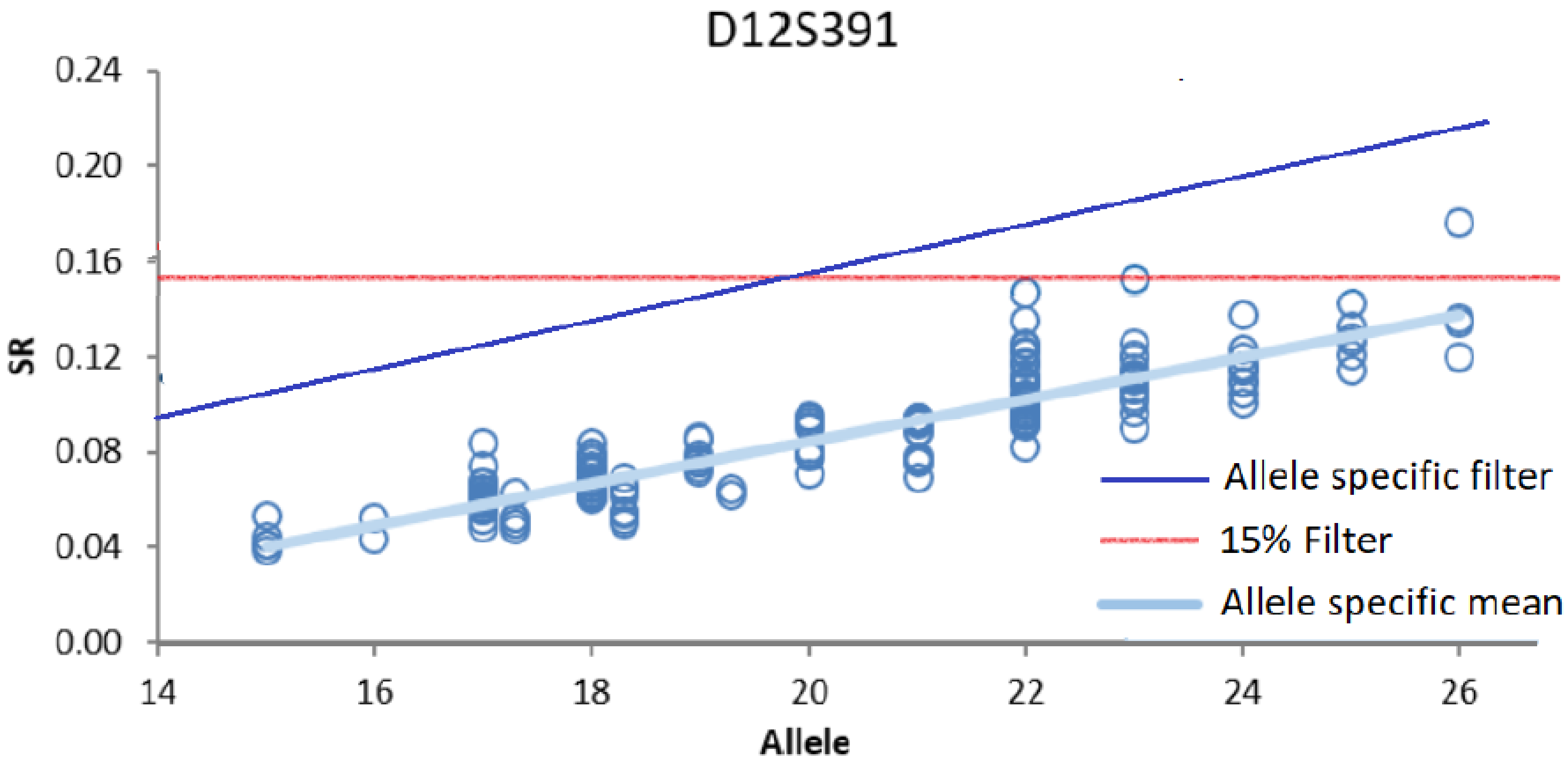
Stutter Modeling – Locus Specific Thresholds

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:

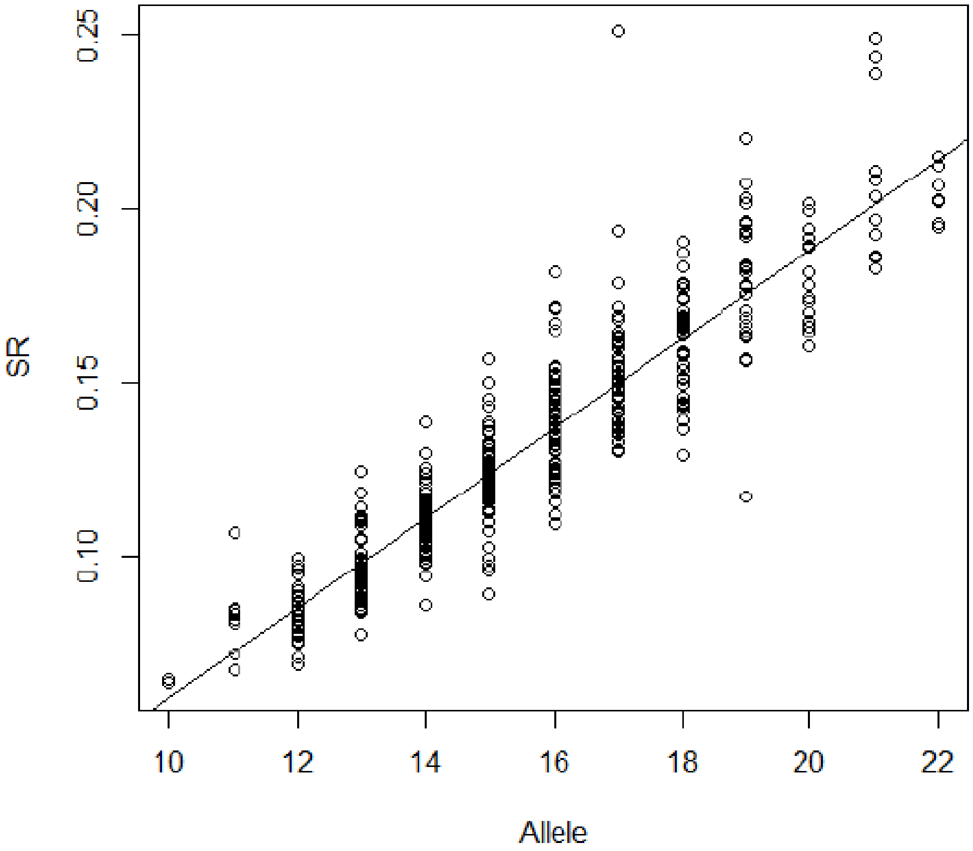
$$SR \sim \text{Allele number} \qquad \Rightarrow \qquad SR = ma + c,$$

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.

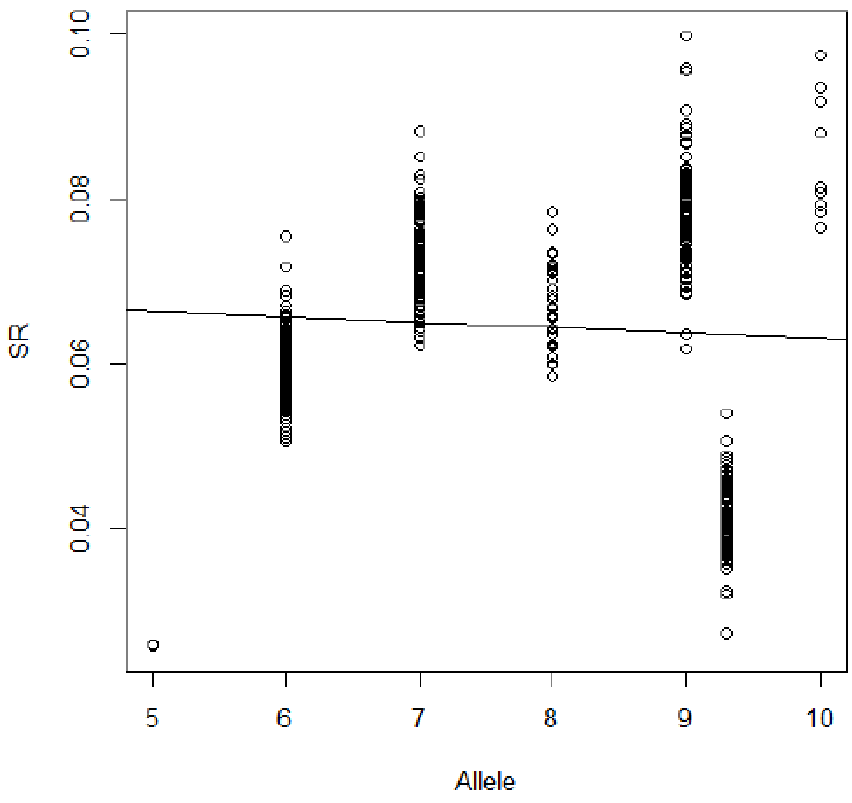
Stutter Modeling – Allele Model

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

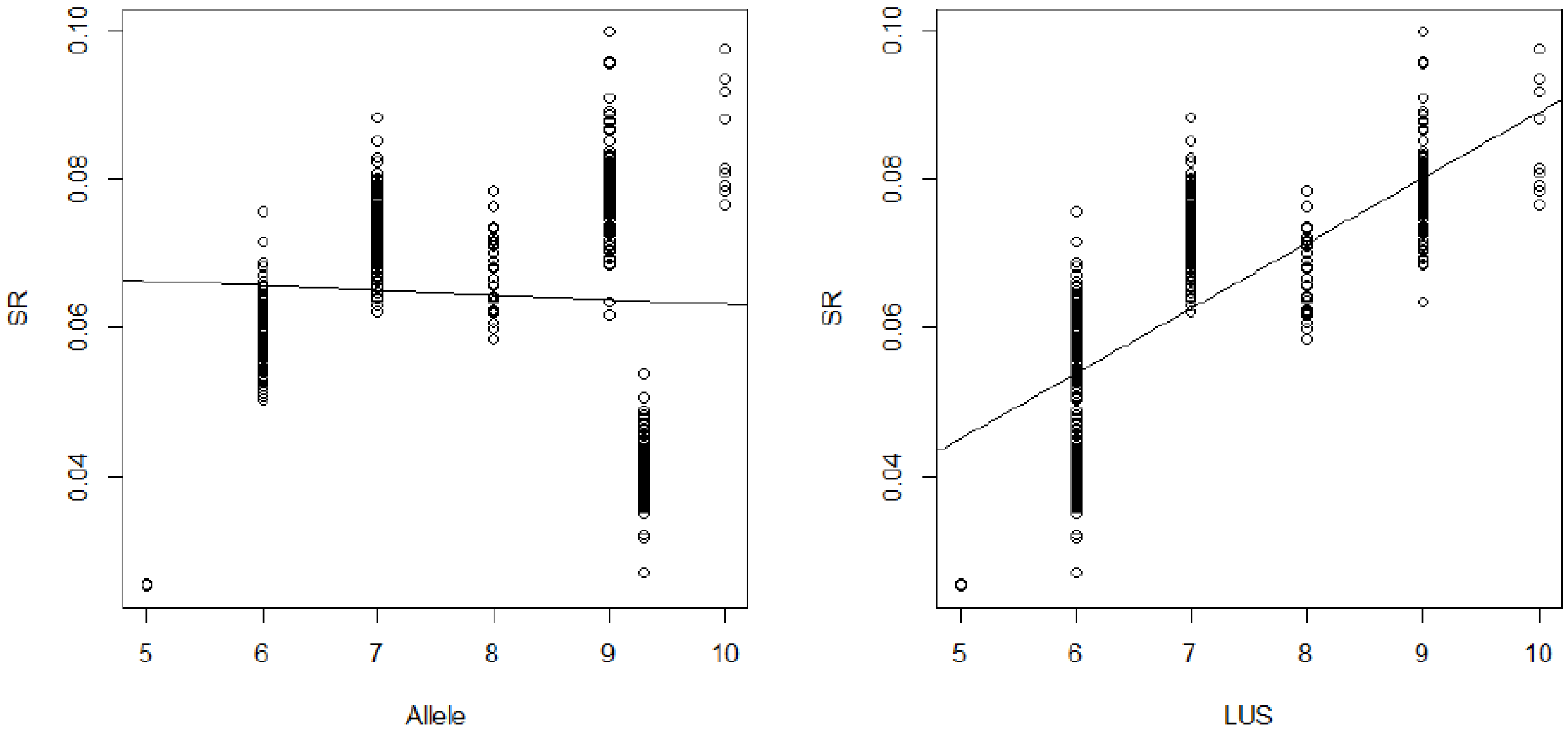
Stutter Modeling – LUS

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

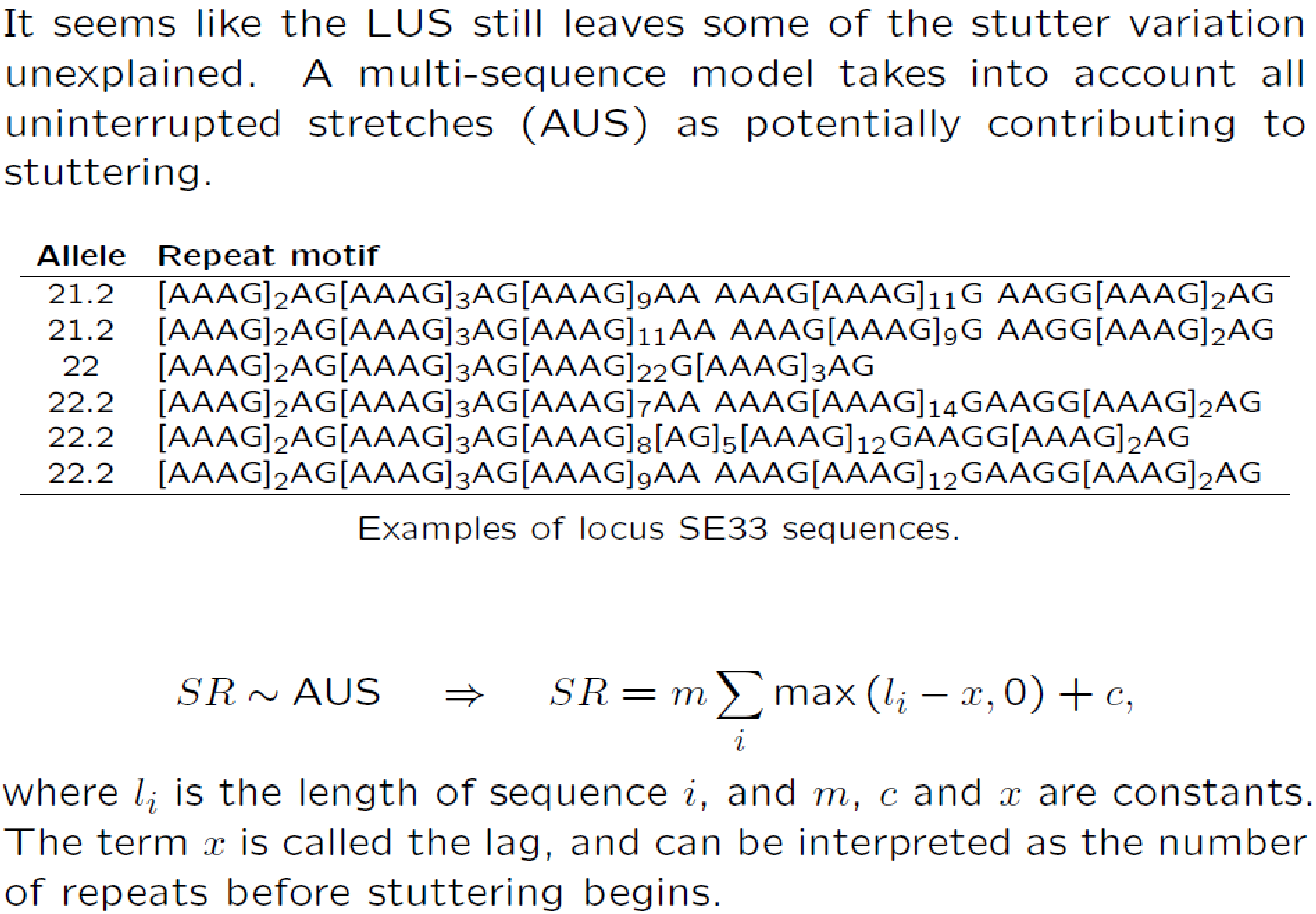
Common TH01 allele sequences.

Stutter Modeling – LUS Model



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.

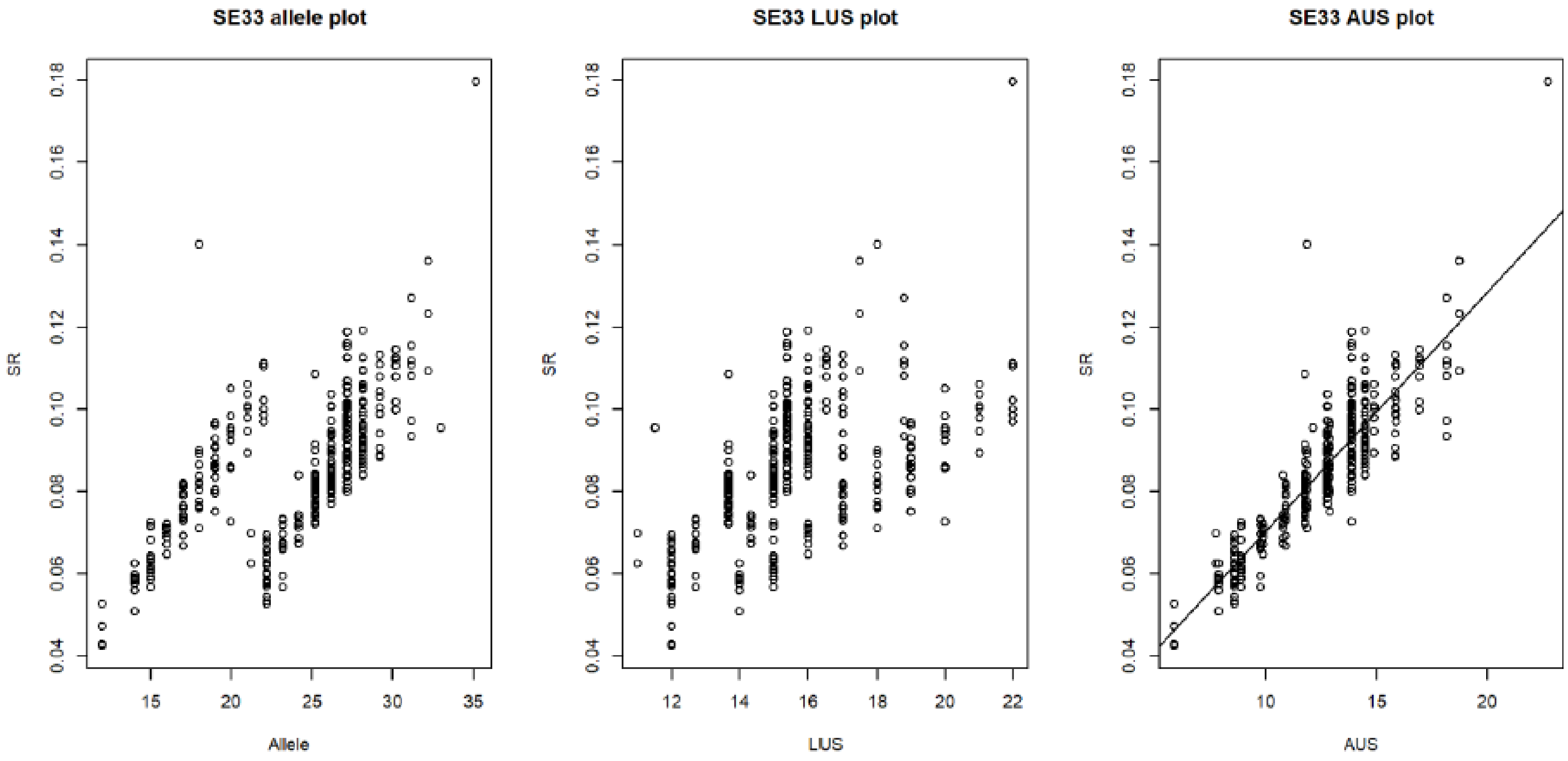
Allele	Repeat motif
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{11}G\ AAGG[AAAG]_2AG$
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{11}AA\ AAAG[AAAG]_9G\ AAGG[AAAG]_2AG$
22	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{22}G[AAAG]_3AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_7AA\ AAAG[AAAG]_{14}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_8[AG]_5[AAAG]_{12}GAAGG[AAAG]_2AG$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

Examples of locus SE33 sequences.

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

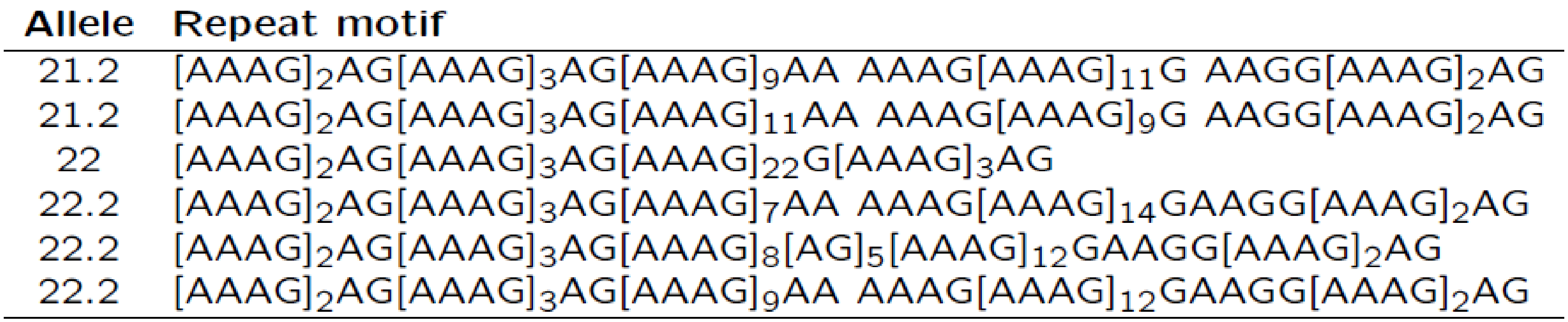
where l_i is the length of sequence i, and m, c and x are constants. The term x is called the lag, and can be interpreted as the number of repeats before stuttering begins.

Stutter Modeling – AUS Model



Stutter Modeling – AUS Model

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

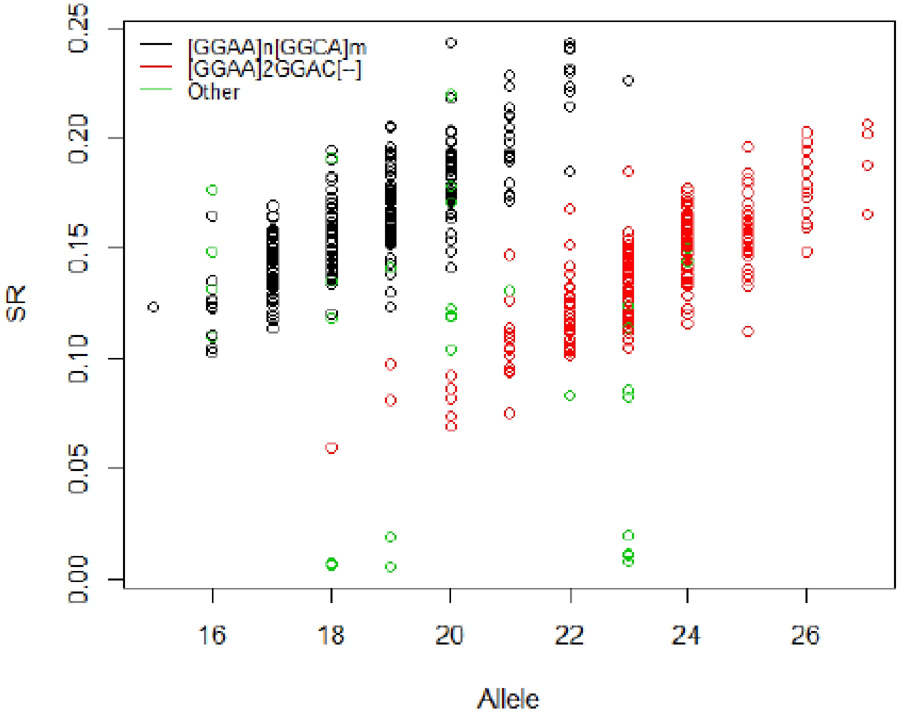


Allele	Repeat motif
21.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_9$AA AAAG[AAAG]$_{11}$G AAGG[AAAG]$_2$AG
21.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{11}$AA AAAG[AAAG]$_9$G AAGG[AAAG]$_2$AG
22	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_{22}$G[AAAG]$_3$AG
22.2	[AAAG]$_2$AG[AAAG]$_3$AG[AAAG]$_7$AA AAAG[AAAG]$_{14}$GAAGG[AAAG]$_2$AG
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

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.

Stutter Modeling

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

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

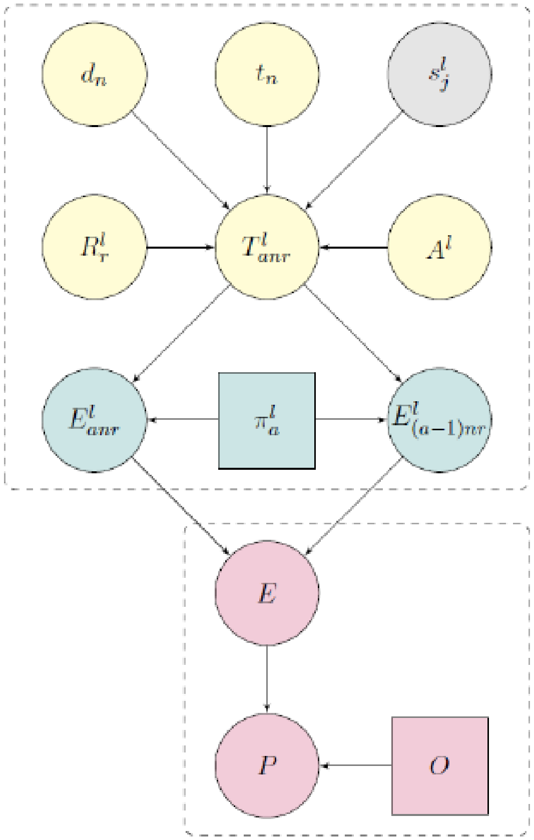
- What about other stutter products?

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.

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).

LR Modeling

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.

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.

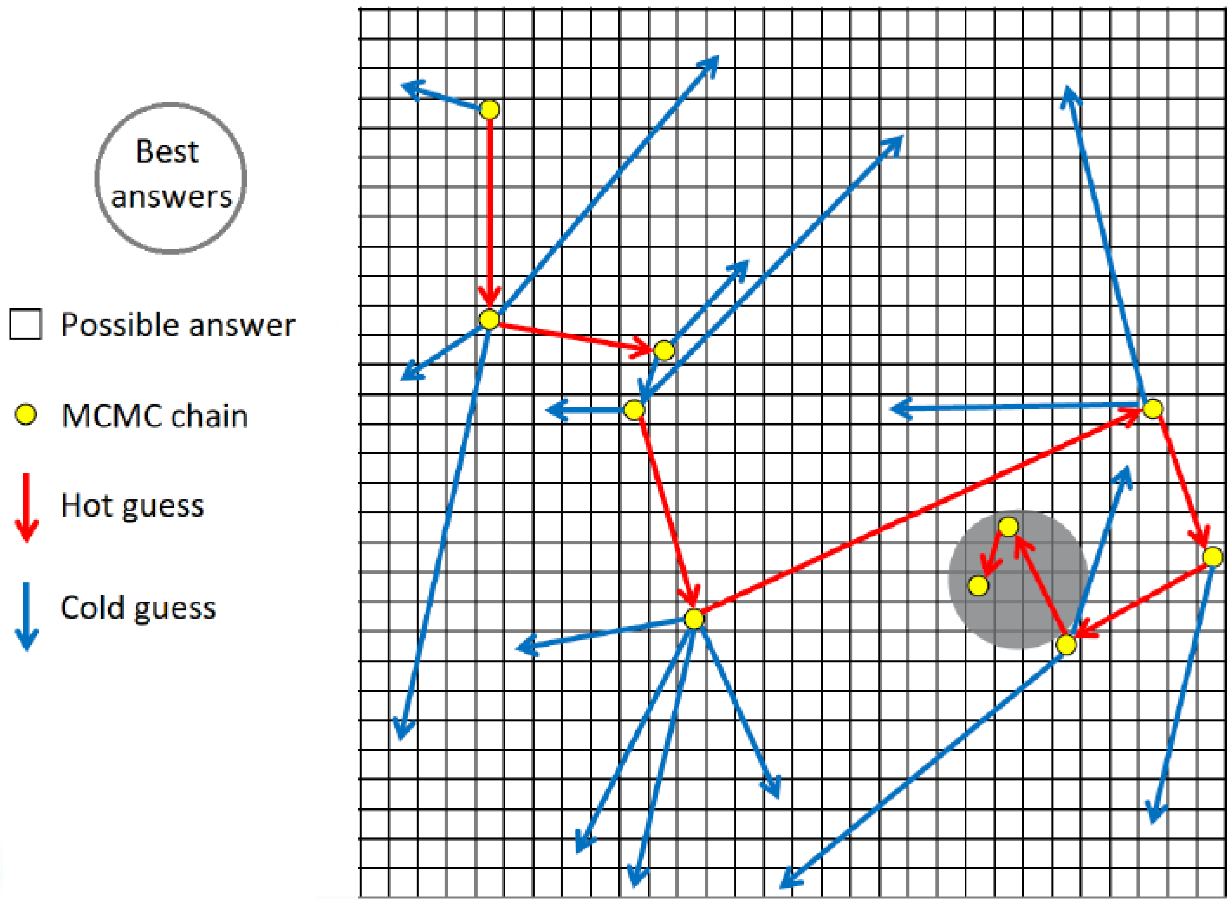
- **Maximization**: Parameters can be chosen that maximize the likelihood of the observations under each hypothesis.

- **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)

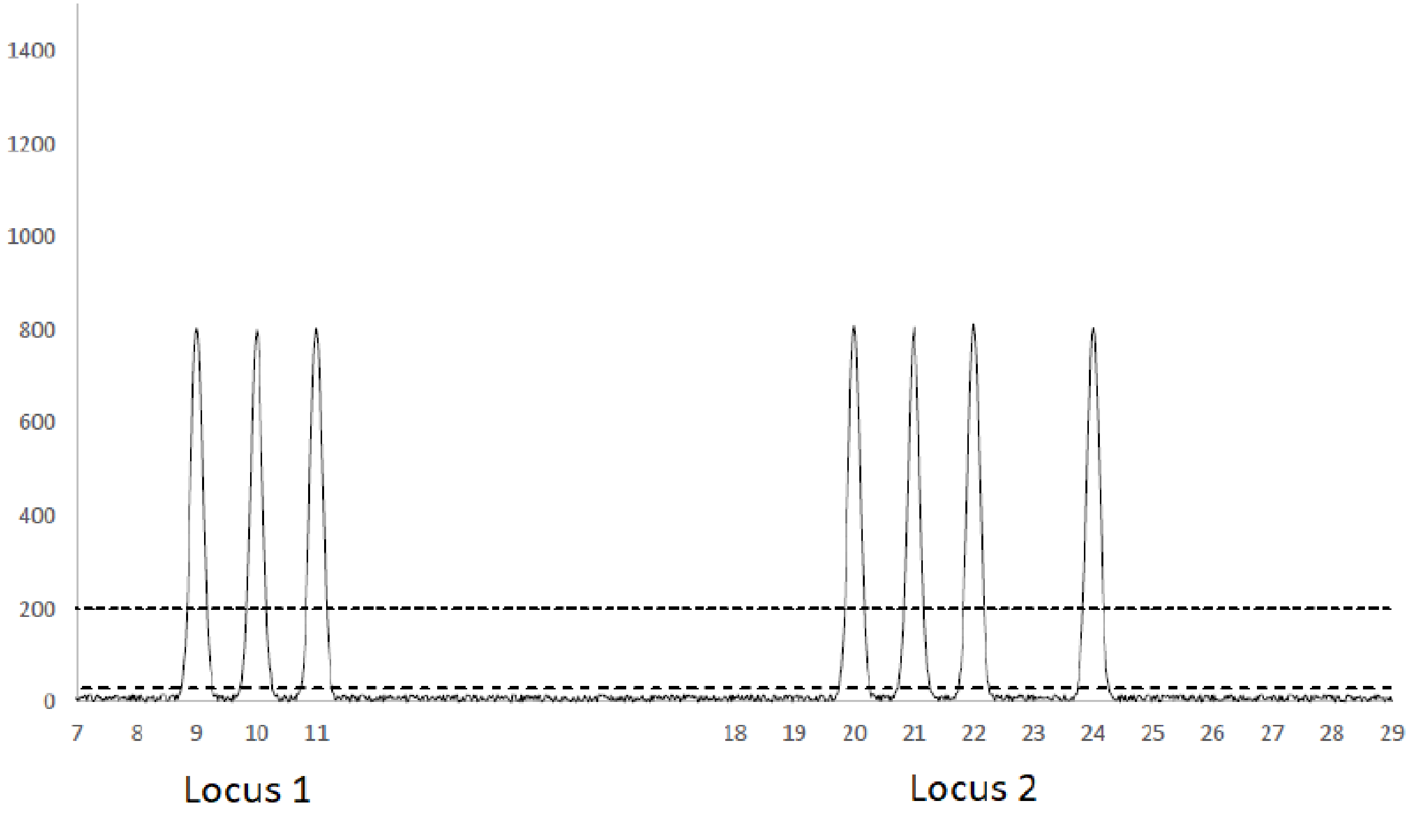
MCMC will start by choosing parameter values at random, eventually leading to more sensible options, until it has reached an equilibrium state.



Expected Peak Heights

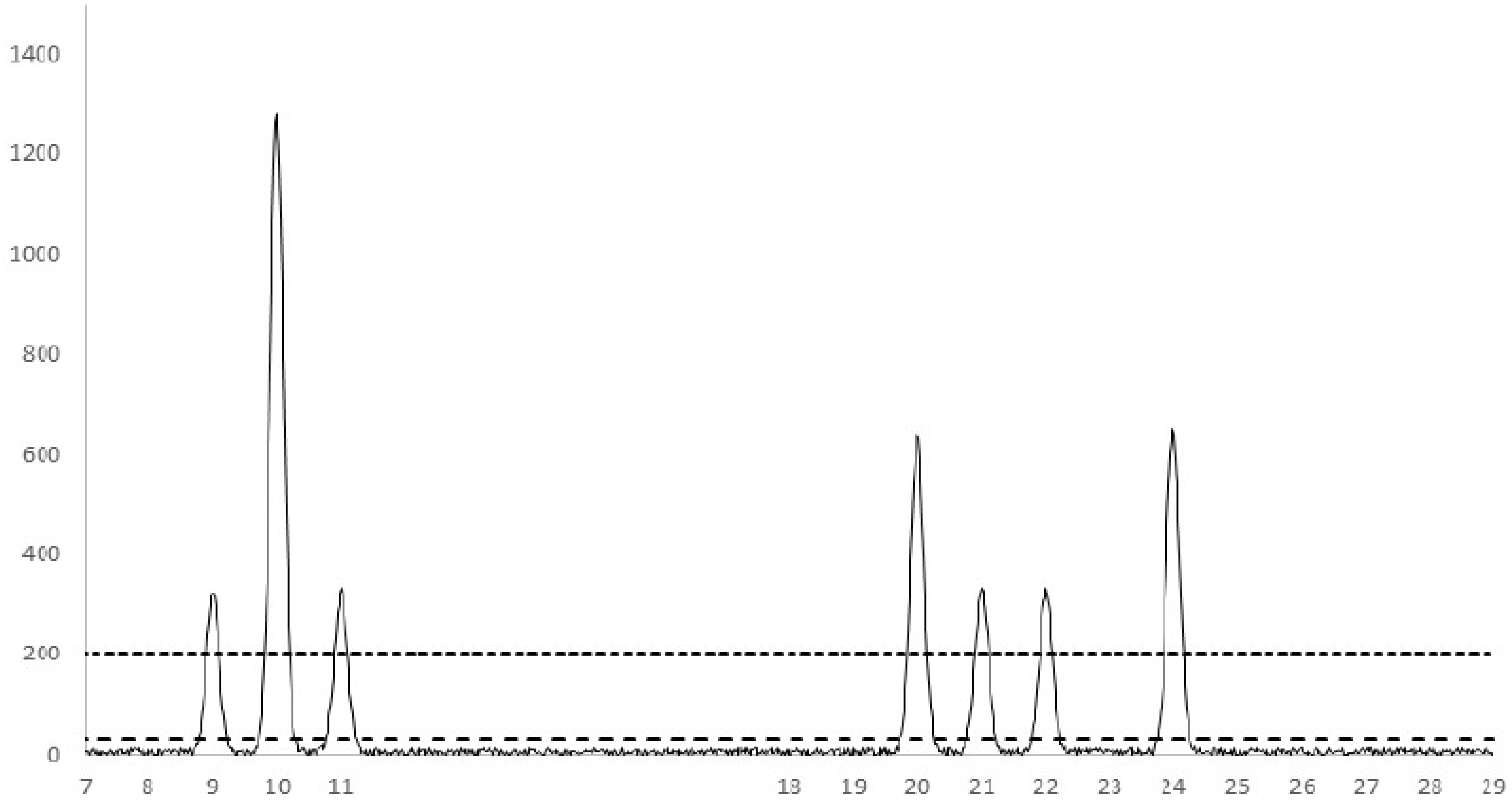
Based on a set of input parameters, an expected profile can be generated.

Step 1: Genotypes are chosen.



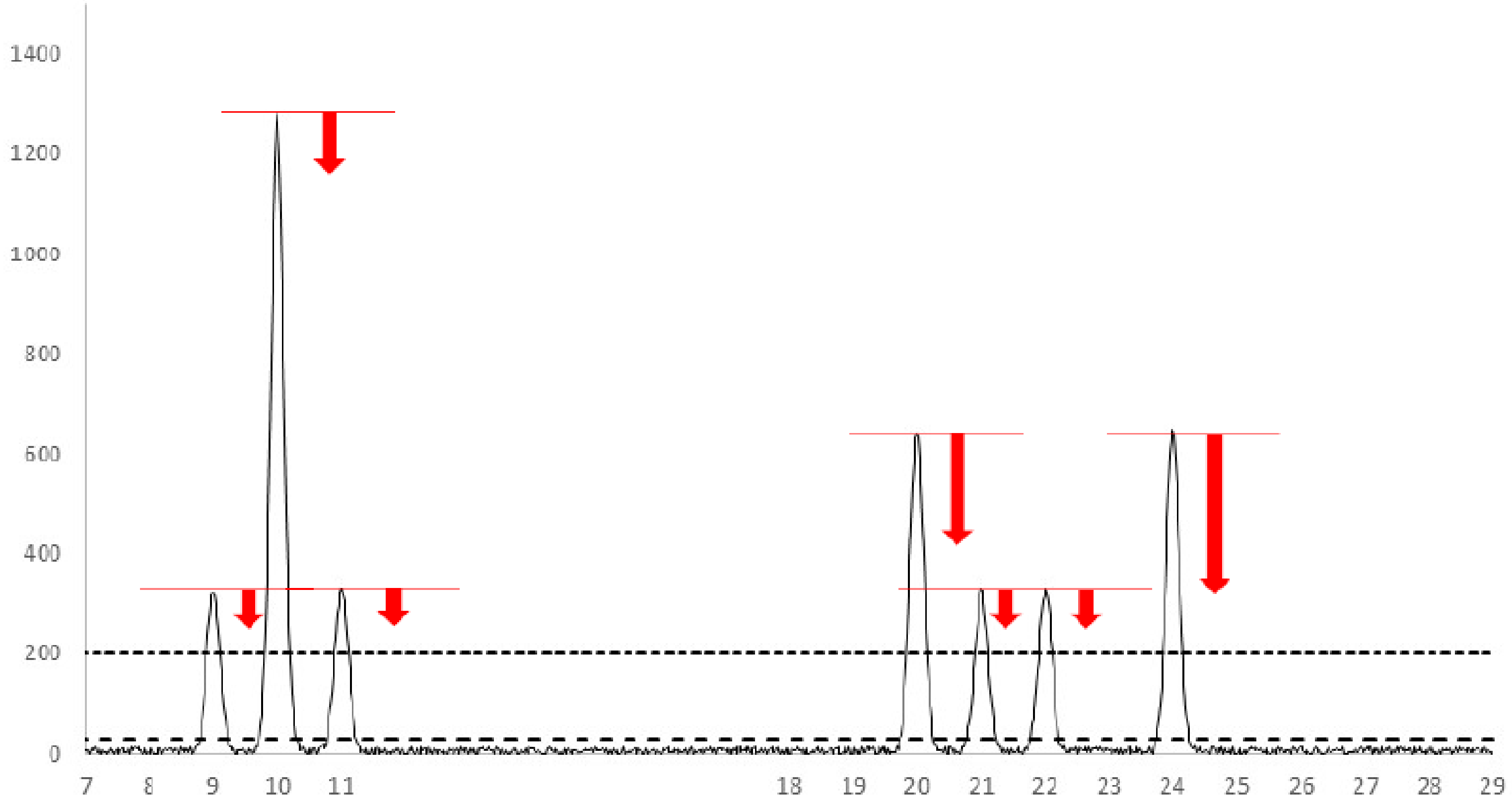
Expected Peak Heights

Step 2: Template amounts per contributor are incorporated.



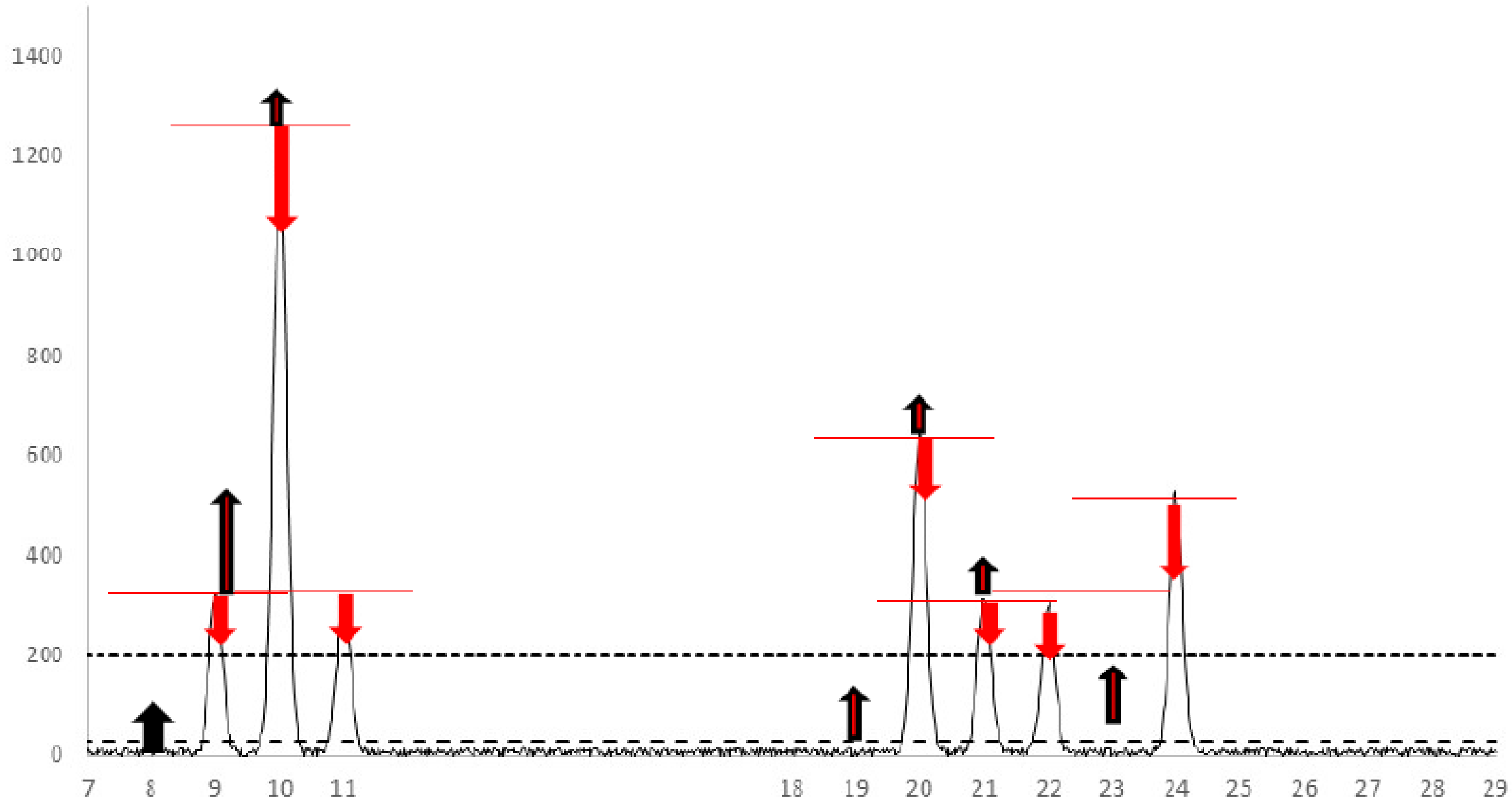
Expected Peak Heights

Step 3: Degradation is taken into account.



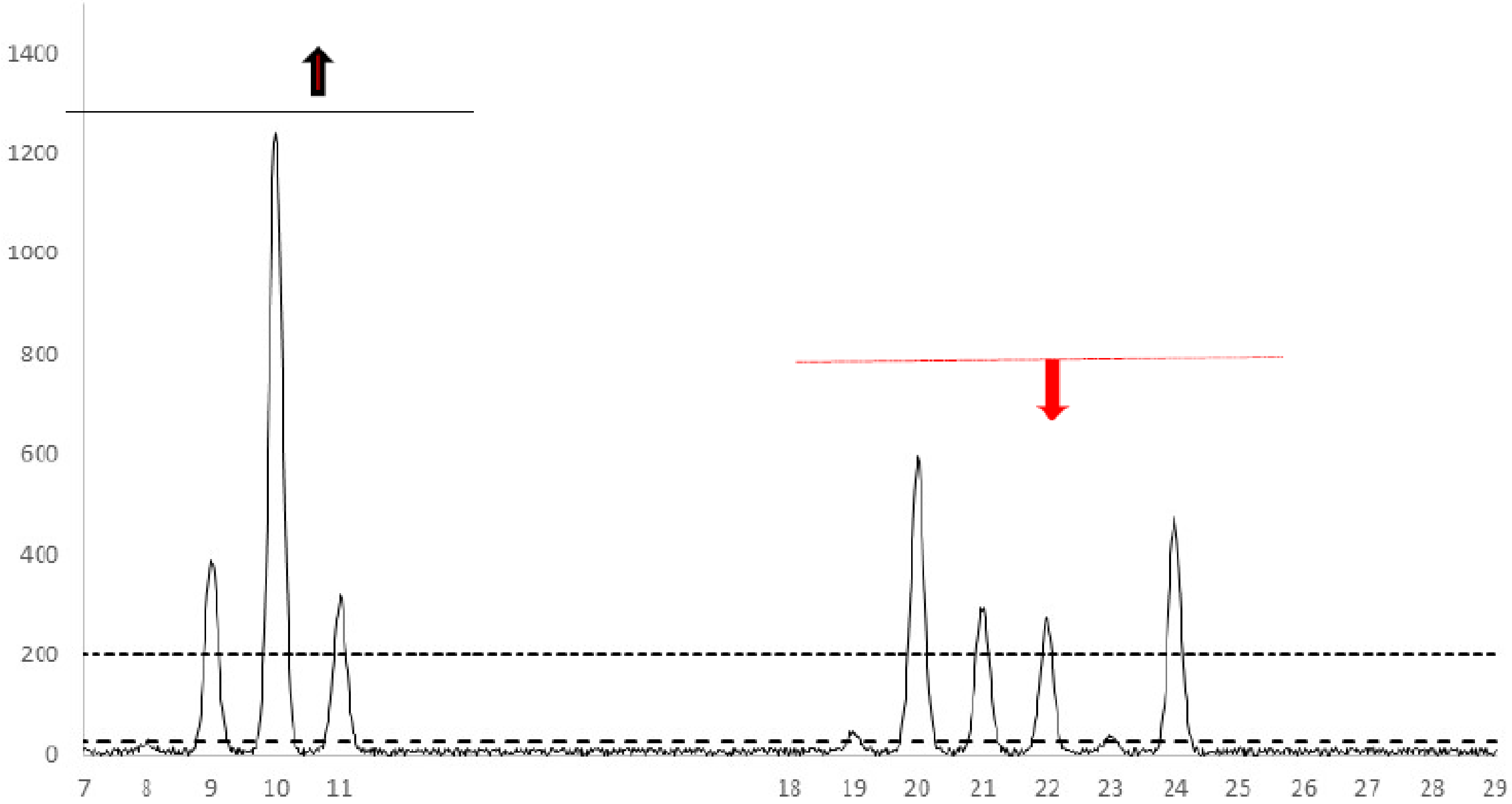
Expected Peak Heights

Step 4: Stutter is taken into account.



Expected Peak Heights

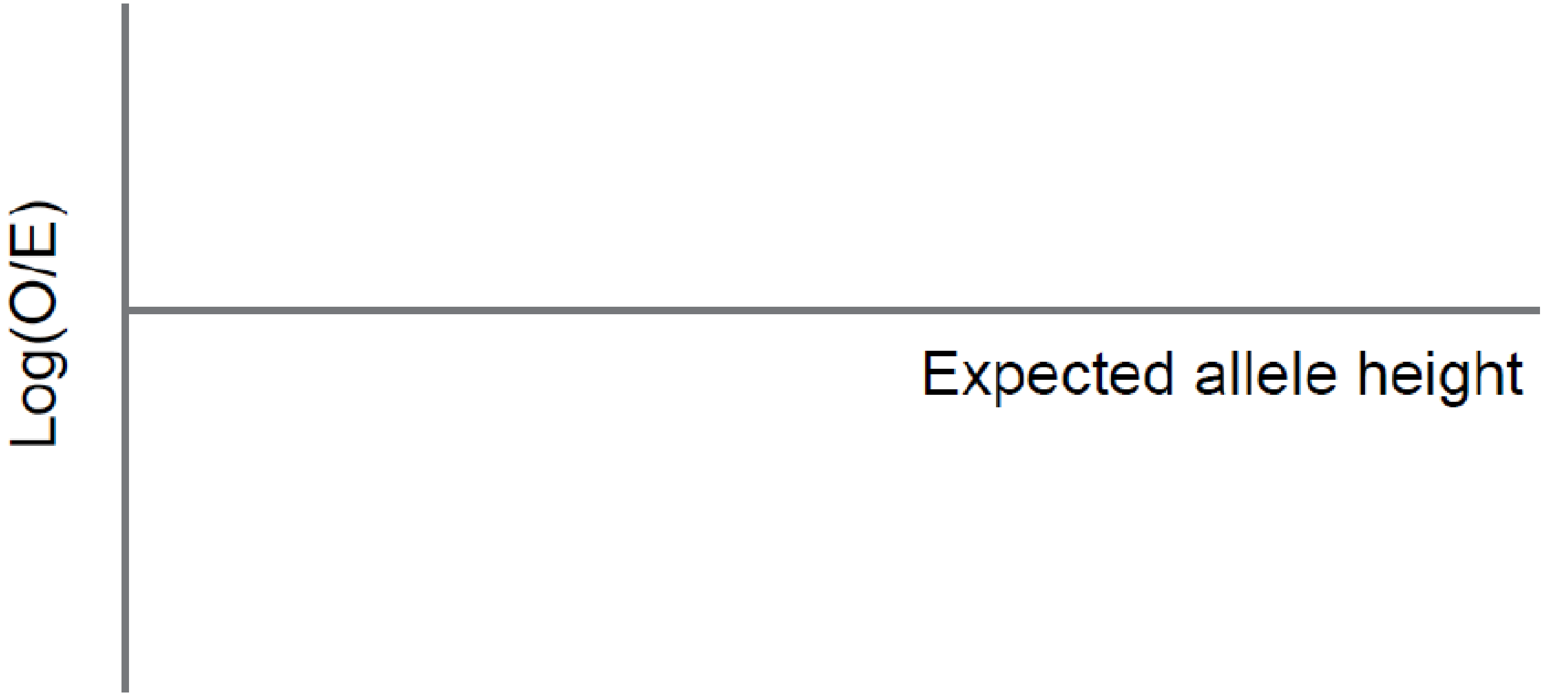
Step 5: Locus specific amplification efficiencies are introduced.



The Perfect Model

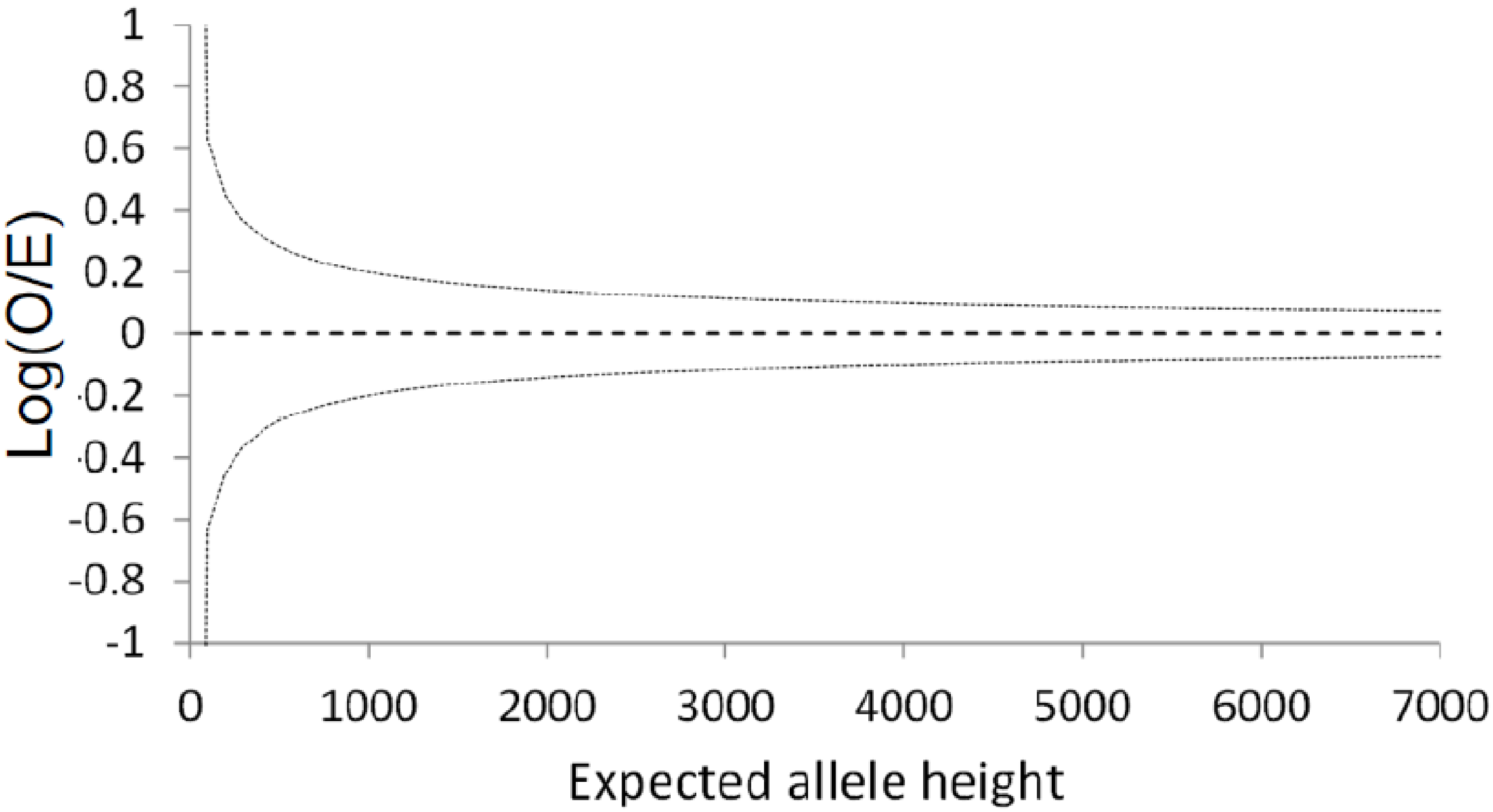
We can now compare our expected profile with the observed STR profile.

What would a perfect model look like?



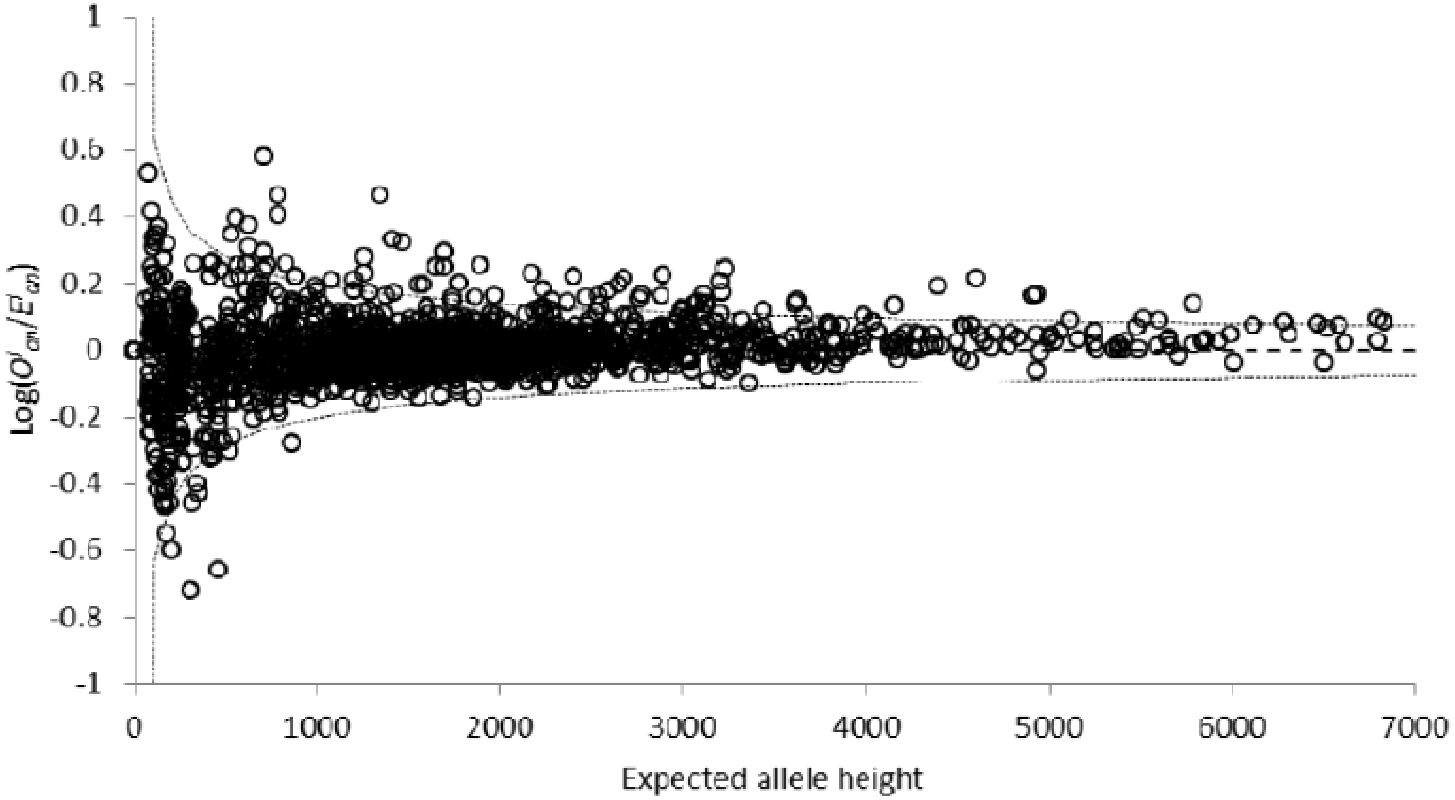
The Perfect Model

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.



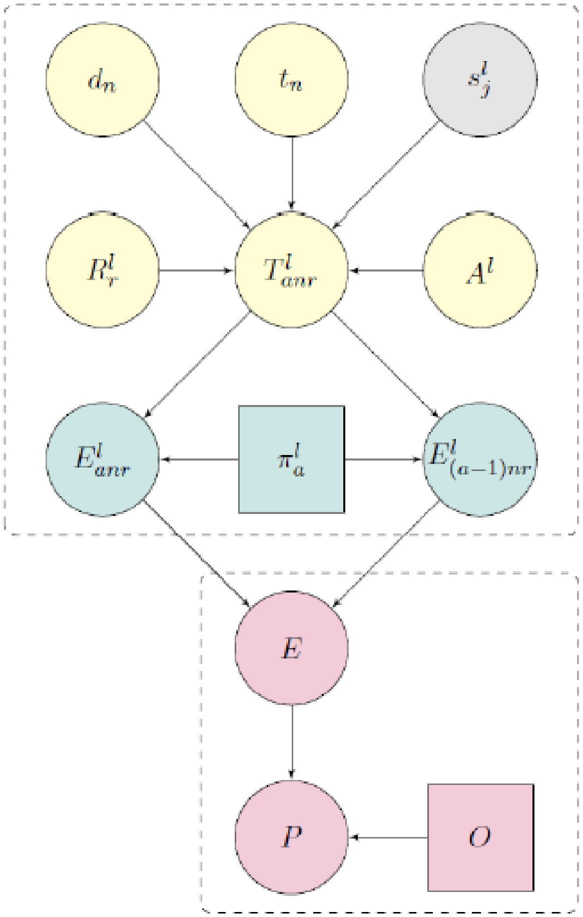
Generating Weights

The probability of obtaining the observed profile 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

Combining all elements leads to an overall continuous model network:



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

Probabilistic Genotyping

The Scientific Working Group on DNA Analysis Methods (SWG-DAM) 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.

Probabilistic Genotyping - 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).

Probabilistic Genotyping

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)?

Probabilistic Genotyping

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).