Forensic Genetics

Module 16 – Session 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

Likelihood Ratio

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

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

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

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

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

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

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 – Binary Model

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 : K + POI (S) and H_d : K + Unknown (U)

LR – Binary Model

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:

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

LR – Binary Model

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

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

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

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

LR – Binary Model

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:

- $\text{LR} = \frac{\Pr(ABCD|AB,CD,H_p)}{\Pr(ABCD|CD,H_d)} = \frac{1}{2p_Ap_B}$;
- $\text{LR} = \frac{\Pr(ABCD|AA,CD,H_p)}{\Pr(ABCD|CD,H_d)} = 0$;

LR – Binary Model

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

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

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

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

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

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

LR – Binary Model

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:

- $\text{LR} = \frac{\Pr(ABCD|AB,CD,H_p)}{\Pr(ABCD|CD,H_d)} = \frac{1}{2p_Ap_B}$;
- $\text{LR} = \frac{\Pr(ABCD|AA,CD,H_p)}{\Pr(ABCD|CD,H_d)} = 0$;
- $\text{LR} = \frac{\Pr(ABCD|AB,H_p)}{\Pr(ABCD|H_d)} = \frac{2p_Cp_D}{6{\cdot}4p_Ap_Bp_Cp_D} = \frac{1}{12p_Ap_B}$.

LR – Semi-continuous Model

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:

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

LR – Semi-continuous Model

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.

LR – Semi-continuous Model

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:

	$\Pr(G_S \to G_C)$	
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.

Estimating Drop-in and Drop-out

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

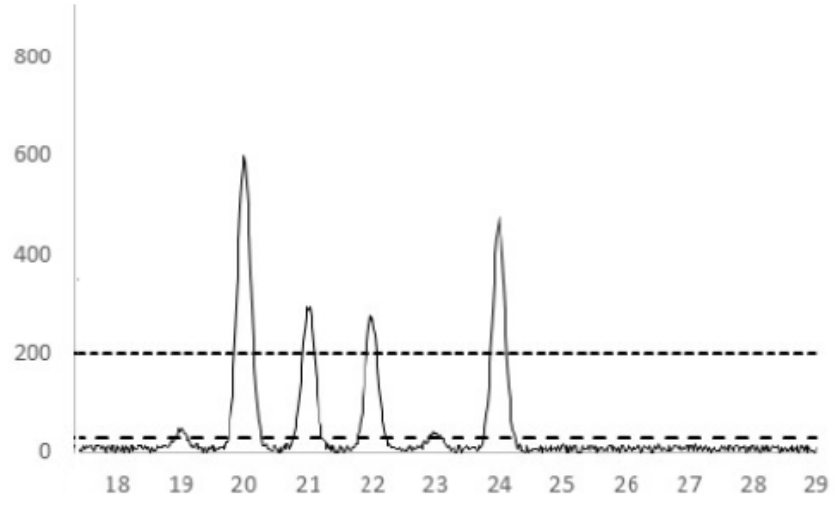
- Several models have been proposed for modeling drop-out probabilities, such as a multidose drop-out model and degradation model. Laboratory trials can be used to choose α when modeling $D_2 = \alpha D^2$, with $0 < \alpha \leq 1$. Instead of assigning probabilities to the drop-out rate they can be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

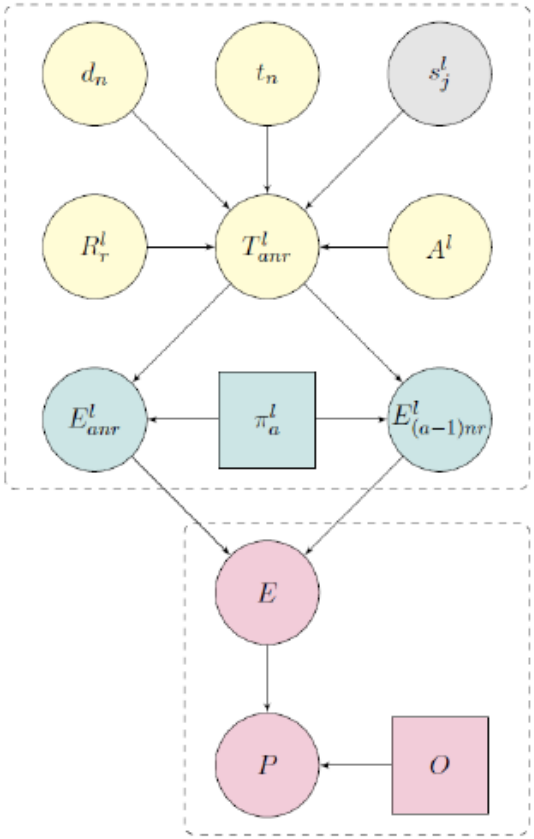
The key point of a fully continuous model is that it considers peak heights as a continuous variable.



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

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).

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:

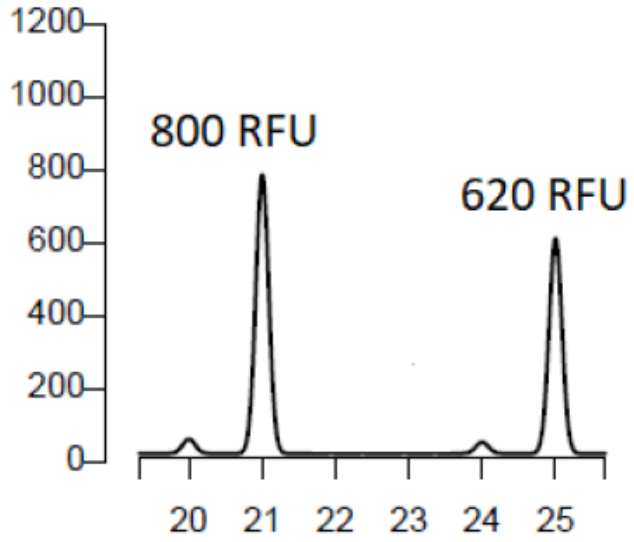
$$\text{Hb}_1 = \frac{O_{\text{HMW}}}{O_{\text{LMW}}}, \qquad \text{and} \qquad \text{Hb}_2 = \frac{O_{\text{smaller}}}{O_{\text{larger}}},$$

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 Heterozygote Balance

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

$$=$$



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

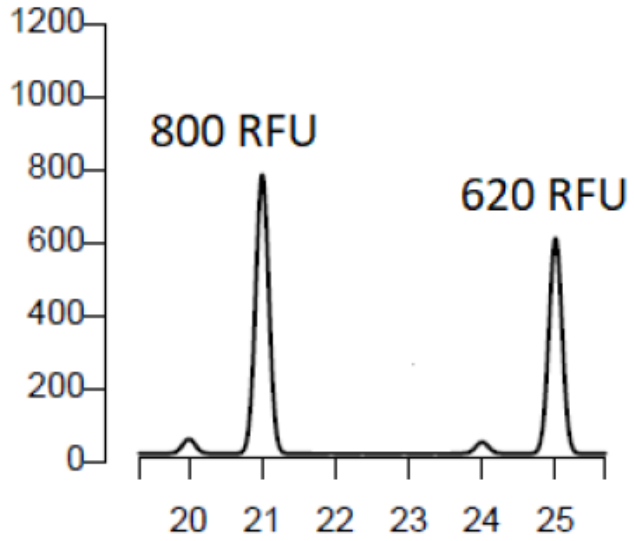
- Hb_2 may be obtained from Hb_1, but not vice versa.

Modeling Heterozygote Balance

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

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

$$= \mathrm{Hb}_2$$

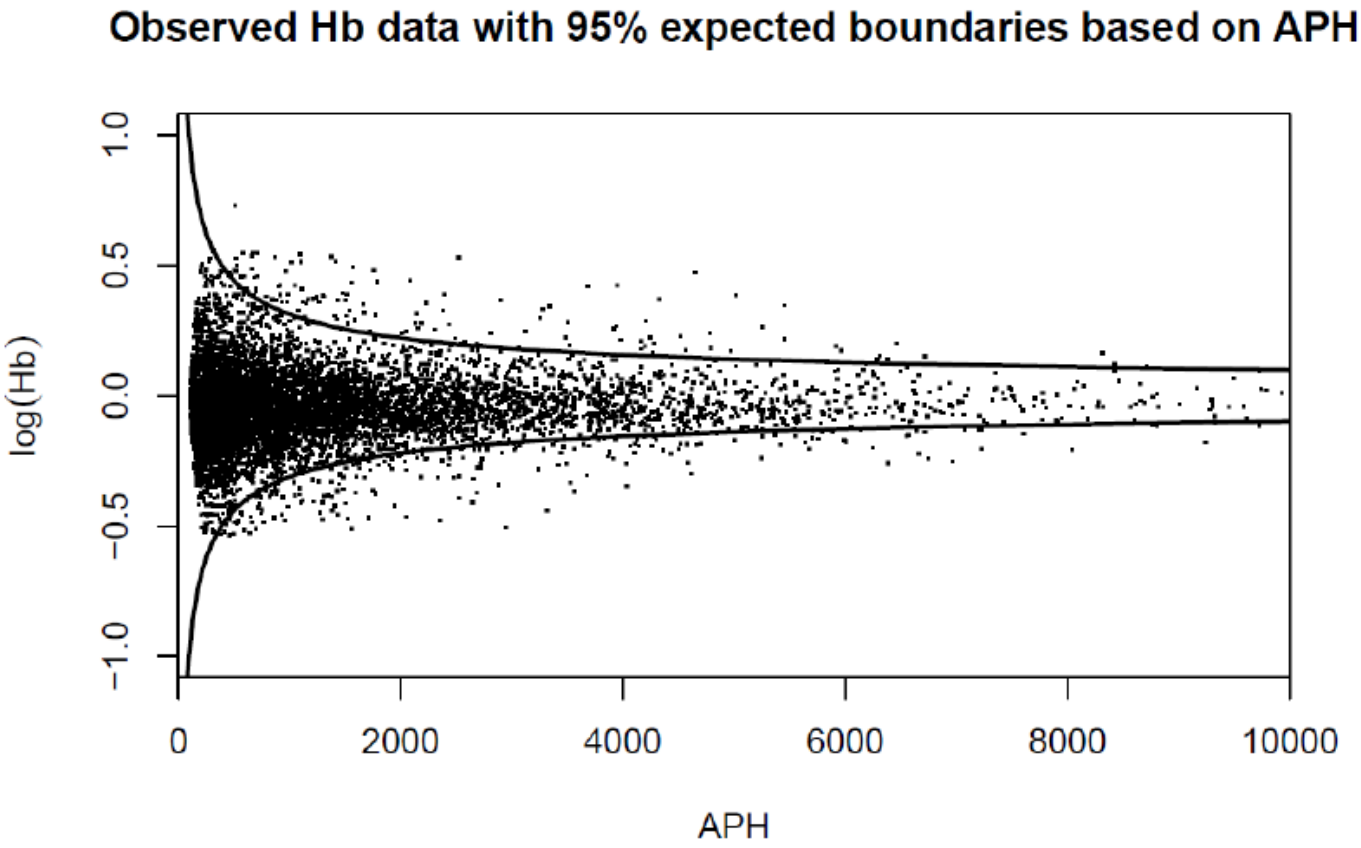


- Hb_1 has the highest information content, because it maintains peak order.

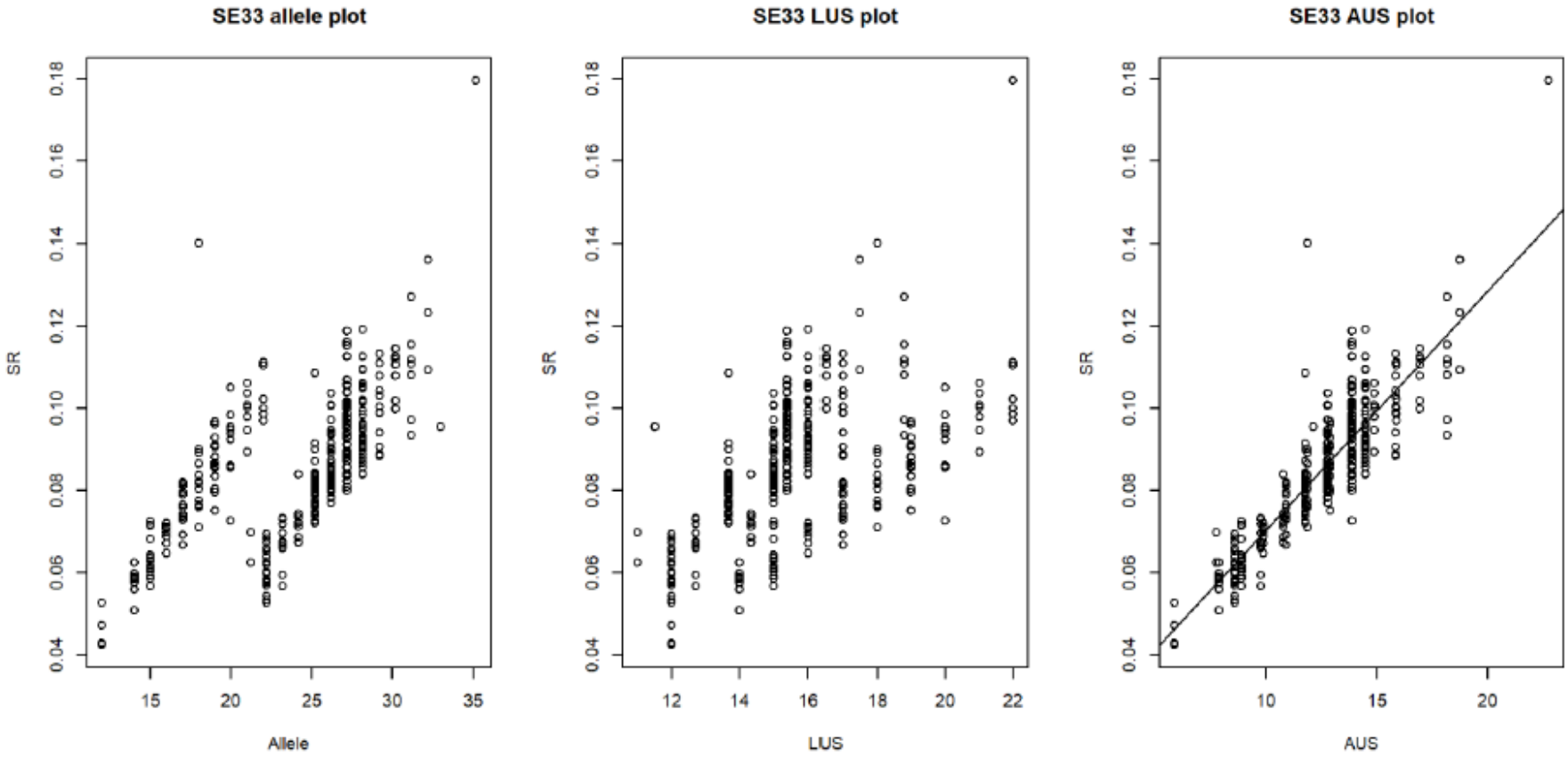
- Hb_2 may be obtained from Hb_1, but not vice versa.

Modeling Heterozygote Balance

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



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

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

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

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

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

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

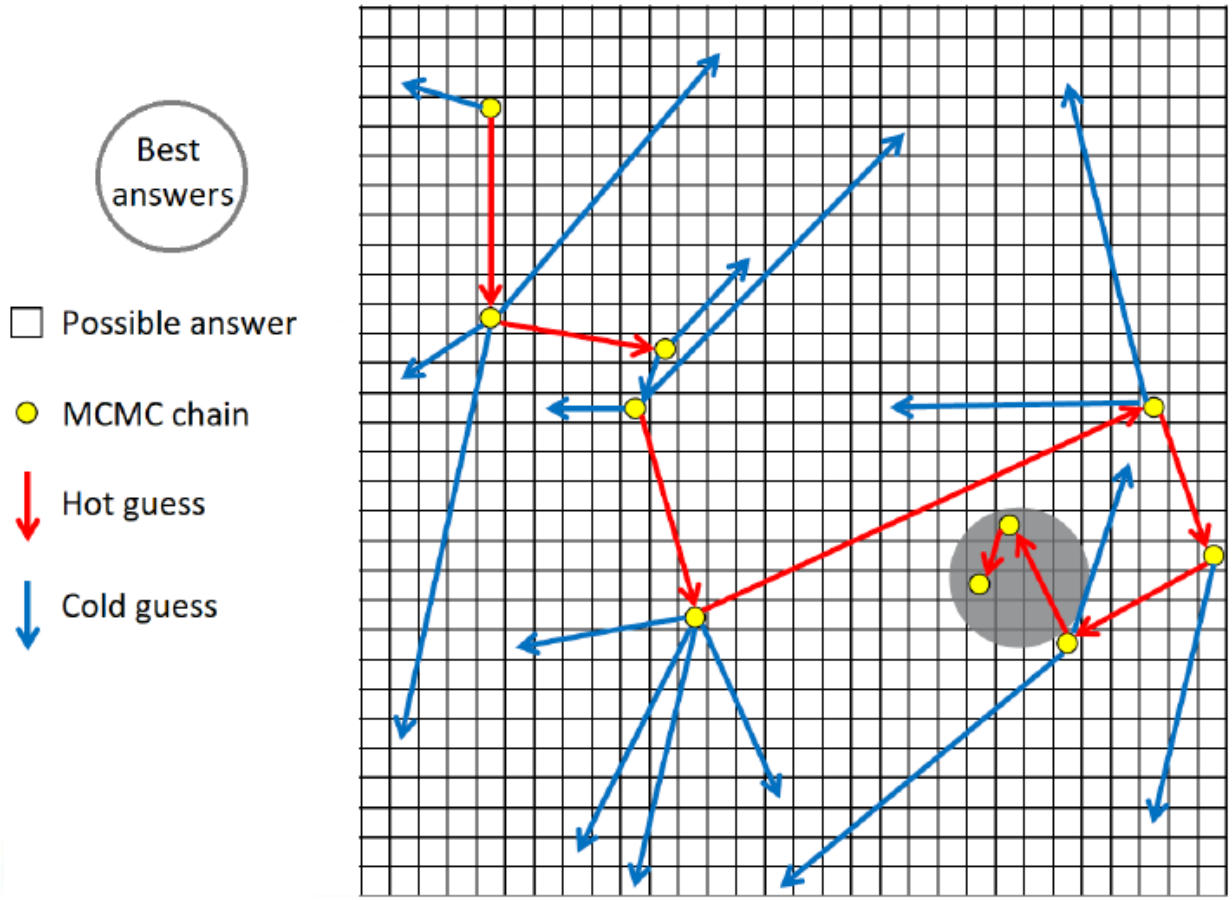
- What about other stutter products?

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.

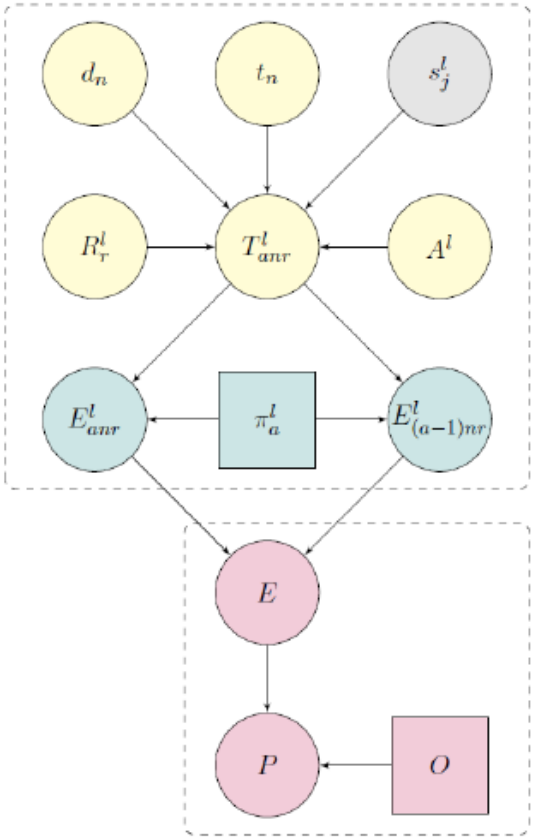
Markov Chain Monte Carlo

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



Continuous Model

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).

Probabilistic Genotyping

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.

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).