AALBORG UNIVERSITY

Statistical Aspects of Forensic Genetics

Models for Qualitative and Quantitative STR Data

Torben Tvedebrink

Department of Mathematical Sciences
Aalborg University

PhD Defence - November 5 2010

Outline

- Introduction to forensic genetics
  - Short Tandem Repeat DNA data
  - Competing hypothesis and likelihood ratios (LRs)
- Models for qualitative data
  - Population stratification and θ estimation
  - Analysis of a single DNA database
- Models for quantitative data
  - DNA mixtures - separation and goodness-of-fit
  - Inclusion of quantitative data in LR
  - Low template DNA and degradation

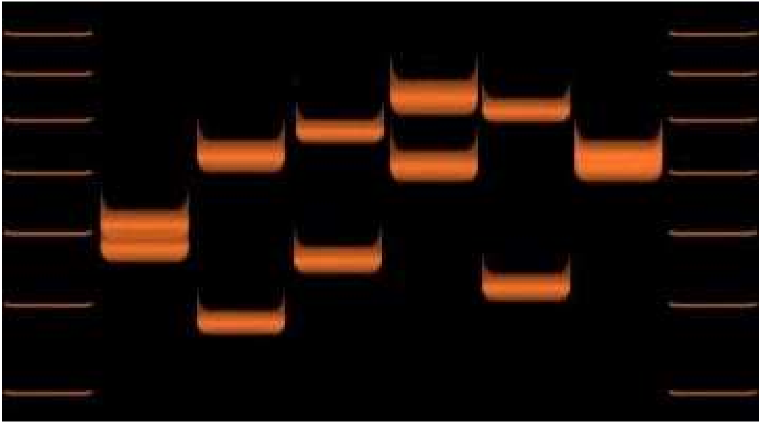
What is a DNA profile?

Most of the human genome is believed to be identical between individuals. Hence, the DNA sequences applicable for identification should be in the remainder of the genome.

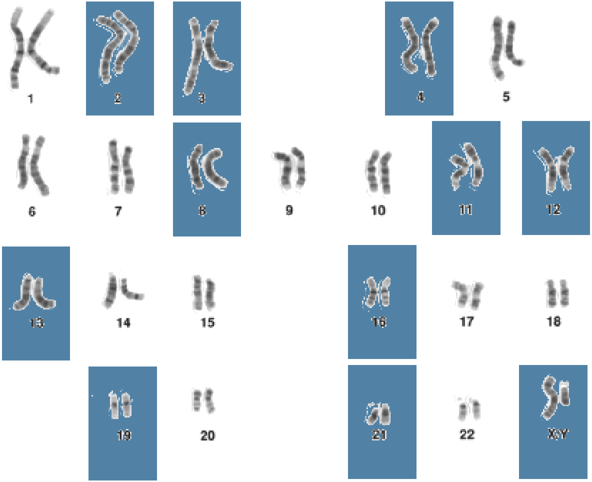
A DNA profile used for forensic purposes consists of the genetic constitution in a few highly polymorphic genetic markers.

The prevailing method for identification is called Short Tandem Repeat (STR). Several commercial produced typing kits are available, however, during my studies I have mainly focused on data obtained by the AmpFℓSTR SGM Plus kit from Applied Biosystems.

What is a DNA profile?

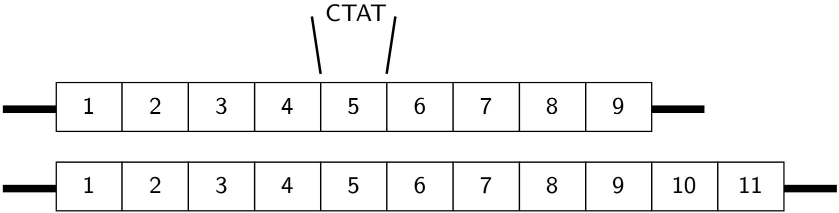


SGM Plus kit

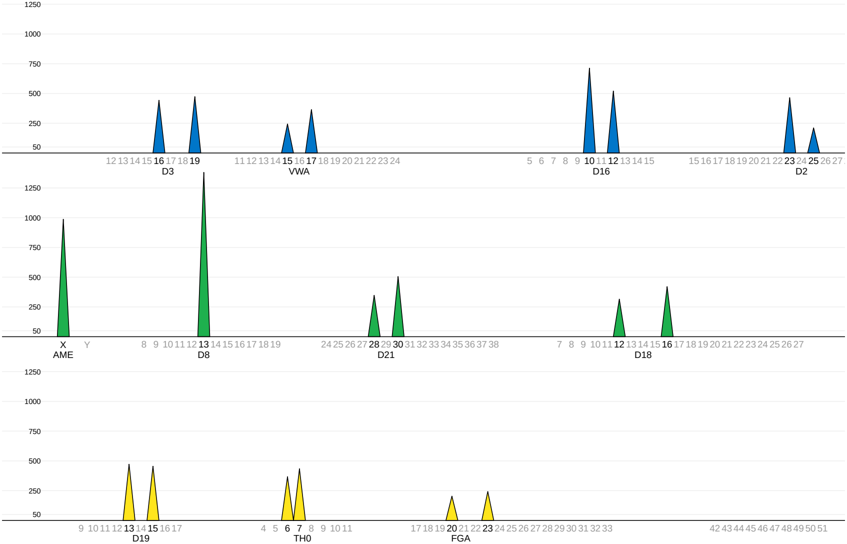


SGM Plus kit

STR alleles are identified by their number of repeats of a given repeat motif. Below the repeat motif is CTAT, which is repeated 9 and 11 times indicating a heterozygous DNA profile (9,11).



SGM Plus kit


1250
1000
750
500
250
50
12 13 14 15 16 17 18 19
D3
11 12 13 14 15 16 17 18 19 20 21 22 23 24
VWA
5 6 7 8 9 10 11 12 13 14 15
D16
15 16 17 18 19 20 21 22 23 24 25 26 27
D2
X Y
AME
8 9 10 11 12 13 14 15 16 17 18 19
D8
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
D21
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
D18
9 10 11 12 13 14 15 16 17
D19
4 5 6 7 8 9 10 11
TH0
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
FGA
42 43 44 45 46 47 48 49 50 51

Likelihood ratio - the central quantity

In forensic genetics, the evaluation of the evidential weight is done by a likelihood ratio approach:

$$LR = \frac{P(\text{Data} \mid \text{Hypothesis } 1)}{P(\text{Data} \mid \text{Hypothesis } 2)}$$

Likelihood ratio - the central quantity

In forensic genetics, the evaluation of the evidential weight is done by a likelihood ratio approach:

$$\begin{aligned} LR &= \frac{P(\text{Data} \mid \text{Hypothesis 1})}{P(\text{Data} \mid \text{Hypothesis 2})} \\ &= \frac{P(\text{DNA evidence} \mid \text{Guilt of suspect})}{P(\text{DNA evidence} \mid \text{Innocence of suspect})} \end{aligned}$$

Likelihood ratio - the central quantity

In forensic genetics, the evaluation of the evidential weight is done by a likelihood ratio approach:

$$\begin{aligned} LR &= \frac{P(\text{Data} \mid \text{Hypothesis 1})}{P(\text{Data} \mid \text{Hypothesis 2})} \\ &= \frac{P(\text{DNA evidence} \mid \text{Guilt of suspect})}{P(\text{DNA evidence} \mid \text{Innocence of suspect})} \end{aligned}$$

Often H_p is used to denote the hypothesis stating the guilt of the suspect/defendant (often called the prosecutors hypothesis) and H_d represents the acquitting of the suspect (defence hypothesis)

DNA evidence

In crime cases the DNA evidence, $\mathcal{E}$, available for evaluation consists of two parts:

- Crime scene data, $\mathcal{E}_c$: Includes the DNA profile obtained from samples at the scene of crime.
- Known/fixed profiles, $\mathbf{K}$: The DNA profiles of known/identified individuals, e.g. the profiles of victim and suspect.

Hence, we have

$$\frac{P(\mathcal{E}|H_p)}{P(\mathcal{E}|H_d)} = \frac{P(\mathcal{E}_c, \mathbf{K}|H_p)}{P(\mathcal{E}_c, \mathbf{K}|H_d)}$$

Example (Single contributor stain)

Assume that an identified suspect's DNA matches that of a crime scene: $\mathcal{E}_c \equiv G_S$. Then $\mathbf{K} = G_S$ and the hypotheses state:

H_p: "The suspect is the contributor of the biological material"

H_d: "An unknown (and to the suspect unrelated) individual is the donor of the biological material"

Example (Single contributor stain) - cont'd

The weight of the evidence is assessed by computing the LR:

$$LR = \frac{P(\mathcal{E}_c, \mathbf{K}|H_p)}{P(\mathcal{E}_c, \mathbf{K}|H_d)}$$

Example (Single contributor stain) - cont'd

The weight of the evidence is assessed by computing the LR:

$$
\begin{aligned}
LR &= \frac{P(\mathcal{E}_c, \mathbf{K} | H_p)}{P(\mathcal{E}_c, \mathbf{K} | H_d)} \\
&= \frac{P(\mathcal{E}_c, G_S | G_S) P(G_S)}{P(\mathcal{E}_c, G_S | G_U) P(G_U)}
\end{aligned}
$$

Example (Single contributor stain) - cont'd

The weight of the evidence is assessed by computing the LR:

$$
\begin{aligned}
LR &= \frac{P(\mathcal{E}_c, \mathbf{K} | H_p)}{P(\mathcal{E}_c, \mathbf{K} | H_d)} \\
&= \frac{P(\mathcal{E}_c, G_S | G_S) P(G_S)}{P(\mathcal{E}_c, G_S | G_U) P(G_U)} \\
&= \frac{P(\mathcal{E}_c | G_S) P(G_S | G_S) P(G_S)}{P(\mathcal{E}_c | G_U) P(G_S | G_U) P(G_U)}
\end{aligned}
$$

Example (Single contributor stain) - cont'd

The weight of the evidence is assessed by computing the LR:

$$\begin{aligned}
LR &= \frac{P(\mathcal{E}_c, \mathbf{K}|H_p)}{P(\mathcal{E}_c, \mathbf{K}|H_d)} \\
&= \frac{P(\mathcal{E}_c, G_S|G_S)P(G_S)}{P(\mathcal{E}_c, G_S|G_U)P(G_U)} \\
&= \frac{\cancel{P(\mathcal{E}_c|G_S)}P(G_S|G_S)P(G_S)}{\cancel{P(\mathcal{E}_c|G_U)}P(G_S|G_U)P(G_U)}
\end{aligned}$$

Example (Single contributor stain) - cont'd

The weight of the evidence is assessed by computing the LR:

$$\begin{aligned} LR &= \frac{P(\mathcal{E}_c, \mathbf{K} | H_p)}{P(\mathcal{E}_c, \mathbf{K} | H_d)} \\ &= \frac{P(\mathcal{E}_c, G_S | G_S) P(G_S)}{P(\mathcal{E}_c, G_S | G_U) P(G_U)} \\ &= \frac{P(G_S)}{P(G_S | G_U) P(G_U)} \end{aligned}$$

Example (Single contributor stain) - cont'd

The weight of the evidence is assessed by computing the *LR*:

$$\begin{aligned} LR &= \frac{P(\mathcal{E}_c, \mathbf{K} | H_p)}{P(\mathcal{E}_c, \mathbf{K} | H_d)} \\ &= \frac{P(\mathcal{E}_c, G_S | G_S) P(G_S)}{P(\mathcal{E}_c, G_S | G_U) P(G_U)} \\ &= \frac{P(G_S)}{P(G_S | G_U) P(G_U)} \\ &= P(G_U | G_S)^{-1}, \end{aligned}$$

where $P(G_U | G_S)$ represents the *rarity* of the particular DNA profile.

Match probability

The STR loci included in the SGM are located on different chromosomes, hence the laws of inheritance suggest that there is statistical independence of the allelic distribution across loci:

$$P(G_U|G_S) = \prod_{l=1}^{L} P_l(G_{U,l}|G_{S,l})$$

Match probability

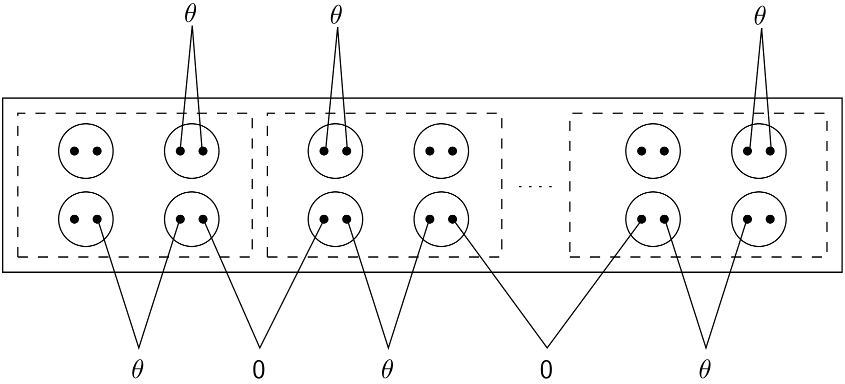
The STR loci included in the SGM are located on different chromosomes, hence the laws of inheritance suggest that there is statistical independence of the allelic distribution across loci:

$$P(G_U|G_S) = \prod_{l=1}^{L} P_l(G_{U,l}|G_{S,l})$$

However, it may be inaccurate to assume that the allelic distribution in a given locus supports independence of alleles:

$$P(A_iA_j) \neq P(A_i)P(A_j)$$

Population stratification



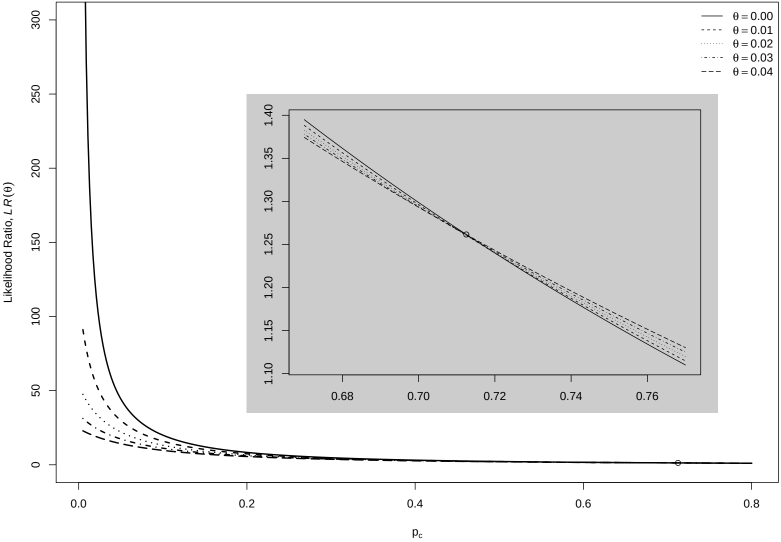
Example - Effect of θ in evidential calculations

Assume that we have a two-person DNA mixture with three alleles observed: A, B and C. The identified victim is $G_V = (A, B)$ while the suspect is $G_S = (C, C)$ for this locus.

Then the likelihood ratio with H_p:(G_V, G_S) and H_d:(G_V, G_U) yields

$$LR = \frac{(1+3\theta)(1+4\theta)}{(7\theta + \{1-\theta\}[2p_a + 2p_b + p_c])(2\theta + \{1-\theta\}p_c)}$$

Example - Effect of θ in evidential calculations

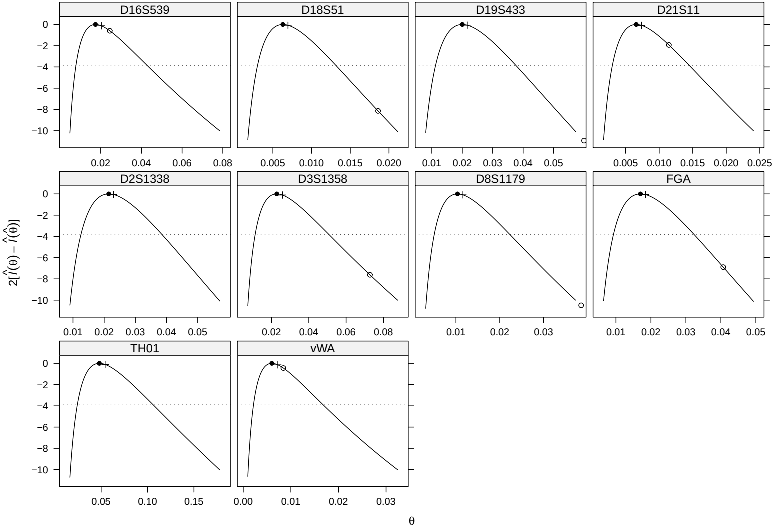


Estimation of θ and confidence intervals

We may estimate θ from data when we have database multiple subpopulations available. By computing the profile log-likelihood an approximative confidence interval may be computed.

The profile log-likelihoods (next slide) are for data obtained from Denmark ($n = 258$), Faroe Islands ($n = 23$) and Greenland ($n = 399$).

Estimation of θ and confidence intervals



Analysis of a single DNA database

The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, made a database with $51,517$ DNA profiles available.

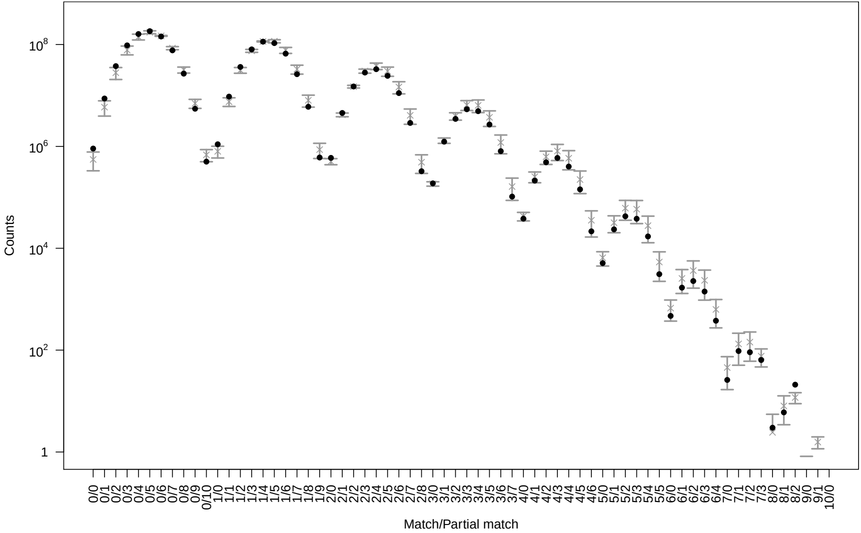
If we make all pairwise comparisons, we end up making $\binom{n}{2} = n(n-1)/2$ comparisons. With $n = 51,517$ profiles this gives $1,326,974,886$ comparisons.

θ-estimation from a single database

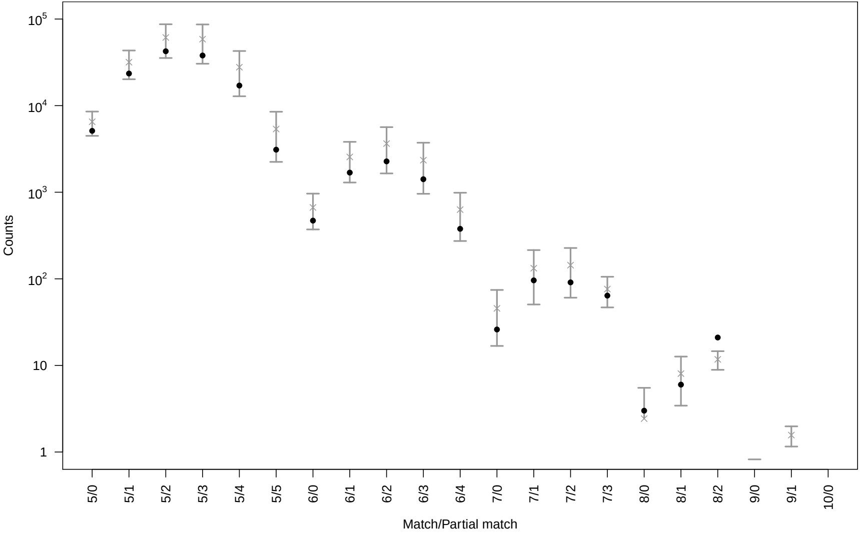
$M_{m/p}$ is the summary statistic showing the number of profiles matching at m loci and partially-matching at p.

M	0	1	2	3	4	5	6	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\iddots$
4	38,094	212,192	487,484	592,929	401,832	143,202	21,490	
5	5,114	23,490	42,459	37,933	17,060	3,100		
6	470	1,685	2,272	1,414	378			
7	26	96	91	64				
8	3	6	21					
9	0	0						
10	0							

θ-estimation from a single database



θ-estimation from a single database

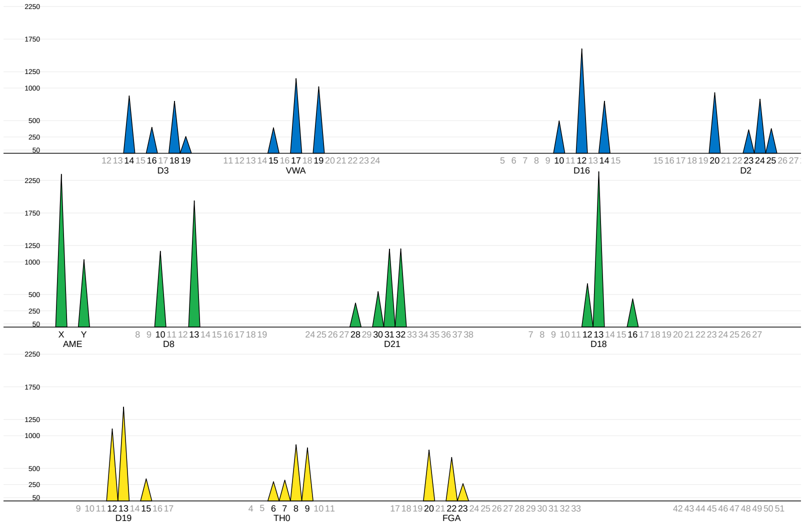


DNA mixtures

If more than one individual contributes to a DNA stain, then the stain is called a DNA mixture. DNA mixtures are more challenging than single contributor stains:

- Uncertainty about number of contributors
- The proportion(s) between the amount of contributed DNA
- The genotypes of the contributors
- ...

Example (DNA mixture)



Example (DNA mixture)

Assume that $\mathcal{E}_c$ originates from a DNA mixture. Let G_V denote the known victim's DNA profile and G_S the identified suspect's profile, then $\mathbf{K} = (G_V, G_S)$.

H_p: "The victim and suspect are the contributors to the stain"

H_d: "The victim and an unknown individual are the contributors to the stain"

Example (DNA mixture) - cont'd

The LR is given by:

$$LR = \frac{P(\mathcal{E}_c | G_V, G_S)}{\sum_{G_U \equiv H_d} P(\mathcal{E}_c | G_V, G_U) P(G_U | G_V, G_S)}$$

where we need to be able to evaluate

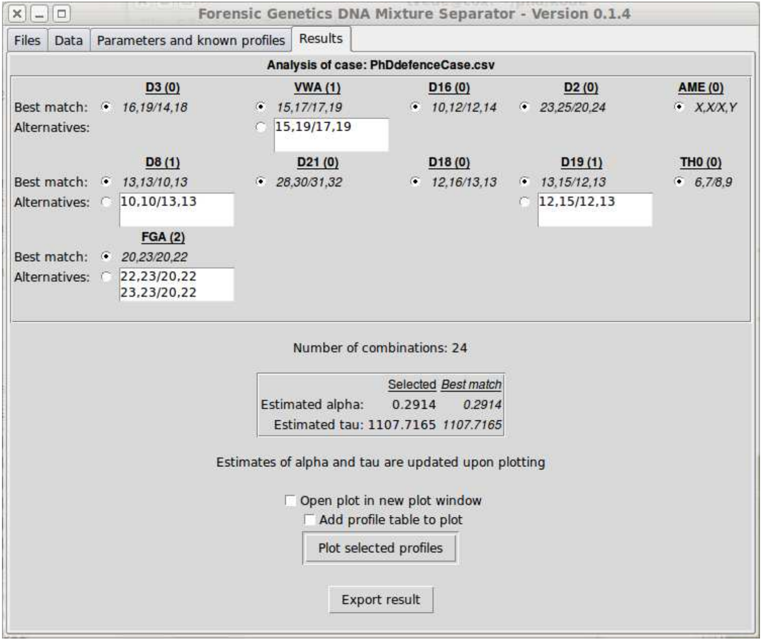
$P(\mathcal{E}_c | G_V, G_S)$ and $P(\mathcal{E}_c | G_V, G_U)$ for some unknown profile G_U

Separation of a DNA mixture

In addition to judging the goodness-of-fit of a proposed combination of DNA profiles, searching for a best set of profiles may be of interest to forensic geneticists.

This facility has been implemented in a R-package `mixsep` with a graphical user interface (GUI):

```
> library(mixsep)
> mixsep()
```


Forensic Genetics DNA Mixture Separator - Version 0.1.4
Files
Data
Parameters and known profiles
Results

Analysis of case: PhDdefenceCase.csv

	D3 (0)	VWA (1)	D16 (0)	D2 (0)	AME (0)
Best match:	● *16,19/14,18*	● *15,17/17,19*	● *10,12/12,14*	● *23,25/20,24*	● *X,X/X,Y*
Alternatives:		○ 15,19/17,19			

	D8 (1)	D21 (0)	D18 (0)	D19 (1)	TH0 (0)
Best match:	● *13,13/10,13*	● *28,30/31,32*	● *12,16/13,13*	● *13,15/12,13*	● *6,7/8,9*
Alternatives:	○ 10,10/13,13			○ 12,15/12,13	

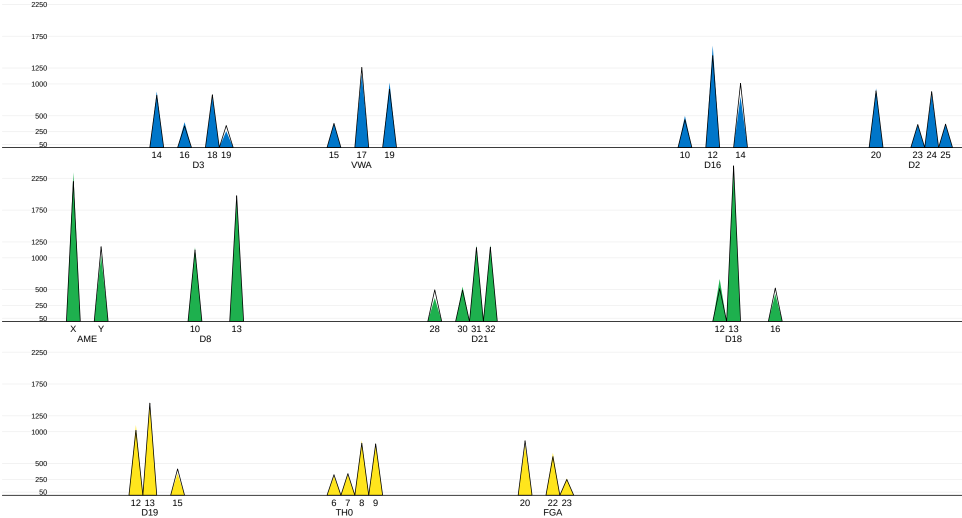
	FGA (2)
Best match:	● *20,23/20,22*
Alternatives:	○ 22,23/20,22 23,23/20,22

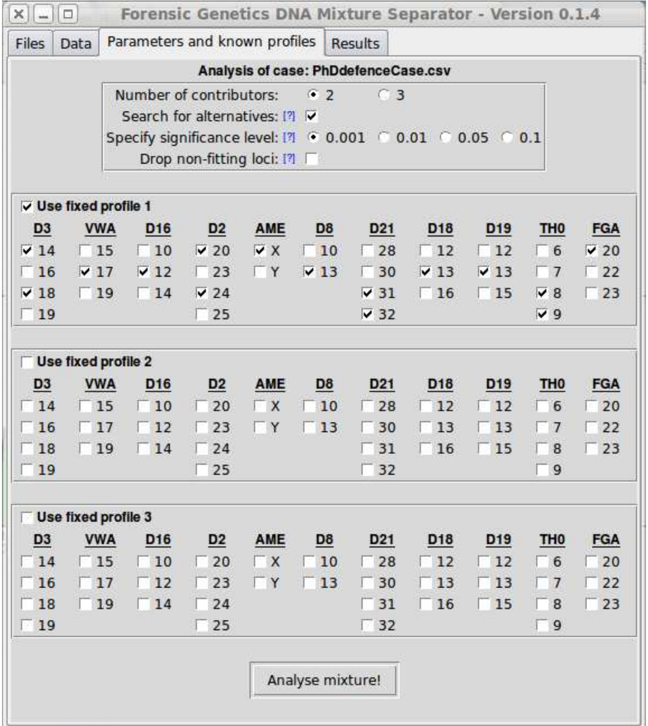
Number of combinations: 24

	Selected	*Best match*
Estimated alpha:	0.2914	*0.2914*
Estimated tau:	1107.7165	*1107.7165*

Estimates of alpha and tau are updated upon plotting

☐ Open plot in new plot window
☐ Add profile table to plot


2250
1750
1250
1000
500
250
50
14 16 18 19
D3
15 17 19
VWA
10 12 14
D16
20 23 24 25
D2
2250
1750
1250
1000
500
250
50
X Y
AME
10 13
D8
28 30 31 32
D21
12 13 16
D18
2250
1750
1250
1000
500
250
50
12 13 15
D19
6 7 8 9
TH0
20 22 23
FGA



Forensic Genetics DNA Mixture Separator - Version 0.1.4

Analysis of case: PhDdefenceCase.csv

Number of contributors: ● 2 ○ 3

Search for alternatives: [?] ☑

Specify significance level: [?] ● 0.001 ○ 0.01 ○ 0.05 ○ 0.1

Drop non-fitting loci: [?] □

☑ **Use fixed profile 1**

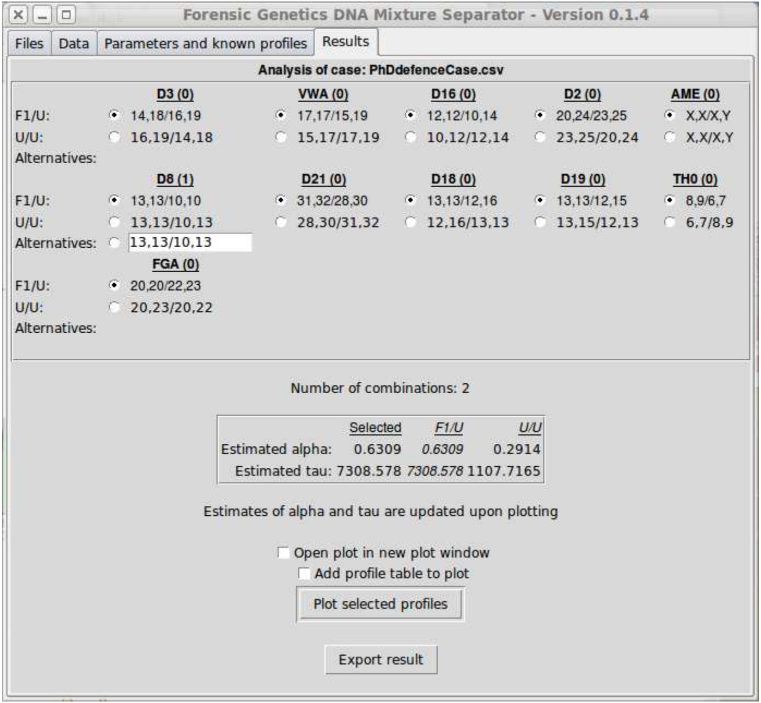
D3	VWA	D16	D2	AME	D8	D21	D18	D19	TH0	FGA
☑ 14	□ 15	□ 10	☑ 20	☑ X	□ 10	□ 28	□ 12	□ 12	□ 6	☑ 20
□ 16	☑ 17	☑ 12	□ 23	□ Y	☑ 13	□ 30	☑ 13	☑ 13	□ 7	□ 22
☑ 18	□ 19	□ 14	☑ 24			☑ 31	□ 16	□ 15	☑ 8	□ 23
□ 19			□ 25			☑ 32			☑ 9	

□ **Use fixed profile 2**

D3	VWA	D16	D2	AME	D8	D21	D18	D19	TH0	FGA
□ 14	□ 15	□ 10	□ 20	□ X	□ 10	□ 28	□ 12	□ 12	□ 6	□ 20
□ 16	□ 17	□ 12	□ 23	□ Y	□ 13	□ 30	□ 13	□ 13	□ 7	□ 22
□ 18	□ 19	□ 14	□ 24			□ 31	□ 16	□ 15	□ 8	□ 23
□ 19			□ 25			□ 32			□ 9	

□ **Use fixed profile 3**

D3	VWA	D16	D2	AME	D8	D21	D18	D19	TH0	FGA
□ 14	□ 15	□ 10	□ 20	□ X	□ 10	□ 28	□ 12	□ 12	□ 6	□ 20
□ 16	□ 17	□ 12	□ 23	□ Y	□ 13	□ 30	□ 13	□ 13	□ 7	□ 22
□ 18	□ 19	□ 14	□ 24			□ 31	□ 16	□ 15	□ 8	□ 23
□ 19			□ 25			□ 32			□ 9	


Forensic Genetics DNA Mixture Separator - Version 0.1.4
Files
Data
Parameters and known profiles
Results

Analysis of case: PhDdefenceCase.csv

	D3 (0)	VWA (0)	D16 (0)	D2 (0)	AME (0)
F1/U:	● 14,18/16,19	● 17,17/15,19	● 12,12/10,14	● 20,24/23,25	● X,X/X,Y
U/U:	○ 16,19/14,18	○ 15,17/17,19	○ 10,12/12,14	○ 23,25/20,24	○ X,X/X,Y
Alternatives:					

	D8 (1)	D21 (0)	D18 (0)	D19 (0)	TH0 (0)
F1/U:	● 13,13/10,10	● 31,32/28,30	● 13,13/12,16	● 13,13/12,15	● 8,9/6,7
U/U:	○ 13,13/10,13	○ 28,30/31,32	○ 12,16/13,13	○ 13,15/12,13	○ 6,7/8,9
Alternatives:	○ 13,13/10,13				

	FGA (0)
F1/U:	● 20,20/22,23
U/U:	○ 20,23/20,22
Alternatives:	

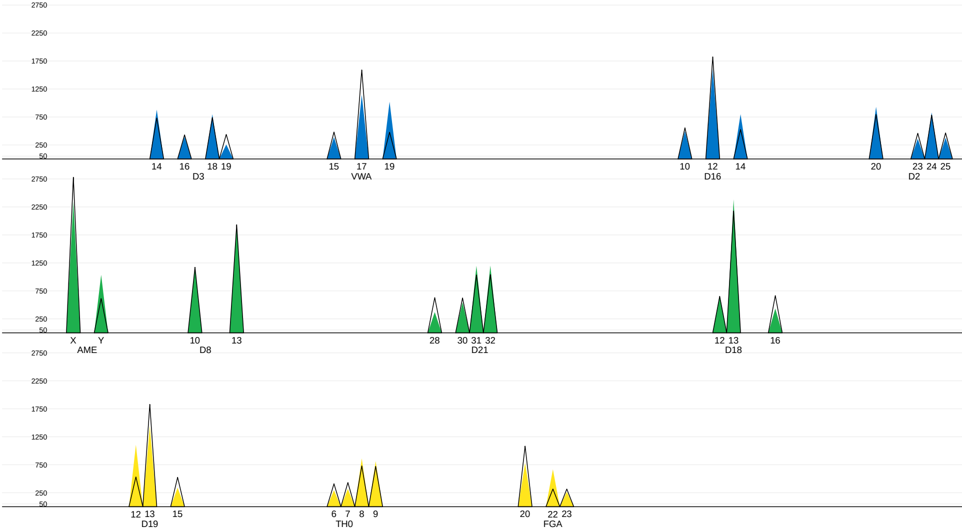
Number of combinations: 2

	Selected	*F1/U*	*U/U*
Estimated alpha:	0.6309	*0.6309*	0.2914
Estimated tau:	7308.578	*7308.578*	1107.7165

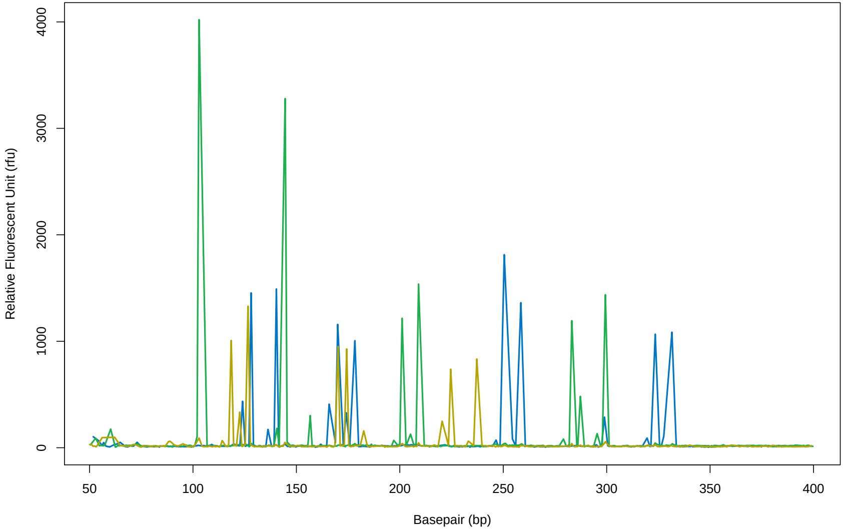
Estimates of alpha and tau are updated upon plotting

☐ Open plot in new plot window

☐ Add profile table to plot


2750
2250
1750
1250
750
250
50
14
16
18
19
D3
15
17
19
VWA
10
12
14
D16
20
23
24
25
D2
2750
2250
1750
1250
750
250
50
X
Y
AME
10
13
D8
28
30
31
32
D21
12
13
16
D18
2750
2250
1750
1250
750
250
50
12
13
15
D19
6
7
8
9
TH0
20
22
23
FGA

Electropherogram (EPG)



Summarising the EPG

There are several ways $\mathcal{E}_c$ can be included in evidence calculations:

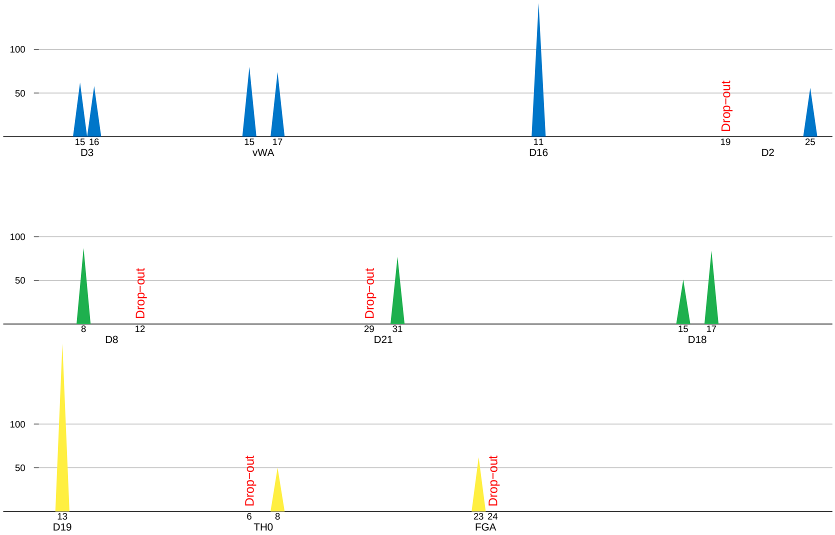
$\mathcal{E}_c$	The entire EPG signal
$\mathcal{E}_c \times \mathbb{I}_{\{x>T\}}(\mathcal{E}_c)$	The part of the EPG signal above T rfu
$\mathbb{I}_{\{x>T\}}(\mathcal{E}_c)$	As above, but discarding peak intensities
...	

Thresholding the EPG

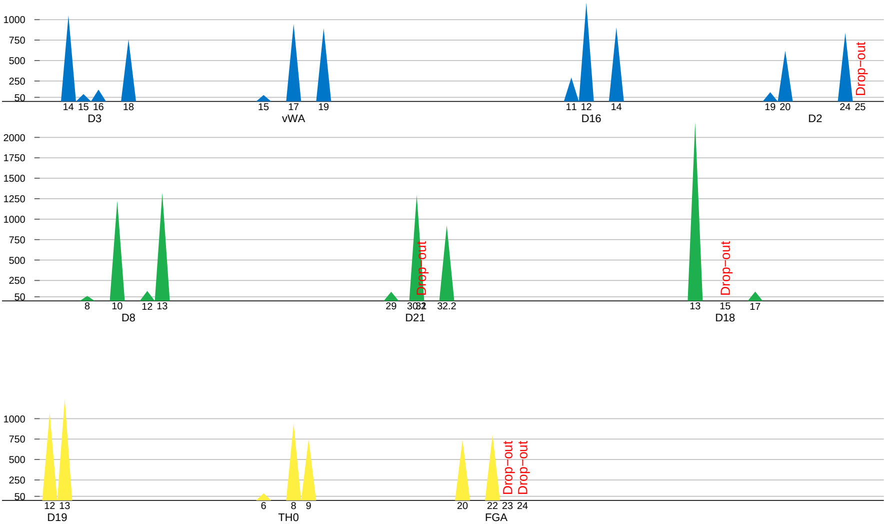
A way of limiting the amount of data obtained from the EPG is to apply a threshold intended to distinguish between noise and true signal. However, this approach introduces other problems:

- Drop-in: Peaks detected above the threshold not ascribed to the contributing DNA profiles.
- Drop-out: When the peak height of a proposed allele is below the threshold, implying that a drop-out probability, $P(D)$, is needed in order to compute the LR.

Low template DNA



Low template DNA



Low template DNA

The probability is primarily relevant under H_p since the this includes the known profile of the suspect. That is,

$$LR \approx \frac{P(D)}{P(G_U|G_S)},$$

i.e. the smaller $P(D)$ the weaker is the evidence against G_S.

Estimating the probability of allelic drop-out

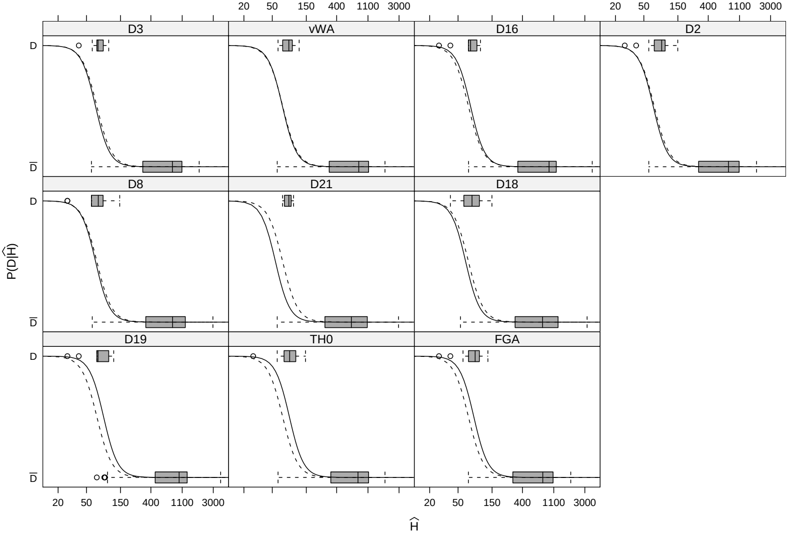
The probability of allelic drop-out can be modelled using logistic regression with a proxy for the amount of DNA as a covariate:

$$\operatorname{logit} P(D; \mathrm{DNA}) = \beta_{0,s} + \beta_1 \log \widehat{H},$$

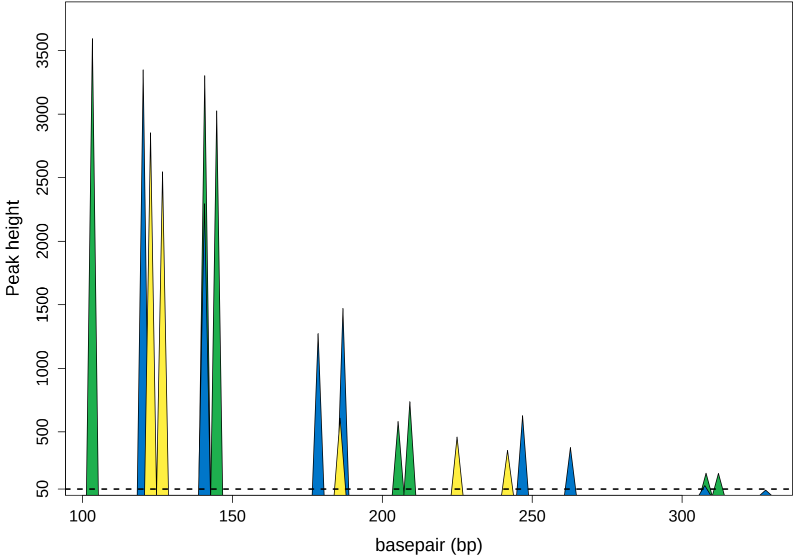
where H is an estimate of the average peak height of a heterozygous allele, hence

$$\mathrm{DNA} \propto \widehat{H} = \begin{cases} H, & \text{Heterozygote allele} \\ 2H, & \text{Homozygote allele} \end{cases}$$

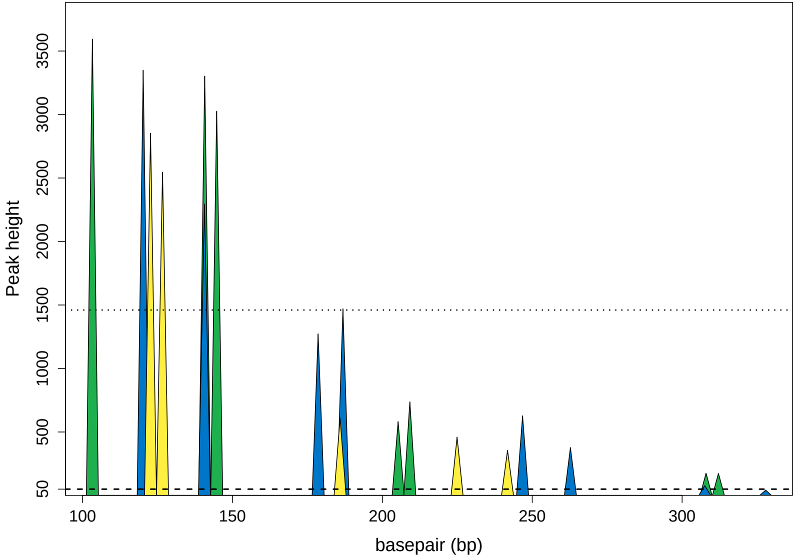
Estimating the probability of allelic drop-out



Damaged and broken DNA fragments



Damaged and broken DNA fragments



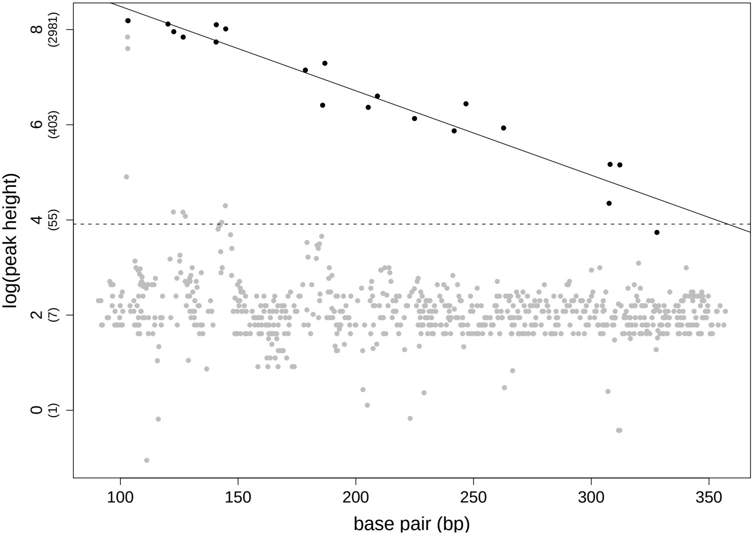
Damaged and broken DNA fragments

For the data producing the plot $H = 1460.41$ rfu. All alleles of the DNA profile is present except allele 24 on D2.

Probability of allelic drop-out **not** taking degradation into account:

$$P(D_{\mathrm{D2}_{24}}; H = 1460.41) = 1.54 \cdot 10^{-6}$$

Modelling the intensity decay



Modelling the intensity decay

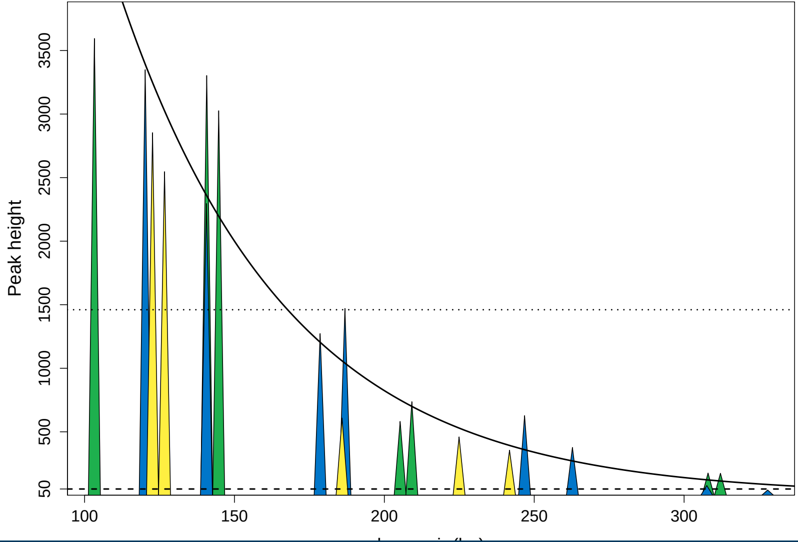
We modelled the intensity decay using a log-linear model

$$\log H(\mathrm{bp}) = \alpha_0 + \alpha_1 \mathrm{bp}$$

Note how this formulation may be substituted into the model for estimating the probability of allelic drop-out:

$$\begin{aligned} \mathrm{logit}\, P(D; H) &= \beta_{0,s} + \beta_1 \log \widehat{H} \\ &= \beta_{0,s} + \beta_1 \log H(\mathrm{bp}) \\ &= \beta_{0,s} + \beta_1 (\alpha_0 + \alpha_1 \mathrm{bp}) \end{aligned}$$

Modelling the intensity decay



Modelling the intensity decay

From before we had that the drop-out probability was $1.54 \cdot 10^{-6}$.

Adjusting for degradation by the fitted solid line:

$$P(D_{\mathrm{D2}_{24}}; H(\mathrm{bp} = 327.87)) = 0.26$$

Since $LR \approx P(D)/P(G_U|G_S)$ this implies that the weight of evidence is increased by more than 10^5 by adjusting for degradation.

Thank you for your attention...