AALBORG UNIVERSITY

On finding the best matching pair of profiles in two-person DNA mixtures and evaluation of the likelihood ratio

Torben Tvedebrink

PhD student, MSc

Department of Mathematical Sciences
Aalborg University

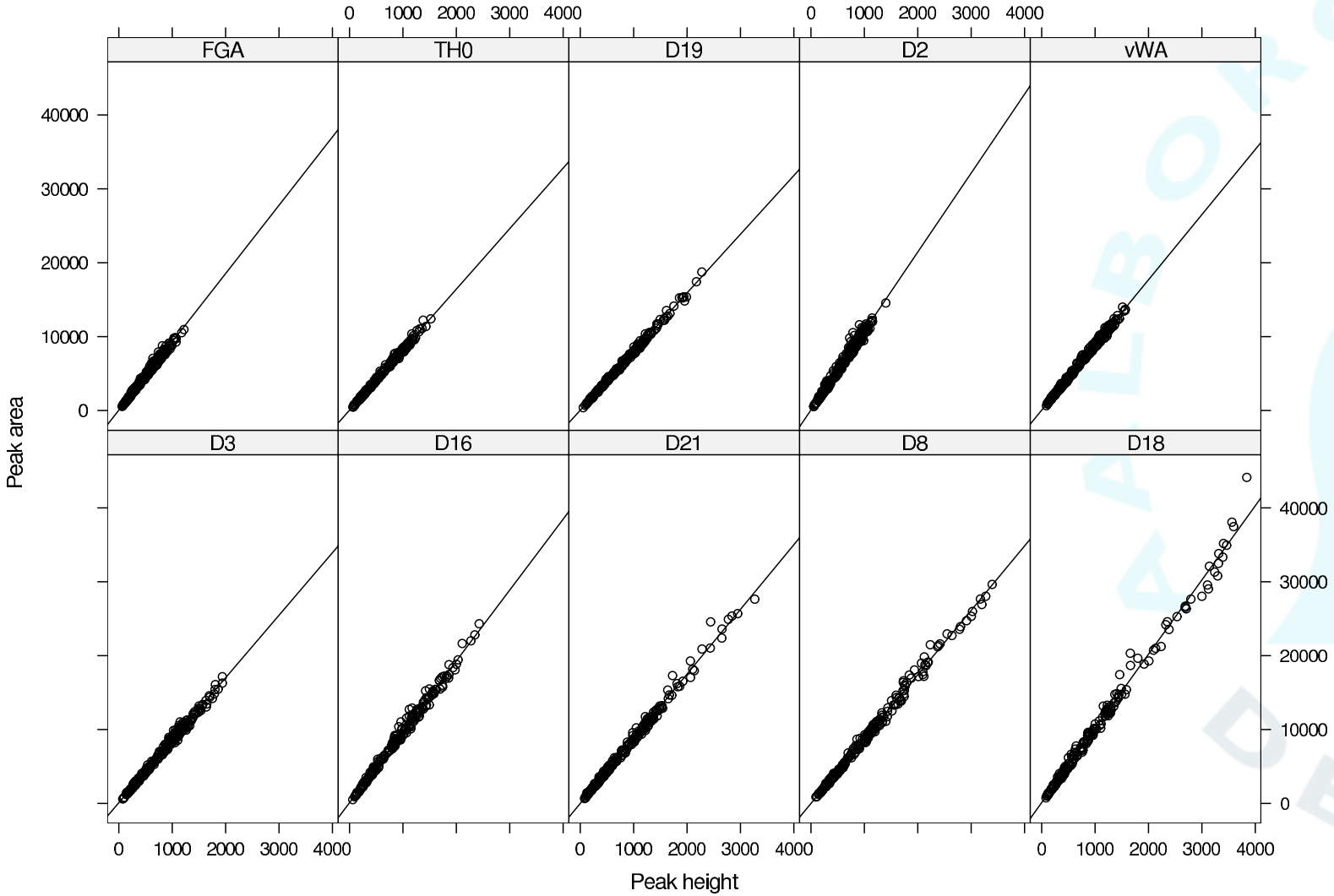
August 22 2008 - Lausanne

Outline

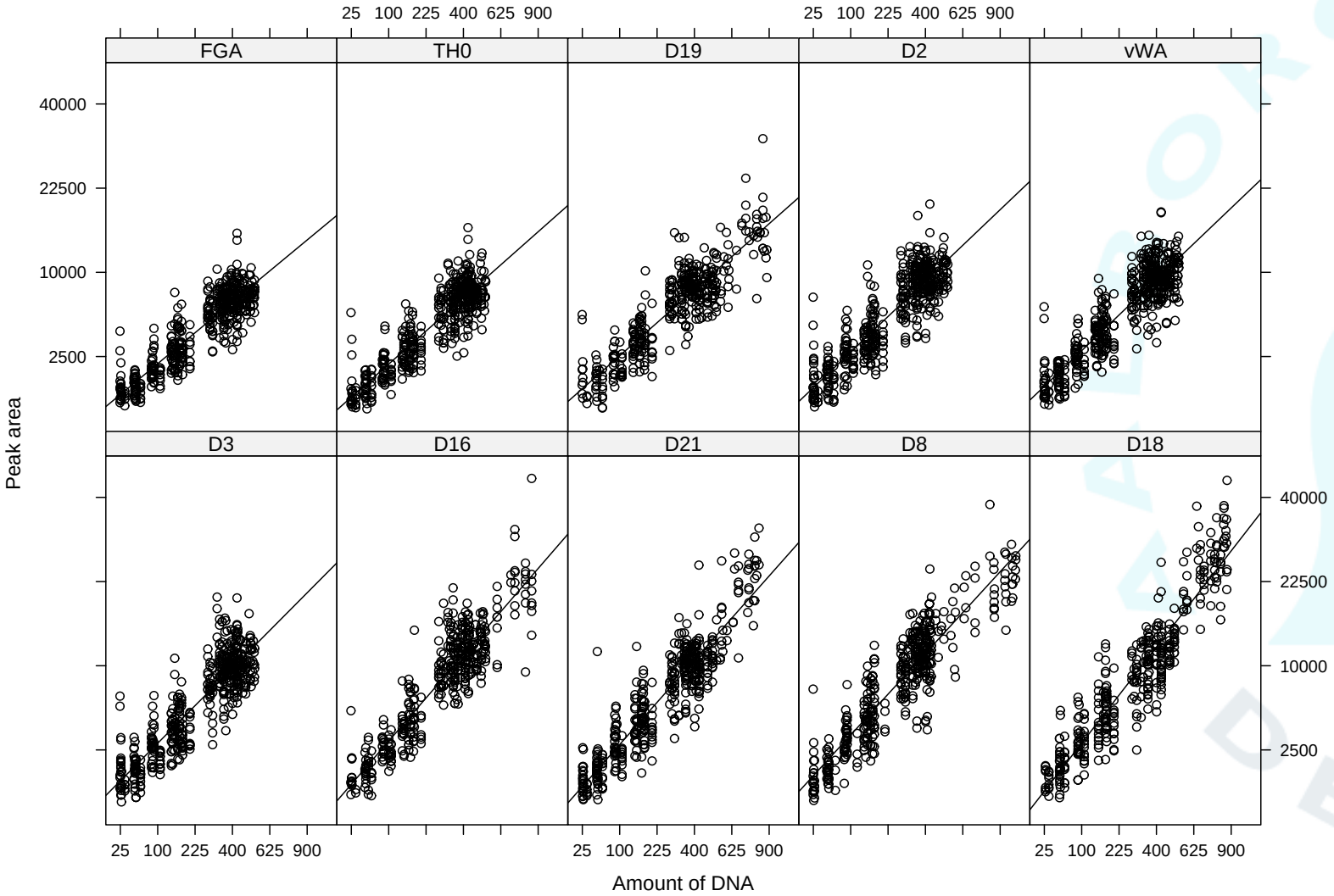
Data exploration

Controlled experiments with pair-wise mixtures of four different profiles in varying ratios conducted at the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen.

Data exploration



Data exploration



Data exploration

From the data exploration we observe the following proportionalities:

- peak height and peak area.
- peak area and amount of DNA.
- mean and variance of peak areas.

Conditional independence of loci

We model the peak areas conditioned on the loci peak area sums, $\mathbf{A}_+ = (A_{s,+})_s \in \mathcal{S}$, where $\mathcal{S}$ is the set of all loci used for discrimination (included in the typing kit).

We assume conditional independence of the peak areas on different loci conditioned on the loci sums,

$$\mathbf{A}_s \perp\!\!\!\perp \mathbf{A}_t | (A_{s,+}, A_{t+}).$$

We note that $\mathbf{A}_+$ is an ancillary statistic as it is fixed for all pairs of profiles and thus contains no information on the profiles.

Normal distributed peak areas

For each locus $s \in \mathcal{S}$ we assume the peak areas follow a conditional normal distribution with the following mean and variance

$$\begin{aligned} \mathbb{E}(\mathbf{A}_s|A_{s,+}) &= \frac{A_{s,+}}{2}\left[\alpha \mathbf{P}_{s,1} + (1-\alpha)\mathbf{P}_{s,2}\right] \\ \mathbb{C}\mathrm{ov}(\mathbf{A}_s|A_{s,+}) &= \tau^2 C_s \mathrm{diag}(\mathbf{h}_s) C_s^\top = \tau^2 W_s, \end{aligned}$$

where $\mathbf{P}_{s,k}$ is an indicator vector with elements 0,1 or 2 indicating whether person k has none, one or two copies of the particular allele.

The α-parameter models the share of person 1 in the mixture and common for all loci.

Furthermore, C_s is given as $I_{n_s} - n_s^{-1}\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top$ where $\mathrm{diag}(\mathbf{h}_s)$ in the covariance ensures proportionality of mean and variance.

Estimators

The estimators of α and τ^2 are found as

$$\begin{aligned}\hat{\alpha} &= \frac{\sum_{s\in\mathcal{S}} \mathbf{x}_0^{s\top} W_s^{-} (\mathbf{A}_s - \mathbf{x}_1^s)}{\sum_{s\in\mathcal{S}} \mathbf{x}_0^{s\top} W_s^{-} \mathbf{x}_0^s} \\ \hat{\tau}^2 &= N^{-1} \sum_{s\in\mathcal{S}} (\mathbf{A}_s - \hat{\alpha}\mathbf{x}_0^s - \mathbf{x}_1^s)^\top W_s^{-} (\mathbf{A}_s - \hat{\alpha}\mathbf{x}_0^s - \mathbf{x}_1^s),\end{aligned}$$

where $\mathbf{x}_0^s = \frac{A_{s,+}}{2}(\mathbf{P}_{s,1} - \mathbf{P}_{s,2})$ and $\mathbf{x}_1^s = \frac{A_{s,+}}{2}\mathbf{P}_{s,2}$ and $N = \sum_{s\in\mathcal{S}}(n_s - 1) - 1 = n_+ - S - 2$.

The matrix W_s^- is the generalised inverse since W_s is not of full rank due to conditioning on the sum $A_{s,+}$.

Sorting the observed peak areas

We assume that $DNA_1 < DNA_2$ such that $\alpha < 0.5$. Furthermore, we sort the observed peak areas $A_{s,(1)} < \cdots < A_{s,(n_s)}$.

This reduces the number of possible combinations for the best matching pair of profiles.

$\mathcal{J}_1$:	$\mathbf{P}_1$	$\mathbf{P}_2$
$A_{s,(1)}$	2	2

$\mathcal{J}_2$:	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$
$A_{s,(1)}$	1	1	2	0	1	0	0	1
$A_{s,(2)}$	1	1	0	2	1	2	2	1

$\mathcal{J}_3$:	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$
$A_{s,(1)}$	2	0	1	0	1	0	0	1
$A_{s,(2)}$	0	1	1	0	0	1	0	1
$A_{s,(3)}$	0	1	0	2	1	1	2	0

$\mathcal{J}_4$:	$\mathbf{P}_1$	$\mathbf{P}_2$
$A_{s,(1)}$	1	0
$A_{s,(2)}$	1	0
$A_{s,(3)}$	0	1
$A_{s,(4)}$	0	1

Example on sorting of areas

Assume that for locus s we observe the following peak areas $\mathbf{A}_s = (100, 200, 300)$. Then there exists 12 possible combinations:

$A_{s,(i)}$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$
100	1	0	1	1	1	0	1	1	1	0	1	0
200	1	1	1	0	1	0	0	1	0	1	0	2
300	0	1	0	1	0	2	1	0	1	1	1	0

$A_{s,(i)}$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$
100	2	0	0	1	0	1	0	2	0	1	0	1
200	0	1	2	0	0	1	1	0	1	1	1	0
300	0	1	0	1	2	0	1	0	1	0	1	1

Example on sorting of areas

Assume that for locus s we observe the following peak areas $\mathbf{A}_s = (100, 200, 300)$. Then there exists 12 possible combinations:

$A_{s,(i)}$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$
100	1	0	1	1	1	0	1	1	1	0	1	0
200	1	1	1	0	1	0	0	1	0	1	0	2
300	0	1	0	1	0	2	1	0	1	1	1	0

$A_{s,(i)}$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$
100	2	0	0	1	0	1	0	2	0	1	0	1
200	0	1	2	0	0	1	1	0	1	1	1	0
300	0	1	0	1	2	0	1	0	1	0	1	1

Greedy algorithm

Algorithm: Find best matching pair of profiles

Let $\mathcal{T} = \emptyset$, $\hat{\alpha} = 0$ and $\hat{\tau}^2 = \infty$.

While $\hat{\tau}^2$ decreases or $\mathcal{T} \neq \mathcal{S}$

For $i \in \{4, 3, 2\}$

For $s \in \mathcal{S}_i = \{s : s \in \mathcal{S} \text{ and } n_s = i\}$

Choose combination $j \in \mathcal{J}_i$ minimising $\hat{\tau}^2$

Set $\mathcal{T} = \{\mathcal{T} \setminus (s, \cdot)\} \cup (s, j)$ and compute $\hat{\alpha}$

Return $\hat{\alpha}$, $\hat{\tau}$ and $\mathcal{T}$.

Properties of the greedy algorithm

The success of the greedy algorithm depends on the presence of loci with four alleles.

The fact that the algorithm is greedy only ensures that it finds the local maximum in each iteration. However, for practical purposes it will in most cases also find the global maximum (this depends on the presence of loci with full information).

Online tool

The greedy algorithm has been implemented as an online tool available at www.math.aau.dk/∼tvede/dna.

Features:

- Upload of csv-files
- Graphical evaluation of the fit
- Estimates of α and τ^2 together with trace
- Fixing a suspect profile
- Testing goodness of fit for each locus
- Generation of alternatives *close* to the best matching profiles

Data

Blue flourescent dye band

Locus	*Allele*			*Height*	*Area*
D3	15	○	●	1802	15410
D3	16	○	●	1939	16282
vWA	14	○		712	6128
vWA	15		●	725	6620
vWA	16	○		626	5637
vWA	17		●	830	7362
D16	10	○		824	7910
D16	11		●	1772	17231
D16	12	○		586	6101
D2	17	○		434	4558
D2	19		●	612	6563
D2	25	○	●	843	9257

Green flourescent dye band

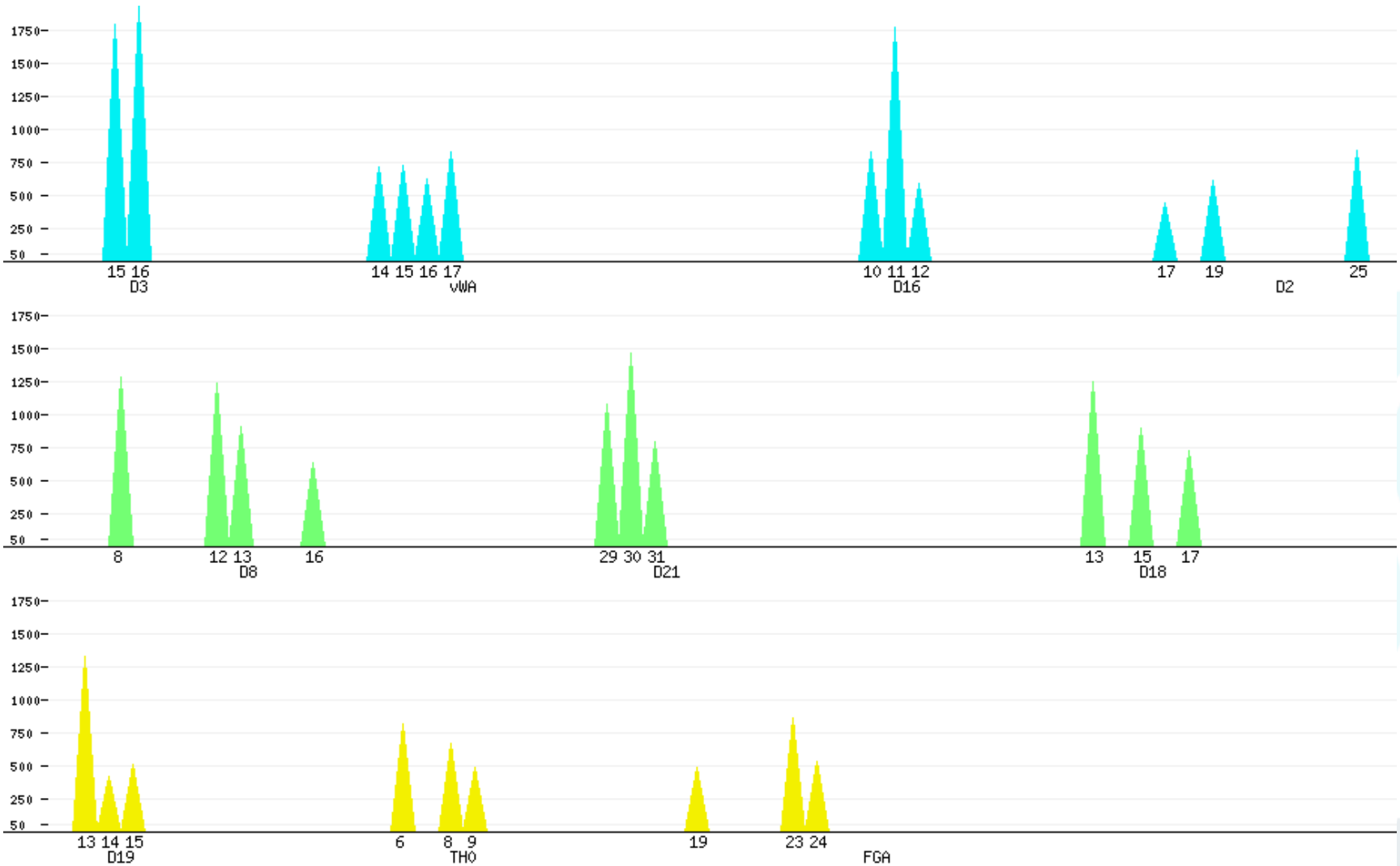
Locus	*Allele*			*Height*	*Area*
D8	8		●	1284	10782
D8	12		●	1232	10359
D8	13	○		903	7891
D8	16	○		638	5291
D21	29		●	1073	9454
D21	30	○		1469	12828
D21	31		●	798	6992
D18	13	○		1247	12302
D18	15		●	899	9104
D18	17		●	726	7549

Yellow flourescent dye band

Locus	*Allele*			*Height*	*Area*
D19	13		●	1332	10534
D19	14	○		416	3478
D19	15	○		504	3968
TH0	6	○	●	820	6739
TH0	8		●	668	5573
TH0	9	○		486	4004
FGA	19	○		490	4415
FGA	23	○	●	865	7968
FGA	24		●	527	5036

Data from a controlled two-person mixture experiment. The true profiles are marked by ○ (profile 1) and ● (profile 2).

Data


1750
1500
1250
1000
750
500
250
50
15 16
D3
14 15 16 17
vWA
10 11 12
D16
17
19
25
D2
8
12 13
D8
16
29 30 31
D21
13
15
17
D18
13 14 15
D19
6
8 9
TH0
19
23 24
FGA

Data

Analysing the mixture using the online DNA mixture separator tool:

http://www.math.aau.dk/∼tvede/dna

Estimating the weight of the evidence

When estimating the weight of evidence in a forensic genetic setting it is common to use a likelihood ratio of two contradictory hypotheses, H_p and H_d.

Data from STR DNA analyses comprises both qualitative (genetic stain, $\mathcal{G}$) and quantitative (peak heights and areas, $\mathcal{Q}$) evidence. Hence, the evidence $\mathcal{E} = (\mathcal{G}, \mathcal{Q})$. Using the definition of conditional probability we have,

$$P(\mathcal{E}|H) = P(\mathcal{G}, \mathcal{Q}|H) = P(\mathcal{Q}|\mathcal{G}, H)P(\mathcal{G}|H).$$

Let $(G', G'') \equiv \mathcal{G}$ be the set of all pairs of profiles G' and G'' consistent with $\mathcal{G}$. Let further $G' : (G', G'') \equiv \mathcal{G}$ be the set of all G' that together with G'' is consistent with $\mathcal{G}$.

Likelihood ratio

The likelihood ratio can then be written as,

$$LR = \frac{P(\mathcal{E}|H_p)}{P(\mathcal{E}|H_d)}$$

Likelihood ratio

The likelihood ratio can then be written as,

$$
\begin{aligned}
LR &= \frac{P(\mathcal{E}|H_p)}{P(\mathcal{E}|H_d)} \\
&= \frac{\sum_{G_U:(G_U,G_S)\equiv\mathcal{G}} P(\mathcal{Q}|G_U,G_S)P(G_U)}{\sum_{(G_{U_1},G_{U_2})\equiv\mathcal{G}} P(\mathcal{Q}|G_{U_1},G_{U_2})P(G_{U_1},G_{U_2})}
\end{aligned}
$$

Likelihood ratio

The likelihood ratio can then be written as,

$$
\begin{aligned}
LR &= \frac{P(\mathcal{E}|H_p)}{P(\mathcal{E}|H_d)} \\
&= \frac{\sum_{G_U:(G_U,G_S)\equiv\mathcal{G}} P(\mathcal{Q}|G_U,G_S)P(G_U)}{\sum_{(G_{U_1},G_{U_2})\equiv\mathcal{G}} P(\mathcal{Q}|G_{U_1},G_{U_2})P(G_{U_1},G_{U_2})} \\
&= \frac{\sum_{G_U:(G_U,G_S)\equiv\mathcal{G}} L(\mathbf{A}|G_U,G_S)P(G_U)}{\sum_{(G_{U_1},G_{U_2})\equiv\mathcal{G}} L(\mathbf{A}|G_{U_1},G_{U_2})P(G_{U_1})P(G_{U_2})}
\end{aligned}
$$

Estimation of LR using simulation

The number of combinations consistent with $\mathcal{G}$ in the denominator of LR equals $1^{S_1}7^{S_2}12^{S_3}6^{S_4}$, where S_i is the number of loci with i observed alleles. For the example this yields 9,029,615,616 terms in the sum.

A possible methodology for estimating the numerator and denominator of $LR = P(\mathcal{E}|H_p)/P(\mathcal{E}|H_d)$ is importance sampling.

Importance sampling

Importance sampling is a method for estimating (often) a mean that is impossible or difficult to compute analytically. Importance sampling increases the probability of sampling the more important points of the sample space for a decrease in the estimate variance.

The denominator of LR may be interpreted as a mean,

$$P(\mathcal{E}|H_d) = \mathbb{E}(f(\mathcal{E}); P) = \sum_{G \equiv \mathcal{G}} L(\mathbf{A}|G)P(G),$$

where $G \equiv \mathcal{G}$ is the set of all combinations consistent with $\mathcal{G}$. However, it is not appropriate to sample $G_i, i = 1, \ldots, M$ using the the allele probabilities, $P(G_i)$, as these does not take the quantitative evidence into account.

Importance sampling

Defining a proposal function, q, we may improve the sampling using importance sampling. Let $q(G) = \prod_{s \in \mathcal{S}} q_s(G_s)$, where

$$q_s(G_s) = \frac{L(\mathbf{A}|G_s, \hat{G}_{-s})P(G_s)}{\sum_{i=1}^{N_s} L(\mathbf{A}|G_{si}, \hat{G}_{-s})P(G_{si})},$$

where $(G_s, \hat{G}_{-s})$ is the particular combination on locus s and fixing all loci $t \in \mathcal{S} \setminus s$ on best match level, and N_s are the number of combinations for locus s.

I.e. $L(\mathbf{A}|G_s, \hat{G}_{-s})$ are 'marginal' likelihood values.

Importance sampling

Using q we may rewrite the expression $P(\mathcal{E}|H_d)$ as,

$$P(\mathcal{E}|H_d) = \sum_{G \equiv \mathcal{G}} \frac{L(\mathbf{A}|G)P(G)}{q(G)} q(G) = \sum_{G \equiv \mathcal{G}} L(\mathbf{A}|G)W(G)q(G),$$

which is just another mean, $\mathbb{E}(f(\mathcal{E})h(\mathcal{E}); q)$. Note that the importance weights $W(G)$ simplifies

$$\begin{aligned} W(G) &= \frac{P(G)}{\prod_{s \in \mathcal{S}} \frac{L(\mathbf{A}|G_s, \hat{G}_{-s})P(G_s)}{\sum_{i=1}^{N_s} L(\mathbf{A}|G_{si}, \hat{G}_{-s})P(G_{si})}} \\ &= \frac{\prod_{s \in \mathcal{S}} \sum_{i=1}^{N_s} L(\mathbf{A}|G_{si}, \hat{G}_{-s})P(G_{si})}{\prod_{s \in \mathcal{S}} L(\mathbf{A}|G_s, \hat{G}_{-s})} = \frac{B}{\prod_{s \in \mathcal{S}} L(\mathbf{A}|G_s, \hat{G}_{-s})} \end{aligned}$$

Importance sampling

This ensures that

$$L(\mathbf{A}|G)W(G) = \frac{L(\mathbf{A}|G)B}{\prod_{s\in\mathcal{S}} L(\mathbf{A}|G_s, \hat{G}_{-s})}$$

is rather constant which induces lower variance of the estimator.

Hence, in order to estimate $P(\mathcal{E}|H_d)$ we need to draw combinations $G_i \sim q(\cdot), i = 1, \ldots, M$, and compute the estimate

$$\hat{P}(\mathcal{E}|H_d) = \frac{1}{M}\sum_{i=1}^{M} L(\mathbf{A}|G_i)W(G_i).$$

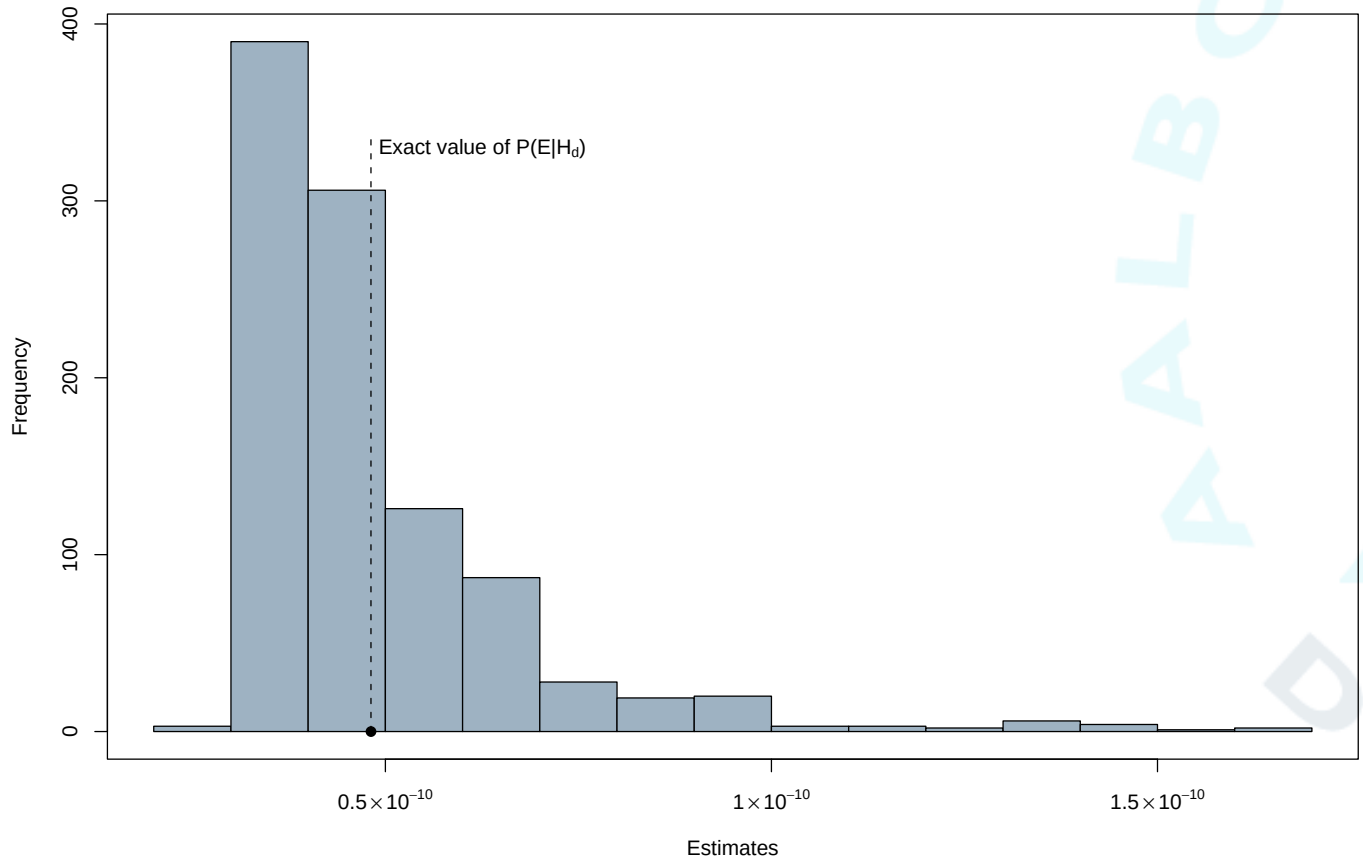
Similar arguments apply for $P(\mathcal{E}|H_p)$ where we use $\hat{G}^{(S)}$ in place of $\hat{G}$ for the restricted set $\{G : G \equiv \mathcal{G}, G_S \subseteq G\}$.

Example - Important sampling

In order to demonstrate the applicability of importance sampling we computed the exact value of $P(\mathcal{E}|H_d)$ for the blue loci (D3, vWA, D16, D2). This reduced the exhaustive list of combinations to $7^1 12^2 6^1 = 6048$ and yielded $P(\mathcal{E}|H_d) = 4.81335 \times 10^{-11}$ when we assume an uniform distribution of the allele probabilities.

Example - Important sampling

We computed 1000 estimates based on 10,000 samples each. The estimates and true value are plotted in the histogram.



Work in progress - Forward sampling

An alternative sampling strategy is based on forward sampling.

1. Sample a list of the loci in random order, $\mathcal{S}^* \sim \pi_{\mathcal{S}}(\cdot)$
2.
   - Sample $s^*_{(1)} \sim q\left(\hat{G}_{-s_{(1)}}\right)$
   - Sample $s^*_{(2)}|s^*_{(1)} \sim q\left(\hat{G}_{-\{s_{(1)},s_{(2)}\}}, s^*_{(1)}\right)$
   - $\vdots$
   - Sample $s^*_{(S)}|s^*_{(1)},\ldots,s^*_{(S-1)} \sim q\left(s^*_{(1)},\ldots,s^*_{(S-1)}\right)$

This method allows for combinations further away from the best matching combination to be sampled compared to the importance sampling scheme.

Conclusion

- The method is based on a statistical model for the peak areas.
- The algorithmic implementation automates the work of the forensic scientist.
- Parameter estimates indicates the goodness-of-fit.
- The free online implementation finds the best matching pair of profiles in a few seconds.
- Fast graphical assessment of the validity of H_p by fixing a suspect profile.
- The model and greedy algorithm are easily extended for more than two contributors.
- The model features the possibility of using sampling techniques for assessing the likelihood ratio.

Acknowledgements

This is work in collaboration with

- Poul Svante Eriksen
  Associate Professor, MSc
  Department of Mathematical Sciences, Aalborg University
- Helle Smidt Mogensen
  Forensic Geneticist, MSc PhD
  Section of Forensic Genetics, Department of Forensic Medicine,
  Faculty of Health Sciences, University of Copenhagen
- Niels Morling
  Professor, MD DMSC
  Section of Forensic Genetics, Department of Forensic Medicine,
  Faculty of Health Sciences, University of Copenhagen