Allelic drop-out in forensic genetics

Importance and estimation

University of Auckland

24th September 2013

Torben Tvedebrink

tvede@math.aau.dk

Department of Mathematical Sciences – Aalborg University, Denmark

AALBORG UNIVERSITY

DENMARK

Forensic genetics

Forensic genetics is the part of forensic medicine that is concerned with the analysis of DNA evidence.

Casework can be grouped into two main categories:

Relationship testing	Crime cases
- Paternity testing	- Burglaries
- Family reunification cases	- Rape cases
	- . . .

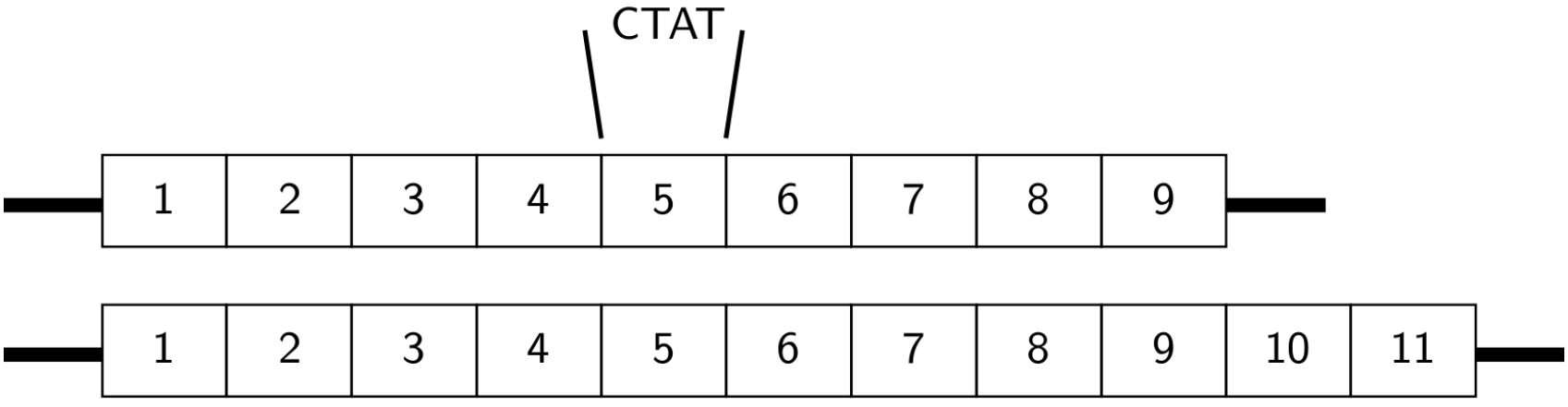
What is a DNA profile?

The majority of the human genome has very little variability between individuals. This makes these regions of limited use for identification purposes. However, specific locations on the genome, e.g. Single Nucleotide Polymorphism (SNP) and Short Tandem Repeat (STR) loci, have a higher levels of polymorphisms. The power of discrimination between individuals and low mutation rates, implies that these regions are applicable for identification and pedigree analysis.

A ’modern’ forensic genetic DNA profile is constituted by the allelic state at 7-16 STR markers. The prevailing method for analysing forensic samples is by capillary electrophoresis, where the genetic profile is read of from a so-called electropherogram (EPG)

STR allele

STR markers are genetic locations where one or more motifs are repeated between a pair of primer sites, which identifies the specific STR marker. An allele represents a variant observed as the number of repeats between the primer sites.



For the example above we denote the genotype (9,11), since only the fragments lengths of the repeated region is measured.

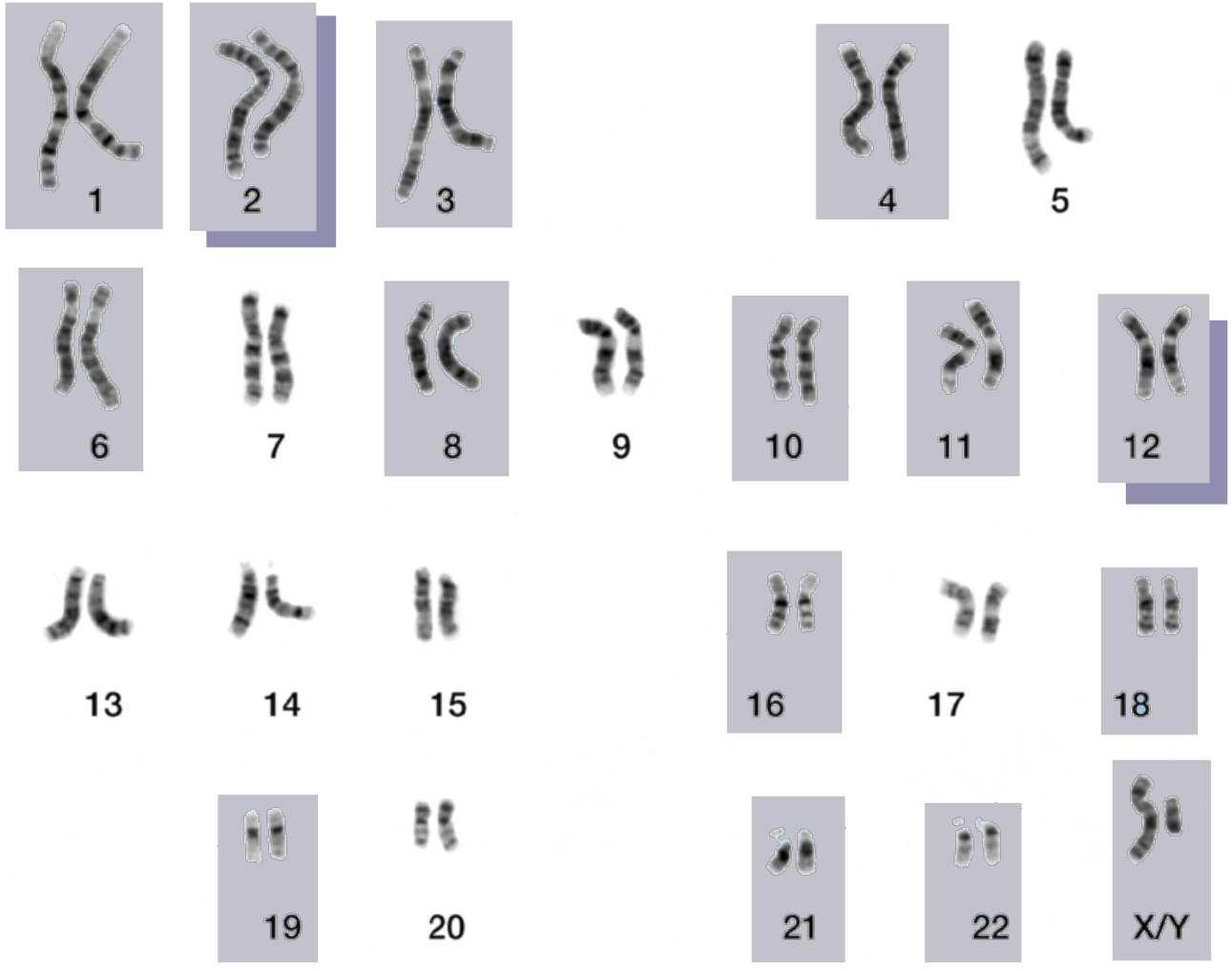
STR markers

The human genome has several STR loci, where one or more DNA motifs are repeated a number of times. However, only a subset of these loci are included in the STR markers used for forensic purposes.

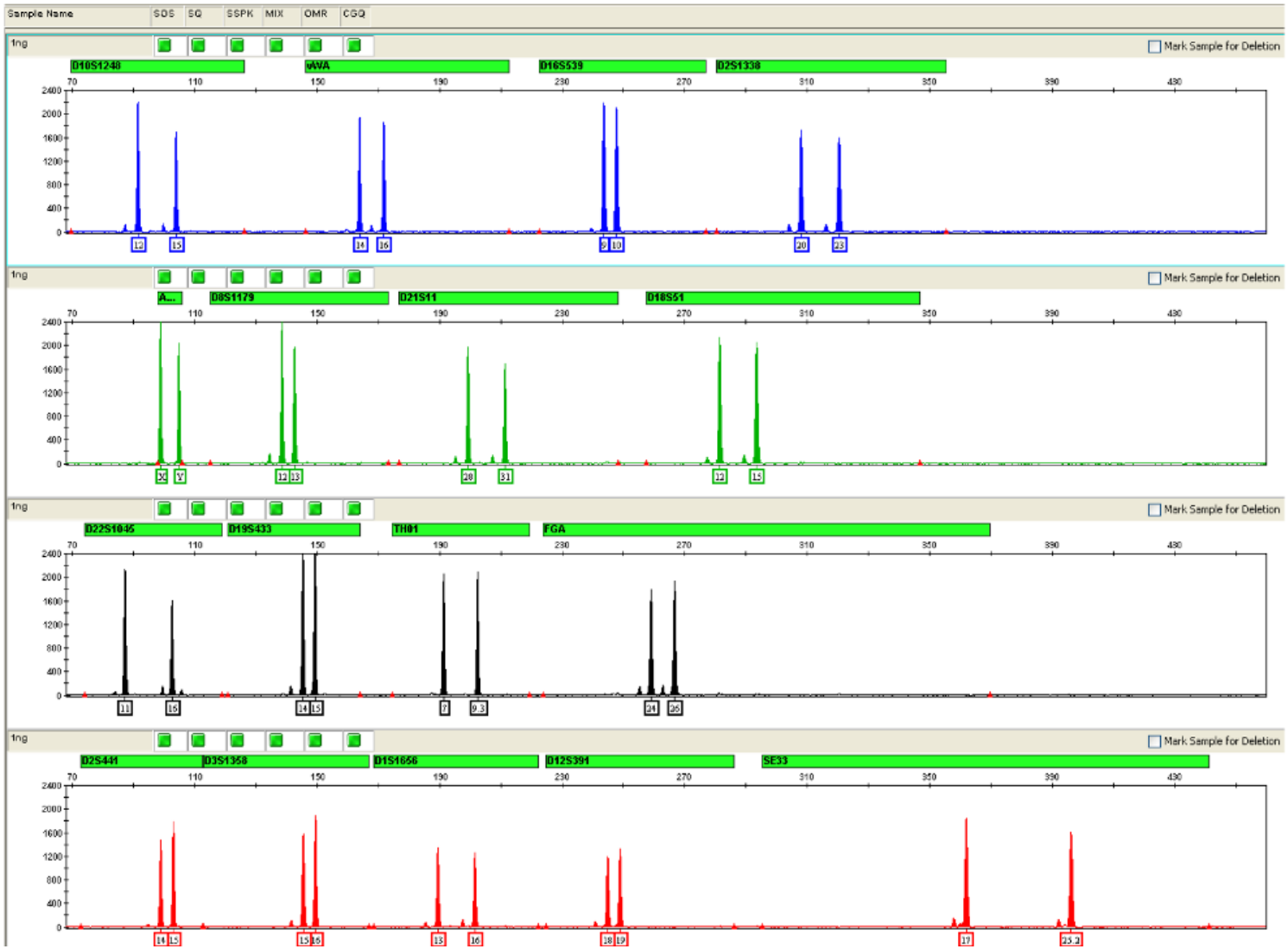
Some key features are:

- Located on different chromosomes (independence assumption)
- Highly polymorphic (many allelic states)
- High power of discrimination (high variation in population)
- Different lengths of primer sites (electrophoresis)
- ...

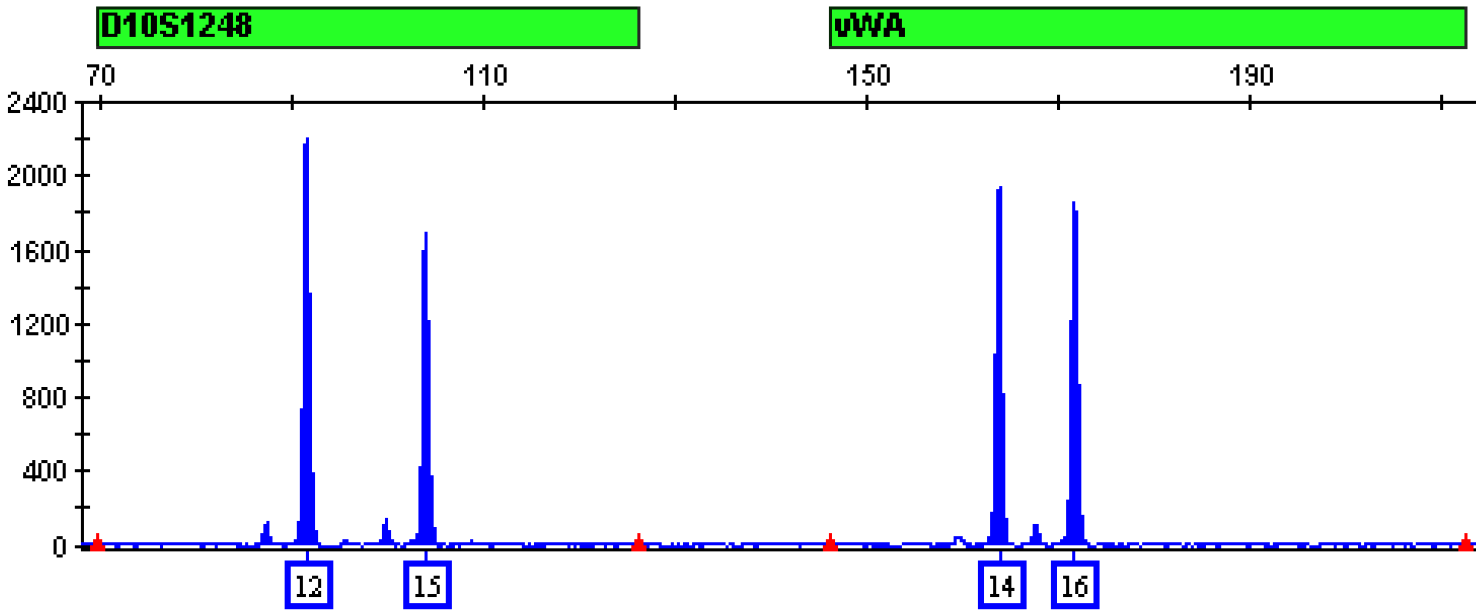
STR markers



STR markers


Sample Name
SDS
SQ
SSPK
MIX
OMR
CGQ
1ng
Mark Sample for Deletion
D10S1248
vWA
D16S539
D2S1338
D8S1179
D21S11
D18S51
D22S1045
D19S433
TH01
FGA
D2S441
D3S1358
D1S1656
D12S391
SE33

STR markers


D10S1248
vWA
70
110
150
190
2400
2000
1600
1200
800
400
0
12
15
14
16

Capillary Electrophoresis

Capillary Electrophoresis (CE) is an old technique and is no longer seen in most other areas genetic analysis. However, due to the challenges from crime scene evidence, forensic genetics has been reluctant in moving towards newer typing technologies.

For example, CE analysis enables

- investigations of limited amounts of DNA,
- identification of mixture proportions in DNA mixtures, and
- detection of degradation of the DNA strand

Likelihood ratio

The principles of evidence interpretation (Evett and Weir, 1998) suggest that the evidence, E, should be (1) considered under at least one alternative hypothesis, (2) evaluated as the conditional on the hypotheses and other relevant information, I.

This principle implies that the evidential weight is assessed using Bayes' theorem,

$$\underbrace{\frac{P(H_p \mid E, I)}{P(H_d \mid E, I)}}_{\text{Posterior odds}} = \underbrace{\frac{P(E \mid H_p, I)}{P(E \mid H_d, I)}}_{\text{Likelihood ratio}} \underbrace{\frac{P(H_p \mid I)}{P(H_d \mid I)}}_{\text{Prior odds}}$$

where H_p and H_d often are refereed to as the prosecutor's and defence's hypotheses, respectively.

Single contributor stain

For a simple one contributor crime scene stain, G_C, a suspect is identified and his DNA profile, G_S, match that of the crime scene, $G_C \equiv G_S$. The evidence is thus $E = (G_S, G_C)$.

In this case, H_p: "The crime scene stain is left by the suspect" while H_d: "An unknown person unrelated to the suspect left the crime scene sample":

$$\begin{aligned} LR &= \frac{P(G_C, G_S|H_p, I)}{P(G_C, G_S|H_d, I)} \\ &= \frac{P(G_C|G_S, H_p, I)}{P(G_C|G_S, H_d, I)} \frac{P(G_S|H_p, I)}{P(G_S|H_d, I)} \\ &= \frac{1}{P(G_C|G_S)}, \end{aligned}$$

where $P(G_C|G_S)$ represents the random match probability.

Crime scene samples

A forensic geneticists are on regular basis challenged by cases where the circumstances of the crime (scene) complicates the analysis:

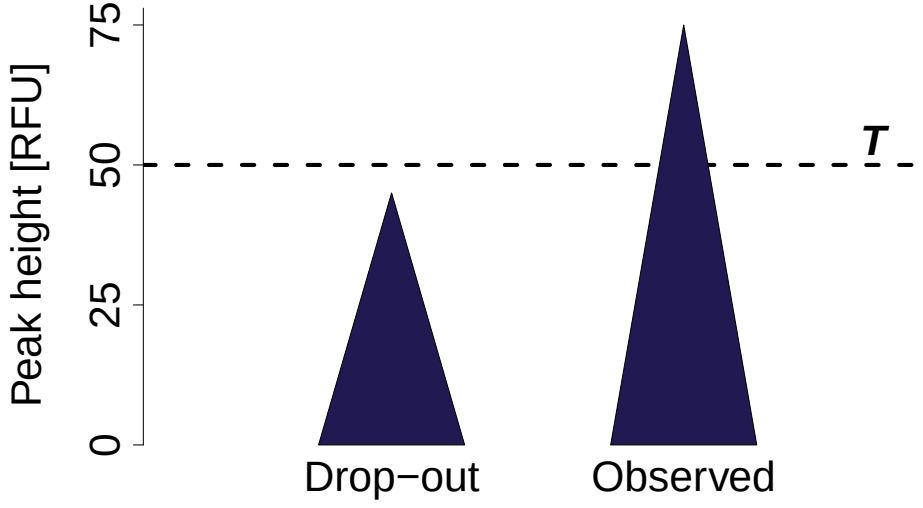
- Extra peaks in the EPG (PCR and CE artifacts)
- More than one contributor to the crime scene stain (DNA mixture)
- Limited amounts of DNA (partial profiles)
- Degraded biological material (damaged DNA strands)
- . . .

Allelic drop-out

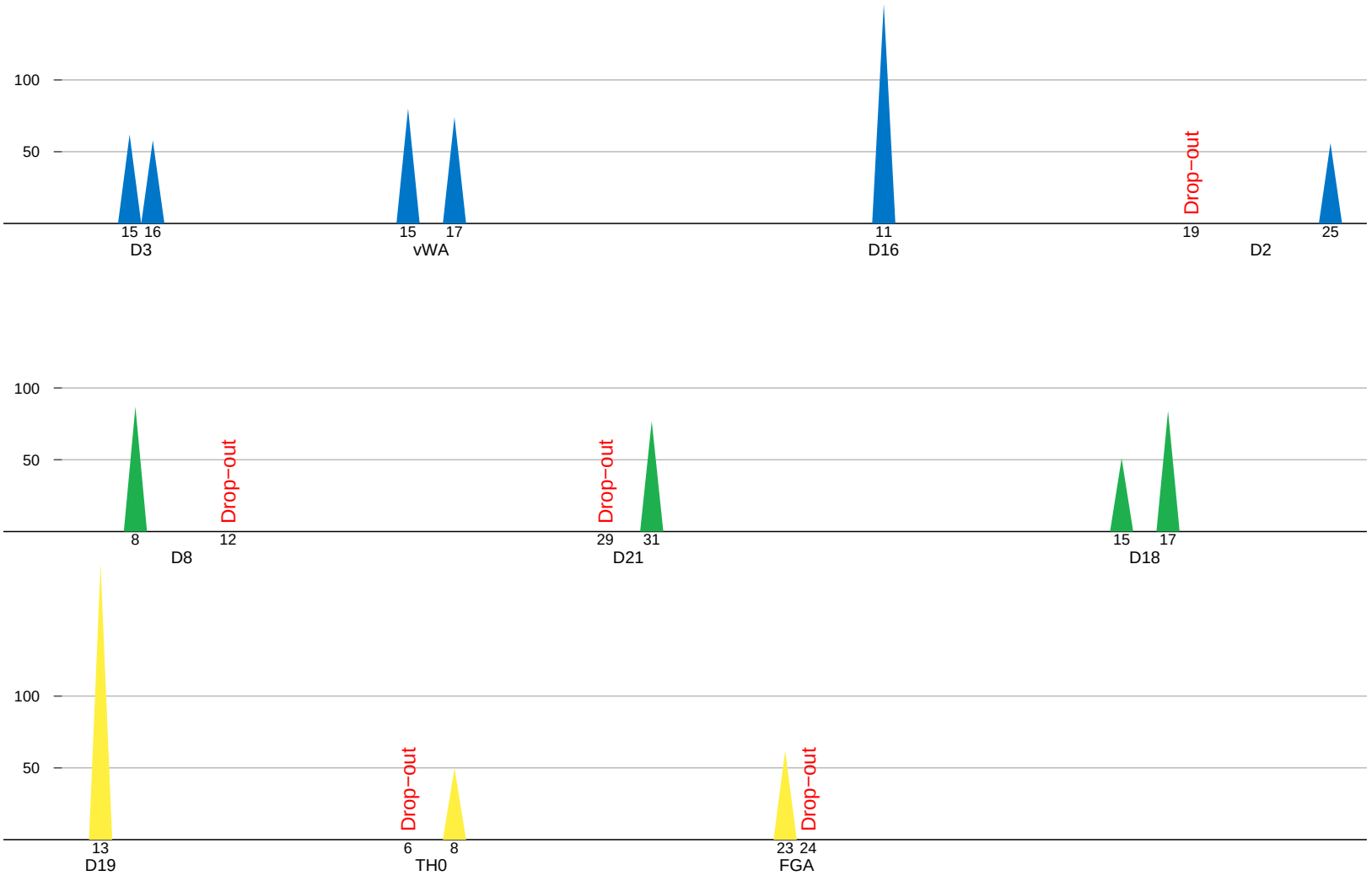
In forensic genetics, the evidential weight should when possible be evaluated by a likelihood ratio, LR.

The exact expression of LR depends on a number of things, and in the case of low-template DNA, also on the drop-out probability, $P(D)$.

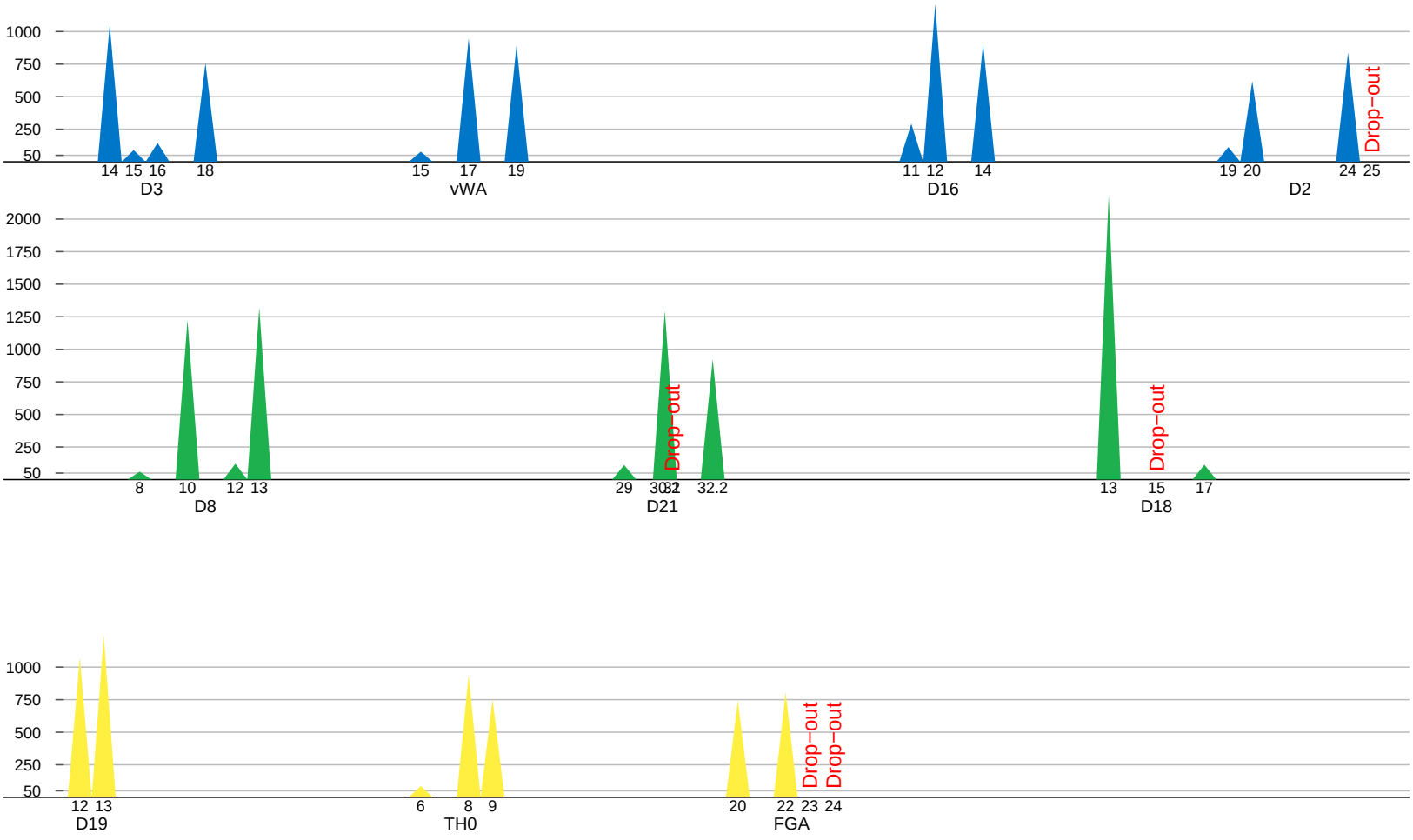
Allelic drop-out occur when alleles of the contributor's DNA profile fail to be detected in the resulting DNA profile. Often, this is equivalent with the peak height, h_i, falling below a detection threshold, T.



Samples with drop-out



Samples with drop-out



Drop-out and evidential weight

In order to incorporate the possibility of allelic drop-out, the *LR*-expression needs an estimate of a drop-out probability $P(D)$.

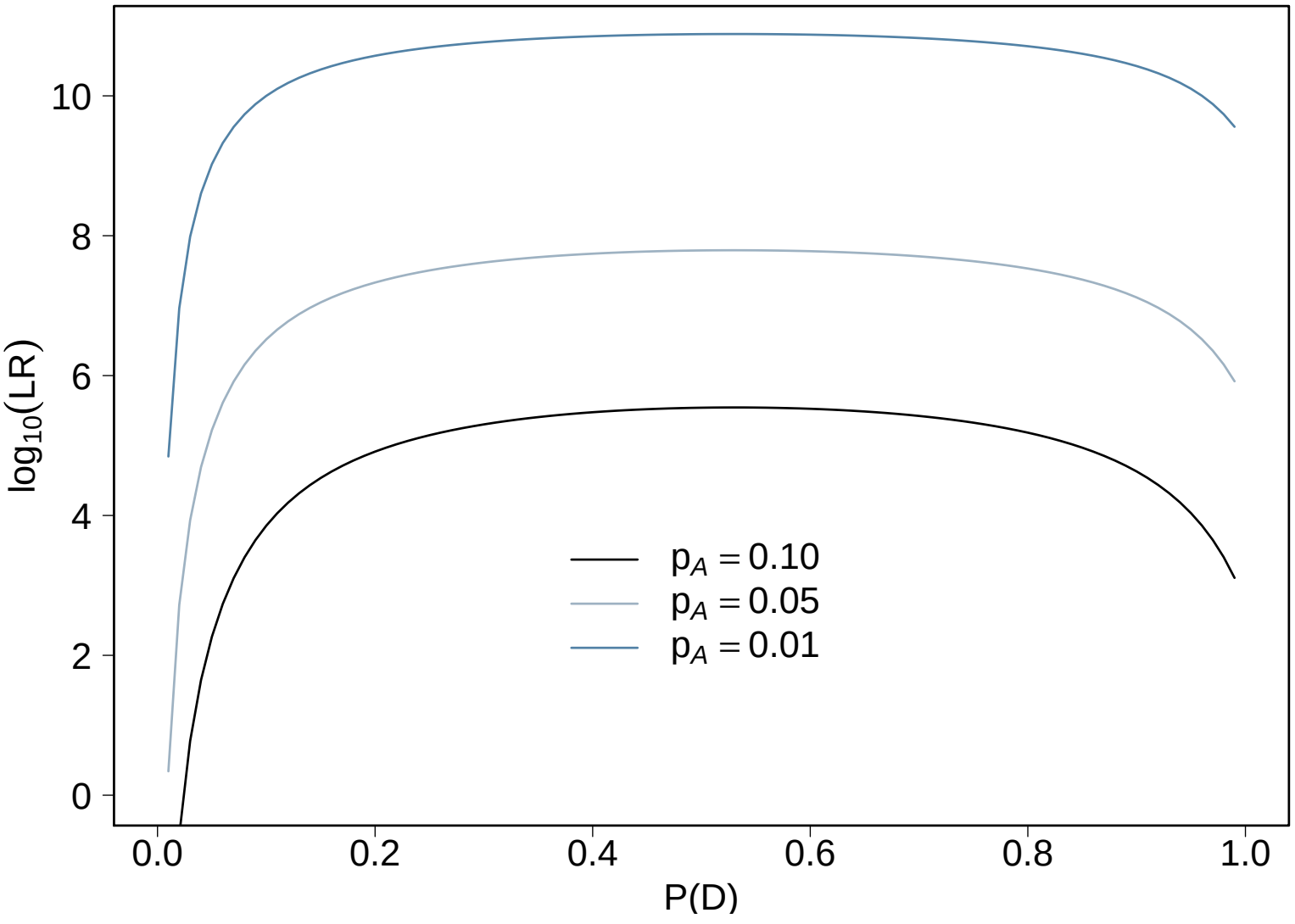
For example, if a suspect is $G_S = (A, B)$, but only allele A is detected in the crime scene evidence, $G_C = (A)$. Hence, under the prosecutors hypothesis allele B must have dropped out:

$$LR = \frac{P(G_C|G_S)}{\sum_{G_U \in \mathcal{G}} P(G_C|G_U)P(G_U|G_S)}$$

$$= \frac{P(D)P(\bar{D})}{P(\bar{D}_2)P(AA|AB) + P(D)P(\bar{D}) \sum_{a \in \mathcal{A} \setminus A} P(Aa|AB)}$$

$$= \frac{P(D)P(\bar{D})}{P(\bar{D}_2)p_A^2 + P(D)P(\bar{D})2p_A(1-p_A)}$$

Why bother? *LR* for ten loci



Properties of the drop-out probability

The drop-out probability should be:

- negatively correlated with the number of DNA templates,
- lower for EPGs with higher peak heights,
- allowed to be profile specific for DNA mixtures,
- ...

The number of observed alleles and $P(D)$

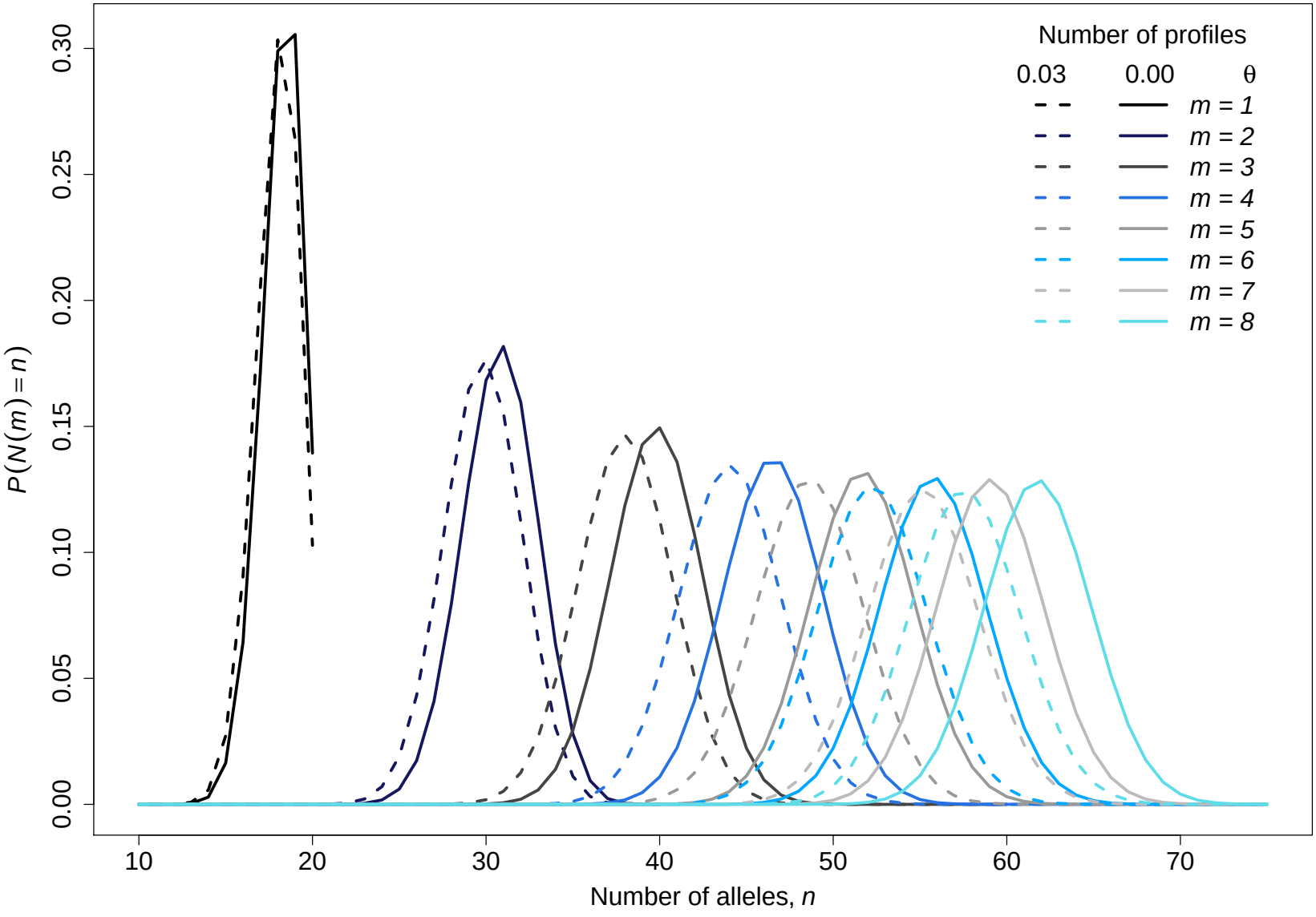
A consequence of allelic drop-out is the fewer alleles with be observed in the EPG.

Hence, early work on allelic drop-out used the number of observed alleles, n_0, to estimate $P(D)$. This was done by a Monte Carlo approach, where the number of alleles was simulated assuming different drop-out probabilities. Hence, the distribution of $P(D)$ was assess by

$$f(P(D)|n_0) \propto f(n_0|P(D))f(P(D)),$$

where $f(P(D))$ was assumed uniform.

The distribution of the number of alleles



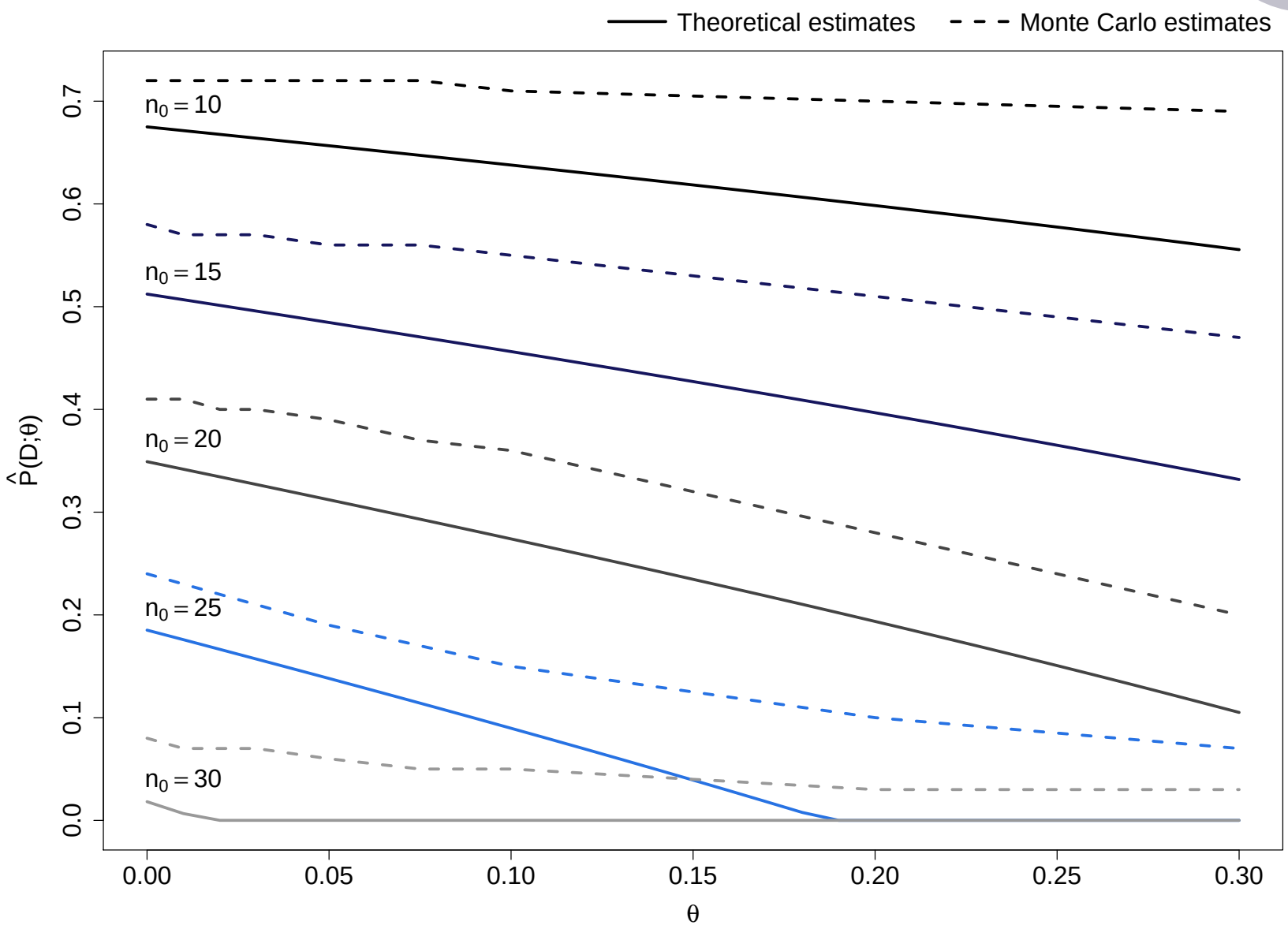
Estimating $P(D)$ based on n_0

For a m-person DNA mixture typed at L STR loci, it is possible to observe 1 through $2mL$ alleles. Let N_{obs} denote the random variable representing the number of observed alleles and $P(D)$ the probability that an allele is missing (the drop-out probability).

Then, allowing for drop-out, the probability of observing n_0 alleles is

$$
\begin{aligned}
P(N_{\text{obs}} = n_0) &= \sum_{i=0}^{2mL-n_0} P(N_{\text{obs}} = n_0, \#D = i) \\
&= \sum_{i=0}^{2mL-n_0} P(\#D = i \mid N = n_0 + i) P(N = n_0 + i) \\
&= [1 - P(D)]^{n_0} \sum_{i=0}^{2mL-n_0} \frac{(n_0 + i)!}{i! n_0!} P(D)^i P(N = n_0 + i)
\end{aligned}
$$

$\hat{P}(D)$ correlated with θ



A plug-in approach to drop-out estimation

The idea we proposed to estimate drop-outs rely on a *plug-in* approach, where a plug-in estimate of the signal intensity is used in a second model.

In general we assume that the peak heights are normally distributed, such that

$$h_i \sim N(\mu n_i, \sigma^2 n_i) \quad \text{where} \quad n_i = \begin{cases} 1 & \text{heterozygous locus} \\ 2 & \text{homozygous locus} \end{cases}$$

The parameter μ is the average peak height of an allele at a heterozygous locus.

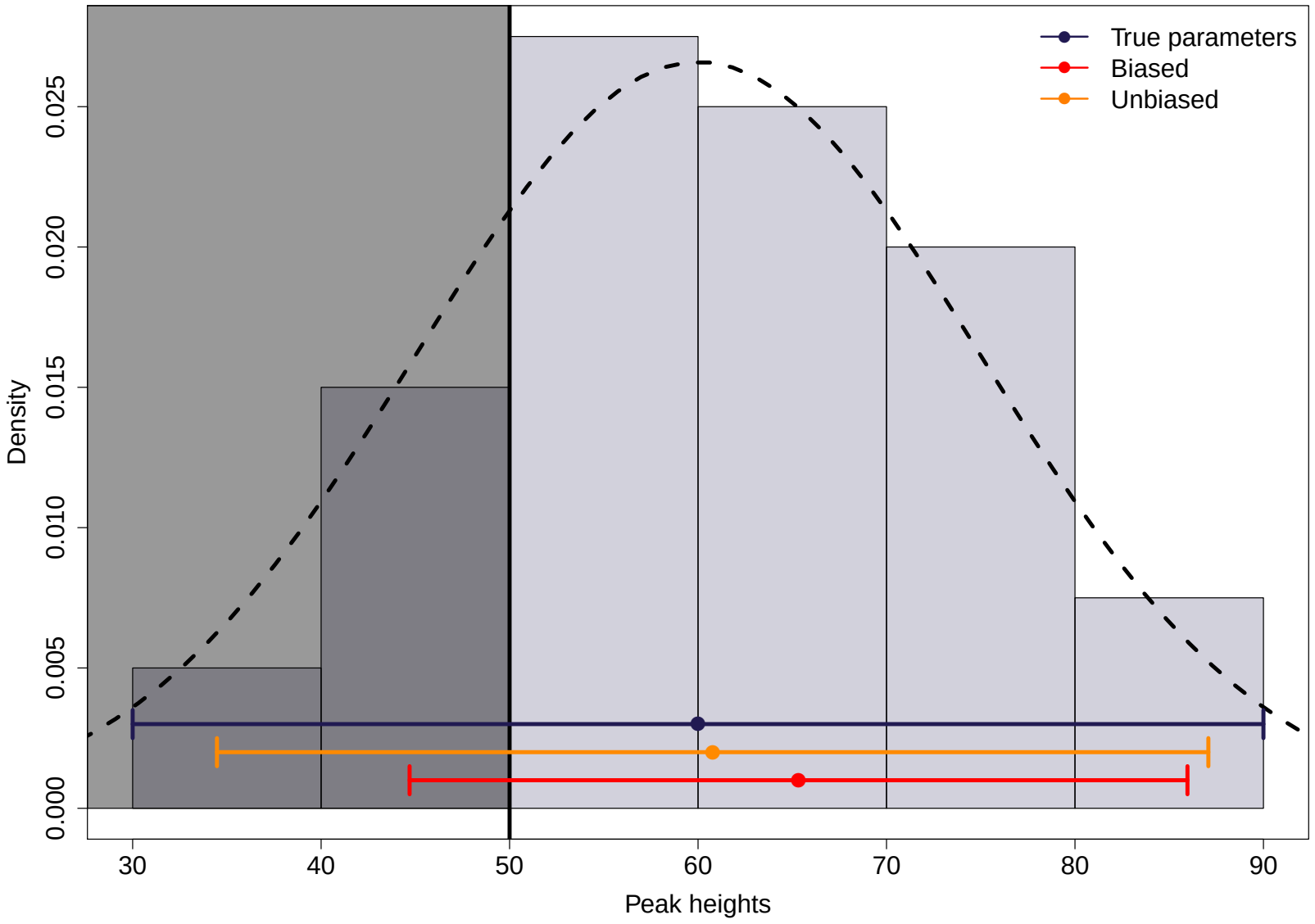
Estimating μ and σ

In order to account for dropped out alleles, the likelihood is given by

$$L(\boldsymbol{h}, \boldsymbol{n}; \mu, \sigma) = \prod_{i=1}^{m_{\text{obs}}} \frac{1}{\sigma\sqrt{n_i}} \phi\left(\frac{h_i - \mu n_i}{\sigma\sqrt{n_i}}\right) \prod_{j=1}^{m_{\text{drop}}} \Phi\left(\frac{T - \mu n_j}{\sigma\sqrt{n_j}}\right),$$

where T is the detection threshold above which alleles are called and below are declared as drop-out.

Truncation example



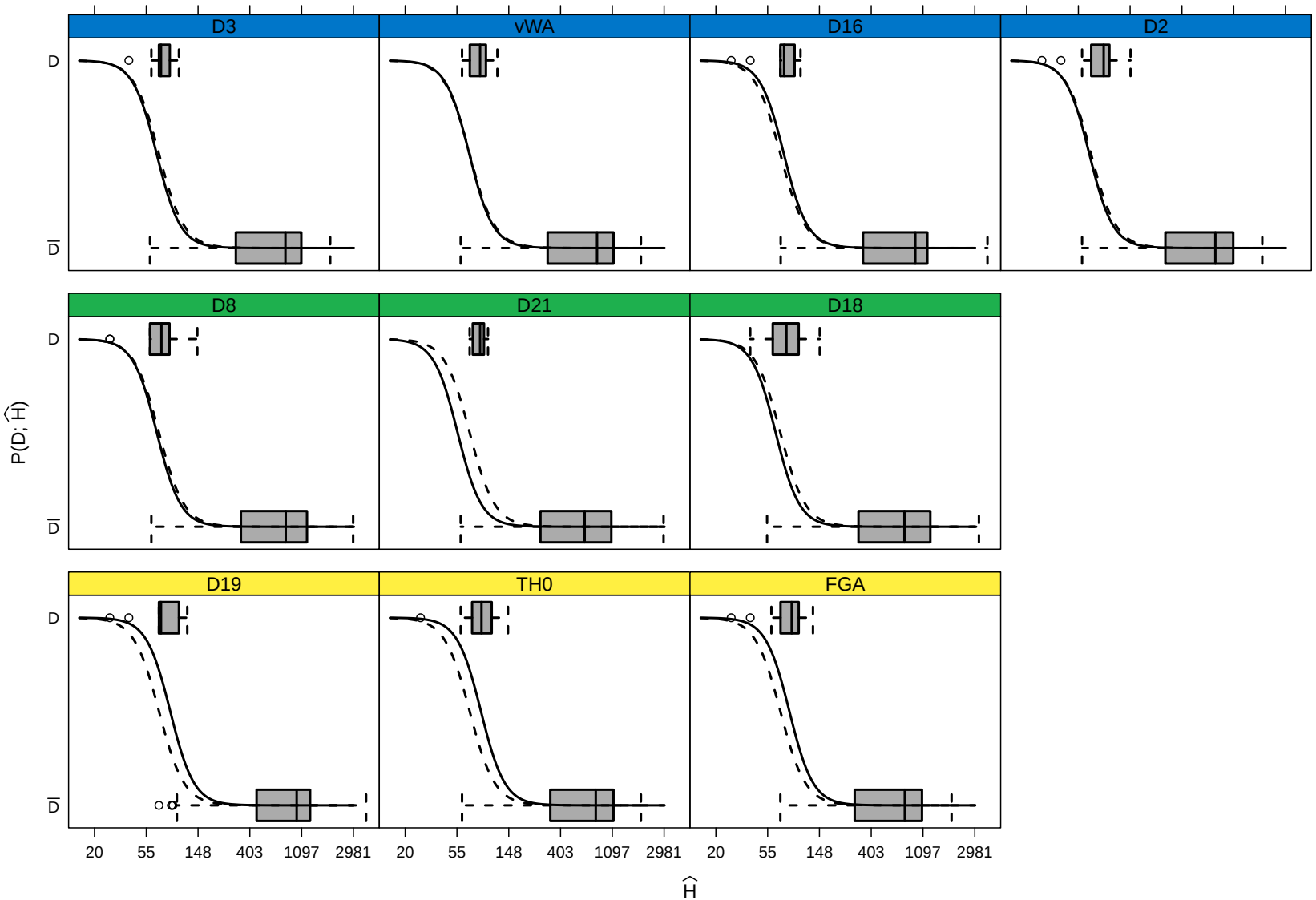
Multi-dose model for drop-out

For some controlled experiments, we used $\hat{H} = \hat{\mu} n_i$ as explanatory variable in the logistic regression

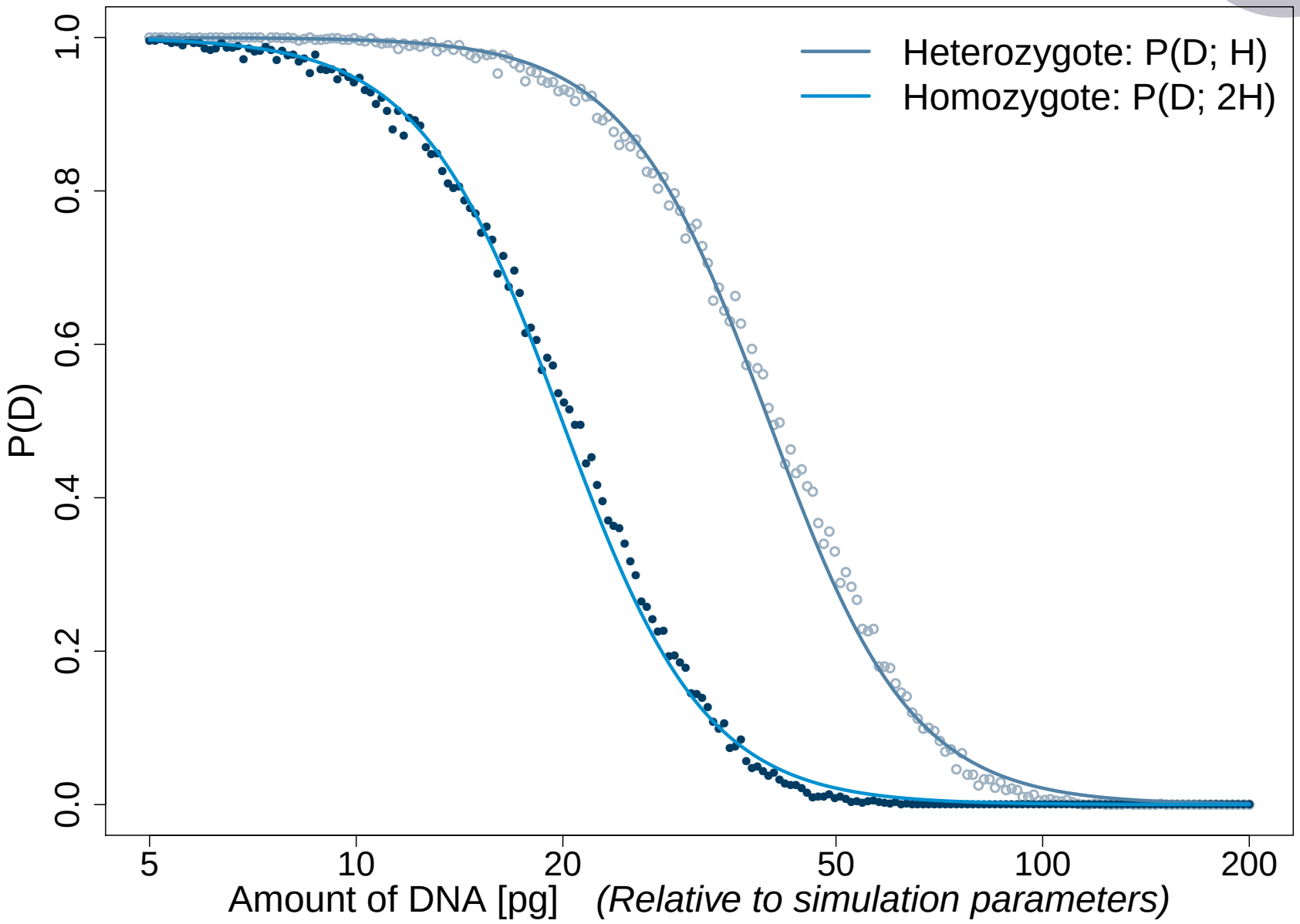
$$\text{logit } P(D; \hat{H}) = \beta_{0,s} + \beta_1 \log \hat{H},$$

where D is an indicator variable, $D = 1$ if $h < T$ and zero otherwise, and β are the regression parameters.

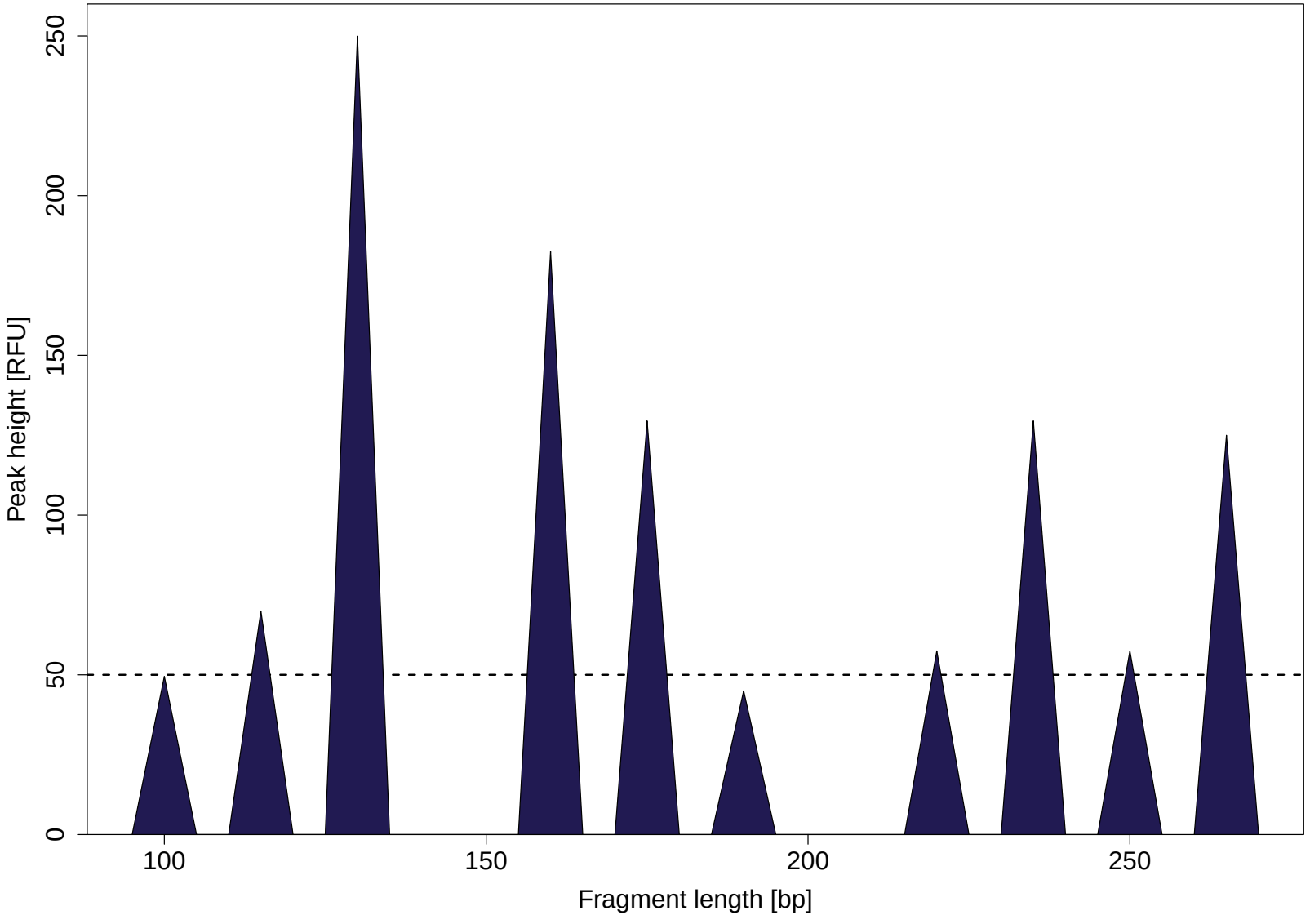
Some drop-out plots



Some drop-out plots

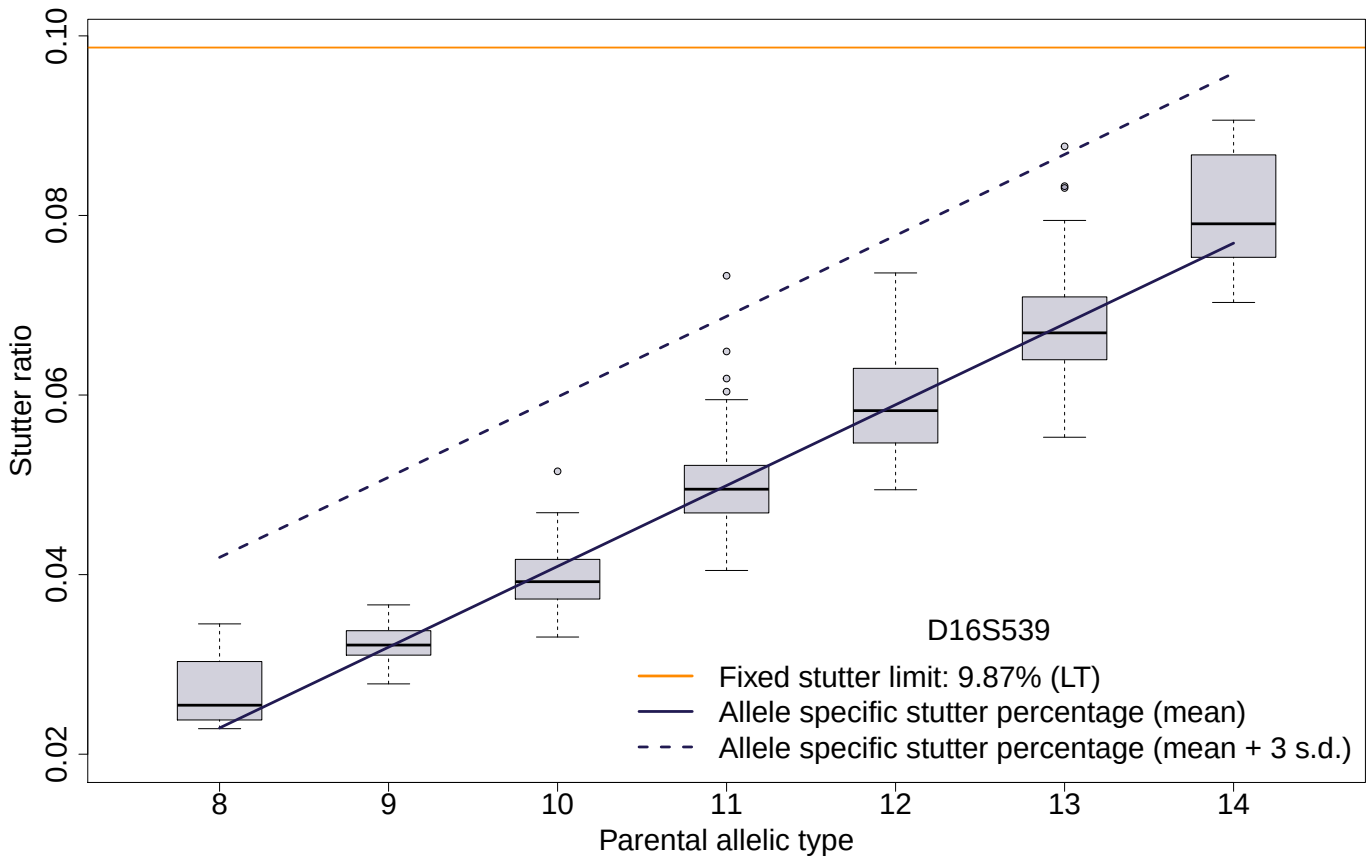


Stutter correction



Compensating for stutter

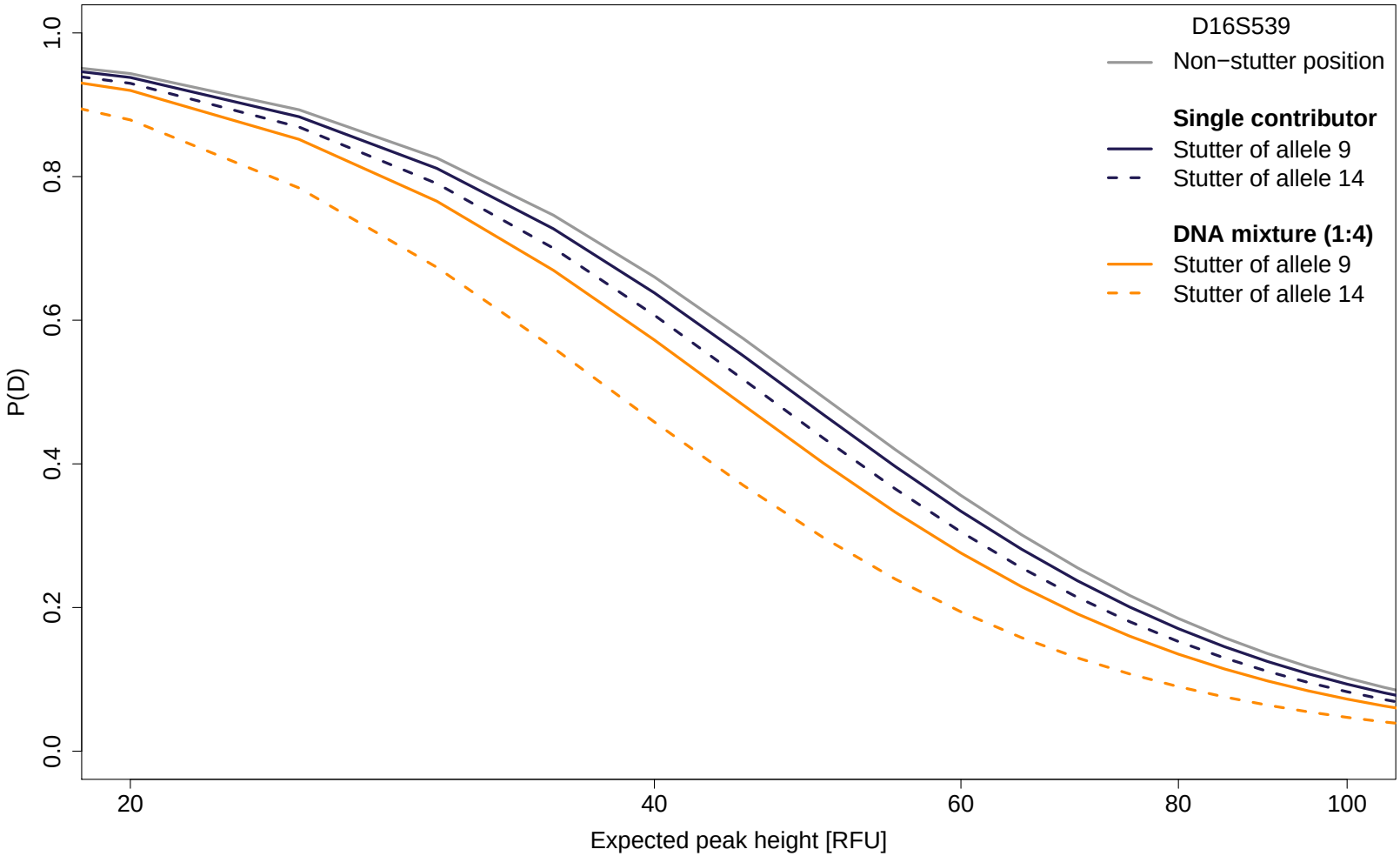
If we expect that the mean peak height of an allele at a heterozygote locus is given by μ, then for an allele in stutter position, we inflate this by a factor $(1 + \nu)$, where ν is the allele specific stutter percentage.



?GATA?

Stutter effect on drop-out probabilities

This implies that peaks in stutter position has a decreased risk of falling below the detection threshold, T.



A handle on the degradation

Degradation of the biological material is believed to cause damages to the DNA strand. One consequence is that the DNA sequence is cleaved, which implies that current STR techniques fail to amplify the DNA sequence.

If we assume it is equally likely that a sequence is cut in two at any position, we find that

$$P(\text{No degradation}) = p^{\text{bp}},$$

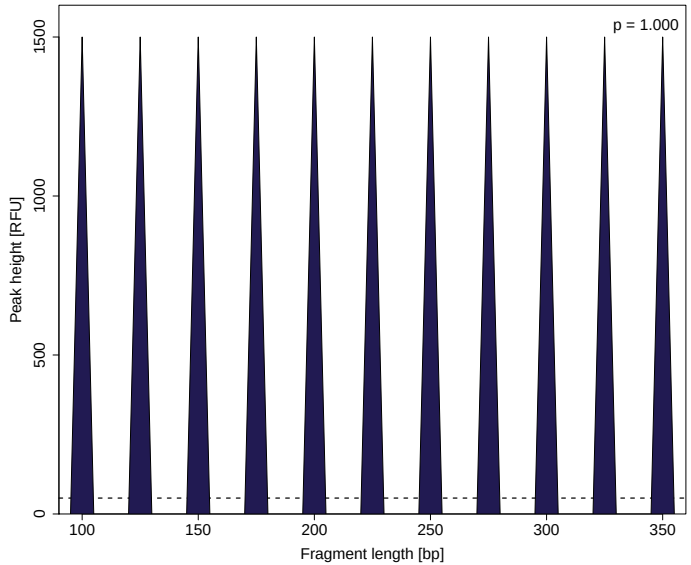
where $p = P(\text{No breakage between a pair of DNA bases})$. Hence, the closer p is to 1, the less is the decay in the peak signals.

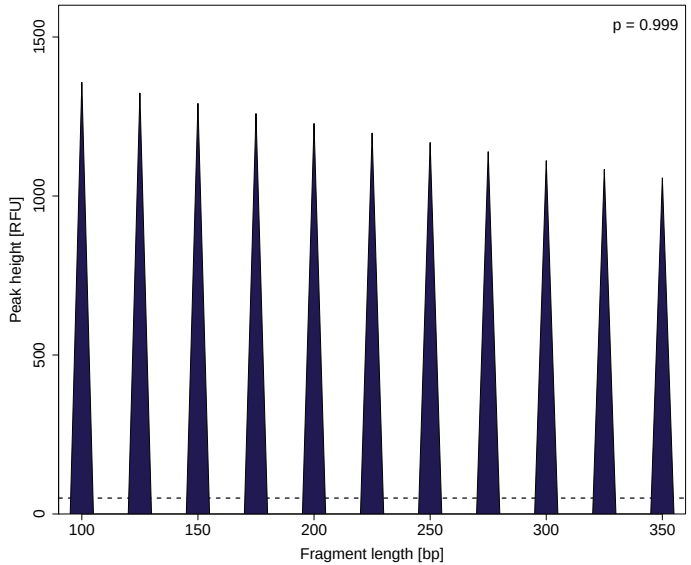
Factoring this effect into the peak height model gives

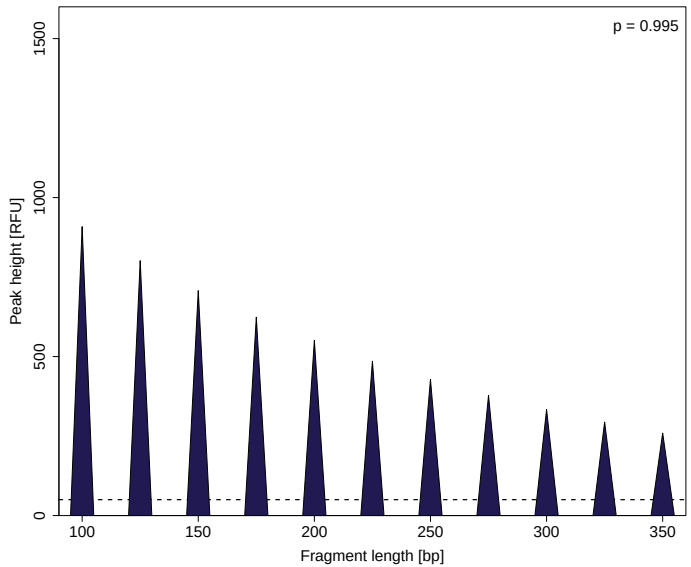
$$h_i \sim N(\mu n_i p^{\text{bp}_i}, \sigma^2 n_i p^{\text{bp}_i})$$

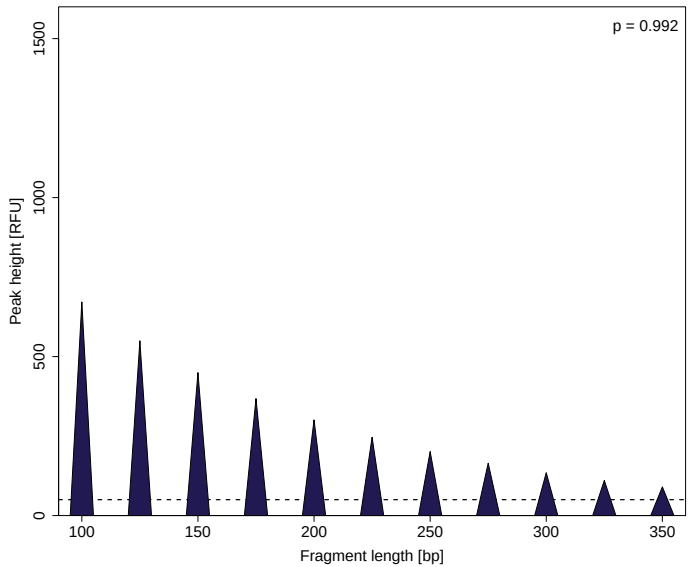
with bp_i being the fragment length of the i^{th} allele.

Degradation – plausible range of *p*

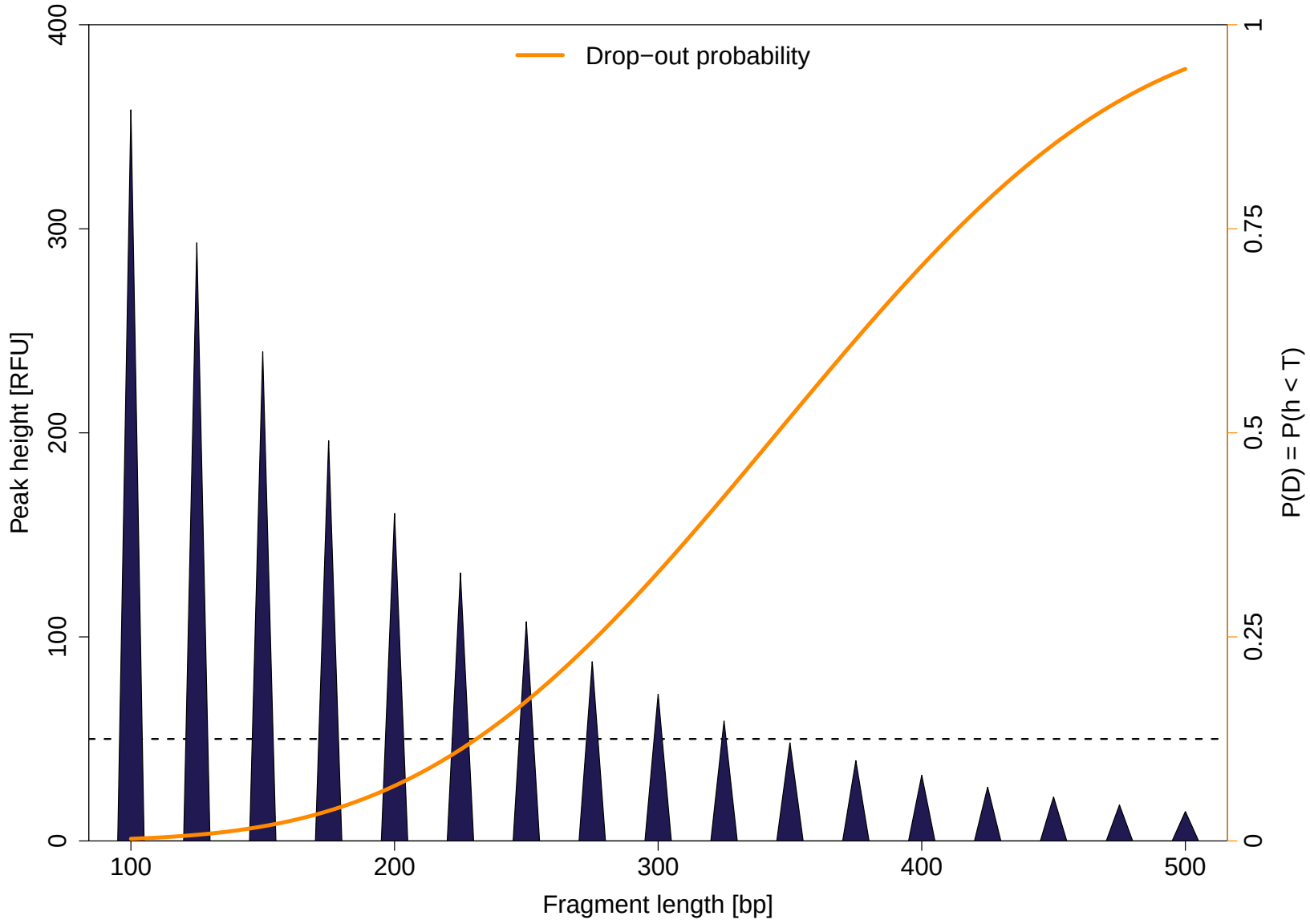








Drop-out probability and degradation



Real crime case data

We analysed 251 samples obtained from real crime cases analysed with the AmpFℓSTR® NGM SElect™ kit (Life Technologies).

The DNA was extracted from fingernail scrapings found under the victim's nails. The victim's DNA profile acted as reference profile, based on which drop-outs and drop-ins were declared.

We investigated whether the cases were subject to detectable degradation, which implies that p is significantly smaller than 1. In 97% of the cases, this was the case.

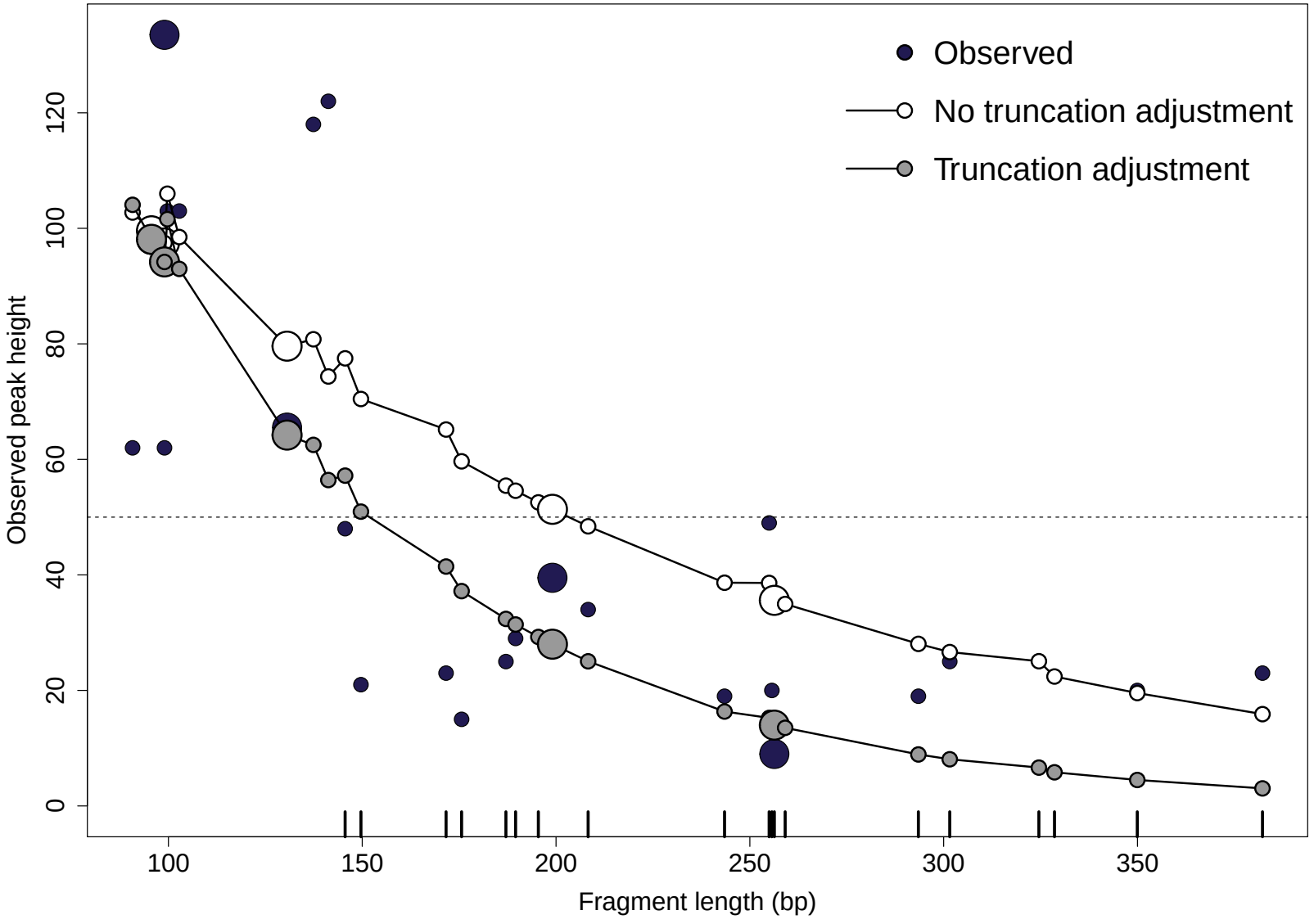
Peak height model

As for the non-degraded samples, the likelihood to be maximised included the sub-threshold samples. However, only the bp_i-value was used:

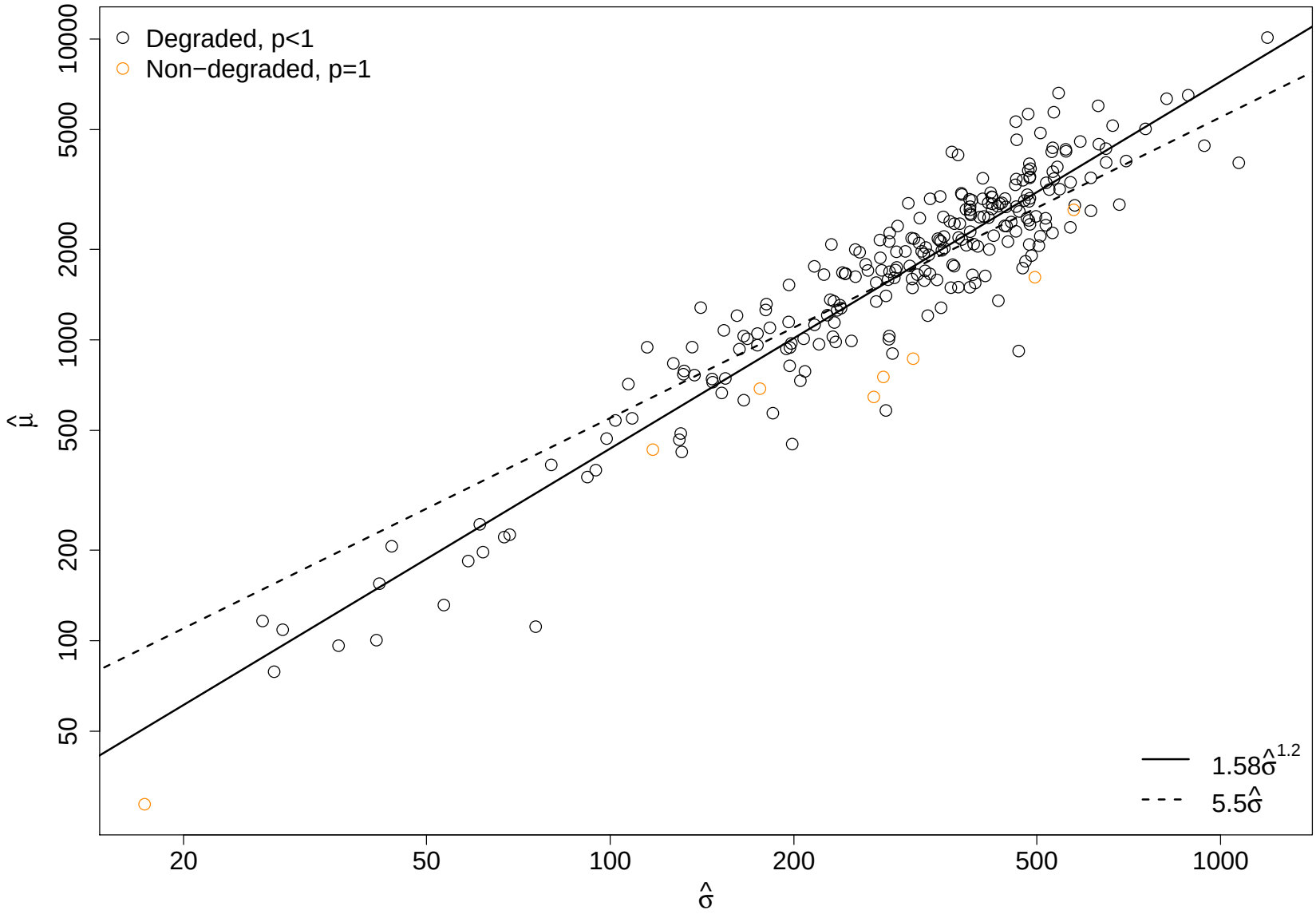
$$L(\boldsymbol{h}, \boldsymbol{n}, \mathbf{bp}; \mu, \sigma) = \prod_{i=1}^{m_{\text{obs}}} \frac{1}{\sigma\sqrt{n_i p^{\text{bp}_i}}} \phi\left(\frac{h_i - \mu n_i p^{\text{bp}_i}}{\sigma\sqrt{n_i p^{\text{bp}_i}}}\right) \prod_{j=1}^{m_{\text{drop}}} \Phi\left(\frac{T - \mu n_j p^{\text{bp}_j}}{\sigma\sqrt{n_j p^{\text{bp}_j}}}\right),$$

Previous analysis of peak height data from STR analysis, indicate that the variance and mean used by proportional. The parameter estimates supported this...

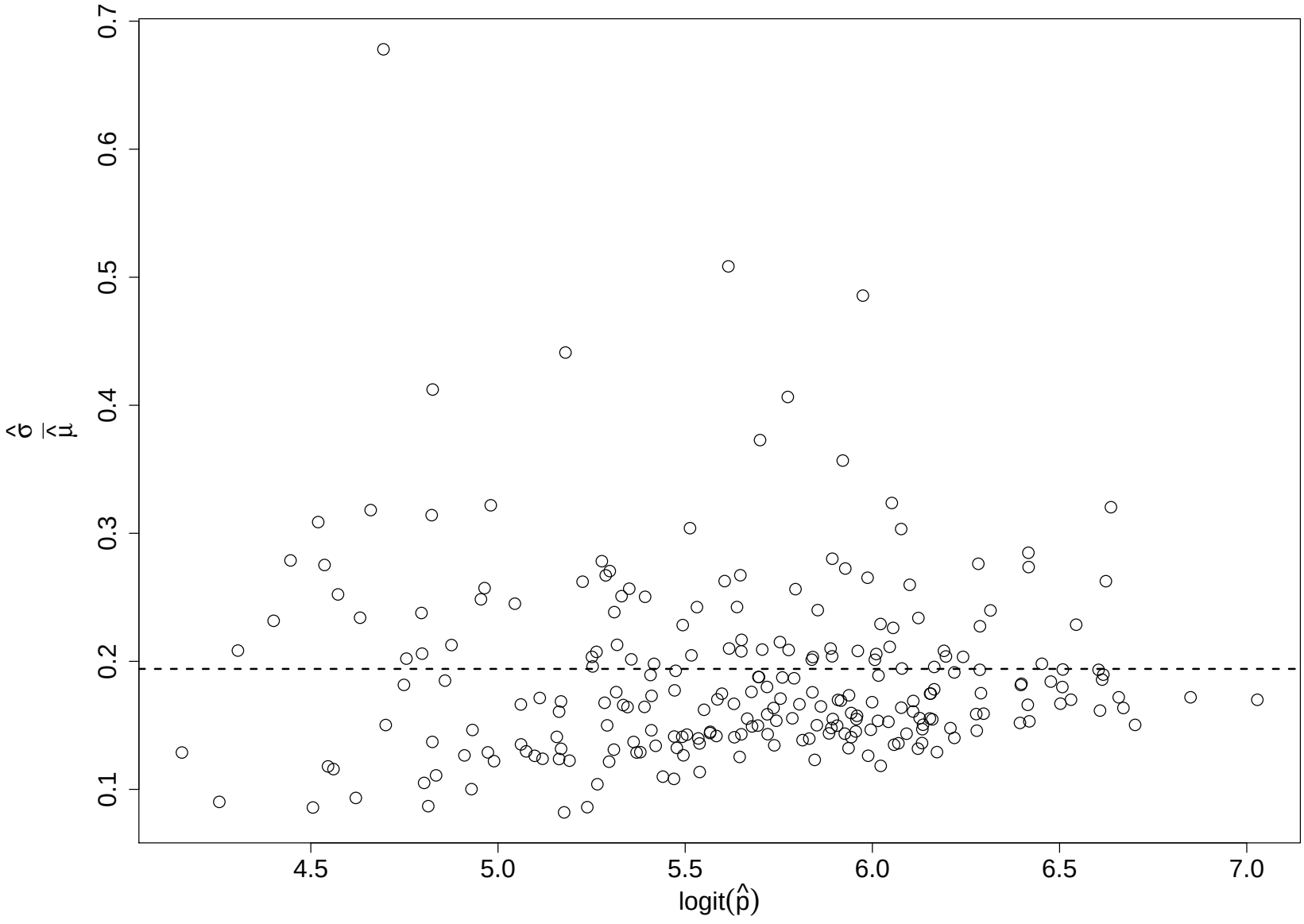
Example of a sample



Parameter estimates



Parameter estimates



Drop-out probability

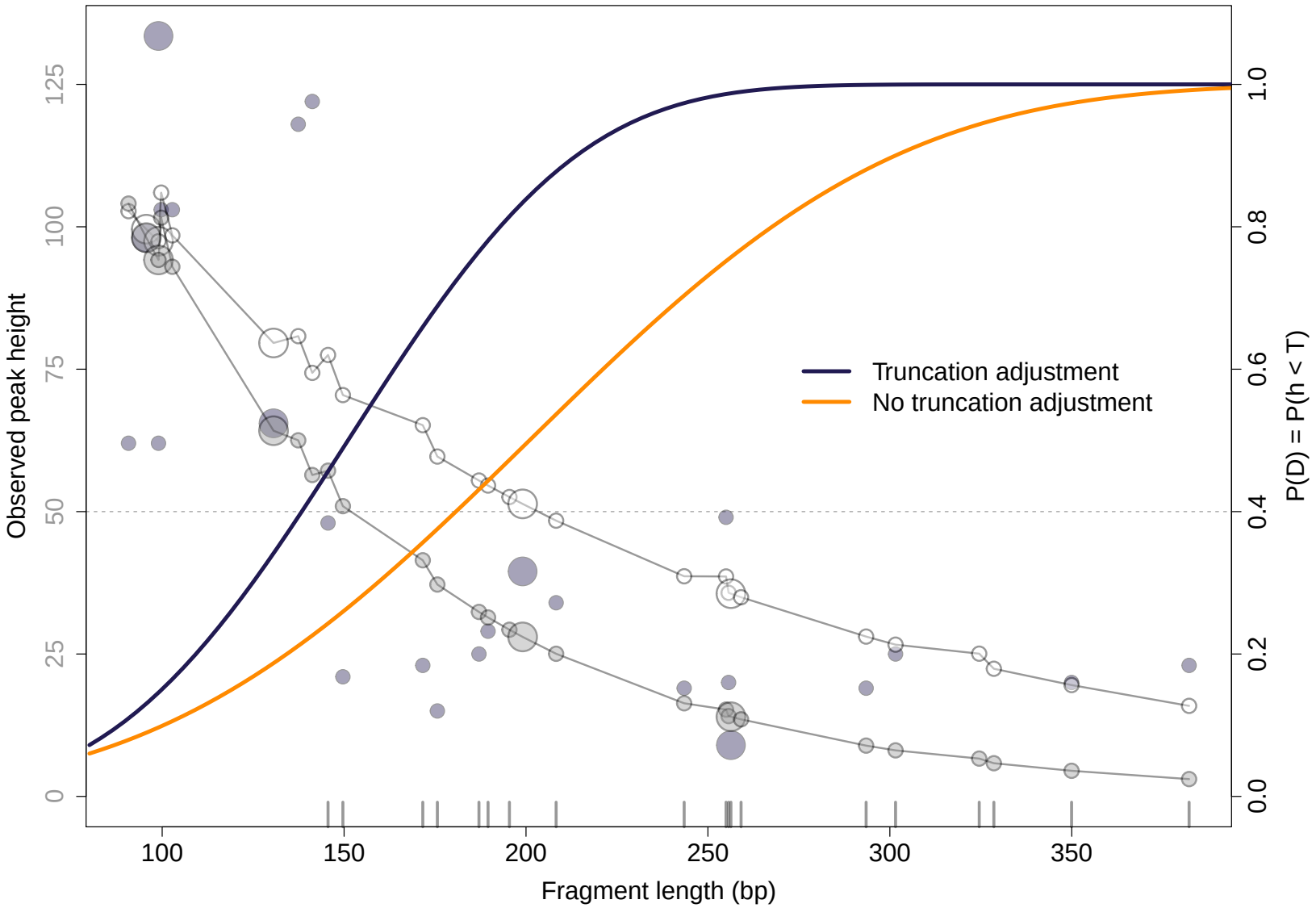
Based on the peak height model, it is possible to estimate the drop-out probability by evaluating

$$\hat{P}(D_i) = P(\hat{h}_i < T) = \Phi\left(\frac{T - \hat{\mu} n_i \hat{p}^{\mathrm{bp}_i}}{\hat{\sigma}\sqrt{n_i \hat{p}^{\mathrm{bp}_i}}}\right),$$

where the cumulative distribution function, Φ, of h_i is used to evaluate the probability.

This approach only depends on the sample itself as no *global* parameters is used when assessing $P(D)$.

Example of a sample – continued



Logistic regression

However, the previous approach does not incorporate potential locus effects, where some loci drop-out more frequently than others.

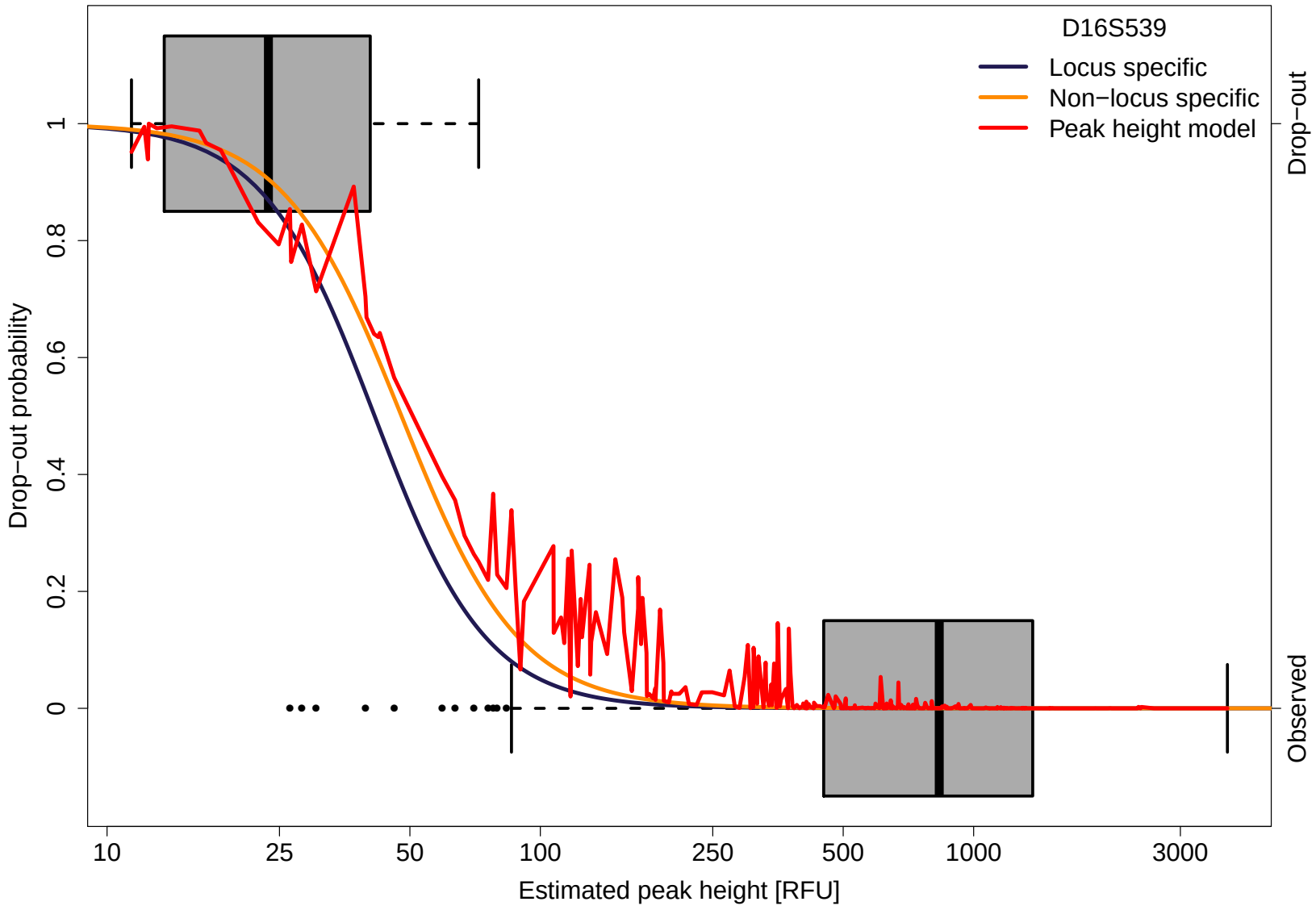
Furthermore, there may be some benefit from "borrowing" power from other samples, e.g. reducing the variance of the estimates by introducing some extra smoothing.

Hence, the expected peak heights were used as explanatory variable in a logistic regression:

$$\log \frac{P(D_i|\hat{H}(\text{bp}_i)_{\text{Trunc}})}{1 - P(D_i|\hat{H}(\text{bp}_i)_{\text{Trunc}})} = \beta_{0,s} + \beta_1 \log \hat{H}(\text{bp}_i)_{\text{Trunc}},$$

where $\hat{H}(\text{bp})_{\text{Trunc}}$ emphasise that this simple expression is only valid when the truncation adjustment is applied.

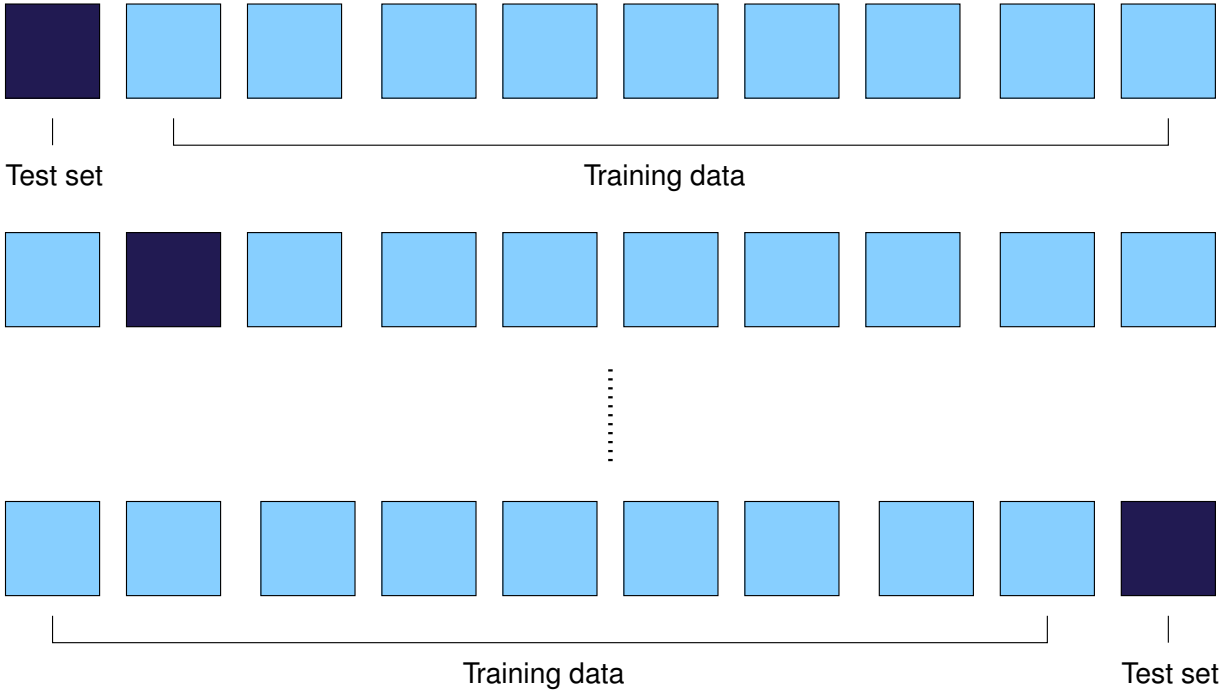
Locus specific?



Robustness and portability

For practical purposes, it may be sufficient to use a non-locus specific version of the logistic regression model, i.e. $\beta_{0,s} \equiv \beta_0$, for all loci.

To determine this, we used 10-fold cross-validation, where the data was randomly split into ten subsets and successively used for training (90% of data) and test data (10% of data):



Brier score

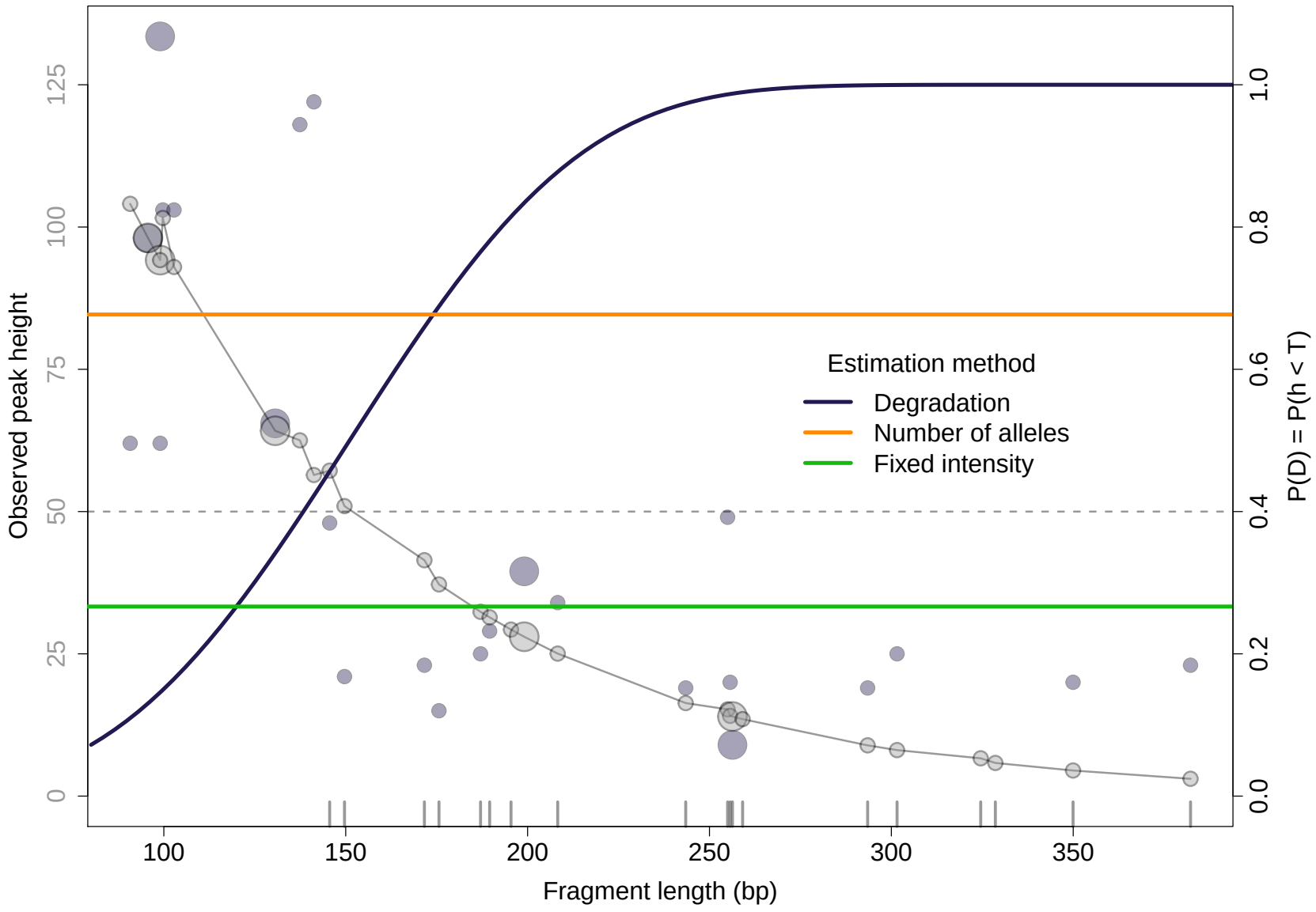
A popular measure of goodness-of-fit for binary outcomes is the Brier score, which measures the mean deviation between D_i and $\hat{P}(D_i)$,

$$B = \frac{1}{n} \sum_{i=1}^{n} \left(D_i - \hat{P}(D_i) \right)^2 .$$

Based on the cross-validation study the non-locus specific logistic regression seems to be the appropriate choice:

Drop-out model, $\hat{P}(D)$	B
Locus specific logistic regression	0.0121
Non-locus specific logistic regression	0.0122
Peak height model	0.0127

Comparing the methods



Summary

By analysing samples from real crime cases, we found that

- estimating the drop-out probability on the number of observed alleles is an inferior method compared to peak height intensity methods.
- it was important to base the expected peak height on both observed and sub-threshold peak heights by adjusting for truncation.
- detectable degradation was present in almost all of the investigated samples, suggesting that $P(D)$ is non-constant across the fragment range
- it for, practical purposes, was sufficient to use non-locus specific logistic regression models to estimate $P(D)$

Thank you, for your attention!

AALBORG UNIVERSITY
DENMARK