Estimating drop-out probabilities of STR alleles

While accounting for truncation, degradation and stutters

ISFG 2013 – Melbourne

Torben Tvedebrink[†,‡]

`tvede@math.aau.dk`

Joint work with M. Asplund[‡], P.S. Eriksen[†], H.S. Mogensen[‡] and N. Morling[‡]

[†] Dept. of Mathematical Sciences, Aalborg University, Denmark

[‡] Section of Forensic Genetics, Dept. of Forensic Medicine Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

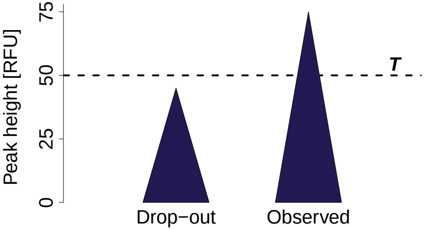
AALBORG UNIVERSITY
DENMARK

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.



Disclaimer

In this talk, I define allelic drop-out as the event where a contributor's allele fail to be detected in the resulting DNA profile.

In our model, we only link the number of DNA templates with the drop-out probability.

Hence, we do not incorporate competing events such as

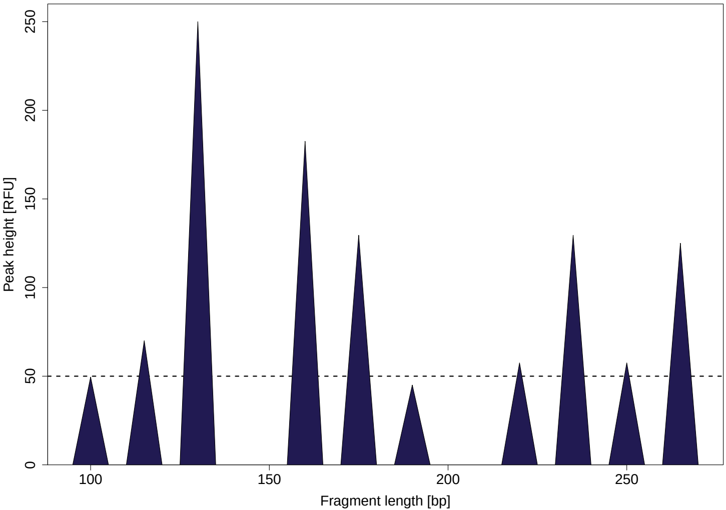
- null alleles or primer site mutations (disables amplification of STR region)
- STR region specific inhibitors (causing severe locus imbalances)
- . . .

Properties of the drop-out probability

The drop-out probability should be:

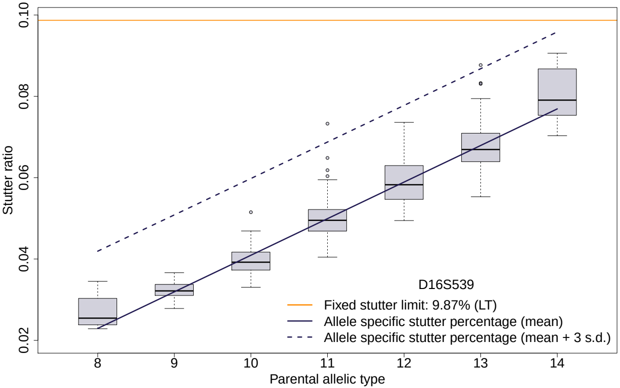
- negatively correlated with the number of DNA templates,
- lower for EPGs with higher peak heights,
- allowed to be profile specific for DNA mixtures,
- ...

Stutter correction



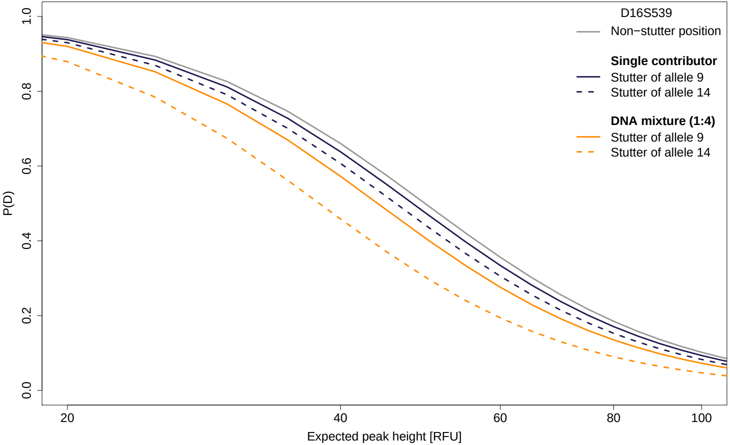
Compensating for stutter

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.



Stutter effect on drop-out probabilities

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



Truncation caused by detection threshold

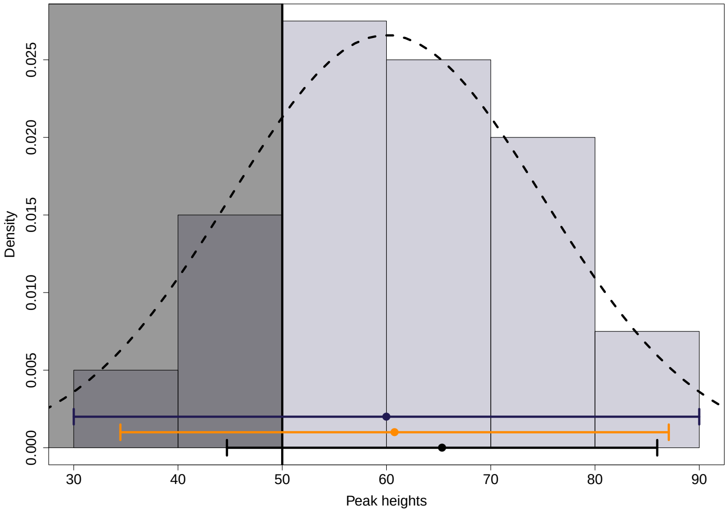
Because of the detection threshold, the peak height observations are truncated at T, e.g. 50 RFU.

If we want to estimate the underlying mean peak height, μ, for a given DNA profile, we need to adjust for this phenomena.

Let us assume that the peak heights follow some probability distribution, e.g. a normal distribution. For the dropped out alleles, all we know about their peak heights, h_i, is that they fall below the detection threshold, T.

However, including this information in the likelihood expression may greatly influence the estimate of μ.

Truncation example



A handle on the degradation

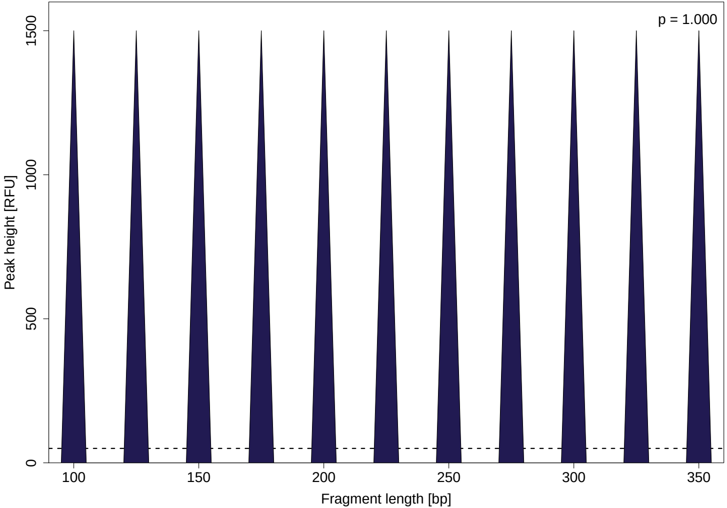
Degradation of the biological material is believed to cause damages to the DNA strand. One consequence is that the DNA sequence is cleaved, which implies that current STR techniques fail to amplify the DNA sequence.

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

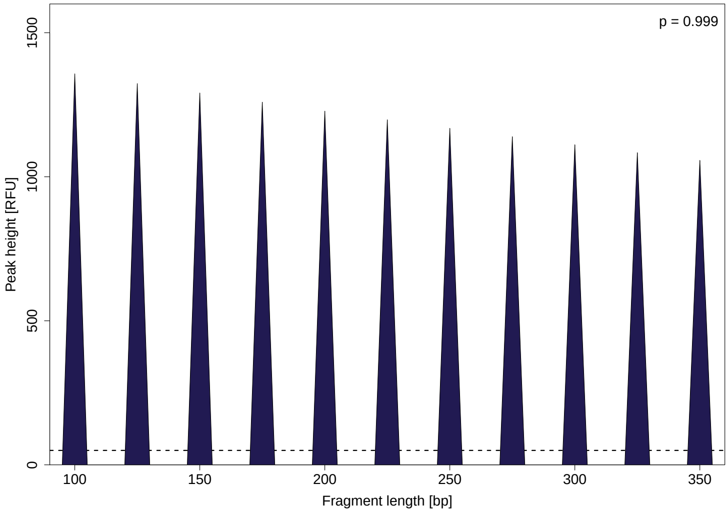
$$P(\text{No degradation}) = p^{\text{bp}},$$

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.

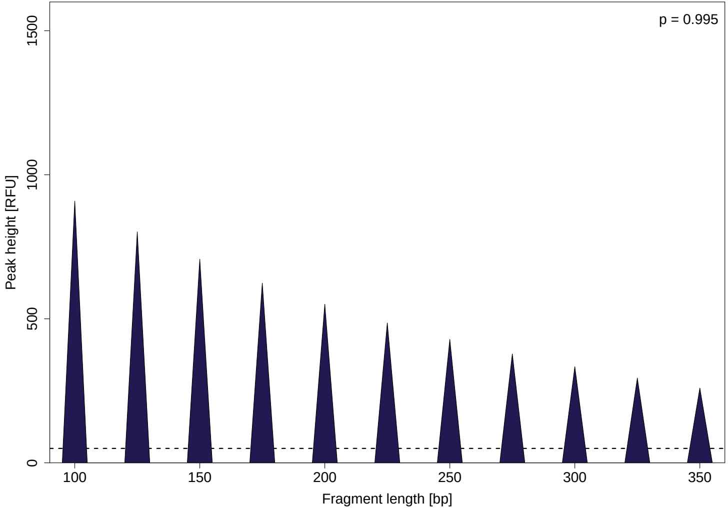
Degradation – plausible range of *p*



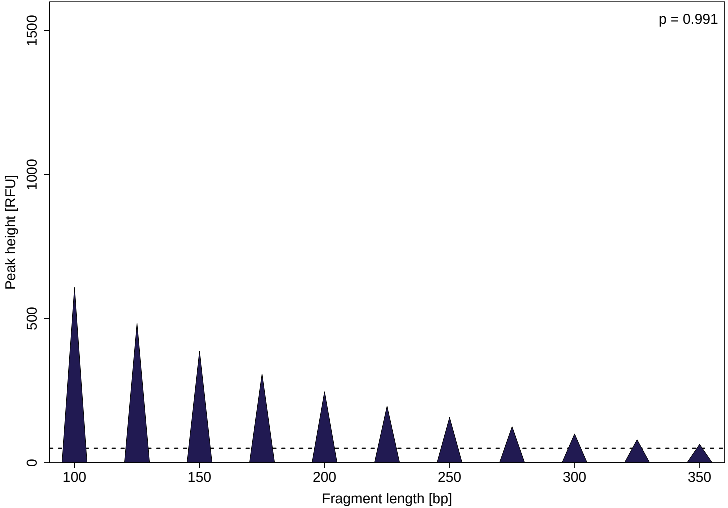
Degradation – plausible range of *p*



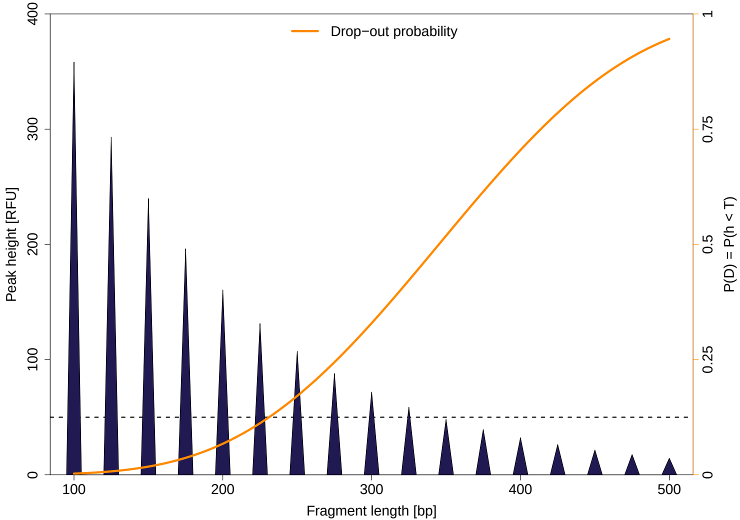
Degradation – plausible range of *p*



Degradation – plausible range of *p*



Drop-out probability and degradation



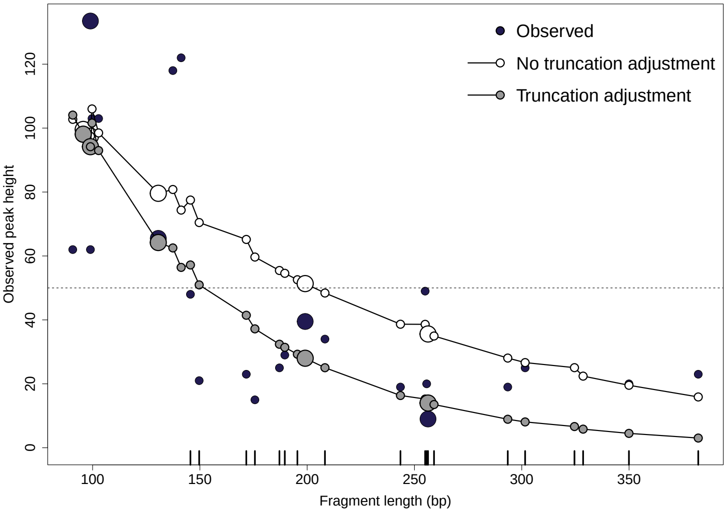
Real crime case data

We analysed 251 samples obtained from real crime cases analysed with the AmpFℓSTR® NGM SElect$^{\mathrm{TM}}$ 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.

Example of a sample



Drop-out probability

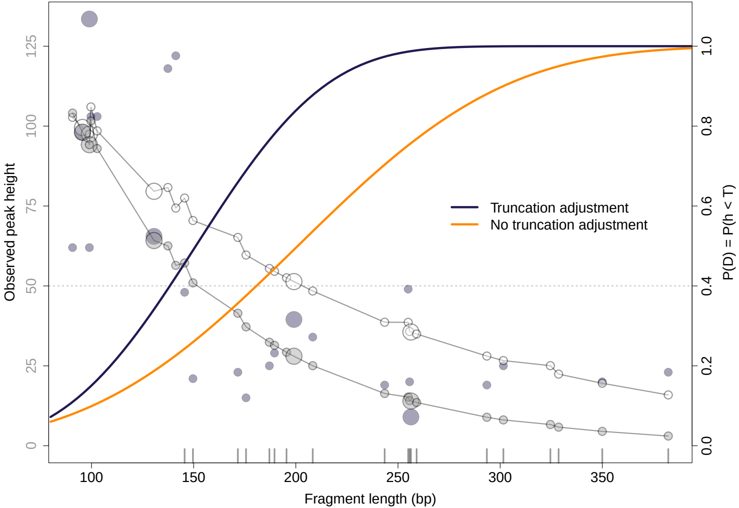
Based on the peak height model, it is possible to estimate the drop-out probability by evaluating

$$\hat{P}(D_i) = P(\hat{h}_i < T),$$

where the cumulative distribution function 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.

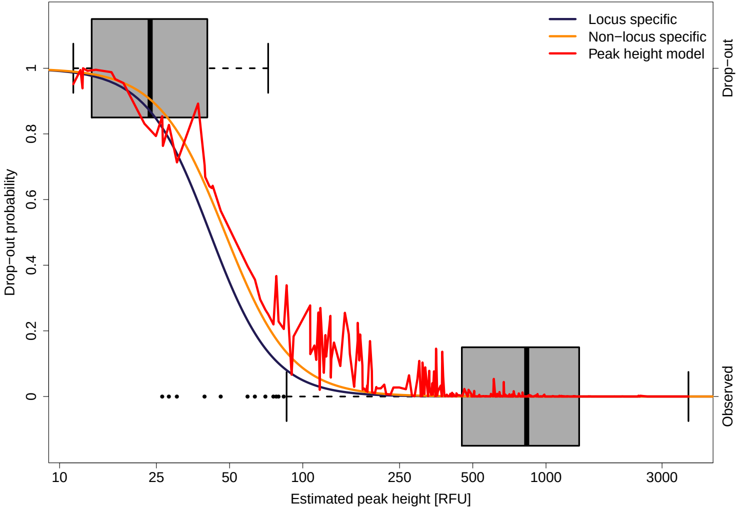
Furthermore, there may be some benefit from “borrowing” power from other samples, e.g. reducing the variance of the estimates by introducing some extra smoothing.

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

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

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

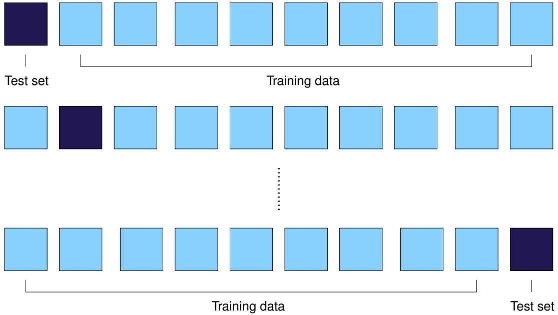
Locus specific?



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 $\hat{P}(D_i)$,

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

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

Summary

By analysing samples from real crime cases, we found that

- 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