A model to estimate dropout probabilities

Torben Tvedebrink
PhD Student, MSc
tvede@math.aau.dk

June 8 2010 - Forensic Summer School - Copenhagen

Handouts *: same contents as presentation slides*

Outline

- Motivation
  - Reasons for allelic drop-out and which causes are modelled
  - Why bother estimating $P(D)$?
- Amount of DNA as covariate
  - How do we estimate the amount of DNA?
- Logistic regression
  - Basic concepts and definitions
  - How to use R for fitting logistic regression models
- Estimating $P(D)$ using logistic regression
  - Model selection and producing plots and tables
- Degraded samples
  - How to adjust the model to handle degraded DNA

Motivation

Reasons for allelic dropout

As discussed previously in this course allelic drop-out might occur in the analysis of DNA samples.

There may be several reasons for allelic drop-out:

- Low amounts of DNA in the sample
- The particular chromosome was not sampled pre-PCR
- The threshold used as detection limit (e.g. 50 rfu)

- Inhibitors affecting some but not necessarily all loci
- Degradation of the biological material
- ...

Why bother?

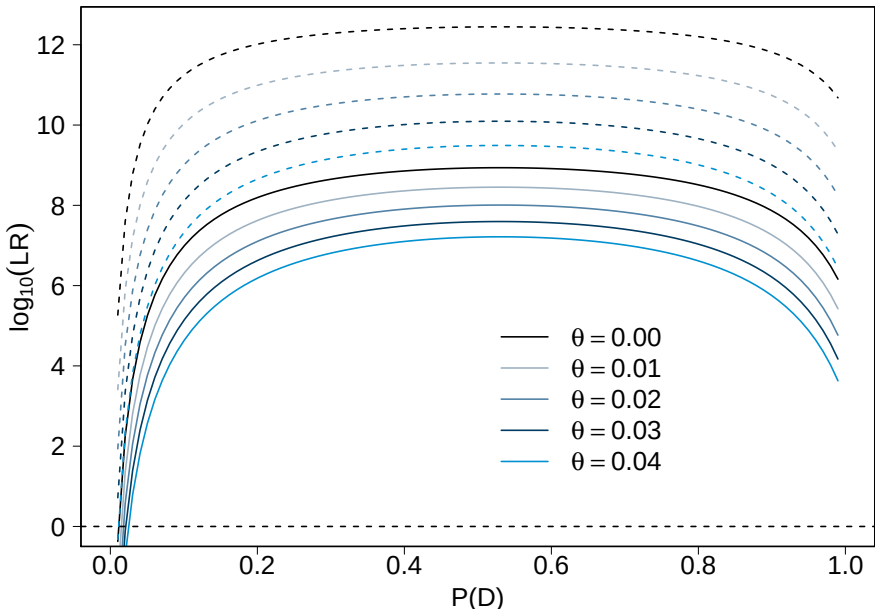
Assume that a suspect's DNA profile is $S = (ab)$ and the observed crime scene stain is $C_s = a$. I.e. if S is the contributor to the stain, then the b allele needs to have dropped out:

$$\begin{aligned}
LR &= \frac{P(E|H_p)}{P(E|H_d)} = \frac{P(C_s, S|H_p)}{P(C_s, S|H_d)} \\
&= \frac{P(C_s|S)P(S)}{\sum_{U \equiv H_d} P(C_s, S|U)P(U)} \\
&= \frac{P(C_s|S)}{\sum_{U \equiv H_d} P(C_s|U)P(U|S)} \\
&= \frac{P(D)P(\bar{D})}{P(\bar{D}^2)P(aa|ab) + P(\bar{D})P(D)\left[P(ab|ab) + \sum_{q \neq \{a,b\}} P(aq|ab)\right]} \\
&= \frac{P(D)P(\bar{D})}{P(\bar{D}^2)\frac{2\theta+(1-\theta)p_a}{1+2\theta}\frac{\theta+(1-\theta)p_a}{1+\theta} + P(\bar{D})P(D)\frac{\theta+(1-\theta)p_a}{1+2\theta}\frac{\theta+(1-\theta)(1-p_a)}{1+\theta}}
\end{aligned}$$

For simplicity we assume that this is the case for all L used for genotyping. Then the overall likelihood ratio is:

$$LR \approx \left(\frac{P(D)P(\bar{D})}{P(\bar{D}^2)\frac{2\theta+(1-\theta)p_a}{1+2\theta}\frac{\theta+(1-\theta)p_a}{1+\theta} + P(\bar{D})P(D)\frac{\theta+(1-\theta)p_a}{1+2\theta}\frac{\theta+(1-\theta)(1-p_a)}{1+\theta}}\right)^L$$

Let $P(a) = 0.1$ (solid), $P(a) = 0.05$ (dashed) and $L = 10$ then LR can be plotted against $P(D)$



Amount of DNA as covariate

Dilution experiments

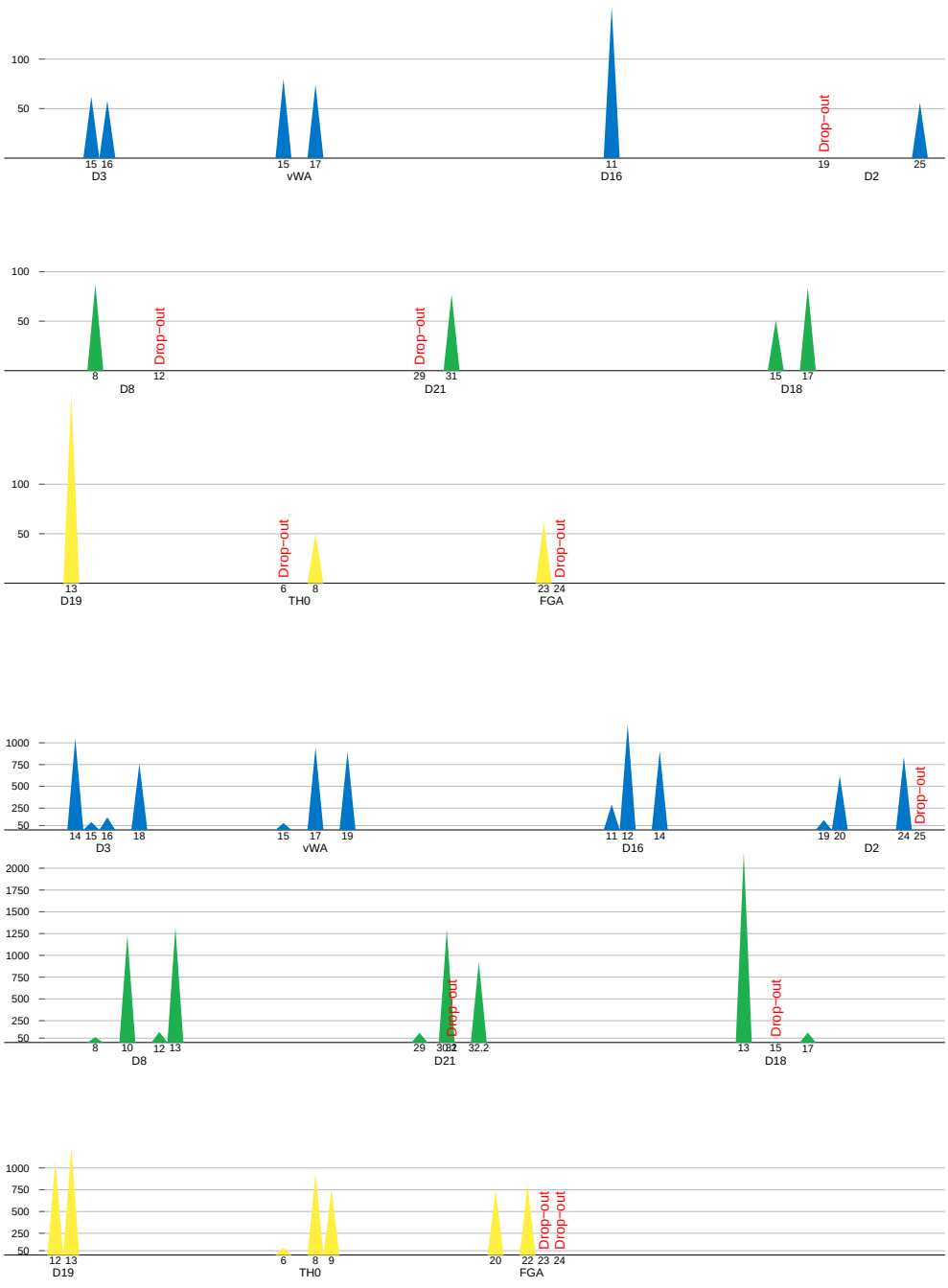
A series of dilution experiments were conducted by The Section of Forensic Genetics here at University of Copenhagen.

Four DNA profiles (cf. below) were serially diluted - pairwise and with water in proportions 1:16, 1:8, 1:4, 1:2 and 1:1.

D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
14,18	17,19	12,14	20,24	10,13	30.2,32.2	13,13	12,13	8,9	20,22
15,16	14,16	10,12	17,25	13,16	30,30	13,13	14,15	6,9	19,23
15,16	15,17	11,11	19,25	8,12	29,31	15,17	13,13	6,8	23,24
16,19	15,17	10,12	23,25	13,13	28,30	12,16	13,15	6,7	20,23

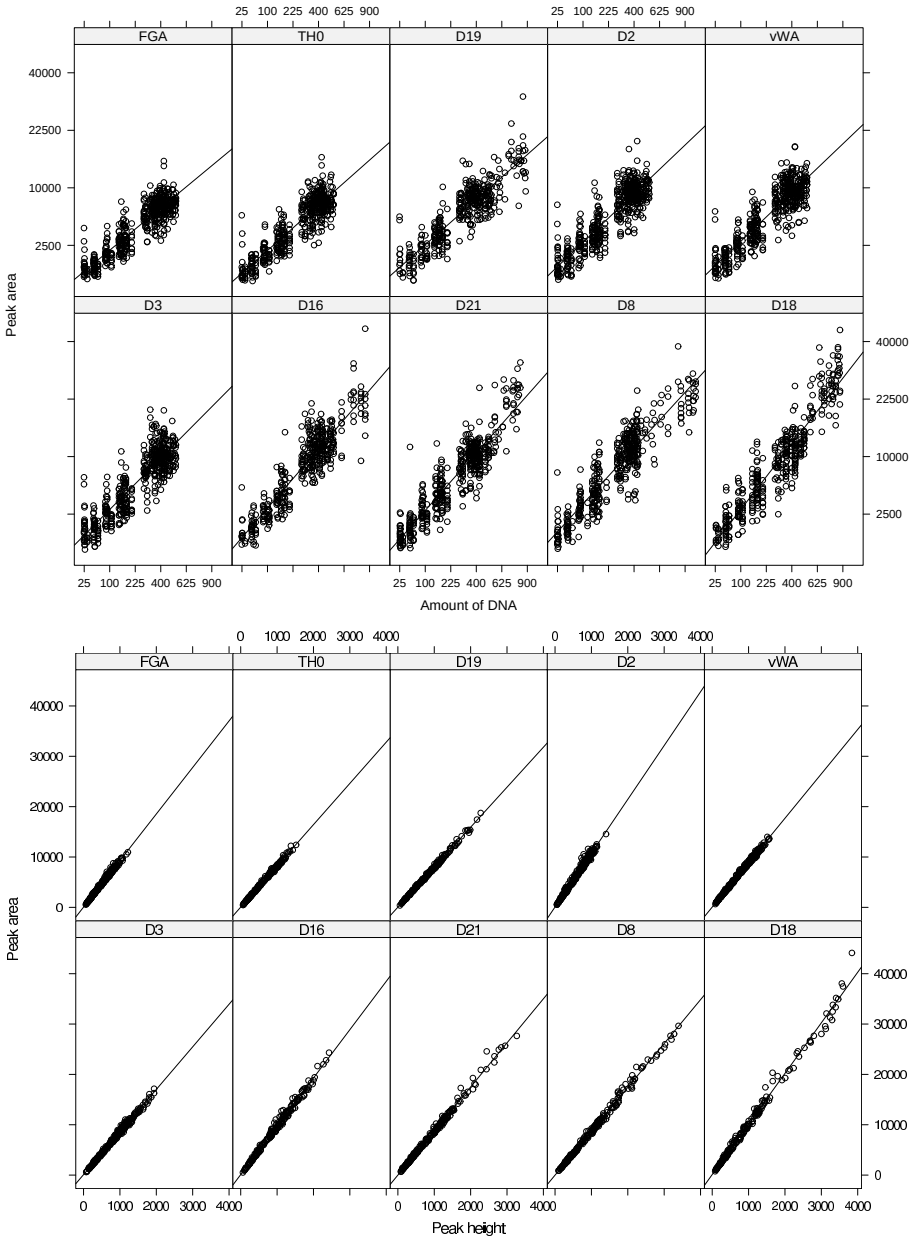
The measured amounts of DNA ranged from 24.6 to 410 pg for “water samples” and from 328 to 528 pg for the DNA mixtures.

Sample plot



Proportionality of peak heights and amount of DNA

It is well known that the peak heights are proportional to the amount of DNA contributed.



Definition H

Let h_i be the i'th observed peak height, n_{het} and n_{hom} the number of observed heterozygote and homozygote peaks.

$$H = \frac{1}{n_{\text{het}} + 2n_{\text{hom}}} \sum_{i=1}^{n} h_i,$$

where $n = n_{\text{het}} + n_{\text{hom}}$.

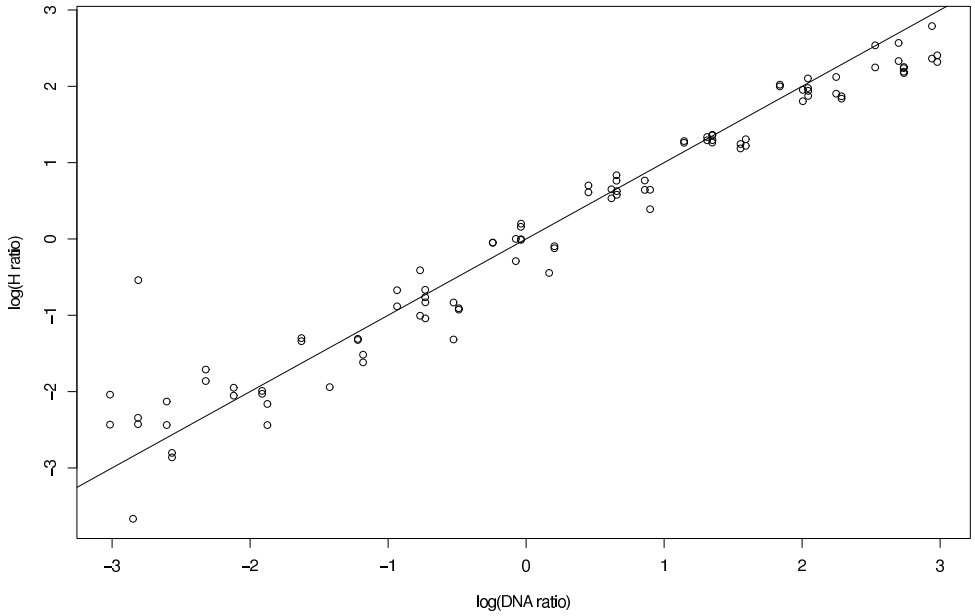
Note:
If no alleles has dropped out then $H = (2L)^{-1} \sum h_i$, i.e. the average peak height when

counting homozygote peaks as two.

If the sample is a DNA mixture, then H_i is only based on those peak height observations where person i is assumed to be the only contributor.

Plot of H versus amount of DNA

Plot of DNA-ratio and H-ratio for DNA mixtures



H as proxy for amount of DNA

The slope of the line in the previous plot was 1.

I.e. we have

$$\frac{H_1}{H_2} = \frac{\alpha H_1}{\alpha H_2} = \frac{\text{DNA}_1}{\text{DNA}_2}$$

for some constant α.

However, we are only interested in finding a *proxy* since in a regression model we have

$$\mathbb{E}(Y|\boldsymbol{X}) = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_p \cdot X_p$$

where $\beta_{\text{DNA}} \cdot \text{DNA} = \alpha \cdot \beta_{\text{DNA}} \cdot H = \tilde{\beta}_{\text{DNA}} \cdot H$

Logistic regression

Bernoulli random variable

Let Y be a random variable taking two possible outcomes, e.g. $\{1, 0\}$, $\{\text{Success}, \text{Failure}\}$, $\{\text{Head}, \text{Tail}\}$, $\{\text{Drop-out}, \text{Not drop-out}\}$, ...

Let $P(Y = 1) = p$ and hence $P(Y = 0) = 1 - p$, then we have

$$\mathbb{E}(Y) = 0 \cdot (1 - p) + 1 \cdot p = p$$

When summing the number of successes in n trials the resulting variable X is binomial distributed:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

where assumptions are that p is fixed for each trial and that the outcomes are mutually independent.

Logistic regression

This restriction is often violated since p will in many experimental designs depend on some covariates!

One way around this is logistic regression where we assume that

$$P(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}$$

where β_j are parameters to be estimated and x_{ij} known values of the j'th covariate for the i'th observation.

Note that this definition ensures $0 \leq \pi(\boldsymbol{x}_i) \leq 1$.

The likelihood function is proportional to

$$L(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{x}) \propto \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i} (1 - \pi(\boldsymbol{x}_i))^{1-y_i}$$

Logit and log odds

Furthermore, the $\text{logit}(p) = \log \frac{p}{1-p}$ gives:

$$\text{logit } P(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i) = \log \frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

I.e. we model percentage-wise change in the odds of the event by the linear term on the right-hand-side.

The logit function is the inverse of logistic function: $\frac{\exp(x)}{1+\exp(x)}$

Logistic regression (cont'd)

Logistic regression is a special case of the larger class of models called *Generalized linear models* (GLMs).

In normal linear regression we have:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

In GLM models we have

$$g\left(\mathbb{E}(Y)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where g is called the link function. The link function specifies the relationship between the linear term of covariates and the mean of the dependent variable:

$$\mathbb{E}(Y) = g^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

Logistic regression in R

In R you can fit GLMs using the `glm`-function: `glm(formula, family, data, ...)` where `formula` specifies the mean structure as in the `lm`-call: `y` $\sim$ `x1 + x2 + x3*x4` $+ \cdots$

For binomial data this is done by setting `family=binomial`

For binomial random variables there are three commonly used link functions (where `logit` is the default):

Name	Link function	R-call
Logit	$g(p) = \log(p/(1-p))$	`binomial(link="logit")`
Probit	$g(p) = \Phi^{-1}(p)$	`binomial(link="probit")`
clog-log	$g(p) = \log[-\log(1-p)]$	`binomial(link="cloglog")`

Estimating $P(D)$

Estimation of $P(D)$ using logistic regression

We used the dilution experiments in order to fit a logistic regression model:

$$\text{logit}P(D;H) = \beta_{0,s} + \beta_{1,s}\log(H),$$

where the s subscript implies that $\beta_{i,s}$ may depend on the locus s.

Furthermore, the reason for using $\log(H)$ rather than H is that:

$$P(D;H=0) = \frac{\exp(\beta_{0,s} + \beta_{1,s}\log(0))}{1+\exp(\beta_{0,s} + \beta_{1,s}\log(0))} = \frac{\exp(-\infty)}{1+\exp(-\infty)} = 1$$

since $\beta_{1,s}$ is negative.

Deviance and model selection

As with any type of regression model - the more covariates the better fit! How to choose one model over an other?

For GLMs the goodness-of-fit of different models is compared using the *deviance*. Let M be a model with p parameters and M_0 a sub-model of $M_0 \subset M$ with $q < p$ parameters, then:

$$D(\boldsymbol{y}; M, M_0) = 2\left(\ell(\boldsymbol{y}; M) - \ell(\boldsymbol{y}; M_0)\right) \underset{\text{approx}}{\sim} \chi^2_{p-q}$$

If the change in deviance D is not greater than one would expect by chance alone, then it is taken as evidence that M_0 (simpler model) is sufficient in order to explain the response relative to M.

.. and in R this is done

If we have fitted the models:

```
intract.fit <- glm(dropout ~ locus*log(H), family=binomial)
maineff.fit <- glm(dropout ~ locus + log(H), family=binomial)
overall.fit <- glm(dropout ~ log(H), family=binomial)
```

Notation: `locus*log(H)` is short for `locus + log(H) + locus:log(H)`.

Then we see that `overall.fit` $\subset$ `maineff.fit` $\subset$ `intract.fit`

Further more if `locus` has 10 levels (e.g. the 10 autosomal SGM Plus loci) then the models has 2×10, 10+1 and 1+1 parameters.

Assess the effect of the interaction of locus and $\log(H)$: `anova(maineff.fit, intract.fit, test="Chisq")`

Assess the effect of locus dependent intercept: `anova(overall.fit, maineff.fit, test="Chisq")`

Fitting the models in R

```
summary(intract.fit)
...
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       29.025      9.600   3.023  0.00250 **
locusvWA         -10.057     11.061  -0.909  0.36326
locusD16           5.861     15.156   0.387  0.69898
locusD2          -14.892     10.282  -1.448  0.14754
locusD8           -9.301     11.361  -0.819  0.41299
locusD21         -13.894     11.272  -1.233  0.21770
locusD18         -14.639     10.524  -1.391  0.16423
locusD19          -5.311     12.547  -0.423  0.67210
locusTH0         -13.969     10.255  -1.362  0.17316
locusFGA          -6.679     11.073  -0.603  0.54637
log(H)            -6.767      2.171  -3.117  0.00183 **
locusvWA:log(H)    2.301      2.487   0.925  0.35487
locusD16:log(H)   -1.131      3.386  -0.334  0.73828
locusD2:log(H)     3.345      2.315   1.445  0.14859
locusD8:log(H)     2.099      2.556   0.821  0.41156
locusD21:log(H)    2.941      2.541   1.158  0.24703
locusD18:log(H)    3.248      2.373   1.369  0.17114
locusD19:log(H)    1.489      2.786   0.534  0.59301
locusTH0:log(H)    3.355      2.302   1.457  0.14505
locusFGA:log(H)    1.742      2.479   0.703  0.48224
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1027.63  on 3343  degrees of freedom
Residual deviance:  425.64  on 3324  degrees of freedom
```

We see that for none of the loci were the interaction term `locus:log(H)` significantly different from zero. Hence, this suggest that the simpler main effects model may be adequate:

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

We fit this model next

```
summary(maineff.fit)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  18.26495    1.74340  10.477   <2e-16 ***
log(H)       -4.34653    0.37814 -11.495   <2e-16 ***
locusvWA      0.16292    0.55130   0.296   0.7676
locusD16      0.48634    0.57009   0.853   0.3936
locusD2       0.04741    0.53404   0.089   0.9293
locusD8       0.01178    0.57121   0.021   0.9835
locusD21     -0.81843    0.58613  -1.396   0.1626
locusD18     -0.19982    0.57542  -0.347   0.7284
locusD19      1.13634    0.63126   1.800   0.0718 .
locusTH0      1.13967    0.54010   2.110   0.0349 *
locusFGA      0.94944    0.52478   1.809   0.0704 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1027.63  on 3343  degrees of freedom
Residual deviance:  434.13  on 3333  degrees of freedom
```

Note that locus D3 is used as reference intercept.

```
anova(maineff.fit, intract.fit, test="Chisq")
Analysis of Deviance Table

Model 1: dropout ~ locus + log(H)
Model 2: dropout ~ locus * log(H)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      3333     434.13
2      3324     425.64  9   8.4846    0.4861
```

The degrees of freedom is 9 since the interaction model has 20 parameters and the main effect model has 11. The difference in deviance of 8.48 is not significant compared to χ^2_9, hence we conclude that the main effects model is sufficient to explain the response.

From the output of `summary(maineff.fit)` we see that only a few of the loci indicated significant departures from $H_0 : \beta_{0,s} = 0$. This may indicate that the overall model is sufficient:

$$\text{logit } P(D;H) = \beta_0 + \beta_1 \log(H)$$

```
summary(overall.fit)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  17.5614     1.6048   10.94   <2e-16 ***
log(H)       -4.1354     0.3529  -11.72   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1027.63  on 3343  degrees of freedom
Residual deviance:  457.13  on 3342  degrees of freedom
```

However, the anova-function is used to test $\beta_{0,s} = 0$ for all loci.

```
anova(overall.fit, maineff.fit, test="Chisq")
Analysis of Deviance Table

Model 1: dropout ~ log(H)
Model 2: dropout ~ locus + log(H)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1      3342     457.13
2      3333     434.13  9   22.998    0.0062 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Again the degrees of freedom is 9: Main effect model has 11 parameters and the overall model 2. However, here the deviance difference (≈ 23) is highly significant compared to χ^2_9. Thus we settle with the main effects model since the overall model does not explain the response sufficiently compared to the main effects model.

Hence the final model is

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

where $\hat{\beta}_1 = -4.35$ and $\hat{\beta}_{0,s}$ are given in the table below:

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
$\hat{\beta}_{0,s}$	18.26	18.43	18.75	18.31	18.28	17.45	18.07	19.40	19.40	19.21

Simulations

In addition to real data one may simulate data based on a model for the data generating process. The drop-out probability depends essentially on the number of copies of the target molecule post-PCR.

By simulating the PCR process we can validate our model further (Simulation procedure similar to Gill et al. (2005)):

(1) Assume there are N chromosomes extracted for typing.

(2) Of these do $n_{(0)}$ carry the specific allele of interest, where

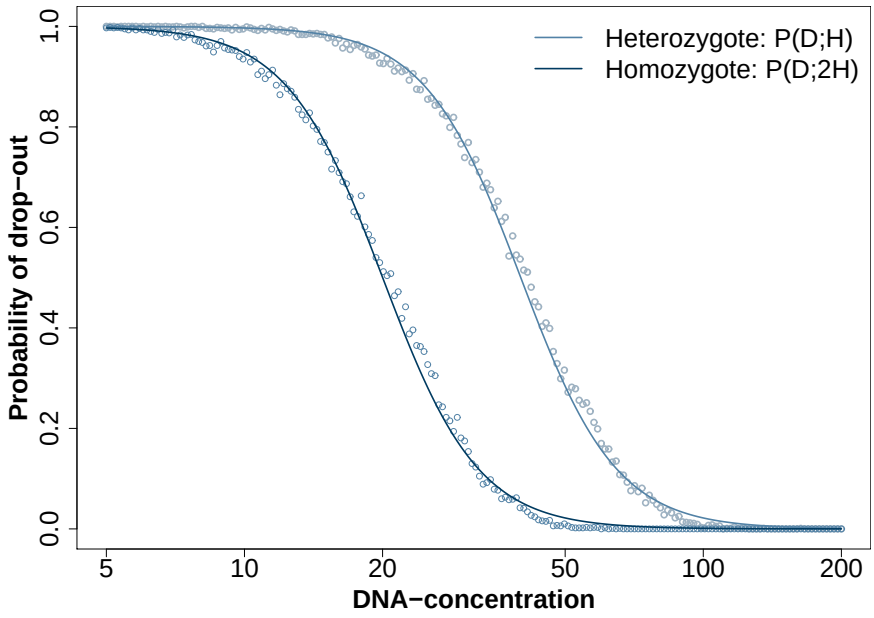
$$\underset{\textit{Heterozygote}}{n_{(0)} = \text{bin}(N, 1/46)} \qquad \text{or} \qquad \underset{\textit{Homozygote}}{n_{(0)} = \text{bin}(N, 2/46)}$$

(3) The PCR process is assumed to be a binomial process: $n_{(i)} = n_{(i-1)} + \text{bin}(n_{(i-1)}, p_{\text{eff}})$, $i = 1, \ldots, C$ cycles

(4) If $n_{(C)}$ + Noise gives reason to peak heights lower than a given threshold we declare a drop-out

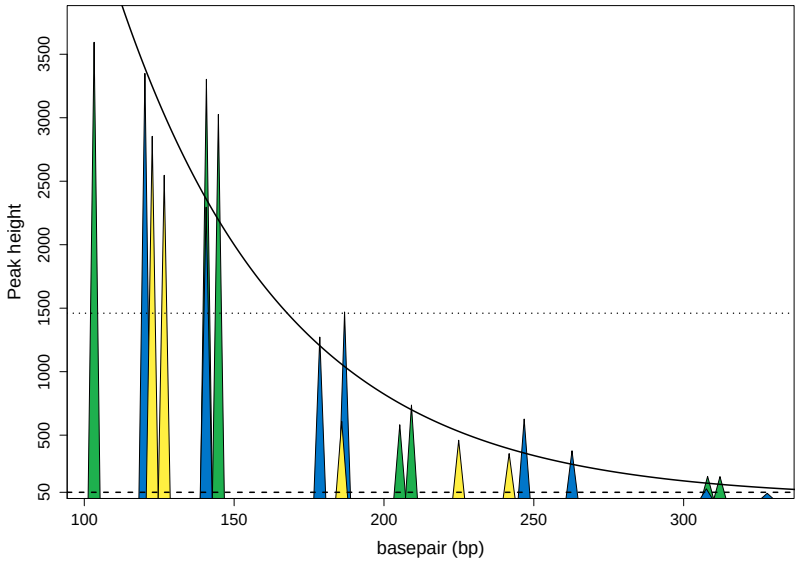
By running (1)-(4) several times with varying initial values N we get an simulated distribution of $P(D)$.

PCR efficiency $p_{\text{eff}} = 0.85$, 50 rfu threshold and $C = 28$ cycles



Degraded samples

How to handle degraded DNA?



Modelling the peak intensity decay

The decay in peak intensities may be modelled using the following approach.

Let p denote the probability that there **isn't** a breakage between two DNA acids.

$$\begin{aligned} P(\text{No degradation}) &= P(\text{No breakage between any acid pair}) \\ &= P(\text{No breakage between a given acid pair})^{\text{bp}} \\ &= p^{\text{bp}} \end{aligned}$$

Hence $P(\text{Degradation}) = 1 - P(\text{No degradation}) = 1 - p^{\text{bp}}$, which implies larger bp gives higher probability for degradation and decay in peak intensities.

Modelling the peak intensity decay

From previous slide the peak height is affected by p and bp:

$$H(\text{bp}) = c \cdot p^{\text{bp}},$$

where c depends, e.g. on the amount of DNA in the sample.

If the sample is "healthy" then $p \approx 1$ and $c \approx H$ which is a measure/proxy for the amount of DNA.

Estimate c and p from data:

$$\log H(\text{bp}) = \log(c \cdot p^{\text{bp}}) = \log(c) + \text{bp}\log(p) = \alpha_0 + \alpha_1 \text{bp}$$

which can be modelled by a normal linear model.

Adjusting the $P(D;H)$ for degradation

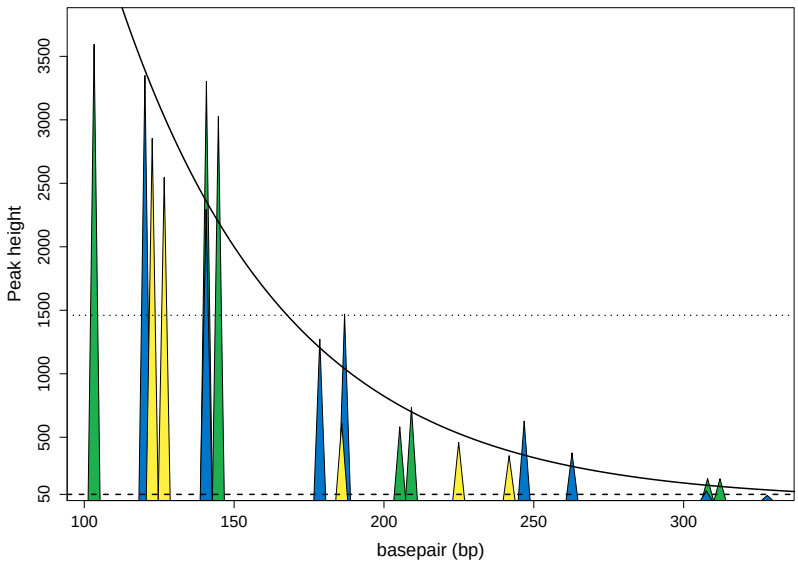
The model for allelic drop-out were derived for "healthy" samples:

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

In order to adjust for degradation insert $\log H(\text{bp}) = \alpha_0 + \alpha_1 \text{bp}$ in the model:

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

Example



For the data producing the plot $H = 1460.41$ rfu. All alleles of the DNA profile is present except allele 24 on D2 ($\text{bp}_{\text{D2}_{24}} = 327.87$).

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

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

Adjusting for degradation by the fitted solid line:

$$P(D_{\text{D2}_{24}}; H(\text{bp} = 327.87) = 85.25) = 0.26$$

References

Tvedebrink T, PS Eriksen, HS Mogensen, N Morling:

- *Evaluating the weight of evidence using quantitative STR data in DNA mixtures.* Applied Statistics (Accepted for publication)
- *Estimating the probability of allelic drop-out of STR alleles in forensic genetics.* FSI:Gen 3 (2009): 222-226.
- *Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out.* Manuscript in preparation.

Gill P, Curran J, Elliot K: *A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci.* Nucleic Acid Research 33 (2005): 632-643.

There are several books on logistic regression. However, James Curran is currently finishing up a book called "Introduction to data analysis with R for forensic scientists" (August 2010) which covers some basic statistics (e.g. logistic regression) and R tutorials.