AALBORG UNIVERSITY

Estimating Y-STR Allele Drop-out Probabilities Adjusting for Locus Imbalance

Mikkel Meyer Andersen[a], Poul Svante Eriksen[a], Jill Olofsson[b], Maria Asplund[b], Helle Smidt Mogensen[b], Niels Morling[b]

a) Department of Mathematical Sciences, Aalborg University, Denmark
b) The Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark

24th World Congress of the International Society for Forensic Genetics, University of Vienna

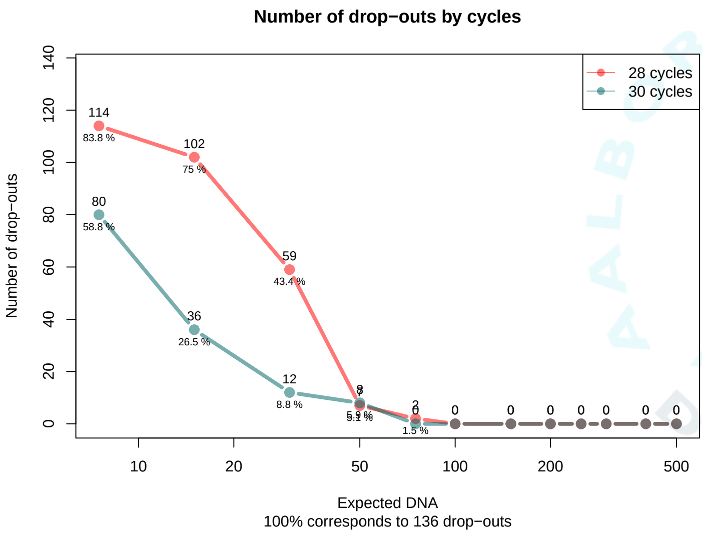
Outline

- Need to correct for drop-outs in various cases (e.g. in analysing and interpreting mixtures)
  - Two persons' mixture with two peaks at all loci but one
  - Do the persons share an allele?
  - Has a drop-out occured?
- Quantify the probability of drop-out
  - Amount of biological mass, i.e. signal strength
  - Locus balances
  - Cycles
- Similar to that of autosomal STR alleles, see e.g. Tvedebrink *et al.* (2009), but some significant differences

Design of experiment

- Balanced design
  - 4 males with known Y-STR profiles
  - 12 dilutions from 7.5 to 500 $pg/\mu l$ expected DNA
  - 28 and 30 PCR cycles
  - Duplicates
  - 192 samples in total
- Unfortunately three samples were found technically too poor to include
- Peak classified as drop-out when peak height was less than 50 RFU

Descriptive view of drop-outs – without bad samples



Modelling

- Simple logistic regression model:

$$p \sim S + L + C$$
$$\text{logit}(p) = \beta_0 + \beta_1 S + \beta_2 L + \beta_3 C$$

  - p: probability of drop-out
  - S: signal strength
  - L: locus
  - C: number of PCR-cycles

- Two- and three-ways interactions

$$\begin{aligned} p &\sim S * L * C \\ &= S + L + C + (S : L) + (S : C) + (L : C) + (S : L : C) \end{aligned}$$

Signal strength

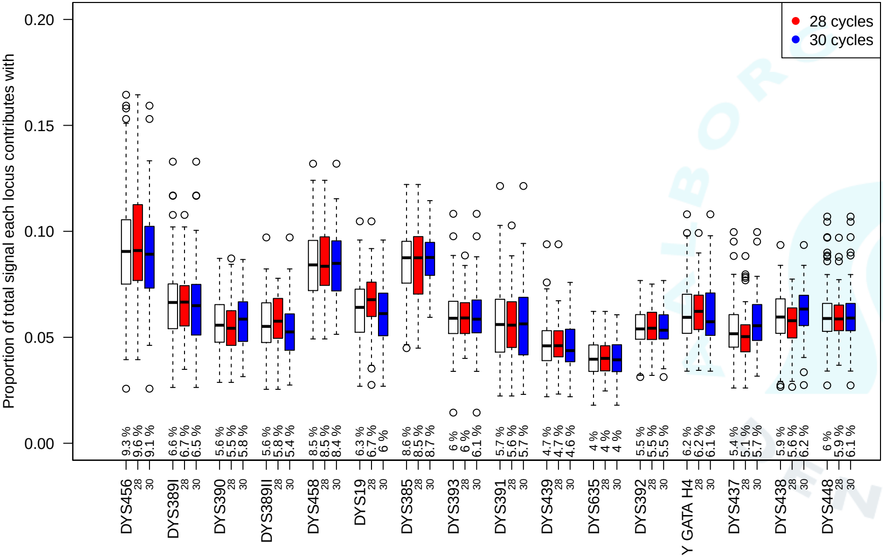
What is signal strength?

Or rather:
How should signal strength be quantified?

Signal strength

- Locus balances for the Yfiler kit
    - Samples with full profiles collected and peak heights compared
- Knowledge of truncated observations

Locus (im)balances for full profiles



Signal strength estimators: dealing with locus balances

- Assume that log peaks heights are normally distributed

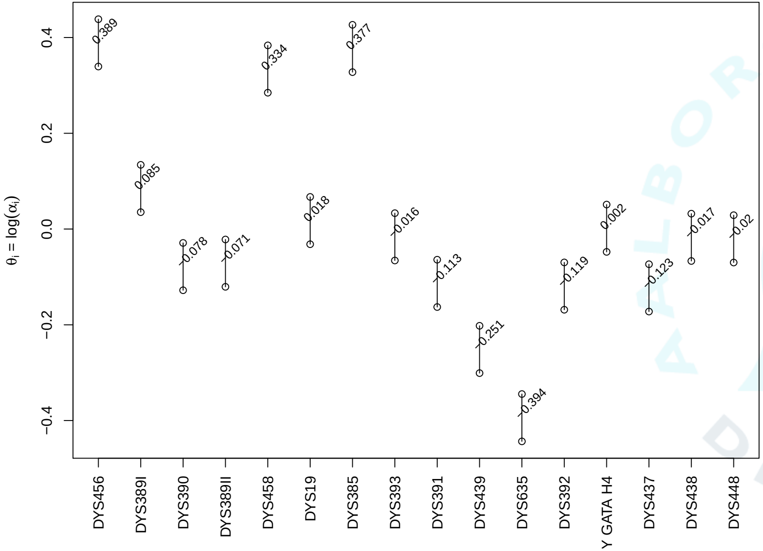
$$\log x_{ij} \sim N(\log \alpha_i + \log S_j, \sigma^2) = N(\theta_i + \log S_j, \sigma^2) \qquad (1)$$

- $i = 1, \dots, r$ refers to locus, hence $\theta_i = \log \alpha_i$ are locus balances
- $j = 1, \dots, n$ refers to sample, hence $\log S_j$ are signal strengths
- Sum contrasts for θ_i: $0 = \sum_i \theta_i = \log(\prod_i \alpha_i)$, s.t. $\prod_i \alpha_i = 1$
- Threatment contrasts for $\log S_j$
- Strategy: use full profiles to fit this model to obtain locus balances θ_i for $i = 1, \dots, r$ and use these later for estimating signal strength in samples with drop-out

Signal strength estimators: dealing with locus balances

- In `R`, this model is easily fitted using the linear model fit function called `lm`
- Fitting done with only full profiles without drop-outs resulting in an adjusted $R^2 = 0.9982$

Locus (im)balances



Notation

Let $I \subseteq \{1, 2, \ldots, r\}$ denote the set of loci that dropped out (i.e. was below a given threshold), $I^C = \{1, 2, \ldots, r\} \setminus I$ the observed loci, and $k = |I|$ the number of drop-outs

Biased estimator

- A simple biased estimator for the signal strength is then

$$\log \hat{S} = \frac{1}{r-k} \sum_{i \in I^C} (\log x_i - \theta_i) \tag{2}$$

- Biased: does not incorporate the fact that we are aware of observing a truncated sample

Signal strength estimators: dealing with truncation

- We observed only a truncated sample, namely

  $$\log x_{ij} \sim N_{\log t}(\theta_i + \log S_j, \sigma^2)$$

  that is, normally distributed truncated below $\log t$ (e.g. $t = 50$ RFU)
- Parameters can be estimated by maximising the likelihood (see e.g. Persson and Rootzen (1977))
- Complex to do by hand (analytically), but R can do it fast

Single term deletions

- Logistic regression with two- and three-ways interactions

$$
\begin{aligned}
p &\sim S * L * C \\
&= S + L + C + (S : L) + (S : C) + (L : C) + (S : L : C)
\end{aligned}
$$

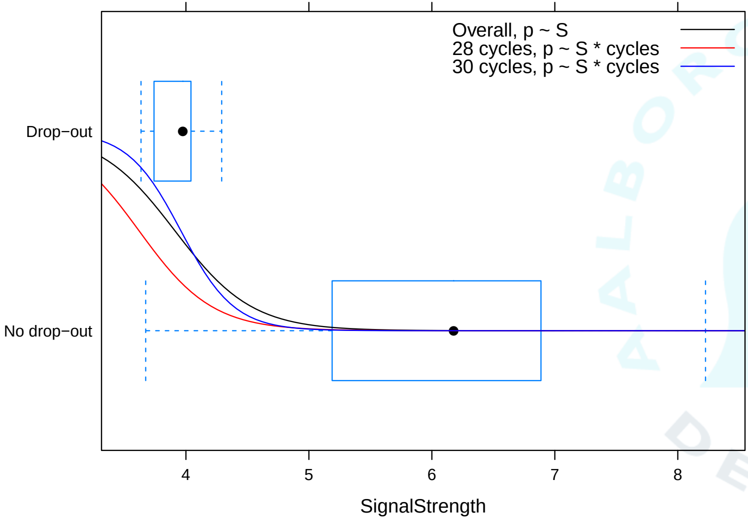
  - p: probability of drop-out
  - S: signal strength
  - L: locus
  - C: number of PCR-cycles
- In R, this model is easily fitted using the generalized linear model fit function called glm

Single term deletions

To test if a single term in the model can be removed, R's drop1 function can be used:

```
DropOut ˜ Locus * Cycles * SignalStrength
                            Df Deviance    AIC    LRT   Pr(Chi)
<none>                           592.55 720.55
Locus:Cycles:SignalStrength 15   631.51 729.51 38.951 0.0006517 ***
```

Logistic regression for DYS456



Tables

- For 28 cycles and certain drop-out probabilities, the average peak height per locus resulting in this probability is listed

P(Dropout)	DYS456	DYS389I	DYS390	DYS389II	DYS458	...
0.0001	853	384	332	265	1122	
0.0005	528	275	247	203	606	
0.0010	430	238	217	181	465	
0.0050	266	170	161	138	251	
0.0100	216	147	142	123	192	
0.0500	132	105	104	93	102	
0.1000	106	90	91	82	77	
0.2000	83	76	78	72	56	
0.3000	71	68	71	66	46	
0.4000	62	62	65	61	39	
0.5000	55	57	61	57	33	
0.6000	49	52	56	53	28	
0.7000	43	48	52	50	24	
0.8000	36	43	47	45	20	
0.9000	29	36	40	40	14	
0.9500	23	31	35	35	11	
0.9900	14	22	26	27	6	

References

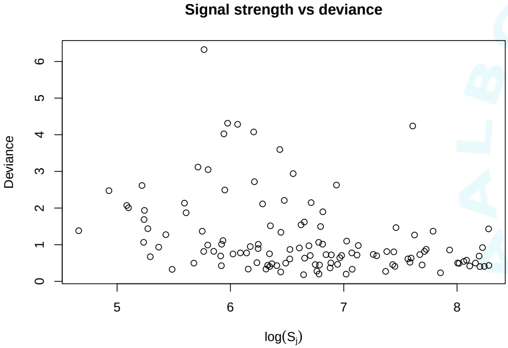
- T Tvedebrink, PS Eriksen, HS Mogensen, N Morling. *Estimating the probability of allelic drop-out of STR alleles in forensic genetics*. FSI:Gen (2009), 3: 222-226
- T Persson, H Rootzen. *Simple and Highly Efficient Estimators for a Type I Censored Normal Sample*. Biometrika (1977), 64: 123-128

DYS385

- As (almost) always, DYS385 introduces problems
- In this experiment, all persons had two alleles at DYS385
- When estimating the locus balances, the peak heights were replaced with the sum of the heights of the two peaks to get just a single height
- In the drop-out analysis when counting the number of drop-outs, DYS385 was divided into two loci, DYS385a and DYS385b, each with $\theta_i = \theta_{i'}/2$ where i' is the index corresponding to DYS385 to respect sum contrasts
- Analysis is then performed as follows:
  - Two peaks: Two contributions to second product in (3)
  - One peak: A contribution to each product in (3) (each with same θ_i)
  - No peaks: Two contributions to first product in (3)

Variance dependence on $\log S_j$?

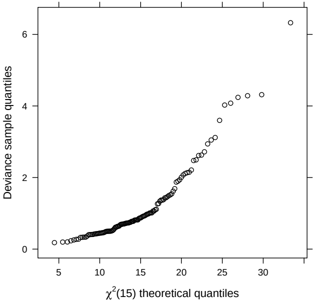
- Deviance for j'th sample $= \sum_i r_{ij}^2$ where r_{ij} is the residual for the j'th sample at the i'th locus



- No need to think that σ^2 depends on $\log S_j$ (for the range of $\log S_j$ values used here)

Variance dependence on j?

- Deviances are expected to follow a $\chi^2(15)$ distribution if the variance does not dependent on j



- Might be improved by allowing σ^2 to vary between samples, e.g. $\sigma^2(j) \sim \Gamma(a)$

Likelihoods: preparation

- Let Φ and ϕ be the cumulative distribution function and probability density function, respectively, of the standard normal distribution
- Remembering that if $x \sim N(\mu, \psi^2)$ then $(x - \mu)/\psi \sim N(0, 1)$ and $f(x; \mu, \psi^2) = \psi^{-1}\phi((x - \mu)/\psi)$ where f is the probability density function of x
- The likelihood of a sample $x_1, x_2, \ldots, x_n$ from $N(\mu, \psi^2)$ is $L\left(\mu, \psi^2; \bigcup_{i=1}^{n}\{x_i\}\right) = \prod_{i=1}^{n} f(x_i; \mu, \psi^2)$
- The likelihood of a sample $x_1, x_2, \ldots, x_{n-k}$ from $N_t(\mu, \psi^2)$ (i.e. $\geq t$), and additional k samples below t can be shown to be proportional to (see e.g. Persson and Rootzen (1977)) $\left[\Phi\left(\frac{t-\mu}{\psi}\right)\right]^k \prod_{i=1}^{n-k} f(x_i; \mu, \psi^2) =$ $\left[\Phi\left(\frac{t-\mu}{\psi}\right)\right]^k \prod_{i=1}^{n-k} \psi^{-1}\phi\left(\frac{x_i-\mu}{\psi}\right)$

Likelihoods

- The likelihood of observing a signal assuming known loci balances θ_i is then given by

$$L\left(\gamma,\sigma;\bigcup_{i\in I^C}\{x_i\}\right) = \prod_{i\in I}\Phi\left(\frac{t-(\theta_i+\gamma)}{\sigma}\right) \tag{3}$$

$$\prod_{i\in I^C}\sigma^{-1}\phi\left(\frac{x_i-(\theta_i+\gamma)}{\sigma}\right) \tag{4}$$

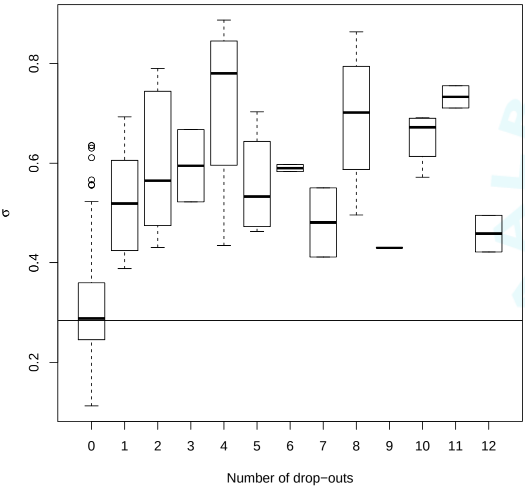
- Ignoring locus balances, i.e. $\alpha_i = 1$ and $\theta_i = 0$, we get

$$\prod_{i\in I}\Phi\left(\frac{t-(\theta_i+\gamma)}{\sigma}\right) = \left[\Phi\left(\frac{t-\gamma}{\sigma}\right)\right]^k \tag{5}$$

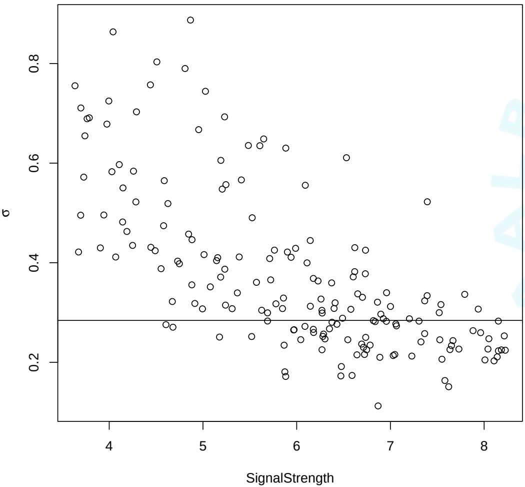
Variance control

- In the likelihood optimisation, both $\log S_j$ and σ are estimated (we assumed the locus balances θ_i to be known – estimated from full profiles)
- σ has previously been estimated in the linear model so comparisons can be made

Variance control



Variance control



Variance control

- This might look terrible, but remember selection bias: when the variance for a sample is large, the probability of drop-out increases
- In the likelihood optimisation, both $\log S$ and σ are already estimated, which can be exploited

Dilutions

Dilution name	Expected DNA ($pg/\mu l$)	Proportion
F1	500	5
F2	400	4
F3	300	3
F4	250	2.5
F5	200	2
F6	150	1.5
F7	**100**	**1**
F8	75	0.75
F9	50	0.5
F10	30	0.3
F11	15	0.15
F12	7.5	0.075