Stochastic filtering of quantitative data from STR DNA analysis

Torben Tvedebrink[a], Poul Svante Eriksen[a],Helle Smidt Mogensen[b], Niels Morling[b]
[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Science, University of Copenhagen, Denmark.

Introduction

The quantitative data observed from analysing STR DNA is a mixture of contributions from various sources. Apart from the true allelic peaks, the observed signal consists of at least three components resulting from the measurement technique and the PCR amplification:

- Background noise (random noise due to the apparatus used for measurements).
- Pull-up effects (more systematic increase caused by overlap in the spectrum, see right picture of Figure 1).
- Stutters (peaks located four basepairs before the true peak - are proposed to originate from primer mispairings [1]).

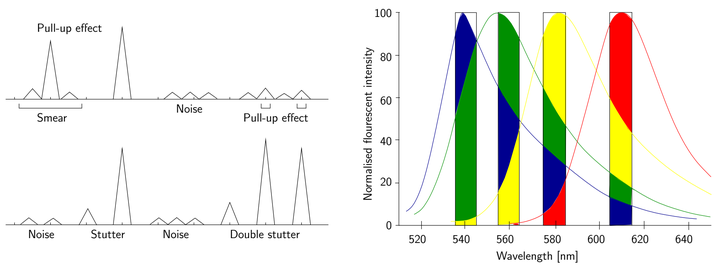

Fig. 1. *Left:* Picture of the non-signal components of a STR DNA trace. *Right:* Fluorescent dye bands where the colours under each curve indicate the amount of spectral overlap between the various dyes.

We present filtering techniques for all three technical artifacts based on statistical analysis of data from controlled experiments conducted at The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark[b].

The filter

In the sections below, we describe the methods used, in filtering the observed data. The samples were prepared as described in [5] and the data were analysed using a threshold of 5 rfu on peak heights and with no stutter filter or any other method of pre-filtering.

Figure 2 shows a flow chart of the filter, and in Section "Noise filter", we describe how, the signal detection limit is determined. This limit is also used as threshold when deleting signals. In Sections "Pull-up filter" and "Stutter filter", the two remaining filters based on regression for pull-up and stutters are presented.

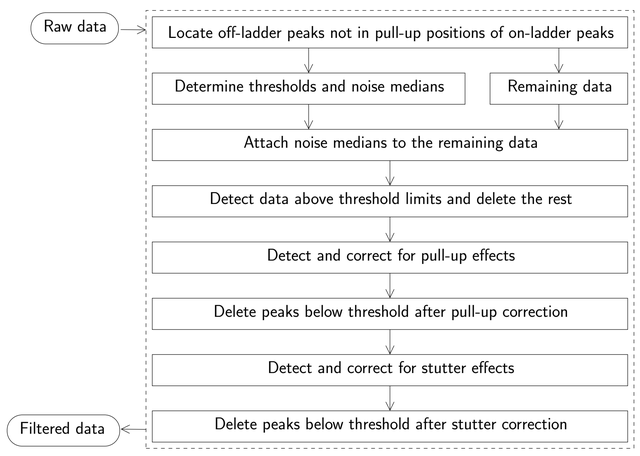

Fig. 2. Flow chart of the filter.

Noise filter

The peak height observations all fell in the interval from 5 rfu and upwards indicating that the noise distribution is a truncated positive valued distribution.

Since the signal comprised both background noise and more systematic components, we removed all peaks on the allelic ladder and also off-ladder alleles in possible pull-up position of the ladder on different fluorescent dye bands. This ensures that the remaining points are pure noise as stutters and true peaks per definition fall on the ladder, and pull-up effects need therefore to be induced by peaks on the ladder.

Graphical inspections of the data indicated that the noise distribution is heavy tailed. We investigated several heavy tailed distributions including the exponential, Fisher-Tippett, Pareto, Rayleigh, and Weibull. However, transforming the peak height observations, h, by $\log(h - 4.5)$ indicated a sufficient fit to normality.

In Figure 3, we stratified the observed peak heights by STR locus and plotted the transformed heights, i.e. $\log(h - 4.5)$, against a standard normal distribution in a QQ-plot. This plot demonstrated that the noise (shifted by -4.5) followed a log-normal distribution with individual mean, μ_s, and variance, σ_s^2, for each locus s. These parameters determine the intercept and slope of the superimposed QQ-line and are found by

$$\sigma_s = \frac{x_{s(90\%)} - x_{s(50\%)}}{z_{(90\%)} - z_{(50\%)}} \quad \text{and} \quad \mu_s = x_{s(50\%)} - \sigma_s z_{(50\%)},$$

where $x_{s(q)}$ and $z_{(q)}$ are the empirical and standard normal q-quantiles, respectively. We used these quantile estimators rather than the ordinary maximum likelihood estimators in order to increase the robustness. Due to filtering, the noise from the true signal, the main interest of the noise distribution, is the upper tail as indicated by the chosen fractiles.

The threshold was determined by three times the standard deviation implying that ca. 99.9% of the noise will be determined as noise by the filter. Hence, the locus specific threshold is found as

$$\text{Threshold for locus } s = \exp(3\sigma_s + \mu_s) + 4.5\,.$$

From Figure 3 it is clear that the fit to normality is poorer for the low values of $\log(h - 4.5)$. However, the observations in this region are not of concern with respect to noise filtering due to their limited height.

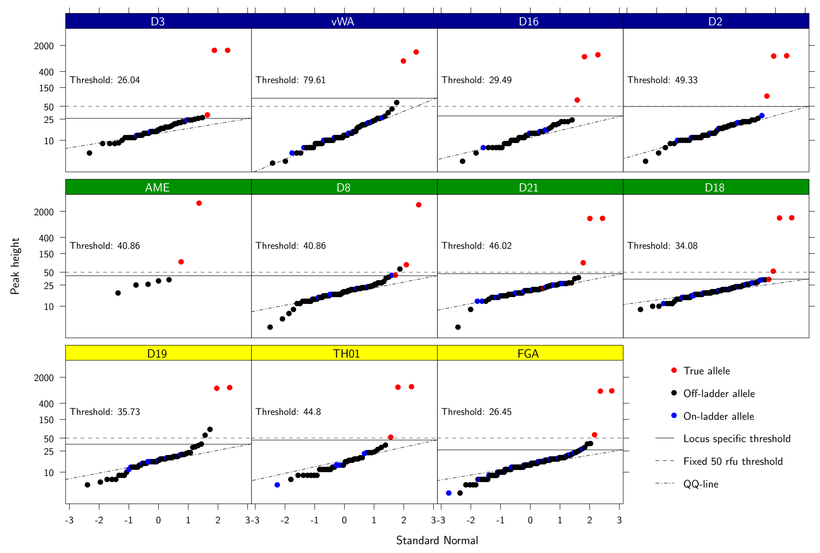

Fig. 3. QQ-plots of the observed signal. Note the different thresholds computed using the locus itself as reference. For this particular sample, the fixed 50 rfu-threshold caused four drop-outs (D3, D8, D21 and D18) and two (D21 and D8) for the locus specific threshold (one true peak at locus D21 had height 21 rfu and was embedded in the noise).

Pull-up filter

We defined pull-ups as peaks not being true alleles or possibly stutters on a different dye band than the parent peak within ± 0.5 basepairs of the parent's basepair.

In Figure 4, we stratified the pull-up observations by transferring and receiving fluorescent dyes. The magnitudes of the observed pull-up effects were in accordance with the spectrum overlap in the right picture of Figure 1.

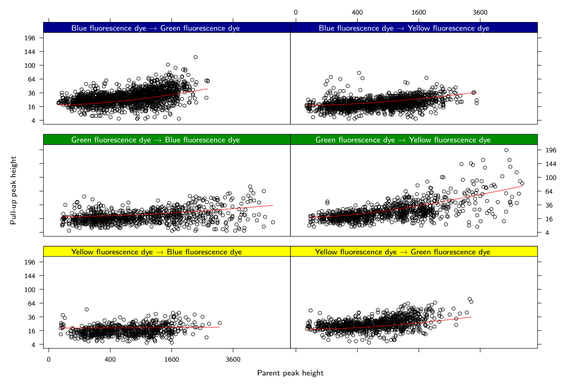

Fig. 4. Pull-up effects stratified by overlapping fluorescent dyes. The superimposed lines indicate the estimated model for the directions of the different panels. The plot is on square-root scale since this is the variance stabilising transformation.

For predictive purposes, we fitted a linear model to the observed pattern in Figure 4. Of the included data-points, only a limited subset comprised *detectable* pull-up peaks, and the remaining observations were background noise in pull-up position. Our model takes this into account by having a noise dependent intercept, $\phi_{\text{Noise},s}$, for locus s. In the formulation of the model, the notation $D \to d$ reflects that the pull-up peak is located on fluorescent dye band d and the parent peak on fluorescent dye band D,

$$\phi_{\text{Pull-up}} = \alpha_{D\to d}\phi_{\text{Noise},s} + \beta_{D\to d}\phi_{\text{Parent}}. \qquad (1)$$

The model is valid for both peak heights and areas replaced for ϕ in (1), where $\phi_{\text{Pull-up}}$ denotes the mean height or area of the pull-up peak and ϕ_{Parent} is the observed peak height or area of the parent peak. The parameter estimates of $\alpha_{D\to d}$ and $\beta_{D\to d}$ for peak heights are presented in Table 1.

Table 1. Parameter estimates for the pull-up model (1) for peak heights.

Dye $\to$ dye	B $\to$ G	B $\to$ Y	G $\to$ B	G $\to$ Y	Y $\to$ B	Y $\to$ G
$\alpha_{D\to d}$	1.047	0.958	1.069	0.955	1.125	1.018
$\beta_{D\to d}\times 10^2$	0.951	0.566	0.319	1.081	0.000*	0.560

Stutter filter

Assuming additivity of the noise and stutter products, we take into account that stutters from small peaks mainly consist of noise. The model for the expected stutter height or area is given by

$$\phi_{\text{Stutter}} = \alpha_s\phi_{\text{Noise},s} + \beta_s\phi_{\text{Parent}} + \gamma_s\tilde{bp}\cdot\phi_{\text{Parent}}, \qquad (2)$$

where $\phi_{\text{Noise},s}$ is the known median of the off-ladder peaks not in pull-up positions on locus s. In the latter term, $\tilde{bp}$ is the basepair deviation from the mean basepair, $\tilde{bp} = bp - \bar{bp}$. By including basepair in the model, we are able to have different stutter percentages within the same locus for different alleles. It also means that β_s can be interpreted as an average stutter effect at a given locus s. The use of ϕ_{Stutter} and ϕ_{Parent} stresses that the model is valid for both peak heights and areas.

The previously observed increase in stutter percentage as a function of allelic number is captured in the positive estimates of γ_s in Table 2 and our estimates are in concordance with the picture in [3, Figures 9-5, 9-6 and 9-7].

Table 2. Parameter estimates for the stutter model (2) for peak heights.

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH01	FGA
α_s	0.935	0.867	0.988	0.911	0.916	0.954	0.870	0.934	1.034	1.015
$\beta_s\times 10^2$	6.929	6.100	5.255	7.525	4.798	6.182	6.531	6.413	1.567	5.751
$\gamma_s\times 10^2$	0.101	0.202	0.218	0.090	0.092	0.082	0.181	0.192	0.067	0.138

In Figure 5, we plotted the stutter peak heights predicted by the model against the observed stutter peak heights. The plot demonstrates that the model in (2) is sufficient in order to describe the stutter behaviour.

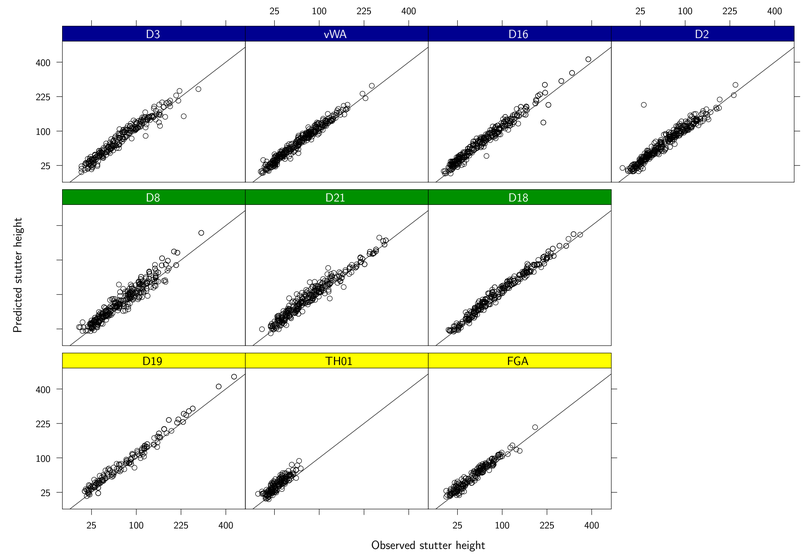

Fig. 5. Predicted stutter peak heights plotted against observed stutter peak heights with the identity line superimposed. The plot is on square-root scale as this is the variance stabilising transformation.

Additional examinations of the data made it clear that backstutters are present in the data. The model for backstutters is based on the same idea as the filters for pull-up and stutters with a noise level and an additional effect from the parental peak, but the details are omitted owing to lack of space.

Results

We have used our filter on 191 two-person mixtures. In Table 3, we have summarised the performance of the overall filter. It is worth emphasising that 179 of the true alleles dropped out and that the stutter filter only let 32 stutters (25 stutters and 7 backstutters) slip trough. In addition to the stutter peaks, another 181 (128 drop-ins, 45 pull-ups and 8 smears) on-ladder peaks were classified as proper peaks by the filter.

Table 3. Results for the overall filter. Smear is peaks ± 3.5 bp from true peak.

Classification	Noise	Signal
True alleles	179	5562
Stutters	3287	25
Back-stutters	2231	7
On-ladder alleles	16798	128
Off-ladder observations	72961	275
Smear positions	19302	545
On-ladder smear	*2460*	*8*
Off-ladder smear	*16842*	*537*
Pull-up peaks	8034	198
On-ladder pull-up peaks	*1753*	*45*
Off-ladder pull-up peaks	*6281*	*153*

The remaining peaks passing the filter were all detected to be off-ladder and thus removed from the analysis afterwards. The data were also analysed following the standard protocol of the Section of Forensic Genetics at University of Copenhagen[b]. Using the technique recommended by the manufacturer, 262 drop-outs were observed together with 26 stutters and 14 drop-ins.

Discussion

Three times the standard deviation was also used in [4] for determining the limit of detection (LOD). However, the parameters μ and σ in [4] were based on negative controls and reagent blank samples, which implies that the parameters were computed for capillaries not containing the actual sample itself. This does not take the possible differences between capillaries within a batch into account.

An advantage of the locus specific threshold is that it enables the case worker to assess the noise level of the sample. Furthermore, in cases where the distribution of the transformed peak heights deviates substantially from normality, the sample may be subject to extensive noise and/or contamination of some sort.

Conclusion

The methodology of regression and distributional analysis of the noise yielded satisfying results in order to deduce a stochastic filter for STR DNA samples.

Comparisons of the results with those based on the recommendations of the manufacturer indicated that the number of drop-outs decreased by approximately 30%. Studies of different data sets supported this improvement and suggests that the methodology of the threshold determination is adequate for the noise filtering of quantitative STR data.

References

[1] Butler, JM. 2005. 'Forensic DNA Typing, 2ed'. *Elsevier Science and Technology*.

[2] Gill, P., Curran, J., Elliot, K. (2005). 'A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci'. *Nucleic Acids Research*, 33(2): 632-643.

[3] Applied Biosystems. 2006. *AmpFℓSTR® SGM Plus® PCR Amplification Kit User's Manual*. Applied Biosystems.

[4] Gilder, JR., Doom, TE., Inman, K., Krane, DE. (2007). 'Run-Specific Limits of Detection and Quantitation for STR-based DNA Testing'. *Journal of Forensic Science*, 52(1): 97-101.

[5] Tvedebrink, T., Eriksen, PS., Mogensen, HS., Morling, N. (2008). 'Evaluating the weight of evidence using quantitative STR data in DNA mixtures - With notes on handling missing data'. *Manuscript in preparation*.