A Signal Model for Forensic DNA Mixtures

Ullrich J. Mönich*, Catherine Grgicak†, Viveck Cadambe§, Jason Yonglin Wu*,
Genevieve Wellner†, Ken Duffy‡, Muriel Médard*

*Research Laboratory of Electronics
Massachusetts Institute of Technology
Email: {moenich, ylwu, medard}@mit.edu

†Biomedical Forensic Sciences
Boston University School of Medicine
Email: {cgrgicak, gwellner}@bu.edu

§Department of Electrical Engineering
Pennsylvania State University
Email: viveck@engr.psu.edu

‡Hamilton Institute
National University of Ireland Maynooth
Email: ken.duffy@nuim.ie

***Abstract*—For forensic purposes, short tandem repeat allele signals are used as DNA fingerprints. The interpretation of signals measured from samples has traditionally been conducted by applying thresholding. More quantitative approaches have recently been developed, but not for the purposes of identifying an appropriate signal model. By analyzing data from 643 single person samples, we develop such a signal model. Three standard classes of two-parameter distributions, one symmetric (normal) and two right-skewed (gamma and log-normal), were investigated for their ability to adequately describe the data. Our analysis suggests that additive noise is well modeled via the log-normal distribution class and that variability in peak heights is well described by the gamma distribution class. This is a crucial step towards the development of principled techniques for mixed sample signal deconvolution.**

I. Introduction

Short tandem repeat (STR) allele signal interpretation is a central tool in forensic analysis, as the number of repeats, i.e. the number of repeated copies of a basic motif, at given loci serve as an individual's DNA fingerprint. The main artifacts that affect the interpretation are stutter, which is an echo at a fixed known distance from the allelic peak, variabilities in the allelic peak heights, and baseline noise [1].

These artifacts are conventionally treated by applying different thresholds to the data. For example, the effects of baseline noise in STR profiles are suppressed by applying a threshold which is called analytical threshold, detection threshold, or minimum distinguishable signal threshold [2]–[6]. Further, a second threshold, the stochastic threshold, may be used as a tool to detect the presence of allelic peaks [7]. The traditional way to counter the effect of stutter is to apply a stutter ratio threshold, where the ratio is calculated by dividing the height of the peak in stutter position by the height of the allelic peak [8]–[11]. Other effects are generally not treated specifically [12].

U. Mönich was supported by the German Research Foundation (DFG) under grant MO 2572/1-1.
This project was partially supported by 2012-DN-BX-K050 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not reflect those of the Department of Justice.

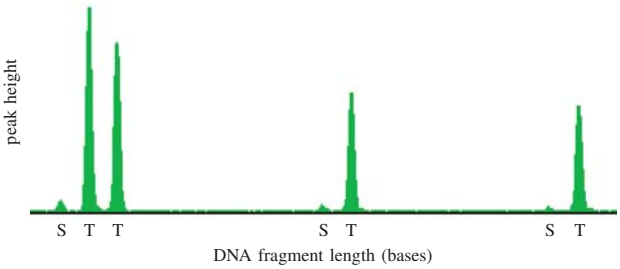

Fig. 1. Segment of an electropherogram signal. True peaks are marked with T and stutter peaks with S.

Applying thresholds during analysis has the drawback that information is lost. For that reason, continuous methods, where fewer or no thresholds are used, have been developed [11], [13], [14]. In these methods the full variability in the peak heights is taken into consideration, which leads to a soft decision instead of a hard limiting.

Noise in STR profiles has been modeled as a normally distributed random variable [15], though a log-normal distribution has also been suggested [6].

Recently, a Gaussian model for noise and allelic peak heights has been proposed for the purpose of determining the likelihood that a given number of individuals contributed to a mixed sample [14]. Although the Gaussian model provided improved identification over previous techniques, [14] did not provide an analysis, independent of determining the most likely number of contributors, to confirm its appropriateness. Here we revisit this premise.

In this work, we derive a signal model for forensic DNA mixtures from empirical data, using the Kolmogorov–Smirnov (KS) and the chi-squared tests to assess the suitability of different distribution classes. We believe that such a signal model is beneficial, as it will enable well-established techniques and methods from signal processing to be used in the analysis and interpretation of DNA profiles.

II. Biological Background Information

The most widely used method for forensic DNA analysis is based on short tandem repeats (STRs). In this method,

the DNA sample is first amplified by polymerase chain reaction (PCR), and then processed by capillary electrophoresis, the output of which is an electropherogram, as shown in Fig. 1. Since there is little information in the peak shape, the electropherogram is further processed by a peak detection algorithm, which outputs a list of peak locations together with the corresponding heights of the peaks, measured in relative fluorescence units (RFUs). The peak location contains information about the fragment length, i.e. the number of repeats, and the peak height information about the amount of DNA in the sample.

The data that was used for the analyses in this paper was generated with the AmpFlSTR Identifiler Plus kit, the GeneAmp PCR System 9700, the 3130 Genetic Analyzer, and the GeneMapper IDX v1.1.1 software from 643 single person measurements. The kit has 15 tetranucleotide STR loci. An injection time of 10 s was used, and the amount of DNA in the samples was varied between 0.008 ng and 0.25 ng.

Artifacts in the generation of the electropherogram include stutter, dye blobs, bleed through, -A, and spikes. For a definition of these terms and further explanations, see for example [1]. We do not model dye blobs, bleed through, -A, and spikes because they can be reliably removed in advance, as it has been done manually for our data.

Stutter is a common artifact that is created during the amplification of the DNA. Due to errors in the PCR process, spurious stutter peaks, close to the allelic peaks, are inserted. For tetranucleotide STR loci, the strongest stutter occurs at a location which corresponds to a fragment length that is 4 base pairs shorter than the fragment length of the allelic peak. Stutter at this location is referred to as $N-4$ stutter. Analogously, $N+4$ stutter denotes stutter that is 4 base pairs longer than the fragment length of the allelic peak.

III. The Signal Model

There are different ways to represent the location of a peak. In an idealized DNA signal, i.e., in a signal with no artifacts or noise, each peak corresponds to an allele that is present in the DNA sample. Hence, we can specify each peak location by a pair (locus, allele name). This representation, which is common in biology, is however not optimal for our purpose of building a signal model.

We choose a vector representation of the measured data, similar to [16], in which we list the signal values at all the possible allele positions in a vector $\underline{y} = (y_1, y_2, \ldots, y_I)^{\mathsf{T}}$ of length I. In our case, since the peak heights are given as non-negative integers, and we have 287 possible allele positions, $\underline{y}$ is a vector in $\mathbb{N}_0^{287}$, where $\mathbb{N}_0$ denotes the set of natural number including zero.

The proposed signal model is given by

$$\underline{y} = \sum_{n=1}^{N} (t_n \underline{\gamma} \circ S\underline{x}_n) + \underline{\eta}, \tag{1}$$

where $\underline{a} \circ \underline{b}$ denotes the component-wise (Hadamard) product of the vectors $\underline{a}$ and $\underline{b}$.

The parameters of the model are: The number of contributors in the mixture N, the genotypes of the contributors $\underline{x}_1, \ldots, \underline{x}_N$, and the DNA amounts of the contributions $t_1, \ldots, t_N$. The genotype vector $\underline{x}_n \in \{0,1,2\}^I$ has a 2 at index i if person n has a homozygote allele at this index, a 1 if person n has a heterozygote allele at this index, and a 0 if person n has no allele at this index. The matrix $S \in \mathbb{R}^{I \times I}$ models stutter. $\underline{\gamma}$ is a random vector that describes the variation in the allelic peak heights and $\underline{\eta}$ a random vector that describes the effect of additive noise. We use the standard assumption that both random vectors have independent components. The distributions of $\underline{\gamma}$ and $\underline{\eta}$ are analyzed in the next section.

We model mixtures with more than one person as a linear superposition of the individual contributions, which is justified by considerations about the involved physical and chemical processes.

IV. Data Analysis

In order to determine the signal model (1), we analyzed the data from 643 single person measurements with a DNA amount ranging from 0.008 ng to 0.25 ng. Knowing the genotype of these samples, we can group the components of the signal vector $\underline{y}$ into three categories:

1) true peak component,
2) stutter component, and
3) noise component.

We call a component y_i of the signal vector $\underline{y}$ a true peak component if the person has either a homozygote allele (double true peak) or a heterozygote allele (single true peak) at index i. We call a component, a stutter component if it is either in $N-4$ or in $N+4$ stutter location of a true peak. Further, all remaining components are called noise components.

The presence of small random errors in the processing of the DNA sample and the measurement can be interpreted as noise. Thus, even if we would not expect a peak at a certain location according to the genotype, it nevertheless can happen that we measure a non-zero value, due to noise.

We start with the analysis of noise in Section IV-A. The true peaks are treated in Section IV-B and stutter in Section IV-C. Since it is well-known that the statistics of the peaks depends on the locus, we do a per locus analysis.

A. Noise

Roughly 80% of the noise components have zero height. The rest of the analysis in this section deals only with the non-zero noise measurement values.

The actual peak heights are given as integers by the software. Hence, we can either model the peak heights as a discrete random variable, or as a continuous random variable that is quantized to integer values. We choose the second approach, because, as we will see, such a model explains the data very well. In the statistical literature quantized data is also known as grouped data.

In order to apply the KS test, the parameters of the reference distributions are obtained from the data by maximum-likelihood (ML) estimators.

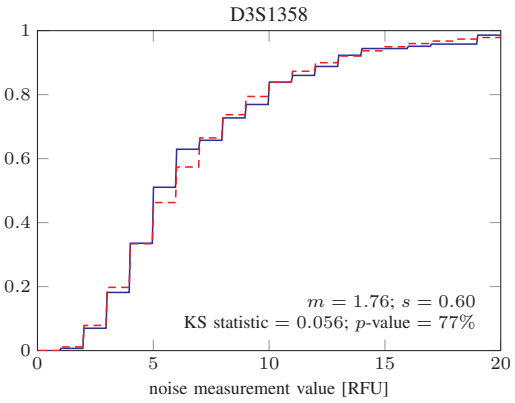

Fig. 2. Empirical CDF of the non-zero noise measurement values (blue) and CDF of a quantized log-normal random variable (dashed red) for locus D3S1358 and a DNA mass of 0.25 ng.

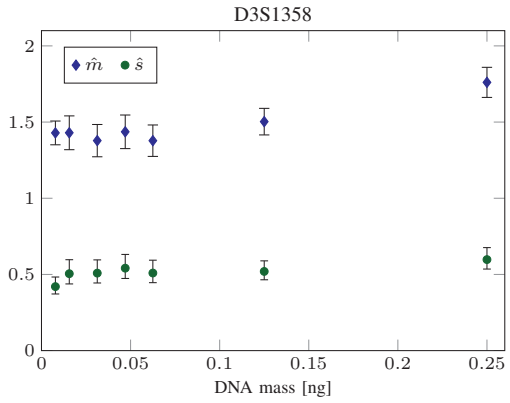

Fig. 3. Maximum likelihood estimates $\hat{m}$ and $\hat{s}$ of the parameters m and s for the non-zero noise measurement values, together with confidence intervals, as a function of the DNA mass for locus D3S1358.

		min p-val.	max p-val.	< 0.05	rej.
non-zero noise	log-normal	59.7%	100%	0	0
	gamma	2.6%	100%	1	0
	normal	0.0%	69.7%	9	5
single true peak	log-normal	14.3%	99.9%	0	0
	gamma	20.5%	99.8%	0	0
	normal	10.8%	94.2%	0	0

TABLE I
KS TEST: MINIMUM AND MAXIMUM p-VALUES OVER THE 15 LOCI, THE NUMBER OF LOCI WITH A p-VALUE SMALLER THAN 5%, AND THE NUMBER OF LOCI FOR WHICH THE HYPOTHESIS IS REJECTED AFTER HOLM-BONFERRONI CORRECTION.

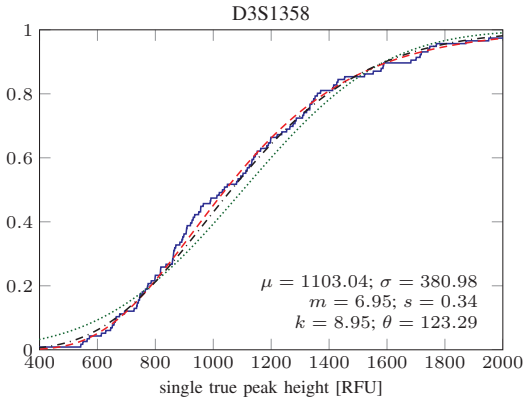

Fig. 4. Empirical CDF of the single true peak heights (blue), CDF of a normal random variable (dotted green), CDF of a log-normal random variable (dashed red), and CDF of a gamma random variable (dash-dotted black) for locus D3S1358 and a DNA mass of 0.25 ng.

In Fig. 2 we see the empirical cumulative distribution function (CDF) of the noise measurement values for the D3S1358 locus and a DNA mass of 0.25 ng in blue and the CDF of a quantized log-normally distributed random variable with parameters $m = 1.76$ and $s = 0.60$ in dashed red. The quantized log-normal cumulative distribution function is given by

$$F^{\mathrm{q}}_{m,s}(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln(\lfloor x \rfloor + 1/2) - m}{s\sqrt{2}}\right)\right],$$

where erf is the error function. The KS statistic is 0.056, which leads to a p-value of 77%. The p-values for all loci range from 59.7% to 100%.

We also perform the test for the gamma distribution class, which gives p-values in the range from 2.6% to 100%, and the normal distribution class, where 9 of the 15 loci have a p-value smaller then 5%.

The p-value is a measure for the quality of a fit. If the p-value is smaller than 5%, the hypothesis that the samples are taken from the reference distribution would be rejected. However, if multiple hypotheses are tested, as in our case for the different loci, the likelihood to witness a rare event increases. The Holm-Bonferroni correction [17] is a method to counteract the problem of multiple testing.

The minimum and maximum p-values over all loci for the log-normal, the gamma, and the normal distribution are shown in Table I for a DNA mass of 0.25 ng. Further, the table shows the number of loci for which the hypothesis is rejected before and after Holm-Bonferroni correction.

Although we estimate the parameters of the reference distribution from the same data that is used for the KS test, and the KS test is known to be conservative for quantized data in the sense that the obtained p-values are too large, we still can exclude the normal distribution, and have an indication that the log-normal distribution might explain the data better than the gamma distribution. Application of Pearson's chi-squared test, the results of which are summarized in Table II, supports this.

The results that have been presented so far, are from data with a DNA mass of 0.25 ng. Since the results for the other DNA masses are similar, and the dependence of the estimated parameters $\hat{m}$ and $\hat{s}$ on the DNA mass is minimal for the DNA mass range from 0.008 ng to 0.25 ng, as shown in Figure 3, we choose to model the additive noise $\underline{\eta}$ as target independent.

B. True Peaks

In the case of the true peaks, we can ignore the effects of quantization, because the peak heights are in the order of hundreds of RFUs.

In Fig. 4 we see the empirical CDF of the single true peak heights in blue, the CDF of a normally distributed random variable with mean μ and variance σ^2

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

in dotted green, the CDF of a log-normally distributed random variable with parameters m and s

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\ln x-m}{s\sqrt{2}}\right)\right]$$

in dashed red, and the CDF of a gamma distributed random variable with parameters k and θ

$$\frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)},$$

where γ is the lower incomplete gamma function and Γ the gamma function, in dash-dotted black for the D3S1358 locus and a DNA mass of 0.25 ng. The parameters of all distributions were determined by the corresponding maximum likelihood estimator. The p-value of the KS test for the normal distribution is 20.6%, the p-value for the log-normal distribution 95.4%, and the p-value for the gamma distribution 67.7%.

The p-values for all loci range from 10.8% to 94.2% for the normal distribution class, from 14.3% to 99.9% for the log-normal distribution class, and from 20.5% to 99.8% for the gamma distribution class. There is no clear preference for one of the three distribution classes, since none of them is rejected by the KS test, as shown in Table I.

Therefore, we also perform the chi-squared test for the true peaks and the different distribution classes. For each DNA mass, distribution class, and locus the procedure is as follows:

1) Choose the initial number of bins according to $\lfloor 1.88 \cdot M^{2/5} + 1/2 \rfloor$, where M denotes the number of samples [18].
2) Get the ML estimate of the parameters from the binned data.
3) Pool the bins to ensure that the theoretical frequency in each bin is larger than or equal to 5.
4) Get an update of the ML estimate of the parameters from the newly binned data.
5) Calculate the p-value based on the chi-squared test.

After having computed the p-values for all loci, we also do a Holm-Bonferroni correction to correct for the multiple loci testing.

In Table II we see the results for a DNA mass of 0.25 ng. Without Holm-Bonferroni correction, for both the gamma and the log-normal distribution class 4 loci have a p-value smaller than 5%, and for the normal distribution class 5 loci loci have a p-value smaller than 5%. With Holm-Bonferroni correction, for the log-normal and normal distribution class the hypothesis is rejected for 2 loci and for the gamma distribution class the hypothesis is rejected for 1 locus.

		min p-val.	max p-val.	< 0.05	rej.
non-zero noise	log-normal	2.7%	86.8%	1	0
	gamma	0.0%	93.6%	5	1
	normal	0.0%	19.4%	13	11
single true peak	log-normal	0.0%	85.2%	4	2
	gamma	0.1%	99.0%	4	1
	normal	0.0%	57.8%	5	2

TABLE II
CHI-SQUARED TEST: MINIMUM AND MAXIMUM p-VALUES OVER THE 15 LOCI, THE NUMBER OF LOCI WITH A p-VALUE SMALLER THAN 5%, AND THE NUMBER OF LOCI FOR WHICH THE HYPOTHESIS IS REJECTED AFTER HOLM-BONFERRONI CORRECTION.

	DNA mass in ng						
	0.008	0.016	0.031	0.047	0.063	0.125	0.25
gamma	1	0	0	0	0	0	1
log-normal	1	0	0	0	0	1	2
normal	8	15	6	0	4	0	2

TABLE III
CHI-SQUARED TEST WITH HOLM-BONFERRONI CORRECTION FOR SINGLE TRUE PEAKS: NUMBER OF LOCI FOR WHICH THE HYPOTHESIS IS REJECTED.

Since the results for the other DNA masses are different, we summarize them in Table III. With Holm-Bonferroni correction, except for the smallest and largest DNA mass, for the gamma distribution class the hypothesis is rejected for none of the loci. The log-normal distribution class gives comparable results. The normal distribution class in contrast has by far the most rejections. For example, for a DNA mass of 0.016 ng the normal distribution class is rejected for all 15 loci.

We further study the dependence of the peak height on the amount of DNA in the sample. The results are shown in Figs. 5 and 6. Both mean and standard deviation increase linearly with DNA mass. Since the standard deviation increases linearly with the DNA mass, we choose a multiplicative model for the variation in the true peak heights, as expressed by the random vector $\underline{\gamma}$ in the model (1).

C. Stutter

We use a linear non-stochastic stutter model similar to the approach in [16].

In order to characterize the amount of stutter two quantities, the stutter ratio and the stutter proportion have been defined in the literature [10]. The stutter ratio is given by $r_\mathrm{s} = h_\mathrm{s}/h_\mathrm{a}$, and the stutter proportion by $p_\mathrm{s} = h_\mathrm{s}/(h_\mathrm{s}+h_\mathrm{a})$, where h_s is the peak height of the stutter peak and h_a the peak height of the allelic peak. In this paper we work with the stutter proportion, since it reflects more naturally the fact that the DNA that accounts for the stutter peak is removed from the allelic peak.

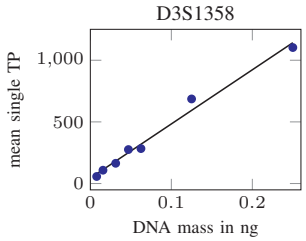

Fig. 5. Mean of the single true peaks versus the DNA mass (blue dots) and linear least squares regression line (black).

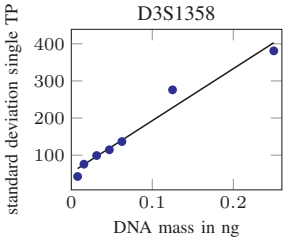

Fig. 6. Standard deviation of the single true peaks versus the DNA mass (blue dots) and linear least squares regression line (black).

If we denote by $\tilde{h}_{\mathrm{a}}$ the hypothetical allelic peak height before the stutter occurred, then h_{s} and h_{a} are given by

$$h_{\mathrm{s}} = p_{\mathrm{s}}\tilde{h}_{\mathrm{a}} \tag{2}$$

and

$$h_{\mathrm{a}} = (1 - p_{\mathrm{s}})\tilde{h}_{\mathrm{a}}, \tag{3}$$

respectively. In our model (1), the linear relationship in (2) and (3) is expressed by a multiplication with the stutter matrix S. Although this model is simple, it is widely used to describe the effects of stutter [10], [13], [16].

It has been observed that the stutter proportion is not constant within a locus. In general, it increases with increasing repeat number. In [10], [11] it was reported that the longest uninterrupted sequence (LUS) in an allele might be more appropriate than the repeat number of an allele to describe the increase of the stutter proportion. Our approach to model stutter is linear in the sense that the stutter peak height is always a fixed proportion of the hypothetical allelic peak height before stutter, according to (2). However, it is flexible in terms of modeling the dependence on the LUS. In principle, if enough data is available to determine the entries of S, the stutter model can be allele based, that is, every allele can have a different stutter proportion if necessary.

V. Conclusion

We proposed a fully quantitative signal model for forensic DNA profiles that models the variability in the allelic peak heights, stutter, and baseline noise. To test the suitability of different probability distribution classes for the noise and the true peak heights, we applied the Kolmogorov–Smirnov and the chi-squared test. Three standard classes of two-parameter distributions, the normal, gamma, and log-normal distribution, were investigated. It turned out that the Gaussian model for noise and allelic peak heights is rejected by several test, and so appears ill suited to the forensics application. Both the gamma and log-normal models, on the other hand, provide good statistical consistency with the data, and so can be employed to succinctly summarize peak-height distributions through a small number of parameters.

Acknowledgments

We would like to thank Desmond Lun and Harish Swaminathan for valuable discussions.

References

[1] J. M. Butler, *Forensic DNA Typing*, 2nd ed. Elsevier Academic Press, 2005.

[2] P. Gill, R. Puch-Solis, and J. Curran, "The low-template-DNA (stochastic) threshold—its determination relative to risk analysis for national DNA databases," *Forensic Science International: Genetics*, vol. 3, no. 2, pp. 104–111, Mar. 2009.

[3] B. Budowle, A. J. Onorato, T. F. Callaghan, A. D. Manna, A. M. Gross, R. A. Guerrieri, J. C. Luttman, and D. L. McClure, "Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework," *Journal of Forensic Sciences*, vol. 54, no. 4, pp. 810–821, Jul. 2009.

[4] R. Puch-Solis, A. Kirkham, P. Gill, J. Read, S. Watson, and D. Drew, "Practical determination of the low template DNA threshold," *Forensic Science International: Genetics*, vol. 5, no. 5, pp. 422–427, Nov. 2011.

[5] C. A. Rakay, J. Bregu, and C. M. Grgicak, "Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and locus drop-out," *Forensic Science International: Genetics*, vol. 6, no. 6, pp. 723–728, Dec. 2012.

[6] J. Bregu, D. Conklin, E. Coronado, M. Terrill, R. W. Cotton, and C. M. Grgicak, "Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis," *Journal of Forensic Sciences*, vol. 58, no. 1, pp. 120–129, Jan. 2013.

[7] Scientific Working Group on DNA Analysis Methods, "SWGDAM interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories," http://swgdam.org/Interpretation_Guidelines_January_2010.pdf, accessed: Sep. 2014.

[8] A. A. Westen, L. J. Grol, J. Harteveld, A. S. Matai, P. de Knijff, and T. Sijen, "Assessment of the stochastic threshold, back-and forward stutter filters and low template techniques for NGM," *Forensic Science International: Genetics*, vol. 6, no. 6, pp. 708–715, Dec. 2012.

[9] J. M. Butler, *Advanced Topics in Forensic DNA Typing: Interpretation*. Academic Press, 2014.

[10] C. Brookes, J.-A. Bright, S. Harbison, and J. Buckleton, "Characterising stutter in forensic STR multiplexes," *Forensic Science International: Genetics*, vol. 6, no. 1, pp. 58–63, Jan. 2012.

[11] J.-A. Bright, D. Taylor, J. M. Curran, and J. S. Buckleton, "Developing allelic and stutter peak height models for a continuous method of DNA interpretation," *Forensic Science International: Genetics*, vol. 7, no. 2, pp. 296–304, Feb. 2013.

[12] Applied Biosystems, *AmpFlSTR Identifiler Plus — User's Guide*, Life Technologies Corporation, 2012.

[13] D. Taylor, J.-A. Bright, and J. Buckleton, "The interpretation of single source and mixed DNA profiles," *Forensic Science International: Genetics*, vol. 7, no. 5, pp. 516–528, Sep. 2013.

[14] H. Swaminathan, C. M. Grgicak, M. Médard, and D. S. Lun, "NOCIt: A high-accuracy computational method for determining the number of contributors in an STR DNA profile," in *24th International Symposium on Human Identification*, Sep. 2013. [Online]. Available: https://www.promega.com/~/media/files/resources/conference%20proceedings/ishi%2024/oral%20presentations/lun-manuscript.pdf

[15] J. R. Gilder, T. E. Doom, K. Inman, and D. E. Krane, "Run-specific limits of detection and quantitation for STR-based DNA testing," *Journal of Forensic Sciences*, vol. 52, no. 1, pp. 97–101, Jan. 2007.

[16] M. W. Perlin, G. Lancia, and S.-K. Ng, "Toward fully automated genotyping: genotyping microsatellite markers by deconvolution," *The American Journal of Human Genetics*, vol. 57, no. 5, pp. 1199–1210, Nov. 1995.

[17] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[18] R. B. D'Agostino and M. A. Stephens, *Goodness-of-fit-techniques*. CRC press, 1986, vol. 68.