COMPUTATIONAL METHODS FOR THE INTERPRETATION OF

FORENSIC DNA SAMPLES

by

HARISH SWAMINATHAN

A dissertation submitted to the

Graduate School-Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

Written under the direction of

Dr. Desmond S. Lun

And approved by

———————————————————
Dr. Desmond S. Lun

———————————————————
Dr. Benedetto Piccoli

———————————————————
Dr. Catherine M. Grgicak

———————————————————
Dr. Muriel Medard

Camden, New Jersey

October 2015

ABSTRACT OF THE DISSERTATION

Computational methods for the interpretation of forensic DNA profiles

by HARISH SWAMINATHAN

Dissertation Director:
Dr. Desmond S. Lun

Interpretation of DNA profiles generated from STRs can be problematic because of dropout, allele overlap and artifacts like stutter. The goal of this research is to develop computational methods for the analysis of STR profiles that are robust to these phenomena and that utilize quantitative peak height information captured in profiles. These methods are expected to improve significantly on existing methods for analysis of STR profiles, particularly in cases of low amounts of template DNA or where there are many contributors. In the first part of our research, we characterized the distribution of signal, noise and stutter peak heights and studied their dependence on template DNA amount. For the second part of our project, we developed a method to identify the number of contributors to a DNA sample. Our method, NOCIt, calculates the *a posteriori* probability on the number of contributors to a forensic sample taking into account signal peak heights, population allele frequencies, baseline noise, allele dropout and stutter. On the experimental samples tested, NOCIt had an accuracy of 83%, while the accuracy of the best pre-existing method was 72%. The accuracies of NOCIt and the best pre-existing method on the simulated profiles were 85% and 73%, respectively. We were able to reduce the running time of NOCIt by developing a faster method based on an importance sampling algorithm. In the third and final part of our research, we developed a computational tool (MatchIt) to directly compute a continuous Likelihood Ratio (LR)

for a person of interest (POI), treating other contributors (if any) as interference. MatchIt also calculates the distribution of the LR along with the p-value, which is the probability a randomly chosen individual results in a LR at least as large as the LR obtained from the POI. We observed that the amount of template DNA from the contributor impacted the LR – small LRs resulted from contributors with low template masses. Moreover, we observed a decrease of p-values as the LR increased. A p-value of 10^{-9}, the lowest possible in our testing, was achieved in all the cases where the LR was greater than 10^8.

DEDICATION

To my parents

ACKNOWLEDGEMENT

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

DNA evidence is widely used for forensic purposes all over the world. DNA profiling technology, though fairly recent, has grown immensely in the past twenty years and reached exceedingly large sensitivity levels. Simultaneously, DNA testing has gained traction among analysts and the common man's confidence in DNA testing efficacy has grown due to its high specificity. Consequently, the combination of its high specificity and sensitivity has enabled DNA to attain precedence over other types of biological evidence like blood typing. Other than being a valuable tool in helping to place a suspect at the scene of the crime, DNA typing has other applications including parentage testing, disaster victim identification and missing persons investigations [1]. DNA evidence has become so influential that when presented in the courtroom, it can play a major part in influencing the outcome of a trial. In light of this trend, it is crucial that we employ the best interpretation practices possible when dealing with forensic DNA evidence. The goal of this research is to develop computational tools and methods that aid in DNA mixture interpretation.

1.1 Short Tandem Repeats

The usage of DNA to generate a profile of an individual and use the profile as if it were a 'unique fingerprint' to identify the individual traces its origins to 1985 when repeated sequences of DNA called *minisatellites,* or *Variable Number of Tandem Repeats* (VNTRs), were

first described by Sir Alec Jeffreys [2]. Since then, *Short Tandem Repeat* or STR testing has replaced VNTRs as the genetic marker of choice because the regions are small enough to be amplified using PCR (polymerase chain reaction). As a result, STRs have gained widespread currency because of the low limit of detection, relatively short turnaround time, ease of automation and the ability to examine multiple markers in a single run [1].

STRs are DNA regions with repeat units ranging in size from 2 to 7 base pairs. They are scattered throughout the human genome and occur on average every 10,000 nucleotides. The STRs used for forensic purposes typically contain repeat units of size 4 (tetra nucleotide repeats) and are located between genes or within the intron regions of genes and thus do not code for genetic variation [1].

At each STR *locus*, or location on the genome, every individual has two alleles - one on each chromosome - that correspond to the *genotype* of the individual (Figure 1.1). *Homozygous* individuals have identical alleles on both chromosomes. If the two alleles are different, the individual is *heterozygous* at that locus.
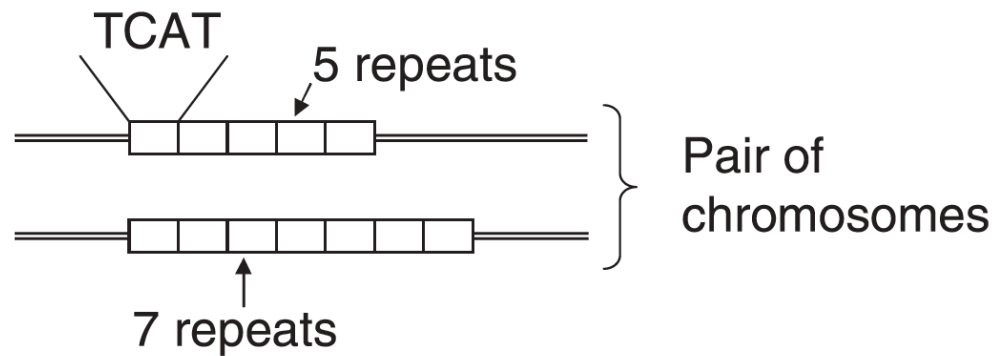
Figure 1.1: STR alleles at a locus on a homologous pair of chromosomes. Figure courtesy: David R. Paoletti, Dan E. Krane, Michael L. Raymer and Traves E. Doom: *Inferring the Number of Contributors to Mixed DNA Profiles.* **IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(1), 2012.**

The value of STRs stems from the fact that genotypes at multiple loci are combined to create the *DNA profile* of an individual. Since the STR loci are polymorphic

(each locus contains a number of different possible alleles) it is very unlikely that two random individuals would share the same genotype across all the loci. Early STR kits using 4 or 6 loci have been superseded by modern kits that work on 16 or even 24 loci, reducing the chances of an adventitious match [3].

1.2 DNA profiling

Rapid advances in DNA profiling technology have enabled the development of a DNA profile from a crime scene sample in just a few hours. A brief description of the profile generation process follows (Figure 1.2).

A biological sample at a crime scene can be in the form of blood, saliva, semen, etc. First, the sample is tested to confirm the presence and type of biological material. Then the DNA present in the sample is extracted to separate it from proteins and other cellular materials. The extracted DNA then needs to be quantified since only a narrow range of DNA concentration, typically 0.5 to 2ng, is suitable for obtaining optimal peak heights in the signal. In the next step, the STR loci of interest are amplified by PCR using fluorescent dye-labeled primers. This results in millions of copies of the alleles at the relevant markers of interest. The DNA fragments are then separated according to their size by capillary electrophoresis. When they move through the capillary, the amplicons are hit by a laser that excites the fluorescent dyes, causing them to emit light. The intensity of the light emitted is measured by a detector and is converted to an electric signal, making up the peaks in the electropherogram. The light intensity is measured in relative fluorescence unit, or RFU, and is proportional to the number of dye-labeled molecules present. An internal size standard is used to calibrate the peaks observed to

their base size based on their time of travel through the capillary. The size is then compared an allelic ladder in order to obtain the allele designation [1].

Thus, a DNA profile is generated from the sample in the form of an electropherogram, with the peaks at the various loci representing the alleles and the heights of the peaks being roughly proportional to the quantity of DNA amplified.

Sample Obtained from
Crime Scene or Paternity
Investigation

Biology

DNA
Extraction → DNA
Quantitation → PCR Amplification
of Multiple STR markers

Technology

Separation and Detection of
PCR Products
(STR Alleles) → Sample Genotype
Determination

Genetics

Comparison of Sample
Genotype to Other Sample
Results

Generation of Case Report
with Probability of Random
Match

If match occurs, comparison of
DNA profile to population
databases

Figure 1.2: Steps in DNA sample processing. Figure courtesy: John M. Butler (2009) Fundamentals of Forensic DNA Typing.

1.3 Artifacts

There are various phenomena associated with creating a DNA profile that can complicate profile interpretation. Biology and technology-related artifacts result in

spurious peaks in the signal in addition to the true alleles in the sample. Incomplete nucleotide addition (*-A*) by the *Taq* polymerase enzyme results in split peaks. Due to spectral overlap of the different colors used for labeling the DNA fragments, failure of the matrix to resolve the dye colors results in '*pull up*' or '*bleed through*' of a peak into another color. *Dye blobs* occur when the dyes get detached from their primers and migrate independently through the capillary [1].

1.3.1 Stutter

The most important artifact associated with DNA profiling is *stutter*. Unlike the previously mentioned artifacts, stutter is intractable because it happens often and is hard to distinguish from alleles. Stutter is an amplification artifact and is observed as additional peaks close to a true allele. Stutter peaks can be seen at one repeat unit less than the allele (*n-4* or *reverse stutter*) or at one repeat unit more than the allele (*n+4* or *forward stutter*). Although stutter could result in peaks more than one repeat unit shorter or longer than the parent allele, this is rare for tetranucleotide repeat sequences.

The *strand slippage model* has been postulated to explain the phenomenon of stutter (Figure 1.3) [4]. According to this model, stutter occurs during PCR when the daughter strand gets separated from the parent strand while it is being copied. When the two strands rejoin, they do not attach at the correct position − the daughter strand is either shifted forward by one repeat unit (resulting in reverse stutter) or is shifted backward by one repeat unit (resulting in forward stutter).

Stutter peaks hinder the signal interpretation process. For example, in a mixture with a major and a minor contributor, it can be hard to distinguish between the minor

contributor's alleles and stutter from the major contributor – the peak in question can be stutter, an allele or a combination of the two (Figure 1.4).

(A) Normal replication
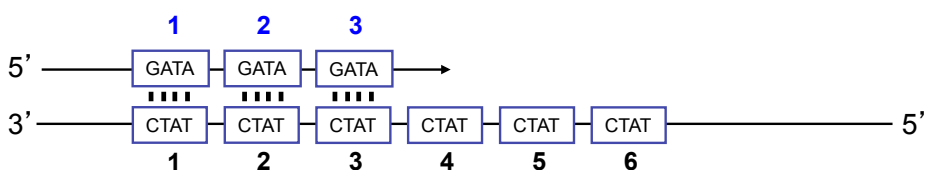
(B) Insertion caused by backward slippage
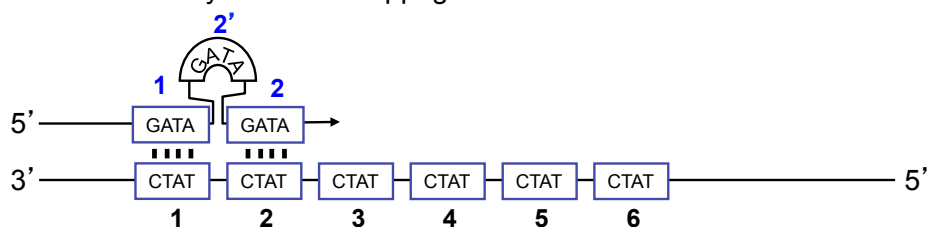
(C) Deletion caused by forward slippage
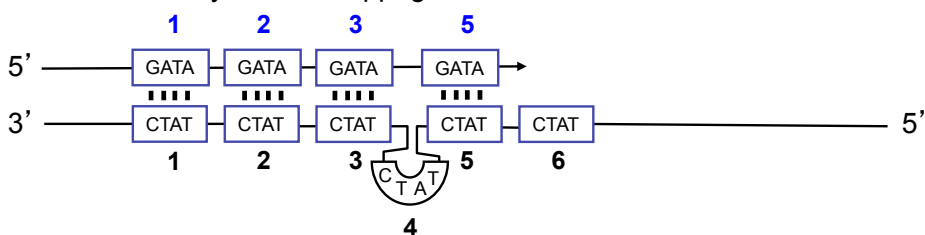
Figure 1.3: Mechanism of stutter product formation. Figure courtesy: John M. Butler (2005) Forensic DNA Typing.
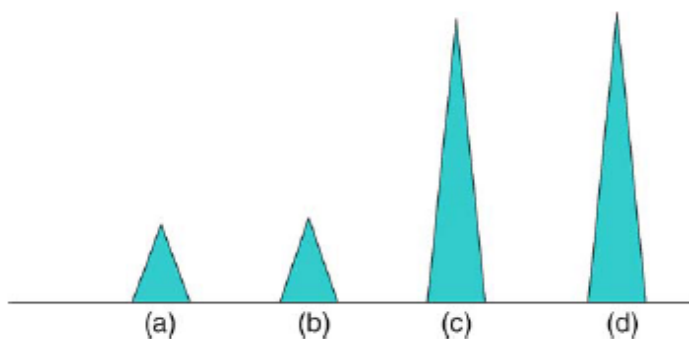
Figure 1.4: Effect of stutter. The major contributor is *cd*. The minor contributor is *ab* (if *b* is not a stutter or an allele with a stutter) or *ac, ad* or *aa* (if *b* is a stutter). Figure courtesy: P. Gill et al./Forensic Science International 160 (2006).

1.4 Low template samples

An increase in the sensitivity of STR typing technology has facilitated the development of profiles even from samples with less than ideal amounts of template DNA. Low template samples have a low signal-noise ratio and can result in profiles where the signal peaks (i.e. the allelic peaks) are ambiguous and hard to detect because of the background noise. These low template samples (containing less than 0.25ng of template DNA) present formidable challenges during interpretation because of additional complexities (Figure 1.5) [5]. As a result, interpretation protocols applied to samples with optimal DNA levels are typically not applicable to low template samples.
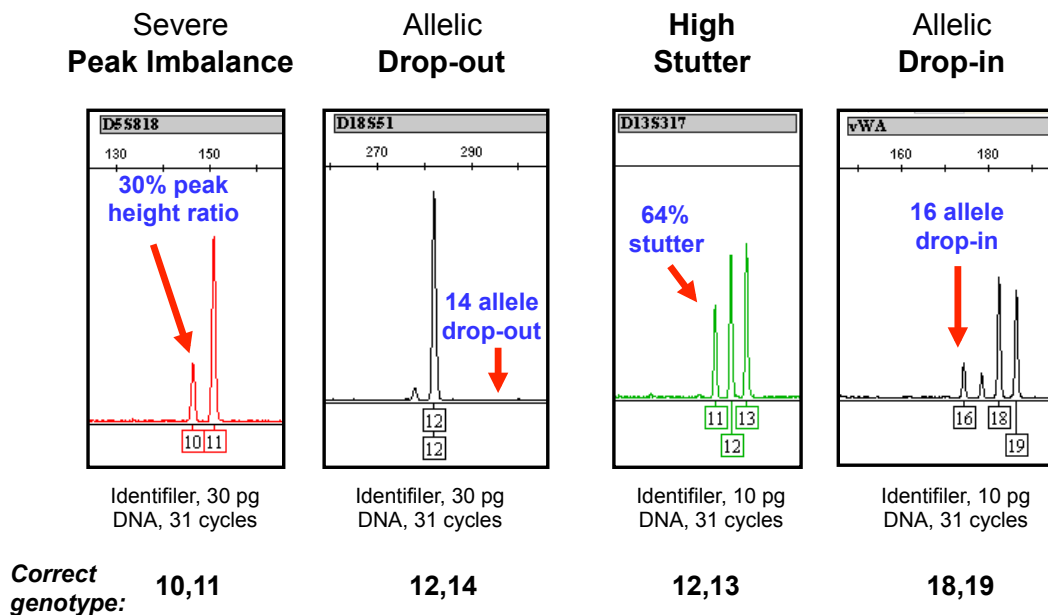
Figure 1.5: Stochastic effects in low template samples. Figure courtesy: John M. Butler (2012) Advanced Topics in Forensic DNA Typing: Methodology.

1.4.1 Elevated stutter

The first problem with low template samples is identifying stutter peaks. While typically the height of a stutter peak is less than 15% of the height of the parent peak [6], for low template samples this number could increase and make it difficult to differentiate between a true allelic peak and a stutter peak.

1.4.2 Heterozygote imbalance and dropout

The second problem associated with profiles of low template samples is *dropout*. For a heterozygous person, the two alleles at a locus show some variation in their heights due to sampling and amplification variability. *Heterozygote Balance* is a statistic for quantifying the variation in the peak heights. One way of defining heterozygote balance (HB) is:

$$HB = \frac{\phi_{smaller}}{\phi_{larger}},$$

where $\phi_{smaller}$ is the height of the smaller allele and ϕ_{larger} is the height of the larger allele. Although a general guideline of $HB \geq 0.6$ is typical [7], heterozygote imbalance in low template samples results in much lower values for HB.

Dropout refers to the phenomenon of non-detection of an allele. It can be thought of as an extreme case of heterozygote imbalance and occurs due to stochastic effects associated with DNA extraction, the PCR process and pipetting.

1.4.3 Drop-in

Another phenomenon that can complicate the interpretation of DNA samples is *drop-in*, which refers to additional alleles in the signal resulting from sporadic contamination.

1.5 Mixtures

In some cases, the sample collected from the crime scene is a mixture containing DNA from two or more individuals. Mixtures are typically identified by the presence of more than two peaks at one or more loci and/or vast differences in the heights of the peaks observed at a locus, although events like trisomy, shared alleles, stutter and dropout can cause errors in this step. If mixture interpretation is carried out by deconvolving the profile into the constituent genotypes, allele sharing between the various contributors to the sample makes it difficult to assign genotypes to the contributors with certainty. Moreover, if one of the contributors is a major, the minor contributor's alleles could be masked by the major contributor's alleles and/or be indistinguishable from stutter from the major (Figure 1.6. Also see Figure 1.4.) [8]. All this hinders the mixture interpretation step.
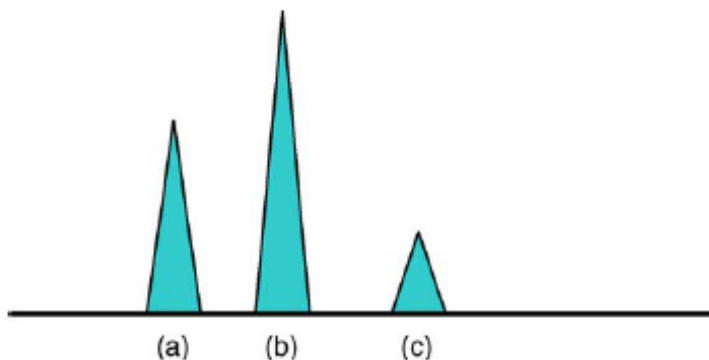
Figure 1.6: A three-allele profile showing masking. The major component is *ab*. The minor component is *bc*, but other combinations such as *ac* and *bb* (with *c* being stutter) are reasonable. Figure courtesy: P. Gill et al./Forensic Science International 160 (2006).

1.6 Match statistic

The SWGDAM interpretation guidelines published in 2010 [9] require laboratories to perform a statistical analysis in support of any inclusion. Calculation of a DNA match

statistic can take one of three approaches: *Random Match Probability, Random Man Not Excluded* and the *Likelihood Ratio.* All three approaches use the frequencies of the alleles observed in the population of interest and are based upon two underlying assumptions about the population genetics model:

➢ *Hardy-Weinberg equilibrium:* the two alleles inherited at a locus from an individual's parents are independent of each other [1].

➢ *Linkage equilibrium:* alleles at different loci are independent of each other [1].

1.6.1 Random Match Probability

The Random Match Probability (RMP) is the frequency with which a particular STR profile would be expected to occur in a population. It can also be thought of as the probability that a randomly picked person from the population would have the STR profile in question.

The match probability is calculated at a locus using the formula p^2 for a homozygous locus and $2pq$ for a heterozygous locus, where p and q are the frequencies for the alleles observed. Due to population substructure, the observed frequency of homozygotes is slightly higher than that predicted by Hardy-Weinberg equilibrium [10]. A correction for population substructure can be carried out by introducing the estimate of co-ancestry θ and calculating the probability as $p^2 + (1-p)\theta$ for a homozygous locus. Once the match probabilities are calculated at all the loci, the individual locus probabilities are multiplied to give the overall RMP for the entire profile.

The RMP is typically calculated for single source samples. It can also be applied to two-person mixtures when the profile of the major (or possibly even the minor)

contributor can be deduced by using a heterozygous balance threshold and an estimate of the mixture ratio, in which case it is called the modified RMP (mRMP). As such, the RMP depends the number of contributors to the sample that is assumed.

1.6.2 Random Man Not Excluded

The Random Man Not Excluded (RMNE) approach is based on the calculation of the CPE (Combined Probability of Exclusion) statistic or its complementary statistic CPI (Combined Probability of Inclusion). The RMNE method seeks to calculate the fraction of the population that would be excluded as a contributor to the sample. It assumes that all the alleles of the suspect are present in the profile, i.e. there is no dropout.

In order to compute the CPE, the Probability of Inclusion (PI) is calculated at every locus by squaring the sum of the frequencies of the alleles observed. The product of the individual locus PI values is the CPI. Then the CPE is calculated as: $CPE = 1 - CPI$.

The RMNE approach has been criticized for its wastage of information in the calculation phase – the peak heights are not utilized and hence all genotypes are given an equal weight. Furthermore, the CPI statistic does not depend upon the genotype of the suspect or on an assumption about the number of contributors to the sample [11].

1.6.3 Likelihood Ratio

The Likelihood Ratio (LR) is the ratio of the probability of observing the evidence under two mutually exclusive hypotheses and informs how much more likely one hypothesis is compared to the other. The LR is defined as:

$$LR = \frac{\Pr(E|H_p, n_p)}{\Pr(E|H_d, n_d)},$$

where E is the evidence in the form of the electropherogram (epg); H_p and H_d are the hypotheses specified by the prosecution and the defense, respectively; and n_p and n_d are the number of contributors specified by the prosecution and the defense, respectively. The numerator is the probability of observing the evidence given the prosecution's hypothesis and the denominator is the probability of observing the evidence given the defense's hypothesis. The evidence shows support for the prosecution's hypotheses if LR > 1; if LR < 1 the defense's hypothesis is supported by the evidence. Specific ways to calculate the probabilities are discussed in Chapter 4.

Unlike the RMNE method, the LR can use information like the number of contributors to the sample, the heights of the peaks observed and the genotype of the suspect [11]. Due to the fact that the RMNE method is restricted to unambiguous profiles, and stutter and dropout can be determined probabilistically while calculating the LR, the LR approach is the recommended method for performing mixture interpretation [8].

1.7 Statement of the problem

The mixture analysis pipeline currently employed in forensic laboratories around the world involve identification of the number of contributors to the sample in question, assessing whether to include a person of interest as a contributor to the sample, followed by the calculation of a match statistic (for example the Likelihood Ratio) in the case of an inclusion.

Along with the qualitative information in the signal, there is also quantitative information in the form of the peak heights. Currently, the heights of the peaks are used

while applying the analytical, stochastic and stutter thresholds [10]. The analytical threshold is used to isolate signal from noise peaks. The stochastic threshold is the threshold above which it is reasonable to assume that dropout of a sister allele hasn't occurred. Peaks in the stutter position below the stutter threshold are considered to be exclusively stutter and removed from further analyses. The application of these thresholds entails a binary view of the process – peaks are either signal or noise; dropout and stutter either have or have not occurred. This binary 'black or white' approach is not representative of reality where events are 'grey' and take place in a continuum. Hence they need to be assessed probabilistically to describe them more accurately. Peak heights are also utilized while calculating the 'restricted' Likelihood Ratio or Random Match Probability (RMP), using a predetermined heterozygote balance threshold and an estimate of the mixture ratio from the peak heights [9]. This approach is also binary and hence not very informative since it assigns values of 0 or 1 to genotypes depending on whether they pass or fail a set of criteria. Moreover for low template samples, the binary approach is problematic because of the issues mentioned previously - reduced peak heights, elevated stutters ratios and reduced peak height ratios. Therefore, the peak height distributions and their variation with input DNA need to be studied in order to optimize mixture interpretation.

An assumption about the number of contributors is needed to calculate a match statistic like the LR. The method commonly used for identifying the number of contributors in U.S. crime laboratories currently is Maximum Allele Count (MAC). This method gives the *minimum number of contributors* needed to explain the profile obtained. The actual number of contributors to the sample might in fact be different because of allele-sharing among the various individuals. In addition, the estimate from MAC may be

inaccurate because of stutter and dropout [12]. Hence, a method to determine the number of contributors that uses the entire information contained in the electropherogram is necessary, especially when working with low template samples.

After the number of contributors to a DNA profile has been identified, mixture interpretation is carried out by deconvolving samples with less than 3 contributors into the constituent genotypes. This step is error prone because dropout and stutter hinder the deconvolution step. Samples containing 3 or more contributors are typically assessed using an allele-centric method, whereby the alleles in the suspect's standard are compared against the observed alleles encountered in the evidence samples. However, this method is problematic, particularly for low-template mixtures, as it becomes increasingly more difficult to ascertain whether all alleles from every contributor have been detected.

Due to the uncertainty associated with them, many complex mixtures are not interpreted and go unused. Even if complex mixtures are chosen for interpretation, not accounting for the peak heights and artifacts would likely lead to misinterpretation and misapplication of the RMNE method. All this makes it important to develop a technique to compute match statistics for a person of interest that incorporates peak heights, artifacts and possible interference from other contributors into the calculation.

1.8 Objectives

The goal of this research is to develop computational tools and methods that aid in mixture interpretation. Mixture interpretation is not a simple process and has some inherent complexity associated with it. While analyzing an evidence profile, ignoring the quantitative information captured in the peak heights in the signal is detrimental to the interpretation of the profile. Hence while analyzing a DNA profile, the quantitative

information obtained, as well as the qualitative information, needs to be taken into account.

Our first aim (Aim 1) is to characterize the distribution of the heights of the different types of peaks observed in the signal - allelic peaks, baseline noise peaks and stutter peaks (reverse and forward) - and how the peak heights vary with input DNA. We will also characterize the rate of dropout and the frequency of observance of stutter and noise peaks.

One piece of information needed while performing mixture interpretation is the number of contributors to the sample. The number of contributors to the sample in question is in most cases unknown, and current probabilistic systems require the analyst interpreting the profile to make an assumption about the number. One of the aims of this research (Aim 2) is to develop a method to identify the number of contributors to a forensic evidence sample that is more accurate than methods currently in use. Unlike existing methods to determine the number of contributors, we plan to include the quantitative peak height information and artifacts such as dropout and stutter in the calculation of the number of contributors.

The ultimate objective of mixture interpretation is to determine whether a person of interest is a contributor to the sample. The final aim of this research (Aim 3) is to develop a method to calculate the Likelihood Ratio (LR) for a person of interest in a continuous manner and to estimate a statistical significance for the LR. Currently, the LR is computed by examining possible combinations of genotypes and using the allele frequencies in the calculation. However, we will not employ deconvolution but directly estimate the probability of the person of interest contributing to the profile, using the quantitative information. Along with the LR, we also plan to compute its distribution

conditioned on the defense hypothesis and to calculate a p-value as a summary statistic of the distribution. *The p-value is the probability that a randomly picked person from the background population has an LR at least as large as the one calculated for the person of interest. The p-value can be used to control the false positive rate of a binary classification system. For example, if suspects whose p-value is below 'x' are classified as contributors (and non-contributors otherwise), the false positive rate of the resulting system is x.

To summarize, in the proposed thesis research, we will address the shortcomings of existing mixture interpretation approaches. Specifically, we will address how to:

➢ Quantitatively model the peak heights and artifacts associated with STR profiling, including dropout and stutter.

➢ Determine the number of contributors while accounting for quantitative peak height information and artifacts.

➢ Calculate match statistics (LR and p-value) for a person of interest that account for quantitative peak height information, artifacts and interference from other contributors.

CHAPTER 2

CHARACTERIZATION OF PEAK HEIGHTS

In this chapter, we describe the research we have performed on modeling the peak heights – specifically characterizing the peak height distributions for the various types of peaks and the variation of the peak heights with input DNA mass.

2.1 Background

The heights of the allelic peaks in a DNA profile increase approximately linearly with the amount of input DNA [13] and thus are an indicator of the mixture ratio and the number of copies of an allele that gave rise to a peak. This information can be valuable while performing mixture interpretation. The traditional binary mixture interpretation approach does not fully utilize the peak height information contained in the signal and would not be suited for samples with low input levels of target DNA that exhibit decreased peak heights, elevated stutter levels and reduced peak height ratios [3]. This makes it important to have a good understanding of the features of peak heights in order to optimize mixture interpretation when dealing with such samples. There is an urgent need for this type of research as mixture interpretation moves into the realm of probabilistic modeling.

2.2 Materials and Methods

All the samples used in this study were kindly provided by the Boston University School of Medicine.

2.2.1 Preparation of samples

To generate the samples used for analysis of peak heights (Table 2.1), high molecular weight DNA was extracted from 27 single source samples using standard organic extraction procedures. The samples were whole blood, dried blood stains or saliva. The blood stains were either on Whatman® paper or cloth swatches. Saliva samples were either whole saliva or dried buccal swabs on cotton.

Injection time (s)	DNA amount (ng)	Number of samples
10	0.008	48
	0.016	45
	0.031	46
	0.047	27
	0.063	42
	0.125	48
	0.25	48
	0.5	9
	1	0
	Total	313

Table 2.1: Single source samples used for analyses of peak heights.

Briefly, the organic extraction consisted of incubating the sample in 300 µg/mL of Proteinase K and 2% v/v SDS (sodium dodecyl sulfate) solution at 37°C for 2 hours to overnight.   Purification was accomplished with phenol/chloroform and alcohol precipitation.   The DNA was dissolved in 50 µl of TE buffer (10 mM Tris, 0.1 mM

EDTA, pH 8.0) at 56°C for 1 hour. Absolute DNA quantification was performed using real-time PCR and the Quantifiler® Duo™ Quantification kit according to the manufacturer's recommended protocol and one external calibration curve [14, 15]. A 7500 Sequence Detection System (Life Technologies, Inc.) was used for C_t (cycle threshold) detection.

The extracted DNA was amplified using the manufacturer's recommended protocol (29 cycles) for AmpFℓSTR® Identifiler® Plus Amplification Kit (Life Technologies, Inc) [16]. Single source samples were amplified using 0.5, 0.25, 0.125, 0.063, 0.047, 0.031, 0.016 and 0.008 ng of DNA by diluting the extracted DNA based on the results of the quantification. The PCR reaction consisted of 15 µL of master mix, the calculated volume of template DNA based on target mass required, and enough Tris-EDTA (TE) buffer (10 mM at pH 8.0) to bring the total reaction volume to 25 µL.

Amplification was performed on Applied Biosystems' GeneAmp® PCR System 9700 using 9600 emulation mode. Positive and negative amplification controls were also run and showed expected results (data not shown). Fragment separation was accomplished by using a 3130 Genetic Analyzer (Life Technologies, Inc.) and a mixture containing appropriate amounts of HiDi (highly-deionized) formamide (8.7 μl/sample) (Life Technologies, Inc.) and GeneScan™-600 LIZ™ Size Standard (0.3 µL/sample) (Life Technologies, Inc.). A volume of 9 µL of that mixture and 1 µL of sample, negative or ladder was added to the appropriate wells. The samples were incubated at 95°C for 3 minutes and snap-cooled at -20°C for 3 minutes. Ten-second injections at 3 kV were performed on each of the samples and run according to the manufacturer's recommended protocol [16].

Fragment analysis was performed using GeneMapper IDX v1.1.1 (Life Technologies, Inc.) using Local Southern sizing and an RFU threshold of 1. A threshold of 1 RFU was used in order to capture all peak height information, i.e. the allelic peaks, baseline noise and stutter peaks, in the signal. Known artifacts such as pull-up, spikes, -A, and artifacts due to dye dissociation were manually removed. A peak was considered pull-up if it was the same size (+/- 0.3 bp) as a larger peak in another color and below 5% of the height of the larger peak. Peaks were determined to be 'spikes' if they were in greater than 2 colors and in the same position. Peaks were considered to be −A if they were one base pair smaller than an allele and peaks determined to originate from dye dissociation had to be in the same position, in the same color channel and be observed in multiple samples. The Genotypes Table, which included the File Name, Marker, Dye, Allele, Size and Height, was exported.

2.2.2 Classification of peaks

All the peaks in the profiles, including those that were not detected (i.e. peaks with a height less than 1 RFU), were separated into one of four categories: true peaks (peaks from alleles present in the contributor to the sample), reverse stutter peaks (peaks in the $n-4$ position of true peaks), forward stutter peaks (peaks in the $n+4$ position of true peaks) and noise peaks (all the other peaks, which had alleles in the frequency table or in the $n-4$ or $n+4$ position of alleles in the frequency table).

2.2.3 Modeling of peak heights

With homozygous samples, it is not always possible to say with certainty whether both alleles have been amplified or whether one of the alleles has dropped out. Hence, only heterozygous loci were used for modeling the peak heights.

Since the height of the stutter peak is dependent upon the height of the parent allele [17], modeling of the stutter peaks was done using the stutter ratio (r):

$$r = \frac{h_s}{h_a},$$

where h_s is the height of the stutter peak and h_a is the height of the allelic peak causing stutter.

The distribution of the peak heights in the four categories were tested using the Kolmogorov-Smirnov (K-S) test in MATLAB® (2014b, The Mathworks, Natick, Massachusetts). The K-S test is a test for the equality of continuous probability distributions that can be used to compare a sample with a reference probability distribution [18]. Only the peaks that were detected were used in the K-S test, i.e. all the peaks that had a height of at least 1 RFU. All loci except AMEL were used for testing the peak height distribution. Each of the 15 loci was tested at 8 DNA masses ranging from $0.008 - 0.5$ng, resulting in a total of $15 \times 8 = 120$ tests. Three continuous distributions were used as possible models for the peak heights – normal, lognormal and gamma.

In order to use the information captured in the peak heights, it is also necessary to understand how the heights of the peaks vary with input DNA. Hence, various models were tested to characterize the variation of allele, noise and stutter peak heights as a function of target DNA amount. In addition, the probability of dropout and the probability of occurrence of stutter and noise were also modeled as a function of target DNA.

Frequency of dropout of allele, stutter and noise peaks was calculated empirically by counting the number of instances in which a peak was not detected and dividing it by the total number of observations. To model the peak heights as a function of DNA mass, the observed value of a variable (such as mean and standard deviation of true peak heights or frequency of dropout) was calculated at all 8 DNA amounts for all the variables in our study. Following that, we performed curve fitting in MATLAB® using various models and chose the model that gave the best fit to the observed values based on the R square metric.

2.3 Results

2.3.1 Peak height distributions

We observed that in all the four categories (true, noise, reverse stutter and forward stutter), and at all 8 DNA masses, the normal distribution could not be rejected at many loci (using a p-value threshold of 0.05), while there were fewer rejections with the lognormal and the gamma distributions. For example, with the true peak heights, we observed that in 11 out of the 120 cases the normal distribution was rejected (lowest p-value: 0.005, highest p-value: 0.992); while the lognormal distribution had 4 rejections (lowest p-value: 0.019, highest p-value: 0.992) and the gamma distribution had 0 rejections (lowest p-value: 0.089, highest p-value: 0.999). The normal distribution had the most rejections for all four types of peaks. For the true peaks, the gamma distribution had the fewest number of rejections. For the other three types of peaks (noise, reverse and forward stutter), the lognormal distribution had the fewest rejections. The number of rejections for all three distributions was highest with the noise peaks (48, 27 and 32 for normal, lognormal and gamma, respectively). The difference between the number of

rejections for the lognormal and gamma distributions was small in all four categories. See

Tables 2.1-2.4 for the complete results of the K-S test on the different types of peaks.

	Normal	Lognormal	Gamma
Min p-value	0.005	0.019	0.089
Max p-value	0.992	0.992	0.999
# rejected	11	4	0

Table 2.2: K-S test results for true peak heights

	Normal	Lognormal	Gamma
Min p-value	0	0	0
Max p-value	0.988	0.964	0.992
# rejected	48	27	32

Table 2.3: K-S test results for noise peak heights

	Normal	Lognormal	Gamma
Min p-value	0.00004	0.03	0.006
Max p-value	0.998	0.99	0.998
# rejected	16	1	3

Table 2.4: K-S test results for reverse stutter peak heights.

	Normal	Lognormal	Gamma
Min p-value	0.001	0.13	0.031
Max p-value	0.988	0.989	0.999
# rejected	3	0	2

Table 2.5: K-S test results for forward stutter peak heights.

Mönich et al performed a G-test on the noise peak heights in which they compared the normal, lognormal and gamma models and obtained similar results − the lognormal distribution was the best fit for the data, while the normal distribution had the most number of rejections [19].

2.3.2 Variation of peak heights and frequency of detection of peaks with input DNA mass

Following our study on the distribution of the peak heights, we developed models to describe the variation of all the variables in our study with respect to DNA mass. For every variable, we tested various models by performing curve fitting in MATLAB® and picked the model that gave the best fit to the observed data. For example, a linear model fit the data well for mean of true peak heights (Figure 2.1), while an exponential model had the best fit for the dropout frequency (Figure 2.2). See Table 2.5 for a list of the models that gave the best fit for each variable.
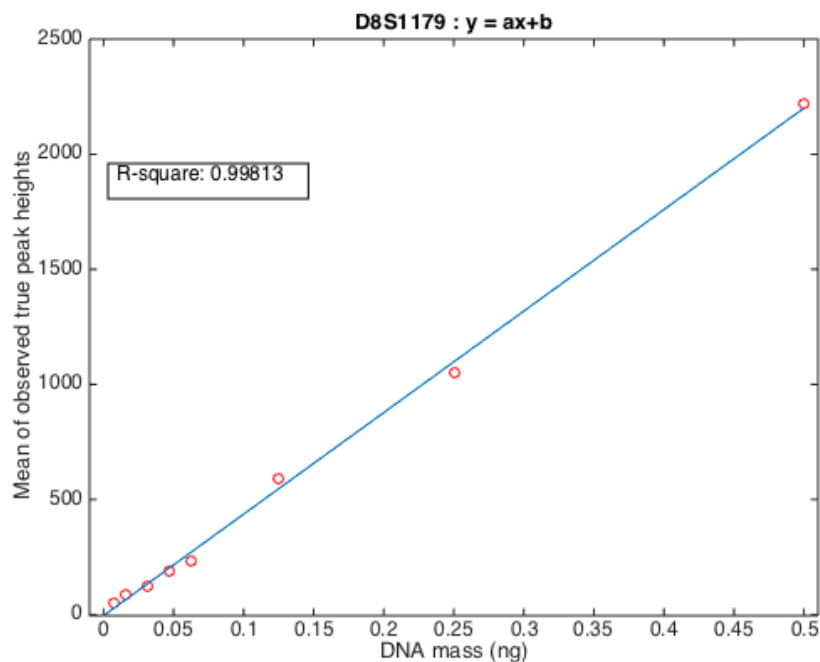
Figure 2.1: Curve fitting in MATLAB for the mean of true peak heights at the locus D8S1179. Using the model y = ax+b had an R-square value of 0.998 and gave the best fit.
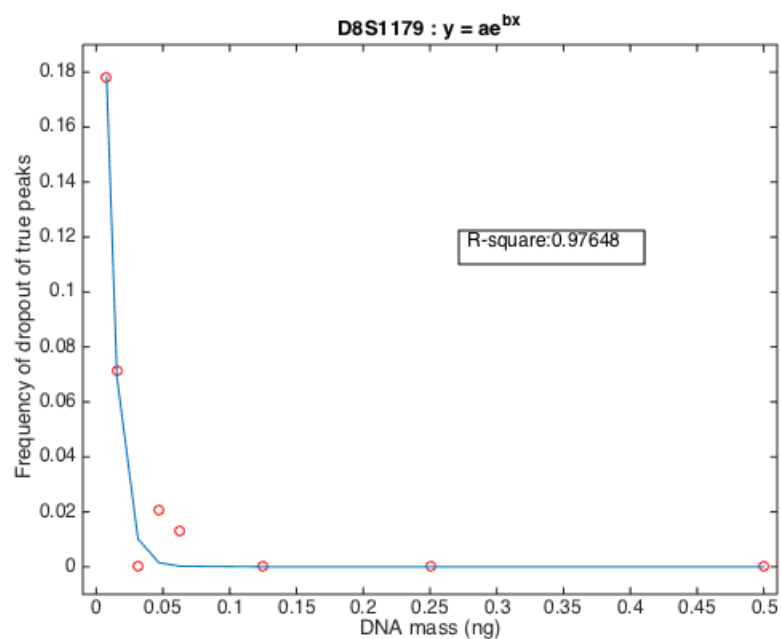
Figure 2.2: Curve fitting in MATLAB for the dropout frequency of true peaks at the locus D8S1179. Using the model y = ae^(bx) had an R-square value of 0.976 and gave the best fit.

Variable	Model description	Model function
Dropout rate of alleles (α)	Exponentially decreasing curve	$\alpha = ae^{bx}$, where x is the DNA mass from the contributor with the allele
Rate of non-detection of stutter (β)	Exponentially decreasing curve	$\beta = ae^{bx}$, where x is the DNA mass in the parent allele that gives rise to stutter
Rate of non-detection of noise (γ)	Constant	Estimated from calibration data
Mean of true peak heights (μ_t)	Line with a positive slope	$\mu_t = ax + b$, where x is the DNA mass in the true peak
Standard deviation of true peak heights (σ_t)	Line with a positive slope	$\sigma_t = ax + b$, where x is the DNA mass in the true peak
Mean of noise peak heights (μ_n)	Line with a positive slope	$\mu_n = ax + b$, where x is the DNA mass that the sample was amplified with
Standard deviation of noise peak heights (σ_n)	Line with a positive slope	$\sigma_n = ax + b$, where x is the DNA mass that the sample was amplified with
Mean of stutter ratios (μ_s)	Exponentially decreasing curve	$\mu_s = ae^{bx} + c$, where x is the DNA mass in the parent allele that gives rise to stutter
Standard deviation of stutter ratios (σ_s)	Exponentially decreasing curve	$\sigma_s = ae^{bx} + c$, where x is the DNA mass in the parent allele that gives rise to stutter

Table 2.6: Models used to describe effect of changes in DNA mass on the variables.

2.4 Discussion

Not utilizing the quantitative information in the signal, i.e. the peak heights, can be detrimental to mixture interpretation, especially in the case of low template samples where it is hard to distinguish between signal and noise peaks. In order to make use of the information contained in the peak heights, it is essential to understand their behavior – what their underlying distribution is and how they are impacted by changes in the input DNA. Our investigation into the peak heights generated from single source samples with known genotypes amplified from a range of low template DNA amounts allows us to answer these questions. Knowledge of the distributions of the heights of the peaks and the

effect of varying the input DNA mass on the peak heights is critical as the forensic community keeps moving towards a probabilistic approach. This knowledge enables us to tackle two crucial problems of mixture interpretation - determining the number of contributors to a sample and calculating a match statistic for a person of interest, the next two aims of our research.

CHAPTER 3

INFERRING THE NUMBER OF CONTRIBUTORS TO A FORENSIC DNA SAMPLE

In this chapter, we discuss our work on NOC*It* (*NOC*: Number Of Contributors) - a computational tool that calculates the *a posteriori* probability (APP) on the number of contributors to a DNA sample.

Existing methods to identify the number of contributors to a forensic DNA sample work on the number of peaks observed and/or allele frequencies. In addition to using the qualitative information contained in the signal, i.e. the allele frequencies, NOC*It* also makes use of the quantitative information, i.e. the heights of the peaks. The peak heights increase with an increase in the amount of input DNA and are an indicator of the mixture ratio and the number of copies of an allele that gave rise to a peak. This is information that could be used in estimating the number of individuals that gave rise to a sample. In addition, NOC*It* accounts for the dropout of alleles and the formation of stutter peaks and baseline noise. NOC*It* works on single source calibration data consisting of known genotypes to compute the APP for an unknown sample.

The method was tested on 278 experimental samples and 40 simulated mixtures consisting of one to five contributors with total DNA mass from 0.016 to 0.25ng. NOCIt correctly identified the number of contributors in 83% of the experimental samples and in 85% of the simulated mixtures, while the accuracy of the best pre-existing method to

determine the number of contributors was 72% for the experimental samples and 73% for the simulated mixtures. Moreover, NOC*It* calculated the APP for the true number of contributors to be at least 1% in 95% of the experimental samples and in all the simulated mixtures.

3.1 Background

The Scientific Working Group on DNA Analysis Methods (SWGDAM) recommends that forensic reports include a statement as to the assumption made about the number, or the minimum number of contributors, to the sample being investigated [9]. The number of contributors to a crime scene sample is generally unknown and must be estimated by the analyst based on the electropherogram obtained. The assumption on the number of contributors affects statistics used to assess the weight of DNA evidence (e.g., the Likelihood Ratio). Benschop et al [20] and Bright et al [21] found that using an incorrect number of contributors can lead to misclassification of a true contributor to a sample as a non-contributor and also increase the chances of an adventitious match. Thus, it is useful to have a good estimate on the number of contributors to the sample.

There are issues associated with the process of generating a DNA profile that hinder the interpretation of a profile. Stochastic effects associated with DNA extraction, the PCR process and pipetting lead to non-detection of alleles (dropout). Further, allele sharing and PCR amplification artifacts like stutter occur frequently and make it difficult to interpret low-template, mixture profiles [22]. These make it difficult to accurately estimate the number of contributors to a sample.

Methods have previously been developed to infer the number of contributors to a DNA sample. The most widely used method is Maximum Allele Count (MAC). This

method seeks to identify the minimum number of individuals who could have contributed to a sample by counting the number of alleles observed at each locus, taking the maximum value over all the loci and dividing it by two. The MAC method may not work well with complex mixtures because of allele-sharing among the contributors [23]. Guidelines for estimating the number of contributors using the total number of alleles observed were established for high template and low template samples by Perez et al [24]. Methods that do not solely rely upon the number of alleles observed but also use the frequencies of the alleles observed in the population, have been created. A probabilistic approach was developed by Biedermann et al [25], employing a Bayesian network, to infer the number of contributors to forensic samples. This method was shown to work better than MAC with degraded DNA and with higher number of contributors. Haned et al [26] extended the work of Egeland et al [27] on diallelic markers to the multi-allelic markers that are commonly used in creating STR profiles to develop a Maximum Likelihood Estimator (MLE) for the number of contributors, taking into account population substructure. This method was also shown to give more accurate results than MAC with higher number of contributors and degraded DNA. A Probabilistic Mixture Model was used by Paoletti et al [28] to infer the number of contributors to a sample based on the frequencies of the alleles observed. This method, like MLE, accounts for correction due to population substructure.

The MAC method fails with complex samples containing three or more contributors because of allele-sharing. Moreover, MAC and the other methods outlined above do not use the peak heights in the signal. Additionally, since they do not incorporate stochastic effects like dropout and artifacts like stutter in their calculation, these methods are not suitable for analyzing low template samples. We have developed

NOC*It* (*NOC*: Number Of Contributors) —a computational tool that calculates the *a posteriori* probability (APP) on the number of contributors to a DNA sample. In addition to using the qualitative information contained in the signal, i.e. the allele frequencies, NOC*It* also makes use of the quantitative information present, i.e. the heights of the peaks. In addition, NOC*It* accounts for the dropout of alleles and the formation of stutter peaks.

3.2   Materials and methods

All the experimental samples described in this section were prepared using the protocol outlined in Section 2.2.1 in Chapter 2.

3.2.1   Calibration of NOCIt

NOCIt uses the quantitative information contained in the signal in the form of peak heights to calculate the probabilities for the number of contributors. This involves characterizing the peak height distribution and the dependence of variables such as probability of dropout, probability of stutter and the peak heights on the input DNA mass. This is done by using single source calibration profiles with known genotypes obtained from samples amplified from a wide range of input DNA masses. Information on the 35 single source samples used to calibrate NOCIt is provided in Table 3.1.

Injection time(s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
	0.008	35		0.008	56		0.008	35
	0.016	35		0.016	53		0.016	35
	0.031	36		0.031	54		0.031	36
5	0.047	32	10	0.047	33	20	0.047	33
	0.063	34		0.063	49		0.063	35
	0.125	35		0.125	53		0.125	35
	0.25	35		0.25	59		0.25	33
	Total	242		Total	357		Total	242

Table 3.1: Samples used for calibration of NOCIt

3.2.2 Modeling of variables

In order to compute the probability of observing the heights of the peaks in the signal, the peak heights were modeled using the Gaussian distribution, a simple distribution that is easy-to-use. This decision was supported by the statistical tests we performed on the peak heights, which were described in Section 2.2.3 in Chapter 2. The different types of peaks (true, noise, forward and reverse stutter) were separated according to their template DNA mass and locus. Based on the Kolmogorov-Smirnov test, no significant difference was found between the Gaussian distribution and the distribution of the peak heights at many loci, indicating that the Gaussian distribution is a good approximation of the distribution of the peak heights. We chose the Gaussian distribution to model the peak heights even though the lognormal and gamma distributions had fewer rejections. The reason for this is that the Gaussian distribution allows one to easily calculate the distribution resulting from a superposition of peaks (such as allele+stutter, allele+allele, etc.), while the other distributions do not. Using this assumption, for a given mass of DNA, the peak heights could be described using their mean and standard deviation computed using the detected peaks in the calibration samples.

In order to use the data from the calibration set, all the variables in our study had to be modeled as a function of DNA mass. The functions used to model the variables were selected after testing various models as explained in Chapter 2. The models used by NOC*It* are given in Table 2.5. The parameters (*a*, and *c*) for the various models were computed at each locus by performing curve fitting MATLAB®.

3.2.3 NOC*It* algorithm

NOC*It* calculates the *a posteriori* probability (APP) on the number of contributors N given a particular evidence sample (electrophoresis profile) E. That is, it calculates $\Pr(N = n|E)$ for $n = 1,2,3, ...$ By Bayes' rule [29], we obtain:

$$\Pr(N = n|E) = \frac{\Pr(E|N = n)\Pr(N = n)}{\Pr(E)},$$

for $n = 1, ..., n_{\max}$. We assume that *a priori* N is uniformly distributed between 1 and $n_{\max}$, the maximum possible number of contributors. Since $\Pr(E)$ is the same for all n, we get: $\Pr(N = n|E) \propto \Pr(E|N = n)$. We calculate the APP according to:

$$\Pr(N = n|E) = \frac{\Pr(E|N = n)}{\sum_{n=1}^{n_{\max}} \Pr(E|N = n)}.$$

Let $\Delta^{n-1} = \{(\Theta_1, ..., \Theta_n) \in \mathbb{R}^n | \sum_{x=1}^{n} \Theta_x = 1, \Theta_x > 0 \; \forall \, x\}$ be the unit $n-1$ simplex and $\boldsymbol{\Theta}$ be the vector with components Θ_x, the mixture proportion of each contributor $x \in \{1, ..., n_{max}\}$. We generate samples $\boldsymbol{\theta^i}$ from $p_{\boldsymbol{\Theta}}$, the probability density function of $\boldsymbol{\Theta}$, which we assume to be uniform over Δ^{n-1}. Let the number of samples used be Y. We have:

$$\Pr(E|N = n) \cong \frac{\sum_{i=1}^{Y} \Pr(E|\boldsymbol{\Theta} = \boldsymbol{\theta^i}, N = n)}{Y}.$$

Let $\boldsymbol{L}$ be the set of all loci in the evidence sample and E_l be the evidence at locus l. Since the STR loci used for forensic analysis are in linkage equilibrium and independent of each other [1], we get:

$$\Pr(E|\boldsymbol{\Theta} = \boldsymbol{\theta^i}, N = n) = \prod_{l \in \boldsymbol{L}} \Pr(E_l|\boldsymbol{\Theta} = \boldsymbol{\theta^i}, N = n).$$

We implement a Monte-Carlo sampling algorithm, generating genotype samples based on the allele frequency distribution. Let Z be the number of genotype samples used and $\boldsymbol{g^j}$ be the jth genotype sample.

$$\Pr(E_l|\boldsymbol{\Theta} = \boldsymbol{\theta^i}, N = n) \cong \frac{\sum_{j=1}^{Z} \Pr(E_l|\boldsymbol{G} = \boldsymbol{g^j}, \boldsymbol{\Theta} = \boldsymbol{\theta^i}, N = n)}{Z},$$

where $\boldsymbol{G}$ is the vector with components G_x, the genotype of each contributor $x \in \{1, \dots, n_{max}\}$.

We have regularly made updates to the way we calculate $\Pr(E_l|\boldsymbol{G} = \boldsymbol{g^j}, \boldsymbol{\Theta} = \boldsymbol{\theta^i}, N = n)$, which is the probability of observing the evidence at a locus, given the mixture ratio and the genotypes of the contributors. A description of our latest model to compute the probability is given in the Appendix. The following is a description of the calculation of the probability used in [12].

For every allele in the genotype of all the contributors, dropout of the allele and the formation of reverse and forward stutter from that allele are simulated by a Bernoulli trial using the probabilities derived from the calibration samples. In the case of homozygous contributors, dropout and stutter are simulated for both the alleles in the contributor. Two assumptions are made with regard to dropout and stutter:

      1.   Dropout and stutter of one allele of a contributor are independent of dropout and stutter of the contributor's other allele.

2.   Dropout and stutter of an allele from a contributor are independent of dropout and stutter of the same allele from another contributor.

Once the alleles that have dropped out and the alleles that given rise to stutter have been established, we compute $\Pr\left(E_l | G = g^j, \Theta = \theta^i, N = n\right)$ for each sample based on the Gaussian distribution assumption using the means and standard deviations for the different types of peaks from the calibration samples.

3.2.4   Testing of NOCIt

1-, 2-, 3-, 4- and 5-person experimental samples were used for testing NOC*It* (Testing Set 1- see Tables 3.2-3.6). These 1-person samples were created using the same protocol described for the samples in the calibration set. The mixtures were created by mixing appropriate volumes of the single source DNA extracts to attain the various ratios specified in Table 3.7. Once mixed, these samples were re-quantified and then amplified using the same target masses used for the single-source samples. In the case of mixtures, the samples were created using various kinds of mixture ratios in such a way that each individual contributed at least two cells' worth of DNA, which corresponds to approximately 0.013ng of DNA [1]. None of the contributors to the Calibration Set were present in Testing Set 1 and none of the contributors to Testing Set 1 were present in the Calibration Set.

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.016	4	10	0.016	4	20	0.016	4
	0.031	4		0.031	4		0.031	4
	0.047	4		0.047	4		0.047	4
	0.063	4		0.063	4		0.063	4
	0.125	4		0.125	4		0.125	4
	Total	20		Total	20		Total	20

Table 3.2: 1-person samples in Testing Set-1

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.031	2	10	0.031	2	20	0.031	2
	0.047	4		0.047	4		0.047	4
	0.063	6		0.063	6		0.063	6
	0.125	8		0.125	8		0.125	7
	0.25	10		0.25	10		0.25	10
Total		30	Total		30	Total		29

Table 3.3: 2-person samples in Testing Set-1

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.047	1	10	0.047	1	20	0.047	1
	0.063	3		0.063	3		0.063	3
	0.125	5		0.125	5		0.125	5
	0.25	7		0.25	7		0.25	7
Total		16	Total		16	Total		16

Table 3.4: 3-person samples in Testing Set-1

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.063	2	10	0.063	2	20	0.063	2
	0.125	5		0.125	5		0.125	5
	0.25	6		0.25	6		0.25	6
Total		13	Total		13	Total		13

Table 3.5: 4-person samples in Testing Set-1

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.063	1	10	0.063	1	20	0.063	1
	0.125	5		0.125	5		0.125	5
	0.25	8		0.25	8		0.25	8
Total		14	Total		14	Total		14

Table 3.6: 5-person samples in Testing Set-1

Number of contributors	Mixture ratios used
2	1:1, 1:2, 1:4, 1:9, 1:19
3	1:1:1, 1:2:1, 1:4:1, 1:9:1, 1:2:2, 1:4:4, 1:9:9
4	1:1:1:1, 1:1:2:1, 1:1:4:1, 1:1:9:1, 1:2:2:1, 1:4:4:1, 1:9:9:1
5	1:1:1:1:1, 1:1:2:1:1, 1:1:4:1:1, 1:1:9:1:1, 1:1:2:2:1, 1:1:4:4:1, 1:2:2:2:1, 1:4:4:4:1

Table 3.7: Mixture ratios of samples in Testing Set-1

The contributors to Testing Set 1 were US Caucasian, Hispanic, Asian or Black. We did not block samples together based on their population into one mixture as we were attempting to mimic evidentiary items. The allele frequencies used in the NOCIt calculation was that of the US Caucasian population published in [16]. Four alleles belonging to five contributors in the calibration samples were not present in the frequency

table. These four alleles were added to the frequency table, each with a frequency of $5/2N$ (corresponding to a value of 0.7%), where N is the number of individuals sampled from, as suggested by the National Research Council (NRC-II 1996) [30].

In Testing Set 1, the 1-person samples contained DNA from 20 different individuals, the 2-person samples contained DNA from 4 different individuals (2 combinations), the 3-person samples contained DNA from 3 different individuals (1 combination), the 4-person samples contained DNA from 4 different individuals (1 combination) and the 5-person samples contained DNA from 5 different individuals (1 combination). The set of contributors was re-used but each sample was unique because variation was introduced through a) using different total DNA masses and b) using different mixture ratios. Thus, the amount of DNA from each contributor varied across the samples. To test the performance of the methods when subjected to diverse scenarios of allele-sharing, we created a set of 40 simulated mixtures using various genotype combinations, total DNA input and mixture ratios (Testing Set 2 – Table 3.8). The mixtures were simulated by adding the signal from the 1-person samples in Testing Set 1 at the 10s injection time in various combinations. The samples were simulated in such a way that no combination of contributors was repeated. Thus, the eight 2-person samples had eight combinations; the ten 3-person samples had ten combinations and so on.

Mixture type	Number of samples	DNA amounts (ng)	Mixture ratios
2-person samples	8	0.08, 0.11, 0.12, 0.14, 0.17, 0.19	1:1, 1:2, 1:3, 1:4, 1:8
3-person samples	10	0.09, 0.12, 0.14, 0.17, 0.20, 0.23, 0.25	1:1:1, 1:1:2, 1:1:3, 1:2:2, 1:2:3, 1:2:4, 1:3:4, 1:4:4, 1:4:8
4-person samples	12	0.11, 0.16, 0.17, 0.19, 0.20, 0.23, 0.25, 0.26	1:1:1:8, 1:1:2:3, 1:2:2:2, 1:2:2:4, 1:2:3:4, 1:2:4:4, 1:2:4:8, 1:3:4:4, 1:3:4:8, 1:4:4:8
5-person samples	10	0.22, 0.25, 0.28, 0.29	1:1:1:8:8, 1:1:2:2:2, 1:1:4:4:8, 1:2:2:2:2, 1:2:3:4:4, 1:2:3:4:8

Table 3.8: Mixtures in Testing Set-2.

An important thing to note is that NOC*It* is meant to *assist the analyst* in interpreting the DNA profile, and not to be used as a standalone tool to pick the number of contributors with the highest probability. The usefulness of NOC*It* stems from the fact that for complex mixtures that are hard to analyze, it can identify the range in which the number of contributors is most likely to lie. MLE, like NOC*It*, also gives a probability distribution on the number of contributors. Hence while analyzing the performance of these two methods, we developed two different ways for assessing their accuracy. One was to term the result as 'accurate' if the number of contributors in the sample had the highest probability (Maximum probability). The other method defined an accurate result as one in which the number of contributors in the sample had a probability of at least 1% (1% probability). The logic behind this is that if a number has a probability of at least 1% then it is quite likely to be the actual underlying number of contributors and therefore cannot be ignored during subsequent steps of the mixture interpretation process. Thus, more than one number of contributors might have to be considered for mixture interpretation if they all have a probability of at least 1% from NOC*It*.

NOC*It* was written in the Java programming language. It takes on average 10 hours to compute the APP on the number of contributors to a sample on a quad core

system with 2 GHz of processor speed with the maximum possible number of contributors $n_{\max} = 5$. To reduce computational running time, we did not compute the probabilities for all n up to 5 for all samples. Our results suggest that the distribution from NOC*It* was unimodal − having a single peak and then decreasing in value. Hence while computing the APP distribution, if the APP for n_0 is less than one-thousandth of the APP for $n_0 - 1$, we stop the calculation at n_0, assuming that the APP for $n > n_0$ is negligible.

The performance of NOC*It* was compared with the MAC and the MLE [26] methods. MAC uses the number of peaks observed in the signal to determine the number of contributors while MLE uses the number of peaks as well as the frequencies of the alleles in the signal. Both methods depend upon the establishment of a threshold to determine the set of true peaks. The threshold is typically chosen by a laboratory based on validation data. NOC*It* on the other hand, does not depend upon the setting of a threshold and works on the entire electropherogram obtained. Two types of thresholds were used for MAC and MLE for comparison purposes: a constant threshold of 50 RFU at all loci, and a variable threshold set as the height of the highest noise peak observed in the calibration data per dye color per DNA amount per time of injection. The average of the variable thresholds was 19, 33 and 52 RFU for the 5, 10 and 20s injection samples respectively. Application of MAC and MLE also uses a stutter threshold to filter out the peaks in the *n-4* position of peaks above the threshold. The stutter filter recommended by the manufacturer [16] was used at each locus to filter out the stutter peaks. We implemented MAC and MLE using the Python programming language.

3.2.5   NOCIt 2.0

As previously mentioned, NOCIt takes ~ 10 hours to compute the APP for up to $n = 5$. To reduce the running time of NOCIt and speed up the process, we created a new version of NOCIt that uses a faster algorithm. This new version (NOCIt 2.0) uses importance sampling, which is a Monte Carlo sampling algorithm in which instead of sampling directly from the target distribution, samples are generated from a different distribution that is easier to sample from [31]. To take into account the fact that the samples have come from the 'wrong' distribution, weights are introduced to adjust the 'importance' of each sample. For the problem at hand, instead of sampling using the allele frequency distribution, we generate samples of the interference genotypes using the peak height distribution observed at the locus. The reason for sampling from the peak height distribution is that this method is faster and requires fewer samples for convergence than the method that samples from the allele frequency distribution. Let $\bar{Z}$ be the number of interference samples used. The probability of observing the evidence at a locus is calculated as:

$$\Pr(E_l | \Theta = \theta^i, N = n) = \frac{\sum_{j=1}^{\bar{Z}} \Pr(E_l | G = g^j, \Theta = \theta^i, N = n) \, w_j}{\sum_{j=1}^{\bar{Z}} w_j},$$

where $w_j = P(g^j)/Q(g^j)$ is the weight of sample j; $P(g^j)$ is the probability of the genotypes under the allele frequency distribution; and $Q(g^j)$ is the probability of the genotypes under the peak height distribution. The signal obtained in the electropherogram consists of alleles and their associated heights. This allows the development of a distribution for all possible alleles (i.e. alleles in the frequency table) based on their respective heights; the probability of an allele is calculated as the ratio of its

height to the sum of the heights of all the alleles at the locus. Subsequently, all the probabilities are normalized such that they add up to 1. Since alleles that are not detected (i.e. alleles with a height less than 1 RFU) could also have been in the genotype of the contributors and dropped out, non-detected alleles were designated the minimum possible height of 1 RFU. $\Pr(E_l | G = g^j, \Theta = \theta^i, N = n)$ is calculated using the method described above for the older version of NOCIt.

3.3 Results

Across all samples in Testing Set 1, the maximum probability form of MLE (constant threshold accuracy: 72%, variable threshold accuracy: 65%) had a higher accuracy than MAC (constant threshold accuracy: 69%, variable threshold accuracy: 63%) with both the constant and the variable thresholds (Figure 3.1). Both MAC and MLE had a higher accuracy with the constant threshold of 50 RFU compared to the variable threshold. While using the 1% probability form as well, MLE had a higher accuracy than MAC with both the constant (accuracy: 84%) and the variable (accuracy: 82%) thresholds. Across all samples in Testing Set 1, NOC*It*, applied using the maximum probability and the 1% probability forms, had a higher accuracy than MAC and MLE. Like MLE, the 1% probability form of NOC*It* (95%) had a higher accuracy than the maximum probability form (83%).
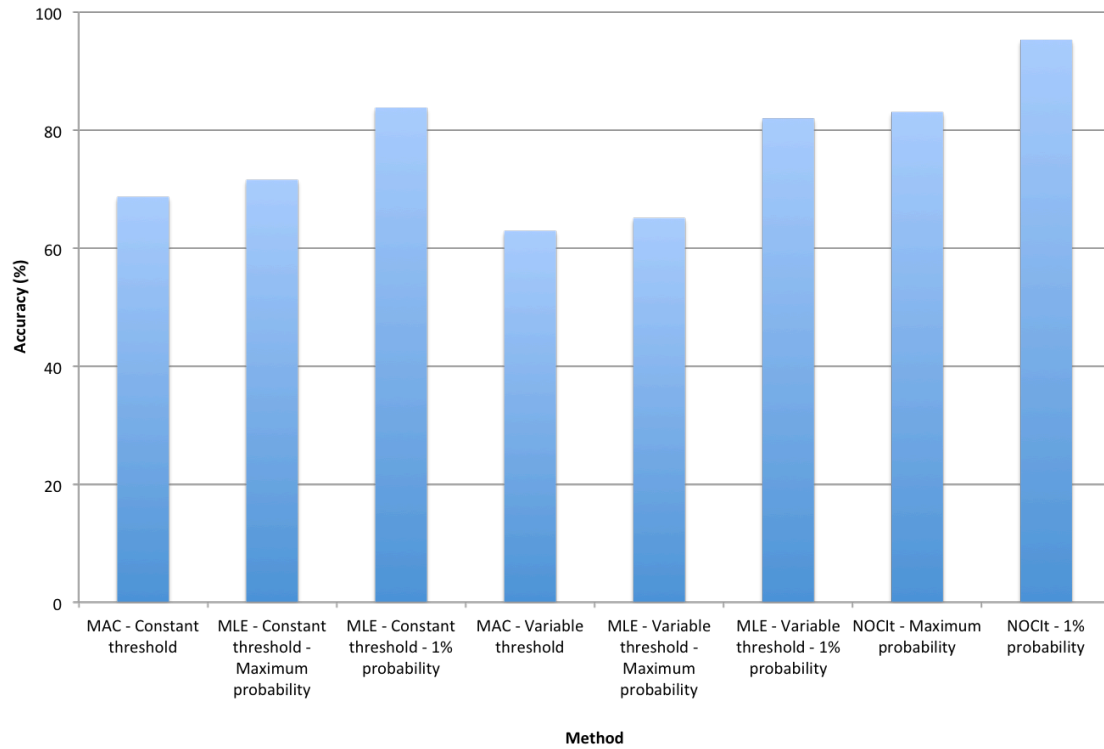
Figure 3.1: The accuracy of MAC, MLE and NOCIt across all samples in Testing Set-1.

The results of the performance of the three methods on Testing Set 1 at the three times of injection are provided in Figure 3.2. Changing the injection time did not have an impact on the performance of the three methods. The 1% probability form of NOC*It* was found to have the highest accuracy at all three injection times.
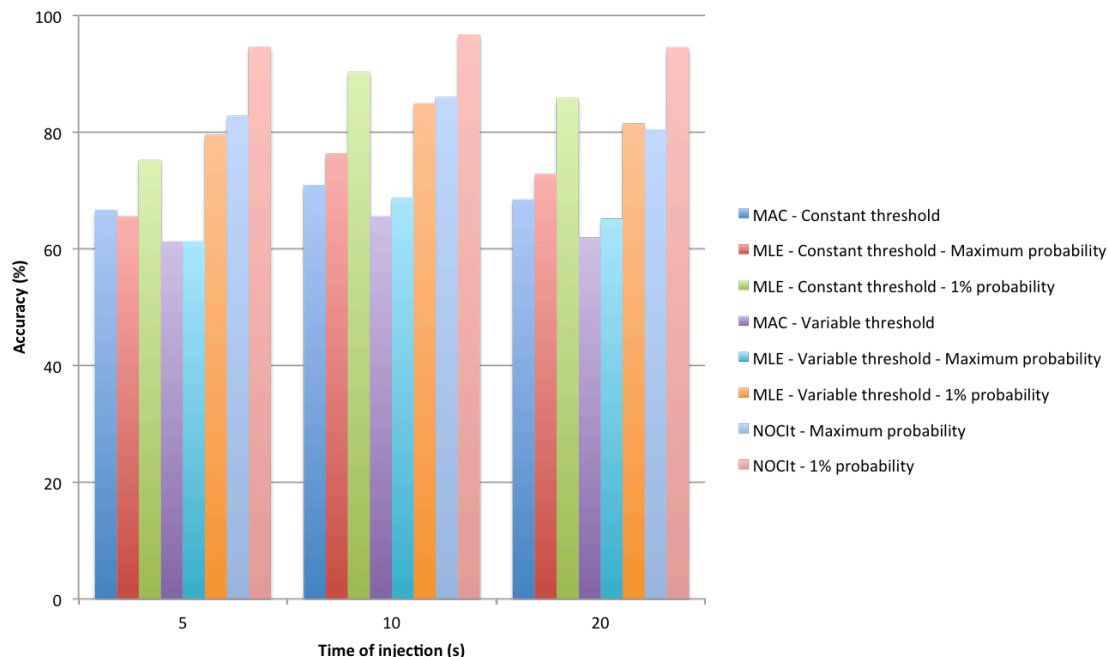
Figure 3.2: The accuracy of MAC, MLE and NOCIt at different injection times.

Figures 3.3 and 3.4 show how the performance of the methods changes as the number of contributors to the sample is increased. The accuracy of MAC and MLE were comparable for the 1-, 2- and 3-person samples, while the accuracy of MLE was higher than MAC for the more complex 4- and 5-person samples (Figure 3.3). As expected, MAC and MLE gave underestimates for mixtures with more than one contributor, due to sharing of alleles between the contributors. We also observed overestimates with both MAC and MLE for the 1- and 2-person samples, due to stutter peaks having a higher than expected height at one or more loci. MLE, unlike MAC, also resulted in overestimates for the 3-person samples (Figure 3.4).
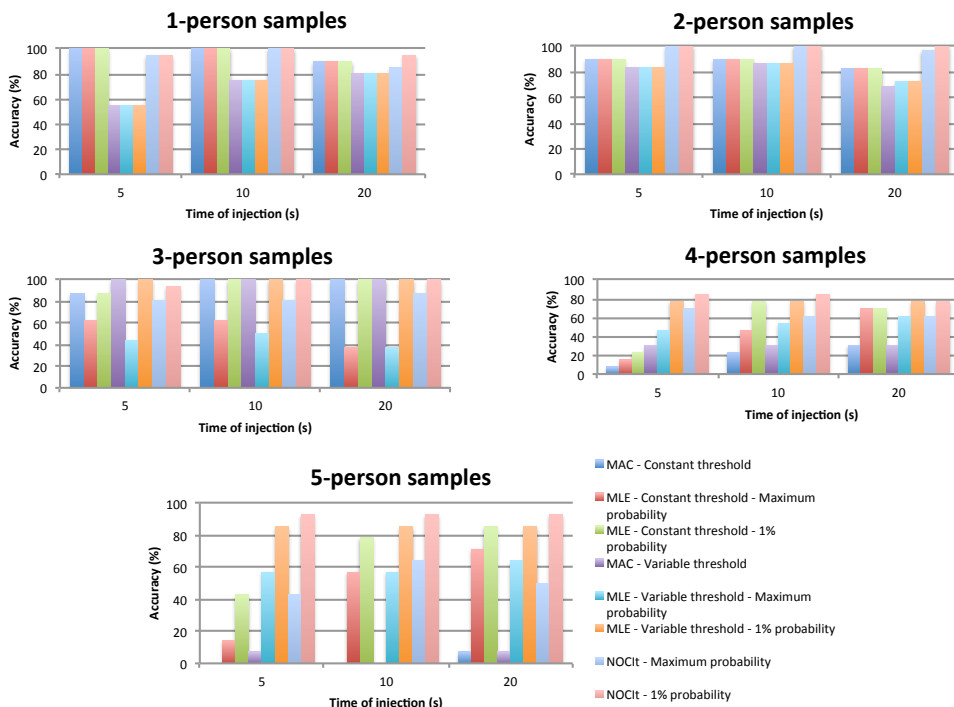
Figure 3.3: The accuracy of MAC, MLE and NOCIt as the number of contributors increases.
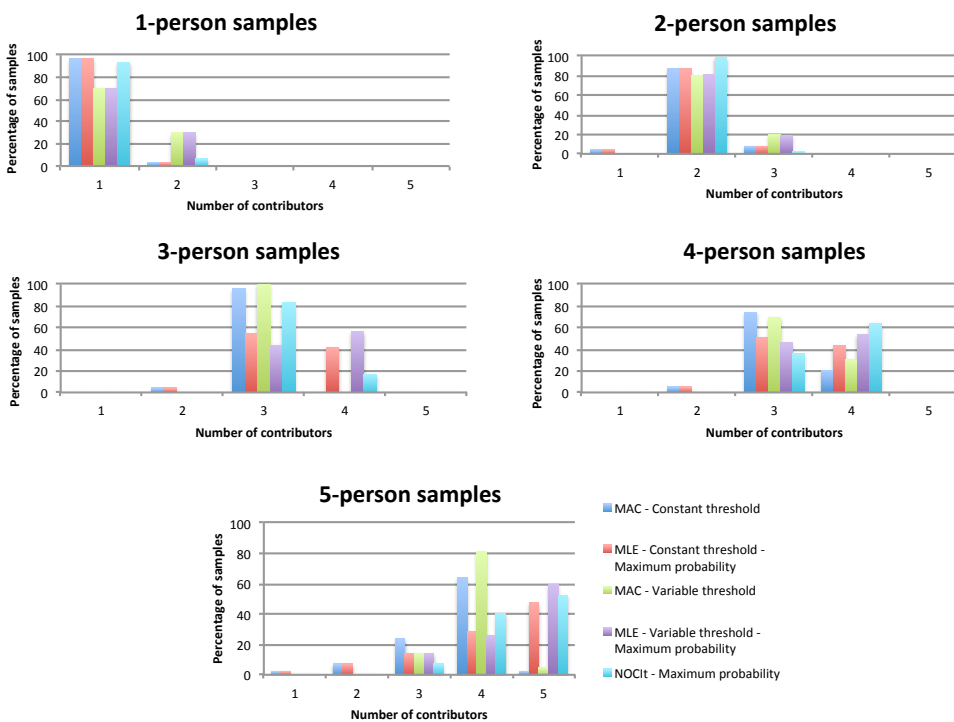
Figure 3.4: The percentage of calls made for each number of contributors by MAC, MLE and NOCIt.

With regards to NOC*It*, the accuracy of the 1% probability method was 82% or higher for the 1-, 2- 3-, 4- and 5- person samples (Figure 3.3). The maximum probability form of NOC*It* had an accuracy that was close to the 1% probability form for the 1- and 2-person samples. The larger difference between the accuracy of the two forms of NOC*It* for the 3-, 4- and 5-person mixtures indicates that in the instances in which NOC*It* does not come up with the highest probability for the actual number of contributors, it still successfully identifies the range in which the number is most likely to lie, even for complex mixtures.

NOC*It* had underestimates with the 4- and 5-person samples. These underestimated samples were characterized by high levels of dropout at multiple loci. There were three cases in which a 5-person sample was called as a 3-person sample. Apart from that, the underestimated samples were called as one less than the actual number of contributors. For the 1-, 2-, and 3-person samples in which NOC*It* overestimated the number of contributors, NOC*It* calculated the number of contributors as one more than the actual number (Figure 3.4). These samples were found to contain elevated levels of reverse and/or forward stutter at one or more loci.

Figure 3.5 shows the performance of the 3 methods on the simulated profiles in Testing Set 2. MAC and MLE were tested using only the constant threshold of 50 RFU. The 1% probability form of NOC*It* had an accuracy of 100%, while the 1% probability form of MLE had an accuracy of 95%. With regards to their maximum probability forms, MLE had an accuracy of 73% while NOCIt had an accuracy of 85%. The accuracy of MAC was 55% with the samples in Testing Set 2.
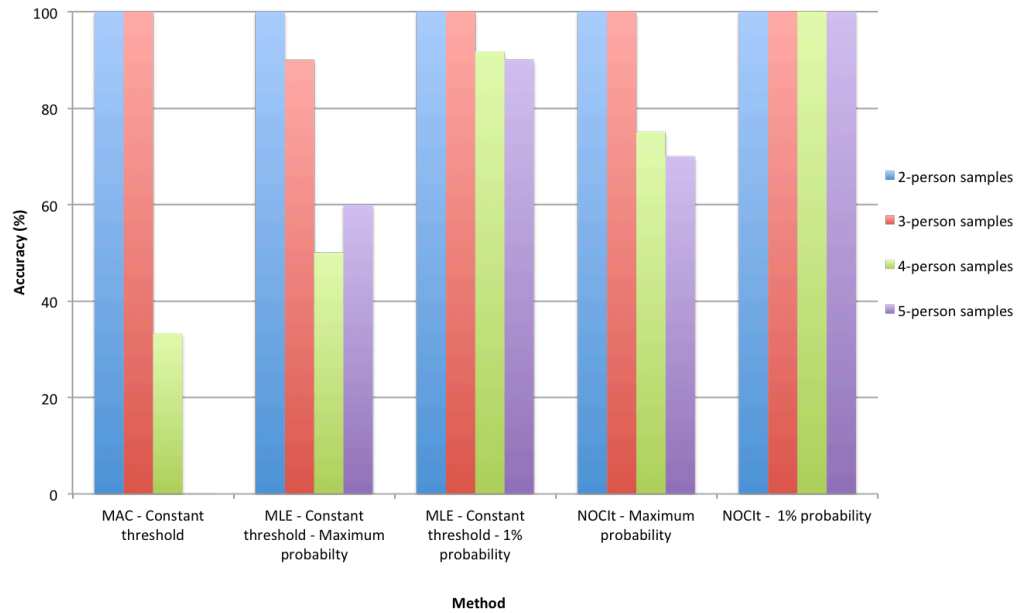
Figure 3.5: The accuracy of MAC, MLE and NOCIt on the simulated profiles in Testing Set-2.

The APP from NOC*It* tended to support one number of contributors much more than the others. The average of the highest APP from NOC*It* across all samples tested was 0.94. The average of the second highest APP across all the samples was 0.057. The third and subsequent highest APP had negligibly low values. For comparison purposes, we also tested the accuracy of NOC*It* on Testing Set 1 using a higher APP threshold of 10% and found the results to be similar to the ones obtained using a 1% threshold (Table 3.9).

Number of contributors	Number of samples	NOC*It* – 1% probability correct answers	NOC*It* – 10% probability correct answers
1	60	58	57
2	89	89	88
3	48	47	47
4	39	32	29
5	42	39	32

Table 3.9: Accuracy of NOCIt using APP thresholds of 1% and 10% on Testing Set-1 samples.

We tested the performance of NOCIt 2.0, an accelerated version of NOCIt using importance sampling, on the samples at the 10s injection. The results were similar to the first version of NOCIt (NOCIt 1.0) just described − out of the 93 samples, NOCIt 1.0 came up with the correct answer in 80 samples, while the number was 81 for NOCIt 2.0. The slight difference in the accuracy can be attributed to modifications in the probability convergence criteria, which were set to a more stringent level for NOCIt 2.0. Importantly, even with the stricter convergence level, the running time of NOCIt 2.0 is ~ 6 hours per sample, down from the ~ 10 hours taken by the previous version of NOCIt 1.0.

3.4 Discussion

On the experimental samples used for testing, both the maximum probability and the 1% probability forms of NOC*It* had a higher accuracy than the MAC and MLE methods using either a constant threshold of 50 RFU or a variable threshold based on the height of the noise peaks in the calibration data set. Similar results were obtained when the accuracy was compared at the three injection times used. These results indicate that using

the quantitative data in the signal, in addition to the qualitative data, results in a better estimate when determining the number of contributors.

In addition to the experimental samples used, the three methods were also tested on 40 simulated mixtures containing between 2 and 5 contributors created by adding the signal from the experimental 1-person samples at the 10s injection time in various combinations. The 1% probability form of NOC*It* had an accuracy of 100% on the simulated mixtures, performing better than the best available method in identifying the number of contributors

Our results suggest that the application of an analytical threshold, and the resulting loss of information about the peaks that do not cross the threshold, is generally detrimental to mixture interpretation when dealing with low template samples. In the case of low template samples, there is a chance that usage of a threshold could lead to dropout of alleles that might otherwise have been observed.

We also found that applying a stutter filter to filter out the stutter peaks might not work all the time, due to stutter peaks having higher than expected peak heights. This was found to be the cause of the overestimates from the MAC and MLE methods.

All three methods used in this study were not affected by changes in the time of injection. We observed that using a constant threshold of 50 RFU resulted in an accuracy higher than the accuracy with a variable threshold based on the height of the noise peaks in the calibration set for both MAC and MLE.

Overall, we found that both MLE and NOC*It* had a higher accuracy with the 1% probability form compared to the maximum probability form, indicating the utility of both the methods in identifying the range in which the number of contributors is likely to lie, even if they do not come up with the highest probability for the actual number of

contributors. This can be a very useful piece of information when analyzing samples with low DNA input and/or samples with a large number of contributors.

We found that the accuracy of MLE and MAC were similar to each other, with MLE having a slightly higher accuracy, when the maximum probability form was used for MLE. Using the 1% probability form of MLE resulted in a bigger difference in the accuracy of the two methods.

The accuracy of MAC was similar to that of MLE for the 1-, 2- and 3-person samples. For the more complex 4- and 5-person mixtures, the accuracy of MLE was higher than that of MAC. As expected, MAC gave underestimates for samples with more than 1 contributor. MAC also had overestimates for some of the 1- and 2-person samples, when the stutter ratio was higher than the expected level. While it is true that MAC gives the minimum number of contributors when the signal-to-noise ratio is sufficiently high, samples that contain elevated stutter could result in an overestimate. The results from MLE were similar to MAC, with underestimates for samples with more than 1 contributor and overestimates for some of the 1-, 2- and 3-person samples.

The underestimates from NOC*It* were characterized by high levels of dropout at multiple loci and were called as one less than the actual number of contributors (apart from three cases in which 5-person samples were called as 3-person samples), while the overestimates from NOCIt were called as one more than the actual number of contributors and were found to contain elevated levels of reverse and/or forward stutter at one or more loci.

The APP from NOC*It* tended to support one number of contributors much more than the others. Thus, though we chose a 1% threshold for our main results, using a 10% threshold resulted in similar accuracy.

With the purpose of speeding up the NOCIt calculation, we created a new version (NOCIt 2.0) of the method that uses importance sampling. We observed that the results from NOCIt 2.0 were similar to the results from NOCIt 1.0. Moreover, as expected, the running time went down from ~11 hours / sample for NOCIt 1.0 to ~ 6 hours / sample for NOCIt 2.0 while testing up to 5 contributors.

CHAPTER 4

MATCH STATISTICS FOR DNA MIXTURES

In this chapter, we describe MatchIt, a method that combines the twin features of a fully continuous model to calculate the LR and its distribution, along with an associated p-value.

In forensic DNA interpretation, the Likelihood Ratio (LR) is often used to convey the strength of a match. Expanding on binary and semi-continuous methods that do not use all of the quantitative data contained in an electropherogram, fully continuous methods to calculate the LR have been created. These fully continuous methods utilize all of the information captured in the electropherogram, including the peak heights. Recently, methods that calculate the distribution of the LR using semi-continuous methods have also been developed. The LR distribution has been proposed as a way of studying the robustness of the LR, which varies depending on the probabilistic model used for its calculation. For example, the LR distribution can be used to calculate the p-value, which is the probability that a randomly chosen individual results in a LR greater than the LR obtained from the person-of-interest (POI). Hence, the p-value is a statistic that is different from, but related to, the LR; and it may be interpreted as the false positive rate resulting from a binary hypothesis test between the prosecution and defense hypotheses.

Our algorithm for computing the LR incorporates dropout, noise and stutter (reverse and forward) in its calculation. As calibration data, MatchIt uses single source

samples with known genotypes and calculates a LR for a specified POI on a question sample, along with the LR distribution and a *p*-value. The method was tested on 306 files representing 1-, 2- and 3- person samples injected using three injection times containing between 0.016 and 1 ng of template DNA. We observed that the amount of template DNA from the contributor impacted the LR – small LRs resulted from contributors with low template masses. Moreover, as expected, we observed a decrease of *p*-values as the LR increased. The lowest possible *p*-value of 10^{-9} was achieved in all the cases where the LR was greater than 10^8. We tested the repeatability of MatchIt by running all samples in duplicate and found the results to be repeatable.

4.1 Background

Until recently, statements of inclusion or exclusion were exclusively used when reporting or presenting DNA comparisons to the trier-of-fact. If a suspect, or other known, is 'included' as a potential contributor to the item of evidence, then the inclusion statement must be accompanied by the calculation of a statistic that conveys the strength of the match [9]. Alternatives to inclusion/exclusion statements have, of late, been adopted, where a verbal scale is used to describe the number obtained [32].

Two protocols for calculating a match statistic are the Random Man Not Excluded (RMNE) approach, based on the Combined Probability of Inclusion (CPI) statistic, and the Likelihood Ratio (LR) approach. The RMNE method seeks to determine the fraction of the population that would not be excluded as a contributor to the profile. During the calculation of the CPI statistic, some information like the genotype of the suspect, the peak heights and the number of contributors to the profile is not utilized [11].

Though RMNE is still employed in practice, this method of evaluation is being replaced with the LR approach [8]. The Likelihood Ratio is defined as:

$$LR = \frac{\Pr(E|H_p, n_p)}{\Pr(E|H_d, n_d)},$$

where E is the evidence in the form of the electropherogram (epg); H_p and H_d are the hypotheses specified by the prosecution and the defense, respectively; and n_p and n_d are the number of contributors specified by the prosecution and the defense, respectively. There is no obligation for the number of contributors to be the same in the prosecution and the defense hypotheses [22]. The numerator is the probability of observing the evidence given the prosecution's hypothesis and the denominator is the probability of observing the evidence given the defense's hypothesis. The evidence shows support for the prosecution's hypotheses if LR > 1; if LR < 1 the defense's hypothesis is supported by the evidence. Unlike the RMNE method, the LR can use information like the number of contributors to the sample, the heights of the peaks observed and the genotype of the suspect [11].

The LR framework can be applied using a binary model that uses only the set of alleles observed in a DNA profile, but not the heights of the peaks [33]. This method is not very informative since it assigns a probability of 0 or 1 to genotypes based on the presence or absence of alleles. Alternatives to the binary model have been proposed that allow for drop-in and/or dropout of alleles [34, 35]. These 'partially continuous' methods use the peak heights to establish probabilities of dropout and drop-in. Unlike binary methods, they can be used to interpret profiles in which one or more of the suspect's alleles are not observed, or when there are incidences of drop-in. Fully continuous methods that employ probabilistic genotyping by modeling the peak heights have also

been created, resulting in the ability to incorporate stutter and noise, or drop-in, into the calculation of the statistic [36, 37, 38, 39, 40]. Fully continuous methods make use of the entire data obtained in the epg, including the qualitative (alleles observed) and the quantitative (peak heights) data. It has been shown that quantitative methods that use peak height information have a higher sensitivity and produce larger LRs than the binary method [13, 40]. The TrueAllele system [36, 37] uses an MCMC sampler to compute a probability for every possible genotype combination based on how well it explains the observed data. The peak heights are linearly modeled with respect to the mixture weights using a multivariate normal distribution. Degradation is modeled as an exponential decay with respect to the allele product length and stutter as a linear function of the allele. Cowell et al [38] model the peak heights using a gamma distribution and employ a Bayesian network for analyzing mixtures that incorporate dropout and a stutter model that is independent of DNA mass and the marker. Puch-Solis et al [39] also use a gamma distribution for stutter and allele heights but differ from [38] by using the total peak height at a locus as a proxy for the DNA mass, estimating parameters conditional on peak heights and jointly modeling stutter and allelic peaks. Both these methods do not take into account drop-in. Taylor et al [40] implement MCMC using the Metropolis-Hastings algorithm to compute the genotype probabilities. Allele peak heights are modeled using an exponential decay with respect to the molecular weight of the allele and stutter peak heights are modeled as a linear function of the allele height. Drop-in is modeled as an exponential decay with respect to the peak height.

In addition to the methods and models that evaluate LRs, computational methods that compute the LR distribution have recently garnered attention [35, 41, 42, 43, 44]. The LR distribution can be used to evaluate the robustness of the model by performing

Tippett tests – "*what is the probability that a non-contributor will give rise an LR greater than 1 (Type I error)?*" [41]. Another statistic that can be obtained from the LR distribution is the *p*-value [35, 42, 43, 44]. The *p*-value is a summary statistic; it is the probability that a randomly picked person from the population gives rise to an LR at least as large as the one observed for the person of interest. It can be used to control the false positive rate of binary classification based on the LR and may be useful when the analyst wants to compute the probability of a random non-contributor giving rise to an LR greater than the one observed for the suspect. While there is controversy surrounding the use of *p*-values [45], several authors have shown that it is a useful statistic that assists in the interpretation of LRs and has other applications like database searching [46] and kinship analysis [10].

4.2 Materials and methods

All the experimental samples described in this section were prepared using the protocol outlined in Section 2.2.1 in Chapter 2.

4.2.1 Calibration of MatchIt

MatchIt employs a continuous method to calculate the LR and hence uses the peak heights in the signal to calculate probabilities. Characterization of the peak heights is accomplished by using single source calibration profiles with known genotypes obtained from samples amplified from a wide range of input DNA masses. The protocol used for calibration of MatchIt is the same as the one used for the calibration of NOCIt. The samples used for calibration of MatchIt are specified in Table 4.1.

Injection time(s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.008	27	10	0.008	48	20	0.008	27
	0.016	27		0.016	45		0.016	27
	0.031	28		0.031	46		0.031	28
	0.047	26		0.047	27		0.047	27
	0.063	26		0.063	42		0.063	27
	0.125	27		0.125	48		0.125	27
	0.25	27		0.25	48		0.25	27
	0.5	13		0.5	9		0.5	0
	1	9		1	0		1	0
	Total	210		Total	313		Total	190

Table 4.1: Samples used for calibration of MatchIt

4.2.2 Testing of MatchIt

A total of 306 1-, 2- and 3-person sample files were used to test MatchIt (Testing Set – Tables 4.2-4.4). These 1-person test samples were created using the same protocol described for the single source samples in the calibration set. The mixtures were created by mixing appropriate volumes of the single source DNA extracts to attain the various ratios specified in Table 4.5. Once mixed, these samples were re-quantified and then amplified using the targets from Tables 4.2-4.4. The 1-person samples contained DNA from 34 different individuals, the 2-person samples contained DNA from 6 different individuals (3 combinations) and the 3-person samples contained DNA from 6 different individuals (2 combinations). When testing the samples injected for 5 sec, the calibration file also consisted of data from 5 sec injections. The same was true for the 10 and 20 sec injection files. None of the contributors to the calibration set were present in the testing set and none of the contributors to the testing set were present in the calibration set.

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.016	5	10	0.016	5	20	0.016	5
	0.031	5		0.031	5		0.031	5
	0.047	5		0.047	5		0.047	5
	0.063	5		0.063	5		0.063	5
	0.125	5		0.125	5		0.125	5
	0.25	5		0.25	5		0.25	5
	0.5	2		0.5	0		0.5	0
	1	1		1	1		1	0
	Total	33		Total	31		Total	30

Table 4.2: 1-person samples in the Testing Set

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
5	0.031	3	10	0.031	3	20	0.031	3
	0.047	4		0.047	4		0.047	4
	0.063	6		0.063	6		0.063	6
	0.125	9		0.125	9		0.125	9
	0.25	12		0.25	12		0.25	12
	0.5	5		0.5	5		0.5	3
	1	5		1	2		1	0
Total		44	Total		41	Total		37

Table 4.3: 2-person samples in the Testing Set

Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples	Injection time (s)	DNA amount (ng)	Number of samples
	0.04	1		0.04	1		0.04	1
	0.05	1		0.05	1		0.05	1
	0.1	9		0.1	9		0.1	9
	0.2	11		0.2	11		0.2	11
5	0.3	3	10	0.3	3	20	0.3	3
	0.4	2		0.4	2		0.4	2
	0.5	1		0.5	1		0.5	1
	0.6	2		0.6	2		0.6	2
Total		30	Total		30	Total		30

Table 4.4: 3-person samples in the Testing Set

Number of contributors	Mixture ratios used to create the samples in the testing set	Mixture ratios used in MatchIt's algorithm
2	1:1, 1:2, 1:4, 1:9, 1:19, 1:49	1:9, 1:4, 3:7, 2:3, 1:1, 3:2, 7:3, 4:1, 9:1
3	1:1:1, 1:2:1, 1:4:1, 1:9:1, 1:2:2, 1:4:4, 1:9:9	3:2:1, 2:1:3, 1:2:3, 6:2:1, 1:6:2, 2:1:6, 9:1:1, 1:9:1, 1:1:9, 5:1:3, 3:5:1, 1:3:5

Table 4.5: Mixture ratios in the Testing Set samples and in MatchIt's algorithm.

4.2.3 Modeling of variables

The models used to characterize the distribution of the various variables are the same as the ones used for NOCIt, as specified in Chapter 3. The peak heights were modeled using a Gaussian distribution. For stutter, since the height of the stutter peak depends upon the height of the parent allele, modeling was done using the stutter ratio. The allele frequencies used in this study were those of the US Caucasian population published in [16], though any population database can be used. Table 2.5 shows the distributions used to model the different variables with respect to the DNA mass.

4.2.4 MatchIt Algorithm

The Likelihood Ratio (LR) is defined as:

$$LR = \frac{\Pr(E|H_p, n_p)}{\Pr(E|H_d, n_d)}.$$

In practice, n_p and n_d can be chosen by the prosecution and the defense to maximize their respective probabilities and there is no necessity for n_p to be equal to n_d. However, we have developed MatchIt to use the same number of contributors in both the numerator and denominator to calculate the LR. It should be noted that the method could be extended to work on different assumptions on the number of contributors. Moving forward, we omit the notation n for the sake of brevity. We note that for purposes of this work, $n_p = n_d$ in all cases presented herein, and we use the known, and thus the true n to test the capabilities of MatchIt.

For this study, we use the following hypotheses for H_p and H_d:

   H_p: The evidence is a mixture of the genotype profile of a suspect (s) and the profiles of $n-1$ other unknown, unrelated contributors, whom we term the 'interference' contributors.

   H_d: The evidence is from n unknown individuals unrelated to the suspect.

In most cases, the value of the LR is very large (or very small) and it is easier to work with log(LR). Hence we have:

$$\log(LR(s)) = \log(\Pr(E|R = s, \boldsymbol{U}^{n-1})) - \log(\Pr(E|\boldsymbol{U}^n)),$$

where $\boldsymbol{U}^i = \{U_1, \dots, U_i\}$ are the random genotypes of i contributors and R is the random genotype of a single contributor, whether it be a true contributor, or non-contributor.

LR numerator calculation:

Our algorithm assumes a constant mixture ratio at all the loci. The mixture ratio specifies the proportion of the total template mass contributed by each contributor to the sample. The underlying mixture ratio of an evidence sample is unknown and needs to be described by a model in order to compute a continuous LR. A constant mixture ratio model assumes that the mixture ratio is the same at all the markers, whereas a variable mixture model accounts for the possibility of the mixture ratio being different at the various markers. Both models are reasonable and are used in existing continuous methods to compute the LR. Perlin et al [37] assign a uniform prior probability for the template mixture weight and construct its probability distribution by drawing individual locus weights using a multivariate normal distribution. Cowell et al [38] and Puch-Solis et al [39] use a constant mixture ratio model and implement a discrete approximation over the interval $(0,1)$ by assigning a uniform prior. Taylor et al [40] use the variable model and assume the mixture weights to be independent across the loci. Since we adopt the constant mixture ratio approach, we integrate over all possible mixture ratios to calculate the probability of observing the evidence:

$$\Pr(E|R = s, \boldsymbol{U}^{n-1}) = \int_{\boldsymbol{\theta} \in \Delta^{n-1}} \Pr(E|\boldsymbol{\Theta} = \boldsymbol{\theta}, R = s, \boldsymbol{U}^{n-1}) \, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}),$$

where $\boldsymbol{\Theta}$ is the vector with components Θ_i, the mixture proportion of each contributor $i \in \{1, \dots, n_{\max}\}$; $\Delta^{n-1} = \{(\Theta_1, \dots, \Theta_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} \Theta_i = 1, \Theta_i > 0 \, \forall \, i\}$ is the unit $n-1$ simplex; and $f_{\boldsymbol{\Theta}}$ is the probability density function of $\boldsymbol{\Theta}$, which we assume to be uniform over Δ^{n-1}. For $n = 1$, Δ^{n-1} consists of the single element $\{1\}$. For mixtures, we implement the integration over Δ^{n-1} by dividing it into equal-sized subsets and representing each subset with its centroid, resulting in a discrete sum. This was done by

performing k-means clustering in Python (Python Software Foundation, Beaverton, Oregon). k-means clustering is an algorithm used to partition observations into a set of clusters by repeated minimization of the distance from an observation to the centroid of its cluster [47]. For $n = 2$, the space was divided into 9 equally sized clusters, while for $n = 3$, 12 clusters were used. See Table 4.5 for a list of the mixture ratios used in MatchIt's algorithm.

Let $\boldsymbol{L}$ be the set of all loci in the evidence sample, E_l be the evidence at locus l, $\boldsymbol{U}_l^{n-1}$ be the genotype of the interference contributors at locus l and s_l be the genotype of the suspect at locus l. The STR loci used for forensic DNA analysis are assumed to be in linkage equilibrium and independent of each other [1]. Hence we obtain:

$$\Pr(E|\boldsymbol{\Theta} = \boldsymbol{\theta}, R = s, \boldsymbol{U}^{n-1}) = \prod_{l \in \boldsymbol{L}} \Pr(E_l|\boldsymbol{\Theta} = \boldsymbol{\theta}, R_l = s_l, \boldsymbol{U}_l^{n-1}).$$

The prosecution's hypothesis states that the profile is made of the suspect's contribution plus the contribution from $n - 1$ other random, unrelated contributors. Since there are many possibilities for the genotype of these interference contributors at each locus and going over each case would take a large amount of time, we calculate $\Pr(E_l|\boldsymbol{\Theta} = \boldsymbol{\theta}, R_l = s_l, \boldsymbol{U}_l^{n-1})$ using importance sampling. Importance sampling is carried out using the same method described for NOCIt 2.0 in Chapter 3.

Let J be the number of interference samples used. Now we obtain:

$$\Pr(E_l|\boldsymbol{\Theta} = \boldsymbol{\theta}, R_l = s_l, \boldsymbol{U}_l^{n-1}) = \frac{\sum_{i=1}^{J} \Pr(E_l|U_{l_i}^{n-1} = u_{l_i}^{n-1}, \boldsymbol{\Theta} = \boldsymbol{\theta}, R_l = s_l) \, w_i}{\sum_{i=1}^{J} w_i},$$

where $w_i = P(U_{l_i}^{n-1})/Q(U_{l_i}^{n-1})$ is the weight of sample i; $P(U_{l_i}^{n-1})$ is the probability of the interference genotypes under the allele frequency distribution; and $Q(U_{l_i}^{n-1})$ is the

probability of the interference genotypes under the peak height distribution. Since $u_{l_i}^{n-1}$ and s_l establish the true peaks in the signal (and by extension the stutter and noise peaks), $\Pr\left(E_l | U_{l_i}^{n-1} = u_{l_i}^{n-1}, \Theta = \theta, R_l = s_l\right)$ is the probability of observing the evidence, given the genotypes of the contributors and the mixture ratio. This calculation is described in the Appendix.

LR distribution and p-*value calculation:*

The *p*-value for the suspect is defined as the probability that a randomly picked person from the population would give rise to an LR at least as large as the one observed for the suspect.

$$p\text{-value}(s) = \Pr(LR(R) \geq LR(s)).$$

Since the denominator of the LR is the same for all the random genotypes R, it is sufficient if we compare the numerator of the LR for R and s.

$$p\text{-value}(s) = \Pr(\Pr(E|R, U^{n-1}) \geq \Pr(E|R = s, U^{n-1})).$$

During testing of MatchIt, we observed that because of floating-point precision, $\Pr(E|R, U^{n-1})$ evaluated to 0 for many of the random genotypes R that fit the data poorly. As a result, we were able to eliminate those genotypes from the *p*-value calculation as a preliminary step. Formally, let $\boldsymbol{R}$ be the set of all genotypes. We define

$$\boldsymbol{R_1} = \{r \in \boldsymbol{R} \mid \Pr(E_l|R_l = r_l) \neq 0 \text{ for all loci } l\}$$

$$\boldsymbol{R_2} = \{r \in \boldsymbol{R} \mid \exists \text{ locus } l \text{ s.t. } \Pr(E_l|R_l = r_l) \simeq 0\}$$

where $\simeq 0$ means "evaluates to 0 using double-precision 64-bit floating-point arithmetic". Thus, we have $\boldsymbol{R} = \boldsymbol{R_1} \cup \boldsymbol{R_2}$ and $\boldsymbol{R_1} \cap \boldsymbol{R_2} = \emptyset$. We see that for all $r \in \boldsymbol{R_2}, \Pr(E|R = r) \simeq 0$. We omit the notation on $\boldsymbol{U}^{n-1}$ for the sake of brevity. We have:

$$p\text{-value}(s) = \Pr(R \in \boldsymbol{R_1}) \sum_{r \in \boldsymbol{R_1}} \Pr(\Pr(E|R = r) \geq \Pr(E| R = s)) \Pr(R = r|R \in \boldsymbol{R_1})$$

$$+ \Pr(R \in \boldsymbol{R_2}) \sum_{r \in \boldsymbol{R_2}} \Pr(\Pr(E|R = r)$$

$$\geq \Pr(E| R = s)) \Pr(R = r|R \in \boldsymbol{R_2}).$$

We see that the second term is 0, provided $\Pr(E| R = s)$ is greater than 0. Hence we get:

$$p\text{-value}(s) = \Pr(R \in \boldsymbol{R_1}) \sum_{r \in \boldsymbol{R_1}} \mathbf{1}(\Pr(E|R = r) \geq \Pr(E|R = s)) \Pr(R = r|R \in \boldsymbol{R_1}),$$

where

$$\mathbf{1}(\Pr(E|R = r) \geq \Pr(E|R = s)) = \begin{cases} 1, \text{if } \Pr(E|R = r) \geq \Pr(E|R = s), \\ 0, \text{ otherwise.} \end{cases}$$

We have:

$$\Pr(R \in \boldsymbol{R_1}) = \prod_{l \in L} \sum_{r_l \in \{r| \Pr(E_l|R_l=r) \neq 0\}} \Pr(R_l = r_l).$$

We compute the p-value using Monte Carlo simulation. We generate M random genotypes $r^1, \ldots, r^M$ according to the distribution $\Pr(R|R \in \boldsymbol{R_1})$ and calculate the p-value as:

$$p\text{-value}(s) = \Pr(R \in \boldsymbol{R_1}) \frac{\sum_{i=1}^{M} \mathbf{1}\left(\left(\Pr(E|R = r^i) \geq \Pr(E|R = s)\right)\right)}{M}.$$

Increasing the value of M increases the accuracy of the p-value computed, but this also increases the run time and a hence a tradeoff has to be achieved between the two. In this study, we have used 1 billion or 10^9 random genotypes to compute the p-value. For a sample, if the term $\sum_{i=1}^{M} \mathbf{1}\left(\left(\Pr(E|R = r^i) \geq \Pr(E|R = s)\right)\right)$ evaluated to 0 (meaning that no genotype was simulated that had a probability greater than the POI's), it was given a value of 1 and the p-value was set to be 10^{-9} as an upper bound, even though the

true value is likely smaller and more samples would be needed to accurately estimate it.

In order to facilitate the computation of the p-value, as an initial step $\Pr(E_l | \Theta = \boldsymbol{\theta}, R_l = g_l, \boldsymbol{U}_l^{n-1})$ is computed for all possible genotypes g_l at all loci l for all values of $\boldsymbol{\theta}$. Once this is done, for the p-value computation, 10^9 genotypes r^i are generated based on the allele frequencies. Since we know $\Pr(E_l | R_l = r_l^i)$ for all loci l, we can compute $\Pr(E|R = r^i)$ as:

$$\Pr(E|R = r^i) = \int_{\boldsymbol{\theta} \in \Delta^{n-1}} \prod_l \Pr(E_l | \Theta = \boldsymbol{\theta}, R_l = r_l^i, \boldsymbol{U}_l^{n-1}) f_\Theta(\boldsymbol{\theta}).$$

LR denominator calculation:

Let $\bar{R}$ be the genotype of an unknown contributor in the defense's hypothesis. The denominator of the LR can be written as:

$$\Pr(E|\boldsymbol{U}^n) = \sum_{\bar{r}} \Pr(E|\bar{R} = \bar{r}, \boldsymbol{U}^{n-1}) \Pr(\bar{R} = \bar{r})$$

Since the number of possible values that $\bar{R}$ can take is large and summing over all of them is computationally intensive, we utilize the random genotypes r^i that are sampled for the p-value computation to compute the denominator of the LR as follows:

$$\Pr(E|\boldsymbol{U}^n) = \Pr(\boldsymbol{R_1}) \frac{\sum_{i=1}^M \Pr(E|R = r^i)}{M}.$$

4.3 RESULTS

Contributors with LR less than 1 have low template DNA masses

Figure 4.1 shows LR vs contributor template mass for all the contributors for 101, 1-, 2-

and 3-person samples at the 10s injection in the testing set. The LRs for the 5s and the 20s samples were similar to the 10s samples and are shown in Figures 4.2 and 4.3. All the log values shown in this study are to base 10. For each sample, the LR was computed for all the contributors. Hence, a single source sample would have 1 LR; a 2-person sample would have 2 LRs, etc. We observed that the amount of template DNA from the contributor impacted the LR from MatchIt, i.e. high LRs corresponded to high template DNA amounts. This effect was previously reported by Taylor et al [40]. While most log(LR) values were greater than 0 (the highest was 102, with a mean of 13), MatchIt computed the log(LR) as less than 0 (or LR < 1) for a contributor in 33 out of the 608 cases. These samples had a low template DNA from the contributor and a high level of dropout and stutter.
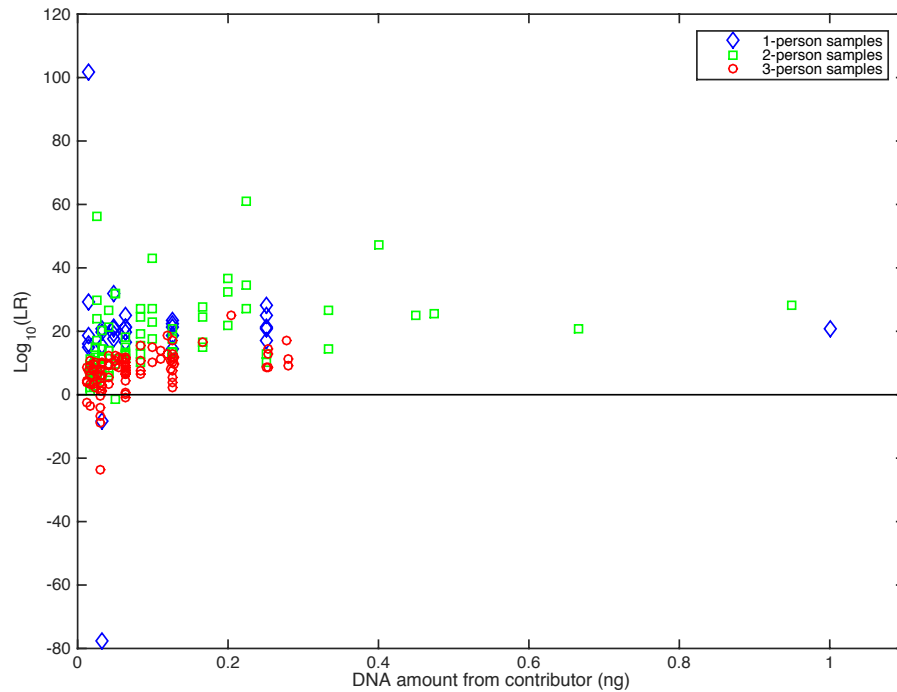

Figure 4.1: LR from MatchIt for the contributors to the 10s injection samples.
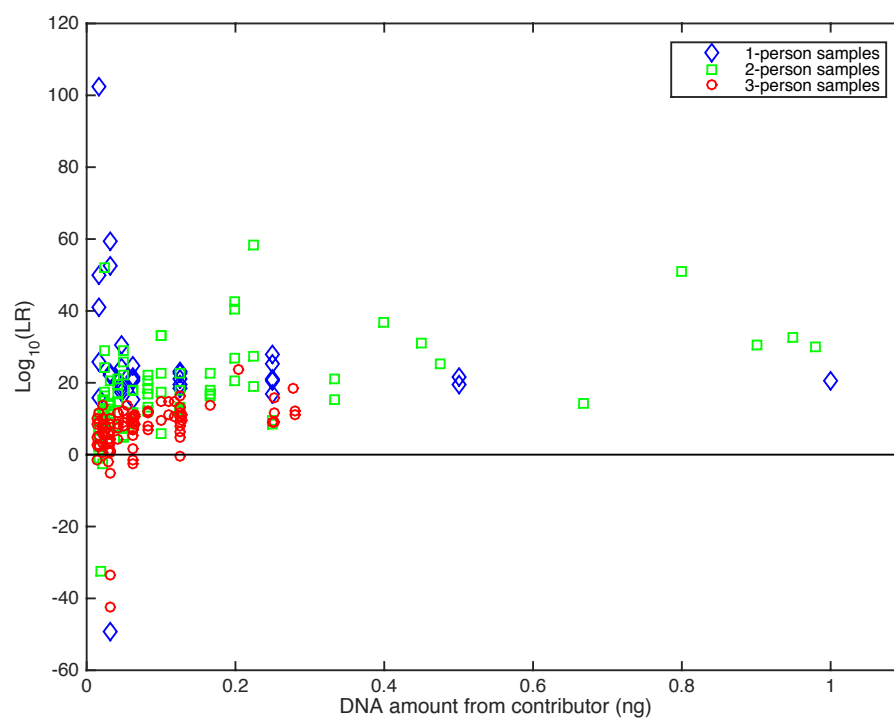
Figure 4.2: LR from MatchIt for the contributors to the 5s injection samples.
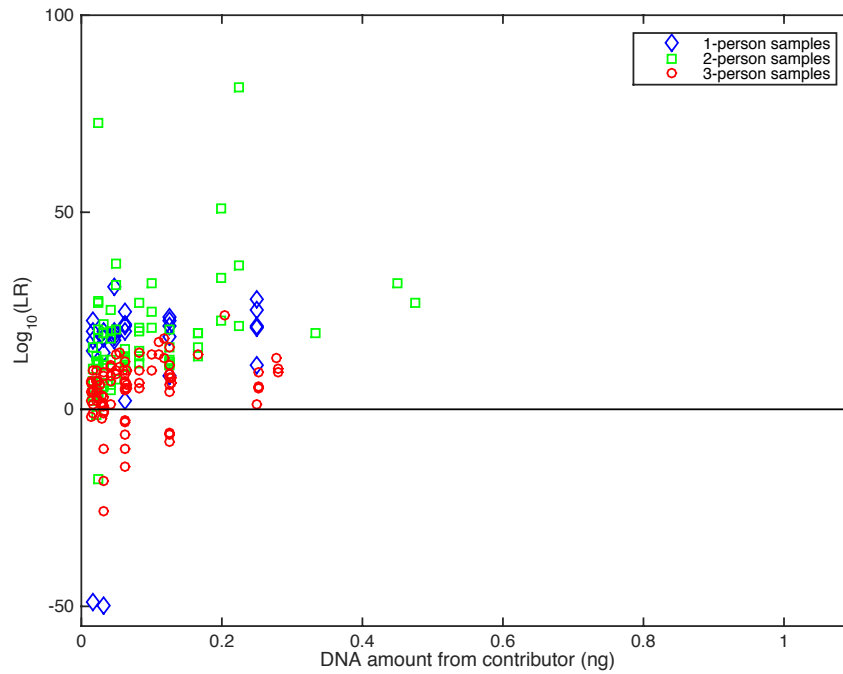
Figure 4.3: LR from MatchIt for the contributors to the 20s injection samples.

For example, a 3-person sample with 0.59 ng of template DNA and a mixture ratio of 1:9:9 resulted in a log(LR) of -23 for the minor contributor (~0.03 ng template mass per qPCR quantification). Figure 4.4 shows the profile of the sample and Table 4.6 shows the genotype of the 3 contributors. One of Contributor 1's alleles completely dropped out at 3 loci: allele 8 at D13S317, allele 20 at D2S1338 and allele 18 at vWA. Moreover, at locus D18S51 the forward stutter (height of 10 RFU) from allele 17 has a larger height than the parent peak (height of 9 RFU). These factors caused MatchIt to compute a low probability for the minor contributor. On the other hand, the alleles of major contributors 2 and 3 (~0.28 ng template mass each) are well represented in the epg resulting in a log(LR) of 9 and 11, respectively. The LR distribution from MatchIt for this sample is seen in Figure 4.5. A vast majority of the 1 billion samples had a log(LR) less

than 0, with a peak at -40. The LRs for the 3 contributors were near the right end of the distribution; the resulting p-values were 0.05 for the minor contributor and 10^{-9} for the two major contributors.
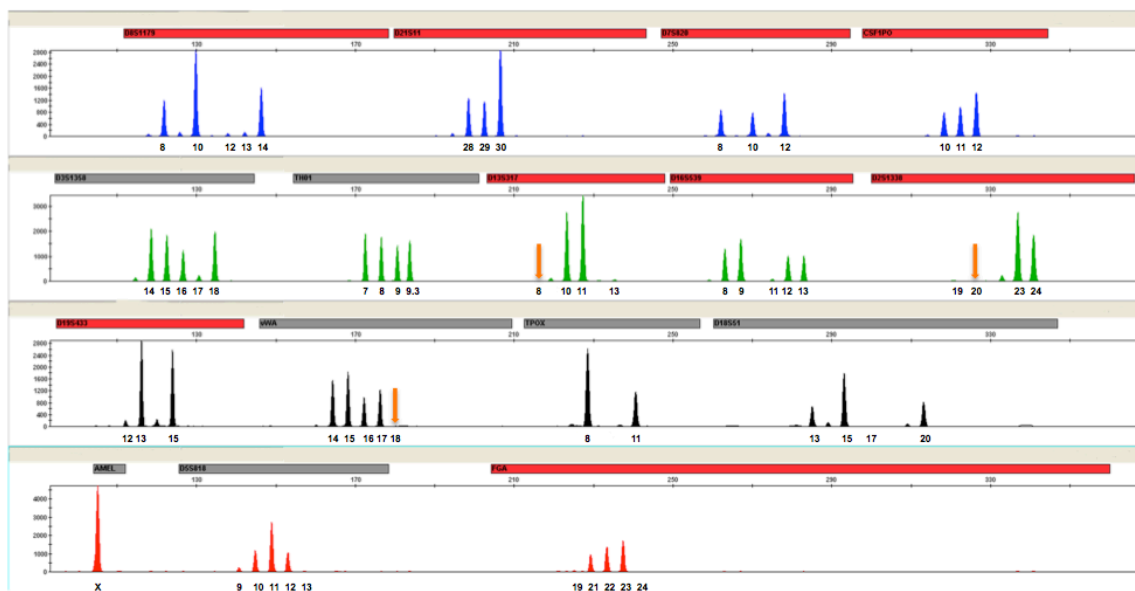

Figure 4.4: Profile of a 3-person sample, amplified from 0.59ng of template DNA and a mixture ratio of 1:9:9.

Locus	Contributor 1	Contributor 2	Contributor 3
D8S1179	12,13	10,10	8,14
D21S11	28,30	30,30	28,29
D7S820	8,10	10,12	8,12
CSF1PO	12,12	11,12	10,12
D3S1358	16,17	14,18	15,16
TH01	8,9	7,9.3	8,9
D13S317	8,13	10,11	10,11
D16S539	11,12	9,12	8,13
D2S1338	19,20	23,24	23,24
D19S433	12,13	13,15	13,15
vWA	15,18	15,17	14,16
TPOX	8,11	8,8	8,11
D18S51	17,20	15,20	13,15
AMEL	X,X	X,X	X,X
D5S818	9,13	11,11	10,12
FGA	19,24	22,23	21,23

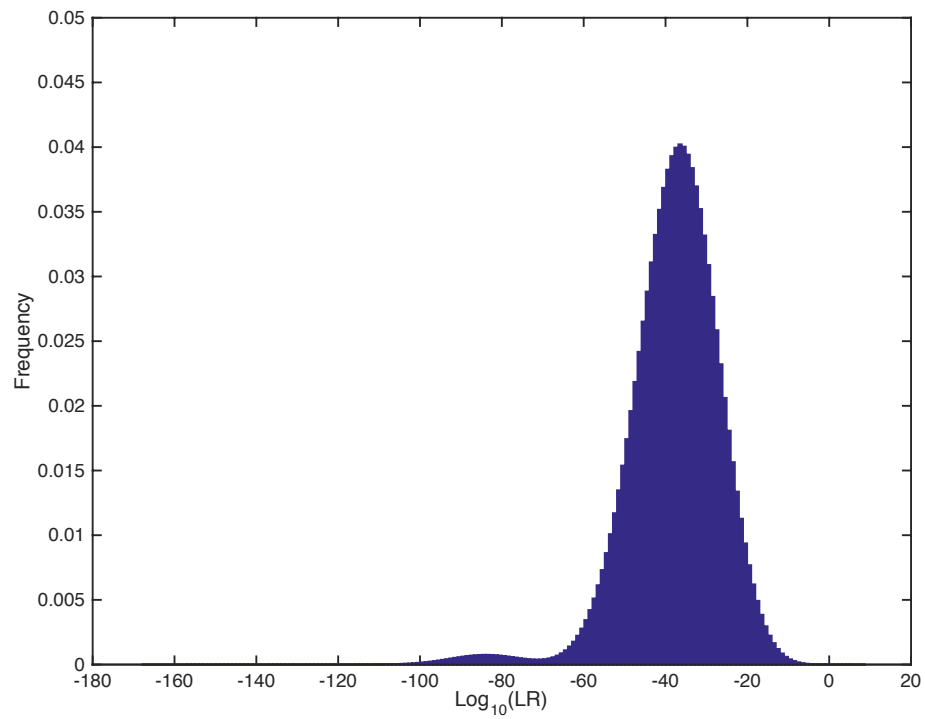**Table 4.6: Genotypes of the contributors to the 3-person sample in Figure 4.4.**

Figure 4.5: LR distribution from MatchIt for the sample in Fig 4.4.

When dropout levels were not high and the alleles were well represented in the epg, even minor contributors had high LRs. Figure 4.6 shows the profile of a 2-person mixture, with 0.25 ng of template DNA and a 1:2 mixture ratio. The genotypes of the 2 contributors can be seen in Table 4.7. All the alleles of the contributors can be seen in the profile and the log(LR) was 15 for both contributors. Figure 4.7 shows the LR distribution for this sample. Most of the samples had a log(LR) less than 0 and the distribution had a peak at -150. The LRs for the contributors to the sample were in the right tail of the distribution; the p-value was 10^{-9} for the major and the minor contributor.
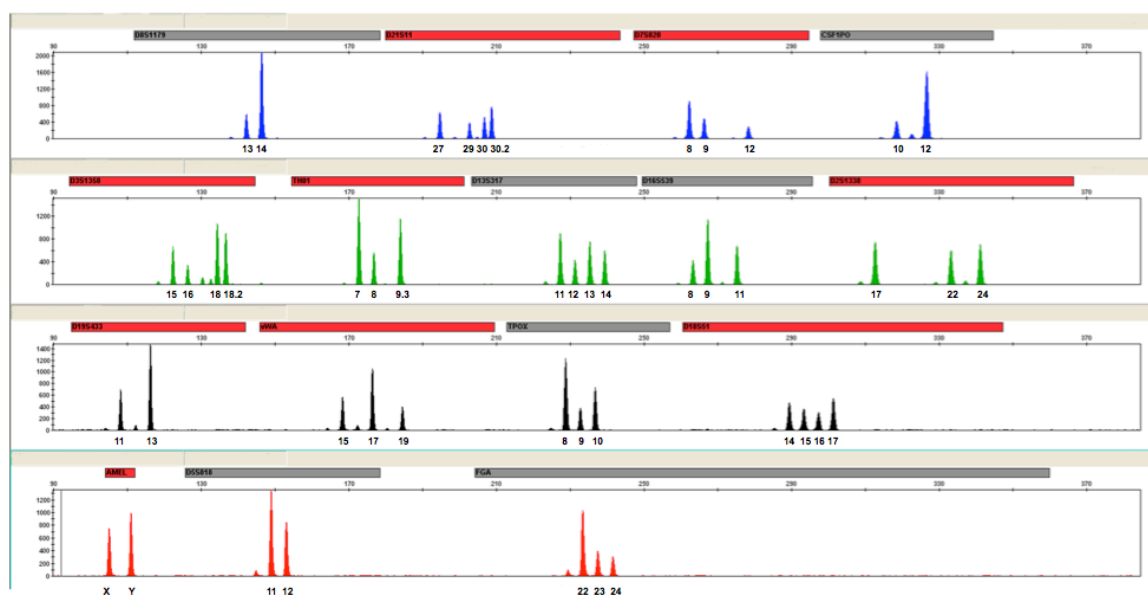

Figure 4.6: Profile of a 2-person sample, amplified from 0.25ng of template DNA and a mixture ratio of 1:2.

Locus	Contributor 1	Contributor 2
D8S1179	13,14	14,14
D21S11	27,29	30,30.2
D7S820	8,12	8,9
CSF1PO	10,12	12,12
D3S1358	16,18	15,18.2
TH01	7,8	7,9.3
D13S317	12,14	11,13
D16S539	8,9	9,11
D2S1338	17,17	22,24
D19S433	13,13	11,13
vWA	17,19	15,17
TPOX	9,10	8,8
D18S51	15,16	14,17
AM	X,Y	X,Y
D5S818	12,12	11,11
FGA	22,24	22,23

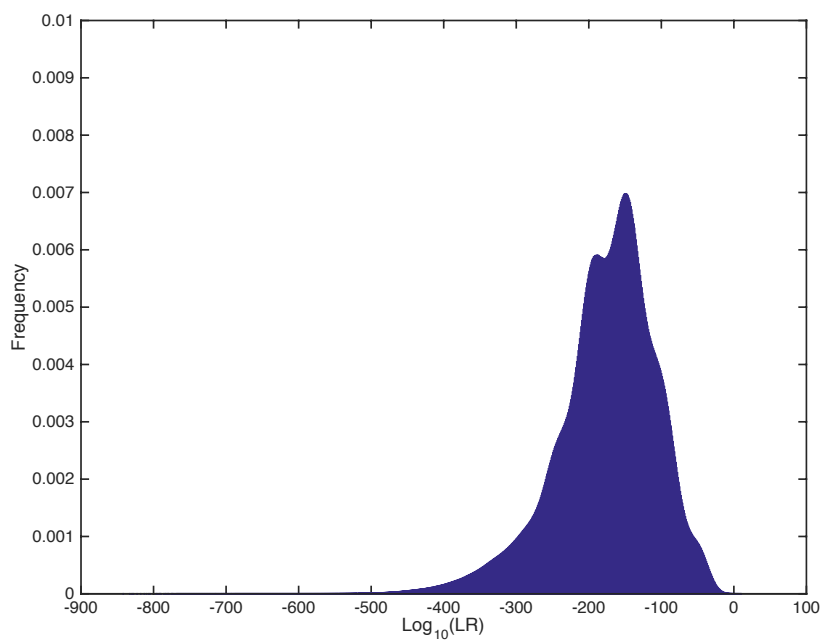**Table 4.7: Genotypes of the contributors to the 2-person sample in Fig 4.6.**

Figure 4.7: LR distribution from MatchIt for the sample in Fig 4.6.

The Random Match Probability (RMP) is an inclusion statistic that stems from a binary approach, i.e. alleles are either present or absent [10]. The statistic is calculated using only the frequencies of the alleles observed. We computed the RMP for the single source samples using an analytical threshold of 50 RFU and the stutter filter and the US Caucasian allele frequencies published in [16]. In our testing set, 33 out of the 94 single source files resulted in a 'match' and hence a RMP was computed for the 33 samples. For all those 33 samples, the LR from MatchIt was similar to 1 / RMP [9] – the absolute value of the difference between log(LR) and log(1/RMP) was less than 1 in 30 out of the 33 samples, with the lowest value being 0.0001. The largest difference observed between log(LR) and log(1/RMP) was 11, with log(LR) being smaller.

Addition of 'interference' contributors reduces the LR for a POI

There were two contributors to the 1-person samples who were also minor contributors in 2-person and 3-person samples in the testing set. We observed that for a given template DNA mass, the LR was highest for the 1-person sample and decreased in the presence of an interference contributor in a 2-person sample with a 1:1 mixture ratio. The LR further decreased with 2 interference contributors in a 3-person sample with a 1:2:1 mixture ratio (Figure 4.8). We did observe that at all three injection times, for some 2-person samples the log(LR) for the minor and major contributors was large (between 43 and 82), and larger than the log(LR) for some 1-person samples. We found that the log(LR)'s were large in these instances because the two contributors gave the best, or close to the best, fit to the signal and the denominator of the LR had a very low probability.
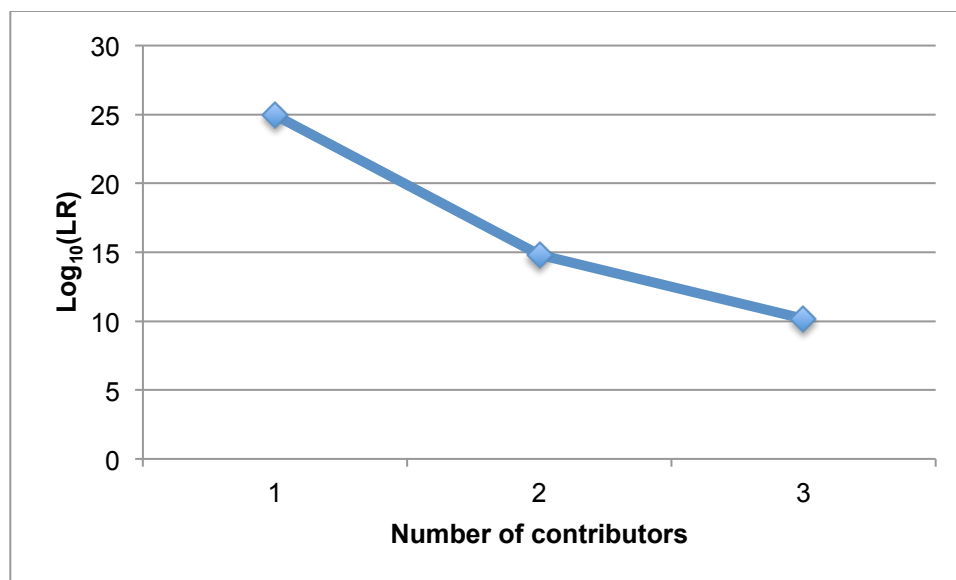
Figure 4.8: The LR for a contributor with 0.0625ng template mass in a 1-, 2- and 3-person sample at the 10s injection

High LRs correspond to low p-values

Figure 4.9 (a) shows how the *p*-values from MatchIt vary with LR for the samples at the 10s injection. In general, as the LR increased in value, the *p*-value decreased, indicating a greater confidence in including the individual as a contributor. All the 1-person samples a the lowest possible *p*-value of 10^{-9}. Even the 1-person samples that had a LR < 1 had a low *p*-value – though their probability was low because of high levels of dropout and stutter, it was still high relative to the population. For the 2-person and 3-person samples, in all cases where the log(LR) was greater than 8, the log(*p*-value) from MatchIt was -9, indicating that MatchIt was able to identify the individual as a contributor with great certainty. Figure 4.9 (b) shows the same plot zoomed in on the log(LR) values between 0 and 9. All the log(*p*-value) points represent 2- and 3-person samples and they lie below the line representing -log(LR). This relationship between the *p*-value and the LR is expected because the *p*-value ≤ 1 / LR according to Markov's inequality [44]. The 5s and the 20s

samples showed a trend similar to the 10s samples (Figures 4.10-11).

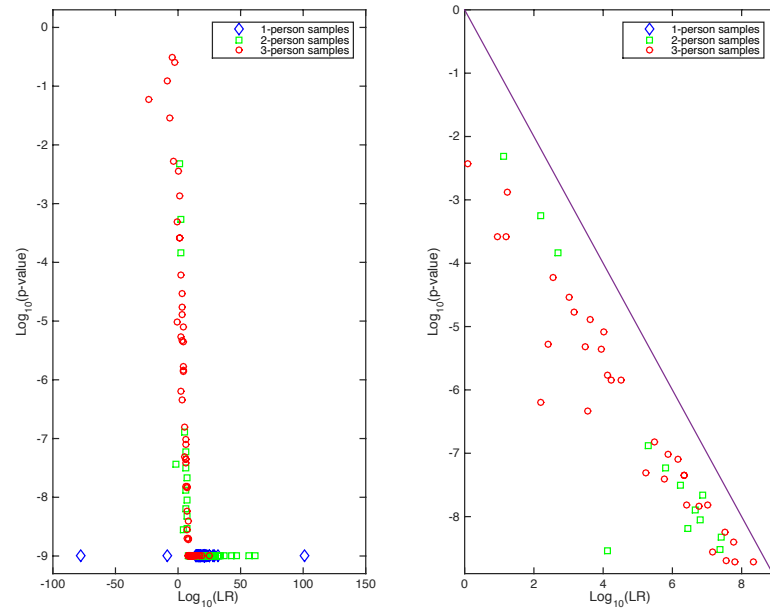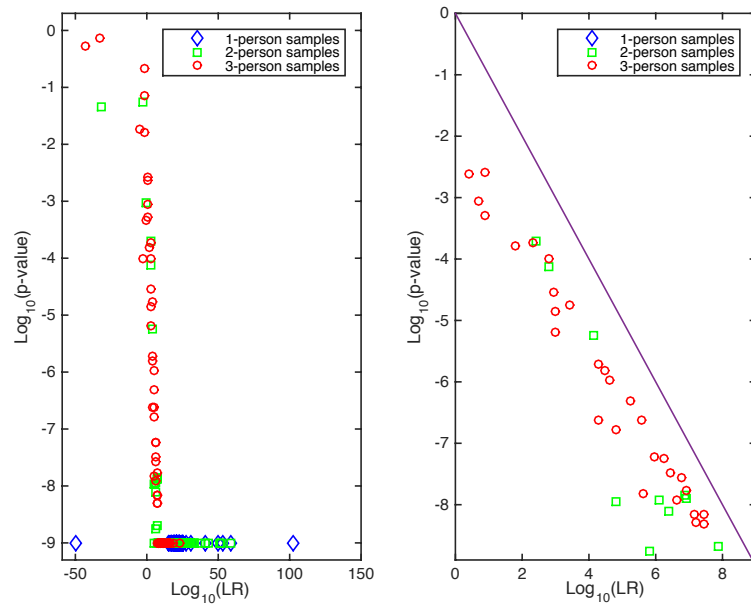**Figure 4.9 :** *p*-values from MatchIt as a function of LR for all the samples at the 10s injection

Figure 4.10: *p*-values from MatchIt as a function of LR for all the samples at the 5s injection
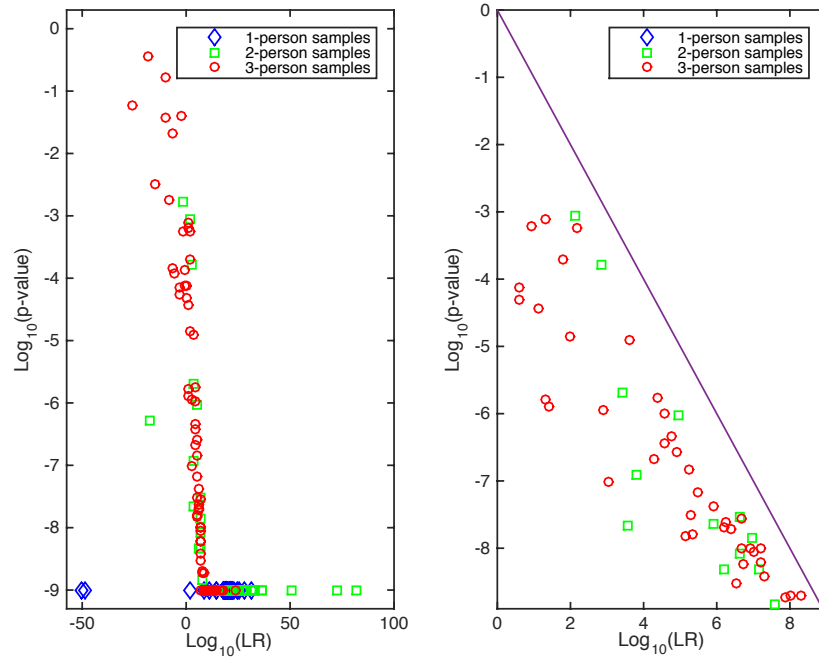
Figure 4.11: *p*-values from MatchIt as a function of LR for all the samples at the 10s injection

LRs for non-contributors are low

As described in Materials and Methods, for each experimental sample in our testing set, $\boldsymbol{R_1}$ contains all the genotypes with a non-0 probability. For 2 of the 306 sample files we used for testing, $\boldsymbol{R_1}$ consisted of only one genotype: that of the true contributor to the sample. They were both 1-person samples with 1ng template DNA mass. For the remaining 304 samples, MatchIt produced the LR distribution by sampling the genotypes of 10^9 non-contributors from the set $\boldsymbol{R_1}$. This resulted in a total of $304{\times}10^9 \approx 3{\times}10^{11}$ random genotypes for which we obtained a LR from MatchIt. The LR distribution for every sample was similar to the ones shown in Figures 4.5 and 4.7, with a majority of the LRs falling below 1. For each sample, we calculated the fraction of the 10^9 genotypes that resulted in a LR > 1 and multiplied it by $\Pr(R \in \boldsymbol{R_1})$ to calculate the Tippet statistic $\Pr(LR > 1|H_d)$. We observed that the average of the Tippett statistic increased as the

number of contributors to the sample increased, indicating that the probability of a random non-contributor having a LR > 1 is low for small n and increases as n increases. The average $\text{Pr}(LR > 1|H_d)$ for the 1-person samples was 8.56×10^{-12} and it increased to 0.001 and 0.002 for the 2- and 3-person samples. The smallest LR observed for the random contributors was 10^{-2581}, while the maximum was 10^{31} (for a 1-person sample in which the true contributor to the sample also had a LR of 10^{31}).

The results from MatchIt are repeatable

One of the consequences of using a sampling algorithm (like importance sampling or Monte Carlo simulation) is that the results of two analyses on the same sample are generally not identical. While increasing the number of samples reduces the run-to-run variation, doing so also increases the running time. Hence a trade-off between the two has to be achieved.

Figure 4.12 shows the results of 2 different runs on the 101, 1-, 2- and 3-person samples at the 10s injection. There is very little variation from run to run; the points lie close to the x=y diagonal line with very few outliers (Pearson correlation coefficient = 0.9764, slope = 0.98, intercept = 0.14). Additionally we tested repeatability of MatchIt on one 2-person sample and one 3-person sample after 5 runs and found that the LRs and *p*-values obtained were repeatable (Figures 4.13-14). Tests for repeatability on the 5s and the 20s samples also produced similar results (Figures 4.15-16).
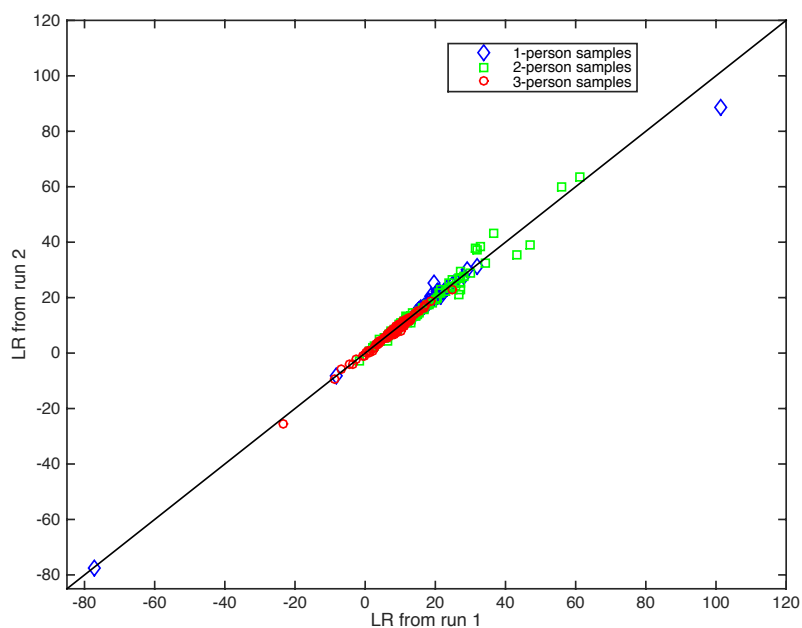
Figure 4.12: The results of 2 different runs on the 101, 1-, 2- and 3-person samples at the 10s injection.
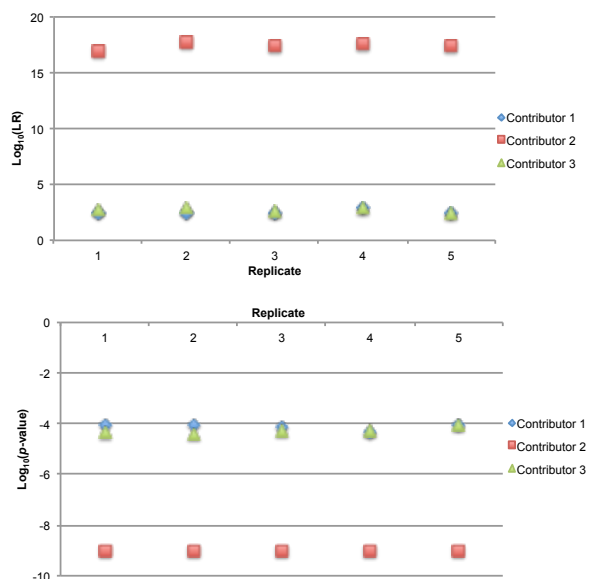

Figure 4.13: LR and *p*-value from 5 runs for the contributors to a 3-person sample amplified from 0.19ng template mass and 1:4:1 mixture ratio.
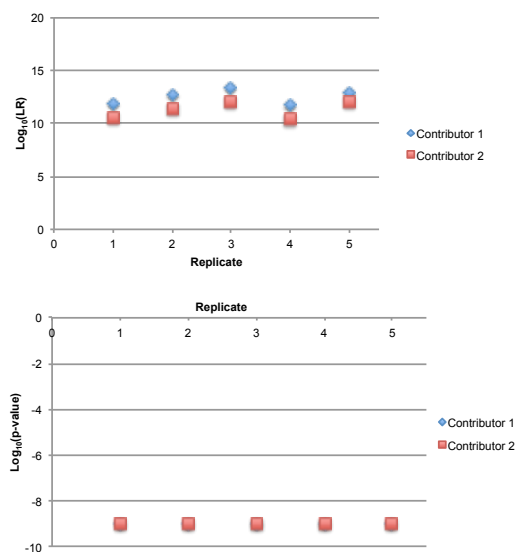
Figure 4.14: LR and *p*-value from 5 runs for the contributors to a 2-person sample amplified from 0.0625ng template mass and 1:1 mixture ratio.
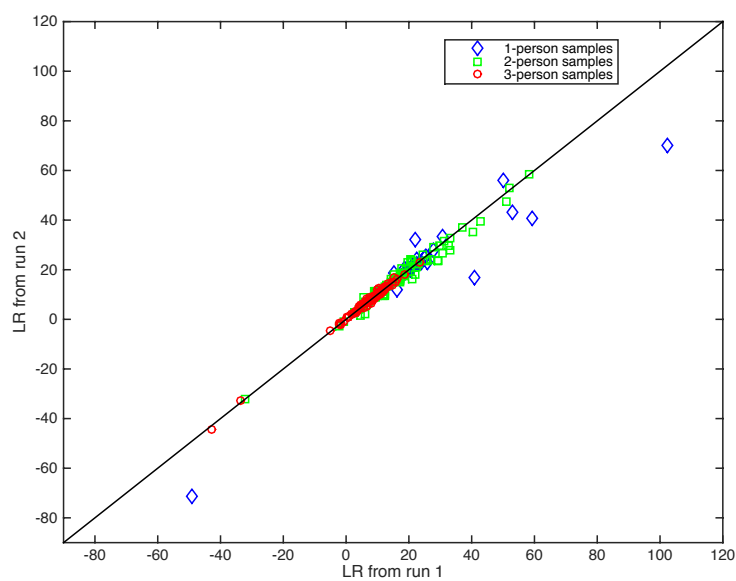

Figure 4.15: The results of 2 different runs on the 1-, 2- and 3-person samples at the 5s injection.
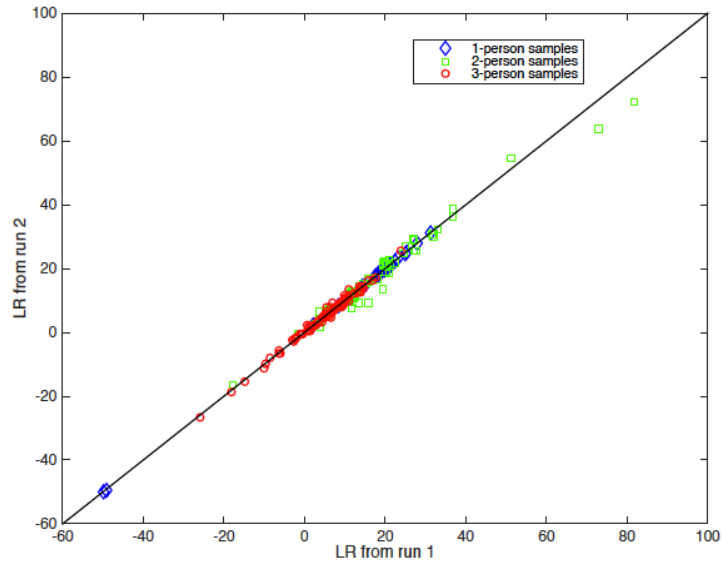
Figure 4.16: The results of 2 different runs on the 1-, 2- and 3-person samples at the 20s injection.

Runtime

The running time of MatchIt increases as the number of contributors increases. For single source samples, the average running time was 7.75 minutes, and it increased to 49.64 and 150.24 minutes for 2- and 3-person samples, respectively. The calculations were done using 8 cores on an Intel E3 3.4GHz processor.

4.4 DISCUSSION

We have developed a computational algorithm - MatchIt - to compute the LR for a POI and the distribution of the LR given an evidence profile. MatchIt also provides a summary statistic of the distribution in the form of the p-value of the LR. There is no 'correct' value for the LR to compare the results with − the LR computed is dependent

upon the model used. Hence, we need to perform tests to check if the results from the model are reasonable, i.e. if MatchIt computes a 'large' LR in the case of a true contributor to the mixture and a 'small' LR for a non-contributor.

MatchIt was tested on 306 sample files containing between 1 and 3 contributors with total template mass ranging from 0.016ng to 1ng, injected using 3 injection times. Across all samples tested, the mean of the LRs was 10^{13}; the lowest and highest LRs were 10^{-77} and 10^{102} respectively for minor (or single) contributors ranging in mass from ~0.013 to 0.33 ng. In instances where MatchIt produced a small LR, the template DNA mass from the contributor was low, resulting in extreme dropout and stutter.

We observed that the presence of additional contributors impacted the LR for a POI. Specifically, we observed that the addition of interference contributors decreased the LR for a POI that contributed the same template mass to a 1-, 2- and 3-person sample, where, the LR was highest for the 1-person sample, followed by the 2-person sample and then the 3-person sample.

For comparison, we calculated a binary inclusion statistic for the single source samples in the form of the RMP. We observed that for samples for which we could calculate a RMP, the LR from MatchIt was close to 1 / RMP, as expected. The smallest difference observed between log(LR) and log(1/RMP) was 0.0001, while the largest difference was 11, with log(LR) being smaller.

In order to compare MatchIt with other published results we focus on 2-person mixture cases. We note that in order to test MatchIt, the template masses used for the 2-person samples set in this studied varied from 0.2 to 1 ng and the mixture ratios varied from 1:1 to 1:49, where the lowest DNA mass of the minor contributed ~0.013 ng of DNA. For all of the 122 2-person mixture files tested in this work, the LR ranged from

10^{-32} to 10^{82}. Taylor et al [40] performed a similar study using an alternative continuous interpretation scheme and observed LRs in the range 10^8 to 10^{23} on 127 artificially constructed 2-person mixtures where the total template mass varied from 0.1 to 0.5 ng and the mixture ratios from 1:1 to 1:5, resulting in a minor contributor mass of at least 0.017 ng. TrueAllele [37] was tested on 101 casework profiles containing between 2 and 4 contributors. The mixture weight for the matched genotype varied from 0.05 to 0.95. The LRs ranged from 10 to 10^{23}, with a mean of 10^{11}.

In addition to calculating the LR, MatchIt also makes available to the user a distribution on the LR. The LR distribution can be useful while performing error rate calculations such as $\Pr(LR > 1|H_d)$ [41]. MatchIt also provides a summary statistic of the LR distribution in the form of the *p*-value. The *p*-value, which can be thought of as the false positive rate, is the probability that a randomly picked person results in a LR greater than the LR observed for the POI. In this study, MatchIt calculated the *p*-value by generating 10^9 random genotypes according to the allele frequency distribution and computing the fraction of those genotypes that have a LR greater than the POI's LR. All the 1-person samples in the testing set had a *p*-value of 10^{-9}, the lowest possible in our testing, indicating that MatchIt was able to identify the individual as a contributor with great certainty. With regard to the 2- and 3-person samples, the *p*-value was 10^{-9} when the LR was greater than 10^8.

One desirable quality of a mixture interpretation method is specificity: How often would a method misidentify a non-contributor? In our testing, MatchIt produced a LR distribution for a test sample by sampling 10^9 genotypes, giving $\sim 3 \times 10^{11}$ random genotypes with a LR from MatchIt. For each experimental sample, we calculated the Tippett test statistic $\Pr(LR > 1|H_d)$ [41]. We observed that the probability of a random

non-contributor resulting in a LR > 1 is low for small n and increases as n increases. The average $\Pr(LR > 1|H_d)$ for the 1-person samples was 8.56×10^{-12} and it increased to 0.001 and 0.002 for the 2- and 3-person samples, respectively. The minimum LR observed for the random contributors (sampled over the set R_1, see Materials and Methods) was 10^{-2581}, while the maximum was 10^{31}. Perlin et al [37] tested TrueAllele by computing the LR for 10,000 random contributors on each of the 101 samples in their testing set and found 133 instances in which the LR was greater than 1. The highest LR was 10^4 and lowest was 10^{-30}. Similarly, Taylor et al observed a maximum LR of 10^4 and 1168 LRs greater than 1 after testing 57, 609 non-contributors on a 3-person mixture [40].

Another desirable aspect of a sampling algorithm is reproducibility from one run to the next on the same sample, i.e. obtaining similar results. MatchIt was run twice on all the samples in the testing set and the results were repeatable (The Pearson correlation coefficient was greater than 0.9 for the samples at all 3 injection times). Additionally, testing on one 2-person and one 3-person sample confirmed the repeatability of MatchIt after 5 runs. TrueAllele variability from run to run was tested on duplicate runs and on average the LR varied by a factor of 2 ($10^{0.305}$) standard deviations [37]. Taylor et al [40] showed their model to give reproducible results over 10 runs on a single 2-person mixture, with a standard deviation of ~3% of the mean.

CHAPTER 5

CONCLUSION

Since the discovery of VNTRs and their initial application for human identification purposes in the 1980s, forensic DNA profiling has grown by leaps and bounds. A vast increase in the sensitivity of the DNA typing method has led to a stage where one can generate a full STR profile at 15 or more loci starting with less than 1ng of DNA. Simultaneously, DNA evidence has gained widespread acceptance by the public and the criminal justice system. Whenever available, DNA evidence is used regularly in the courtroom and plays a decisive part in the trial, leading to a situation where the conviction or exoneration of the suspect hinges on the method used for interpreting the DNA evidence. This bestows a great responsibility on the DNA analyst to not only perform mixture interpretation in a thorough and rigorous fashion. As Buckleton and Curran aptly put it in [11]: "*There is a considerable aura to DNA evidence. Because of this aura it is vital that weak evidence is correctly represented as weak or not presented at all.*"

While DNA evidence can be a valuable investigative tool to have, complex mixtures are not interpreted in many US laboratories due to the complications and uncertainty associated with mixture analysis. Even if a mixture is picked for analysis, due to the non-availability of interpretation standards, the analyst interpreting the DNA profile makes subjective decisions regarding the number of contributors and the calculation of the Likelihood Ratio. Findings by the NIST MIX 13 study demonstrate the complexities associated with mixture interpretation and highlight the need to produce

processes that give rise to consistent results – both within and between laboratories.

The goal of this research is to reduce the uncertainty associated with mixture analysis, aid the analyst in making decisions along the analysis pipeline and make mixture interpretation a more consistent process.

Current methods for mixture interpretation do not use the full data available in the signal and waste some information, which can be detrimental to the interpretation process. Generally, the heights of the peaks are used only while applying thresholds and while calculating match statistic using a 'restricted' approach [9]. While this could work with simple mixtures and samples with optimal template DNA, this procedure is problematic for low template samples because of reduced peak heights, elevated stutters ratios and reduced peak height ratios [3]. It has also been shown that quantitative methods that use peak height information can produce useful match statistics at lower DNA levels than qualitative methods [13]. It is, therefore, imperative to understand completely the electropherogram that makes up the DNA profile in order to work with it during the mixture interpretation process.

To address this, we studied the distribution of the heights of the various types of peaks (true, noise, forward and reverse stutter) in single source samples with known genotypes amplified from a range of low-to-medium template DNA masses. We tested the peak heights using the Kolomogorov-Smirnov test against three distributions – normal, lognormal and Gamma. We observed that while the normal distribution had more rejections than the lognormal and the gamma distributions for all four types of peaks, all three were reasonable distributions to describe the data. We also examined how the peak heights varied with the target DNA mass, and developed models to describe the variation. Studying the characteristics of the peak heights constituted the first aim of our research.

While performing mixture interpretation on a sample, an assumption about the number of contributors to the sample is necessary. An assumption of an incorrect number of contributors can lead to non-reasonable match statistics that has adverse effects on the mixture interpretation process [20, 21]. The underlying number of contributors to the sample might be hard to identify based on the sample's profile. The method commonly used for identifying the number of contributors in forensic laboratories currently is Maximum Allele Count (MAC). This method only gives us the *minimum number of contributors* needed to explain the profile obtained. The actual number of contributors to the sample might in fact be very different because of sharing of alleles between the various individuals [23]. A number of methods that also use the allele frequencies in addition to the number of alleles observed have been developed [25, 26, 28].

To satisfy the second aim of our research, we developed a tool to determine the number of contributors to a DNA sample − NOCIt. In addition to using the allele frequencies, NOC*It* works upon the heights of the peaks observed in the signal. NOC*It* also accounts for commonly observed artifacts like dropout, noise and stutter. NOC*It* was tested on 278 experimental samples and 40 simulated mixtures containing between 1 and 5 contributors. NOC*It* correctly identified the number of contributors in 83% of the experimental samples and 85% of the simulated mixtures. Moreover, in 95% of the experimental samples and in all the simulated mixtures, NOC*It* came up with a probability of at least 1% for the underlying number of contributors, highlighting the utility of the method in identifying the region in which the number of contributors is likely to be in for complex samples that can be hard to analyze. The accuracy of NOC*It* was higher than that of two of the existing methods (Maximum Allele Count and Maximum Likelihood Estimator [26]) to identify the number of contributors. Our results indicate

that using the quantitative peak height information in the signal, in addition to the qualitative information, results in a better estimate on the number of contributors to the stain.

One of the issues associated with NOCIt is the large running time – it takes ~10 hours to analyze a profile up to 5 contributors. We were able to reduce the running time of NOCIt by developing a new method to compute the APP on the number of contributors using an importance sampling algorithm. Instead of sampling the genotypes using the allele frequencies, the importance sampling method samples the genotypes based on the peak height distribution observed in the signal. The new version of NOCIt had an accuracy similar to the previous version, and as anticipated we were able to reduce the running time of the software – the importance sampling algorithm takes ~ 6 hours to run on a profile for up to 5 contributors.

Most forensic laboratories use an analytical threshold (such as 50 or 100 RFU) on the epg to filter out the baseline noise peaks – only peaks above the threshold are considered to be 'reliable' allelic peaks and used in subsequent analyses. While this methodology is suitable for samples amplified from an optimal DNA mass, there is a significant risk of losing true allelic peaks with low template samples that contain reduced peak heights. Erroneously removing allelic peaks would jeopardize the mixture interpretation protocol and potentially lead to incorrect decisions to include or exclude an individual and/or misleading match statistics. To illustrate the effect of application of an analytical threshold and the ensuing loss of data, we are working on an algorithm for NOCIt that incorporates the application of an analytical threshold into its probability calculation. Through analyzing a sample with NOCIt by applying different levels of

thresholds (for e.g. 0, 50 and 100 RFU), we would obtain an improved understanding of the effect of thresholding on low template profiles.

Other areas for future work on NOC*It* include testing it on mixtures with related contributors, samples obtained from touched items, samples with known contributors and samples with contributors from a population that is different from the one used for allele frequency data.

After the number of contributors to a DNA profile has been identified, mixture interpretation is carried out by resolving the sample into the constituent genotypes. Typically, in U.S. crime laboratories mixture interpretation is carried out only for samples with up to 4 contributors, because the number of possible genotypes increases quickly with increasing number of contributors. Resolving mixed DNA samples is prone to error as excluding a potential genotype results in incorrect interpretation of the sample's profile. The Likelihood Ratio (LR) statistic is the preferred method of choice to measure of the weight of evidence in forensic DNA mixture interpretation. In addition to 'binary' methods (which assign a probability of 0 or 1 to genotypes based on presence or absence of alleles) [33] and 'partially continuous' methods (which incorporate probabilities for dropout and drop-in) [34, 35] 'fully continuous methods' to compute the LR have been developed [36, 37, 38, 39, 40]. Fully continuous methods do not waste any information in the signal and employ probabilistic genotyping by modeling the peak heights. Computational methods that compute the LR distribution have recently garnered attention [35, 41, 42, 43, 44]. The p-value is a summary statistic of the LR distribution that provides the probability that a randomly picked person from the population would give rise to an LR at least as large as the one observed for the person of interest [35, 42, 43, 44]. It can be used to control the false positive rate of the

classification system of including/excluding suspects as contributors and will be useful when the analyst wants to compute the probability of a random non-contributor giving rise to an LR greater than the one observed for the suspect.

As the third aim of our research, we have brought together a continuous method (MatchIt) to compute a continuous LR and its distribution, along with a *p*-value. To date, the only systems available to compute both the LR and *p*-value used semi-continuous models. Instead of deconvolving the sample into the constituent genotypes, Match*It* directly computes the LR for a specified person. MatchIt also computes a *p*-value to add meaning to the LR, whose significance might otherwise be hard to understand. The method was tested on 306 1-, 2- and 3-person sample files containing between 0.016 and 1 ng of DNA, injected for 5, 10 and 20 sec. We found that the amount of template DNA from the contributor had an impact on the LR – small LRs arose from contributors with low template masses, indicating that high levels of dropout and stutter could decrease the probability of the evidence under the prosecution's hypothesis even for true contributors. We observed that the presence of additional contributors impacted the LR for a POI - the addition of interference contributors decreased the LR for a POI that contributed the same template mass to a 1-, 2- and 3-person sample, where, the LR was highest for the 1-person sample, followed by the 2-person sample and then the 3-person sample. Since we used 10^9 samples to calculate the *p*-value, the lowest possible *p*-value that can be achieved is 10^{-9}, and this was obtained in all the cases where the LR was greater than 10^8. The results from MatchIt were found to be repeatable after duplicate runs on all the samples in the testing set.

There is no universally accepted way to compute the LR. When implementing fully continuous methods, deciding which model to use to describe the data is not trivial,

and it is desirable to choose a model that best approximates the underlying distribution. The methods that have been published in literature use different, but arguably reasonable, models to compute the LR - for e.g., Puch-Solis et al. [39] use a gamma distribution to model the peak heights, while Perlin et al. use a normal distribution [37]. Since the probability of the two hypotheses being compared depends upon the model used, the LR changes under different modeling assumptions. This can be problematic when a suspect is included under one modeling assumption but excluded under another. It has been suggested that probabilistic approaches to interpretation may be able to aid in obtaining consistent results in the form of the match statistic. We plan to investigate this premise by evaluating the impact of changing a continuous model on the LR and its distribution.

Although theoretically the MatchIt algorithm can handle any number of contributors, practical implementation limits the maximum number of contributors that can be evaluated to 3. Future work involves decreasing the running time to increase the maximum number of contributors that can be evaluated. Future work also involves developing MatchIt to handle different numbers of contributors in the prosecution and the defense hypotheses, account for known (and possibly related contributors) and testing the robustness of the method to the population allele frequencies used.

To satisfy our goal of enhancing the mixture interpretation process we have developed two computational tools: NOCIt, to determine the number of contributors to a DNA sample; and MatchIt, to calculate a LR and its distribution, along with a p-value. Both tools use the qualitative and the quantitative information captured in the signal and perform their calculation using sound models that were selected after testing on experimental data. We believe that the computational tools we developed in this research

would be beneficial to DNA analysts and aid them in the mixture interpretation process. We envision both NOCIt and MatchIt being used in forensic laboratories not only in the US but also all over the world. They would fit into the mixture interpretation pipeline performing the roles specified in Figure 5.1.
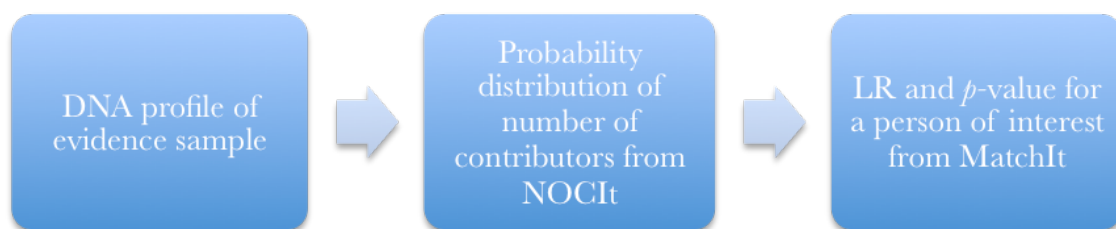

Figure 5.1: Role of NOCIt and MatchIt in the mixture analysis pipeline.

APPENDIX

The problem at hand is to compute $\Pr(E_l|\boldsymbol{G}=\boldsymbol{g},\boldsymbol{\Theta}=\boldsymbol{\theta},N=n)$, which is the probability of observing the evidence at a locus l, given the mixture proportions, the genotypes of the contributors and the number of contributors. Each of these terms is described in detail below:

N is a positive integer that represents the number of contributors.

$\boldsymbol{G}$ is a set with components G_i, the genotype of each contributor $i \in \{1, \dots, n\}$:

$$\boldsymbol{G} = \{\mathsf{G}_1, \mathsf{G}_2, \dots \mathsf{G}_n\}.$$

Each genotype G_i is a set containing two elements:

$$\mathsf{G}_i = \left\{A_{i_1}, A_{i_2}\right\},$$

where $A_{i_1} \in \mathcal{A}$ and $A_{i_2} \in \mathcal{A}$ are the alleles of contributor i at locus l. $\mathcal{A}$ is a finite set of real numbers that represents all possible alleles (i.e. numbers of repeats), e.g., 9, 9.3, 10, etc., for the STRs at locus l.

$\boldsymbol{\Theta}$ is an n-dimensional vector with components Θ_i, the mixture proportion of each contributor $i \in \{1, \dots, n\}$:

$$\boldsymbol{\Theta} = \{(\Theta_1, \Theta_2, \dots \Theta_n): \textstyle\sum_{i=1}^{n} \theta_i = 1 \land \Theta_i \in \mathbb{R}_{>0} \forall i = 1, \dots n\},$$

where $\mathbb{R}_{>0}$ refers to the set of real numbers greater than 0.

To describe the signal E_l, we first define $\boldsymbol{x}$ to be the set that contains all repeat values of STRs at locus l that correspond to allele, reverse stutter and forward stutter positions:

$$\boldsymbol{x} = \mathcal{A} \cup \{a - 1: a \in \mathcal{A}\} \cup \{a + 1: a \in \mathcal{A}\}.$$

Let d be the cardinality of set $\boldsymbol{x}$: $d = |\boldsymbol{x}|$. Now, we can represent the evidence E_l as a d-

dimensional vector $\boldsymbol{H} \in \mathbb{Z}_{\geq 0}^{d}$ that consists of the peak heights observed in the signal, which are non-negative integers:

$$E_l = \boldsymbol{H} = (h_x : x \in \boldsymbol{x} \wedge h_x \in \mathbb{Z}_{\geq 0}),$$

where h_x is the height of allele x.

Next, we define two sets $\widehat{\boldsymbol{a}}$ and $\boldsymbol{m}$ that the signal E_l depends upon. For a given genotype sample $\boldsymbol{g}$, let $\widehat{\boldsymbol{a}}$ be the set of all alleles in the genotypes of the contributors:

$$\widehat{\boldsymbol{a}} = \bigcup_{i=1}^{n} g_i.$$

Let the total template DNA mass be M. Each allele $a \in \widehat{\boldsymbol{a}}$ in the genotype of the contributors has some mass m_a associated with it, that is the sum of the template masses of all the contributors who have that allele:

$$m_a = \sum_{1 \leq i \leq n : a \in g_i} \mathrm{M} * \Theta_i * \rho_i,$$

where

$$\rho_i = \begin{cases} 1, \text{if contributor } i \text{ is heterozygous at locus } l; \\ 2, \text{if contributor } i \text{ is homozygous at locus } l. \end{cases}$$

We define $\boldsymbol{m}$ to be the set containing all the masses of the alleles:

$$\boldsymbol{m} = \{m_a : a \in \widehat{\boldsymbol{a}}\}.$$

It is sufficient to condition the evidence E_l upon the set of alleles $\widehat{\boldsymbol{a}}$ and their masses $\boldsymbol{m}$, since the information captured in them is enough to compute the probability of observing the peak heights, even though the genotypes $\boldsymbol{g}$ and the mixture proportions $\boldsymbol{\theta}$ contain additional information.

$\Pr(E_l | \boldsymbol{G} = \boldsymbol{g}, \Theta = \boldsymbol{\theta}, N = n)$

$\quad = \Pr\left(\boldsymbol{H} = \boldsymbol{h} \middle| \boldsymbol{G} = \boldsymbol{g}, \Theta = \boldsymbol{\theta}, N = n, \widehat{\boldsymbol{A}} = \widehat{\boldsymbol{a}}, \boldsymbol{M} = \boldsymbol{m}\right)$

$$= \Pr\left(H = h \middle| \widehat{A} = \widehat{a}, M = m\right) \tag{1}$$

Let $\widehat{r}$ and $\widehat{f}$ be the sets containing the alleles in the reverse and forward positions of observed alleles in $\widehat{a}$, respectively:

$$\widehat{r} = \{a - 1 : a \in \widehat{a} \wedge h_a > 0\};$$

$$\widehat{f} = \{a + 1 : a \in \widehat{a} \wedge h_a > 0\}.$$

In our model, we make the simplifying assumption that when forward stutter is present along with one or more alleles, the effect of forward stutter is masked by the allele(s) and its contribution to the peak height is negligible. Hence, a combination of allele and forward stutter is effectively treated as being exclusively allelic in origin. One way to refine our model would be to remove this simplifying assumption and incorporate forward stutter into the calculation even when it is combined with one or more alleles.

We define the following sets as part of our model:

$a = \widehat{a} \setminus \widehat{r}$ is the set containing alleles that are in the genotype of the contributors and have no stutter effects (the contribution of forward stutter to an allelic peak is ignored);

$r = \widehat{r} \setminus (\widehat{a} \cup \widehat{f})$ is the set containing alleles with only reverse stutter contributions;

$f = \widehat{f} \setminus (\widehat{a} \cup \widehat{r})$ is the set containing alleles with only forward stutter contributions;

$t = \widehat{a} \cap \widehat{r}$ is the set containing alleles with allelic and reverse stutter contributions (again, the contribution of forward stutter in the presence of alleles is ignored);

$u = (\widehat{r} \cap \widehat{f}) \setminus \widehat{a}$ is the set containing alleles with reverse and forward stutter contributions;

$b = x \setminus (\widehat{a} \cup \widehat{r} \cup \widehat{f})$ is the set containing all the baseline noise alleles.

All the alleles that give rise to the signal can be classified under one of the six sets defined above. The set x can be expressed as the disjoint union of these six sets:

$$x = a + r + f + t + u + b.$$

Since the set x comprises of all the alleles giving rise to the signal, we have:

$$\Pr(H = h | \widehat{A} = \widehat{a}, M = m)$$

$$= \Pr(\bigcup_{x \in x} H_x = h_x | \widehat{A} = \widehat{a}, M = m)$$

$$= \Pr(\bigcup_{a \in a} H_a = h_a | \widehat{A} = \widehat{a}, M = m) \Pr(\bigcup_{x \in (x \setminus a)} H_x = h_x | \bigcup_{a \in a} H_a = h_a, \widehat{A} =$$

$$\widehat{a}, M = m) \tag{2}$$

$\Pr(\bigcup_{a \in a} H_a = h_a | \widehat{A} = \widehat{a}, M = m)$ is the probability of observing the heights of peaks that contain only allelic contributions and no stutter effects. The probability of the heights of these peaks depends only upon the mass of the peak and not upon the other alleles or their masses. Hence we have:

$$\Pr(\bigcup_{a \in a} H_a = h_a | \widehat{A} = \widehat{a}, M = m)$$

$$= \prod_{a \in a} \Pr(H_a = h_a | M_a = m_a).$$

The peak heights in the signal are expressed in RFUs as integers and hence take discrete values. However, we have modeled the peak heights using the normal distribution, which is a continuous distribution. Therefore, we approximate the probability of observing a peak at a particular height as the density of the normal distribution at that height. Let p and P be the PDF and the CDF of the normal distribution, respectively. If $\alpha > 0$ is the height of an allelic peak with mass m_α, then the probability of the peak having height α is approximated as:

$$P(\alpha + 0.5; \mu_{m_\alpha}, \sigma_{m_\alpha}) - P(\alpha - 0.5; \mu_{m_\alpha}, \sigma_{m_\alpha})$$

$$= \int_{\alpha-0.5}^{\alpha+0.5} p(\alpha; \mu_{m_\alpha}, \sigma_{m_\alpha}) \, d\alpha$$

$$\cong p(\alpha; \mu_{m_\alpha}, \sigma_{m_\alpha}),$$

where μ_{m_α} and σ_{m_α} are the mean and standard deviation that correspond to a mass of m_α. The dependence of the mean and standard deviation on the DNA mass of the allele is specified in Table 2.5. Let q_{true} be the dropout probability function. This dropout function depends upon the DNA mass of the allele and is described in Table 2.5. Now we can compute the probability of the height of an allelic peak as follows:

$$\mathrm{Pr}\,(H_a = h_a | M_a = m_a)$$

$$= \begin{cases} p(h_a; \mu_{m_a}, \sigma_{m_a})\,(1 - q_{true}(m_a)), if\ h_a > 0; \\ q_{true}(m_a), otherwise. \end{cases} \quad (3)$$

Here, we are calculating the probability of dropout using the DNA mass of the allele if the peak is not detected. If the peak is detected, we calculate the probability of observing the height of the peak under the normal distribution assumption, using the mean and standard deviation corresponding to the DNA mass of the allele. The latter case also involves calculating the probability of not having a dropout.

$\mathrm{Pr}\,(\cup_{x \in (x \backslash a)} H_x = h_x | \cup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m)$, the second term in equation **(2)**, is the probability of observing the heights of all the peaks other than the ones in a. This can be written as:

$\mathrm{Pr}\,(\cup_{x \in (x \backslash a)} H_x = h_x | \cup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m)$

$= \mathrm{Pr}(\cup_{r \in r} H_r = h_r \,|\, \cup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m) \times$

$\mathrm{Pr}(\cup_{f \in f} H_f = h_f \,|\, \cup_{r \in r} H_r = h_r, \cup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m) \times$

$\mathrm{Pr}(\cup_{x \in (x \backslash (a \cup r \cup f))} H_x = h_x \,|\, \cup_{f \in f} H_f = h_f, \cup_{r \in r} H_r = h_r,)$

$$\cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m) \tag{4}$$

$\Pr\left(\cup_{r\in r} H_r = h_r \mid \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$ is the probability of observing

the reverse stutter peaks. Conditioned upon the allele heights, the stutter peak heights are

independent of each other:

$$\Pr\left(\cup_{r\in r} H_r = h_r \mid \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \prod_{r\in r} \Pr\left(H_r = h_r \mid \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right).$$

We have modeled stutter peak heights using the stutter ratio. Hence the height of

a stutter peak depends upon the height of the parent peak. For a reverse stutter allele r,

$r + 1$ is the parent allele causing reverse stutter. Let q_{stut} (which depends upon the mass

of the parent allele and is described in Table 2.5) be the rate of non-detection of stutter.

To calculate the probability of a stutter peak height, we make the same approximation

that we did for the allelic peaks, with the modification that the stutter peak height H_r is

coming from a distribution scaled according to the height of the parent peak:

$$H_r \sim \mathcal{N}(\mu h_{r+1}, \sigma h_{r+1}).$$

We calculate the probability of observing the reverse stutter peak heights as

follows:

$$\Pr\left(H_r = h_r \mid \cup_{x\in x} H_x = h_x, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \Pr\left(H_r = h_r \mid H_{r+1} = h_{r+1}, M_{r+1} = m_{r+1}\right)$$

$$= \begin{cases} p\left(h_r; h_{r+1}\mu_{m_{r+1}}, h_{r+1}\sigma_{m_{r+1}}\right)\left(1 - q_{stut}(m_{r+1})\right), if\ h_r > 0; \\ q_{stut}(m_{r+1}), otherwise. \end{cases} \tag{5}$$

The term $\Pr\left(\cup_{f\in f} H_f = h_f \mid \cup_{r\in r} H_r = h_r, \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$ in (4)

corresponds to the probability of observing the forward stutter peak heights. We notice

that the forward stutter heights do not depend upon the reverse stutter heights and are independent of each other conditioned on the heights of the alleles. Hence we get:

$$\Pr\left(\cup_{f\in f} H_f = h_f \mid \cup_{r\in r} H_r = h_r, \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \prod_{f\in f} \Pr\left(H_f = h_f \mid \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right).$$

For a forward stutter allele f, $f-1$ is the parent allele causing forward stutter. We calculate the probability of observing the height of a forward stutter peak in a manner similar to that used for the reverse stutter heights:

$$\Pr\left(H_f = h_f \mid \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \Pr\left(H_f = h_f \mid H_{f-1} = h_{f-1}, M_{f-1} = m_{f-1}\right)$$

$$= \begin{cases} p\left(h_f; h_{f-1}\mu_{m_{f-1}}, h_{f-1}\sigma_{m_{f-1}}\right)\left(1 - q_{stut}(m_{f-1})\right), if\ h_f > 0; \\ \qquad\qquad q_{stut}(m_{f-1}), otherwise. \end{cases} \qquad (6)$$

$$\Pr\left(\cup_{x\in(x\setminus(a\cup r\cup f))} H_x = h_x \mid \cup_{f\in f} H_f = h_f, \cup_{r\in r} H_r = h_r, \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

is the third term in equation (**4**) and represents the probability of observing the heights of all the peaks not in a, r and f. This probability is independent of the stutter peak heights and hence those terms can be eliminated:

$$\Pr\left(\cup_{x\in(x\setminus(a\cup r\cup f))} H_x = h_x \mid \cup_{f\in f} H_f = h_f, \cup_{r\in r} H_r = h_r, \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \Pr\left(\cup_{x\in(x\setminus(a\cup r\cup f))} H_x = h_x \mid \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right).$$

$x \setminus (a \cup r \cup f)$ can be divided into three disjoint sets: b, which contains the noise alleles; u, which contains alleles that have reverse and forward stutter effects; and t, which contains allele and reverse stutter effects.

$$\Pr\left(\cup_{x\in(x\setminus(a\cup r\cup f))} H_x = h_x \mid \cup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \Pr\left(\bigcup_{t \in t} H_t = h_t \mid \bigcup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right) \times$$

$$\Pr\left(\bigcup_{u \in u} H_u = h_u \mid \bigcup_{t \in t} H_t = h_t, \bigcup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right) \times$$

$$\Pr\left(\bigcup_{b \in b} H_b = h_b \mid \bigcup_{u \in u} H_u = h_u, \bigcup_{t \in t} H_t = h_t, \bigcup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right) \quad (7)$$

Notice that the heights of the noise alleles b depend only upon the template DNA mass M and are independent of the heights of other peaks. Let q_{noise} be the probability of not detecting a noise allele. Hence we have:

$$\Pr\left(\bigcup_{b \in b} H_b = h_b \mid \bigcup_{u \in u} H_u = h_u, \bigcup_{t \in t} H_t = h_t, \bigcup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \prod_{b \in b} \Pr\left(H_b = h_b \mid \mathcal{M} = M\right)$$

$$= \begin{cases} p(h_b; \mu_M, \sigma_M)\,(1 - q_{noise}), & if\ h_b > 0; \\ q_{noise}, & otherwise. \end{cases} \quad (8)$$

Next, we look at the second term in equation **(7)**. The set u is a combination of two events: reverse stutter and forward stutter. For an allele $u \in u$, $u + 1$ is the parent allele causing reverse stutter at u and $u - 1$ is the parent allele causing forward stutter at u. We assume that the two events reverse and forward stutter occur independent of each other. If a peak is observed at u, then both or any one of the two events might have occurred. On the other hand, if a peak is not observed, then there is a dropout of both events. Therefore we get:

$$\Pr\left(\bigcup_{u \in u} H_u = h_u \mid \bigcup_{t \in t} H_t = h_t, \bigcup_{a \in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\right)$$

$$= \prod_{u \in u} \Pr\left(H_u = h_u \mid H_{u+1} = h_{u+1}, M_{u+1} = m_{u+1}, H_{u-1} = h_{u-1}, M_{u-1} = m_{u-1}\right)$$

$$
=
\begin{cases}
p\big(h_u; h_{u+1}\mu_{m_{u+1}} + h_{u-1}\mu_{m_{u-1}}, h_{u+1}\sigma_{m_{u+1}} + h_{u-1}\sigma_{m_{u-1}}\big)\big(1 - q_{stut}(m_{u+1})\big)\big(1 - q_{stut}(m_{u-1})\big) \\
\quad + p\big(h_u; h_{u+1}\mu_{m_{u+1}}, h_{u+1}\sigma_{m_{u+1}}\big)\big(1 - q_{stut}(m_{u+1})\big)q_{stut}(m_{u-1}) \\
\quad + p\big(h_u; h_{u-1}\mu_{m_{u-1}}, h_{u-1}\sigma_{m_{u-1}}\big)\big(1 - q_{stut}(m_{u-1})\big)q_{stut}(m_{u+1}), if\ h_u > 0; \\
q_{stut}(m_{u+1})q_{stut}(m_{u-1}), otherwise.
\end{cases}
\tag{9}
$$

The other term to compute in (7) is $\Pr\big(\bigcup_{t\in t} H_t = h_t \mid \bigcup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\big)$. t is the set containing alleles that have allele and reverse stutter effects. The probability of observing the height of an allele $t \in t$ depends on the mass of allele t and the mass and height of the parent allele $t + 1$ causing reverse stutter.

$$\Pr\big(\bigcup_{t\in t} H_t = h_t \mid \bigcup_{a\in a} H_a = h_a, \widehat{A} = \widehat{a}, M = m\big)$$

$$= \prod_{t\in t} \Pr\big(H_t = h_t | M_t = m_t, M_{t+1} = m_{t+1}, H_{t+1} = h_{t+1}\big)$$

To compute the probability of observing some height H_t, we use the same reasoning as the one used for H_u:

$$\Pr\big(H_t = h_t | M_t = m_t, M_{t+1} = m_{t+1}, H_{t+1} = h_{t+1}\big)$$

$$
=
\begin{cases}
p\big(h_t; \mu_{m_t} + h_{t+1}\mu_{m_{t+1}}, \sigma_{m_t} + h_{t+1}\sigma_{m_{t+1}}\big)\big(1 - q_{true}(m_t)\big)\big(1 - q_{stut}(m_{t+1})\big) \\
\quad + p\big(h_t; \mu_{m_t}, \sigma_{m_t}\big)\big(1 - q_{true}(m_t)\big)q_{stut}(m_{t+1}) \\
\quad + p\big(h_t; h_{t+1}\mu_{m_{t+1}}, h_{t+1}\sigma_{m_{t+1}}\big)\big(1 - q_{stut}(m_{t+1})\big)q_{true}(m_t), if\ h_t > 0; \\
q_{true}(m_t)q_{stut}(m_{t+1}), otherwise.
\end{cases}
\tag{10}
$$

Equations (1) through (10) enable the calculation of the probability of observing the peak heights in the signal, given the genotype and the mixture proportions of the contributors.

BIBLIOGRAPHY

1. J.M. Butler, Fundamentals of forensic DNA typing, first ed., Associated Press, 2009.

2. A.J. Jeffreys, V. Wilson, S.L. Thein, Hypervariable 'minisatellite' regions in human DNA, Nature. 314 (1985) 67-73. DOI: 10.1038/314067a0.

3. P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, T. Egeland, Genotyping and interpretation of STR-DNA: Low-tempalte, mixtures and database matches – Twenty years of research and development, Forensic Sci Int: Genet. 18 (2015) 100-117. DOI: 10.1016/j.fsigen.2015.03.014.

4. P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res. 24(14) (1996) 2807-2812.

5. P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. Forensic Science International 112 (2000) 17-40.

6. P.M. Schneider, R. Fimmers, W. Keil, G. Molsberger, D. Patzelt, W. Pflug, T. Rothämel, H. Schmitter, H. Schneider, B. Brinkmann, The German Stain Commission: recommendations for the interpretation of mixed stains, Int J Legal Med. 123 (2009) 1-5. DOI: 10.1007/s00414-008-0244-4.

7. M. Bill, P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy, J. Buckleton, PENDULUM – a guideline-based approach to the interpretation of STR mixtures, Forensic Science International. 148 (2005) 181-189.

8. P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawxzak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Science International 160 (2006) 90-101.

9. Scientific Working Group on DNA Analysis Methods (SWGDAM) (2000) Short Tandem Repeat (STR) interpretation guidelines. Forensic Science Communications 2.

10. J.M. Butler, Fundamentals of forensic DNA typing: Interpretation, first ed., Associated Press, 2014.

11. J. Buckleton, J. Curran, A discussion of the merits of random man not excluded and likelihood ratios, Forensic Sci Int Genet. 2 (2008) 343-348. DOI: 10.1016/j.fsigen.2008.05.005.

12. H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci Int Genet. 16 (2015) 172-180. DOI: 10.1016/j.fsigen.2014.11.010.

13. M. W. Perlin, A. Sinelnikov, An Information Gap in DNA Evidence Interpretation, PLoS one. 4(12) (2009). DOI: 10.1371/journal.pone.0008327.

14. C.M. Grgicak, Z.M. Urban, R.W. Cotton, Investigation of Reproducibility and Error Associated with qPCR Methods using Quantifiler Duo DNA Quantitation Kit. J Forensic Sci. 55 (2010) 1331-1339. DOI: 10.1111/j.1556-4029.2010.01460.x.

15. M.C. Cicero, C.M. Grgicak, Examination into the Applicability and Stability of a Single External Calibrator for Forensic DNA Quantification. In *NorthEastern Association of Forensic Scientists*, Cromwell, CT, 2013.

16. Applied Biosystems, AmpFlstr® Identifiler® Plus PCR Amplification Kit Users' Manual, first ed., 2006.

17. J. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci Int: Genet. 7 (2013) 296-304.

18. Chakravarti, Laha, and Roy, Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, 1967.

19. U.J. Mönich, K. Duffy, M. Médard, V. Cadambe, L.E. Alfonse, C. Grgicak, Probabilistic characterization of baseline noise alleles in STR profiles, Forensic Sci Int: Genet. 19 (2015) 107-122.

20. C.C.G. Benschop, H. Haned, L. Jeurissen, P.D. Gill, T. Sijen, The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures, Forensic Sci Int: Genet. 19 (2015) 92-99.

21. J. Bright, J. M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci Int: Genet. 12 (2014) 208-214.

22. J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci Int Genet. 1 (2007) 20-28. DOI: 10.1016/j.fsigen.2006.09.002.

23. D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures, J Forensic Sci. 50 (2005) 1361-1366.

24. J. Perez, A.A. Mitchell, N. Ducasse, J. Tamariz, T. Caragine, Estimating the number of contributors to two-, three- and four-person mixtures containing DNA in high template and low template amounts, Croat Med J. 52 (2011) 314-326. DOI: 10.3325/cmj.2011.52.314.

25. A. Biedermann, S. Bozza, K. Konis, F. Taroni, Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method, Forensic Sci Int Genet. 6 (2012) 689-696. DOI: 10.1016/j.fsigen.2012.03.006.

26. H. Haned, L. Pene. J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the Number of Contributors to Forensic DNA Mixtures: Does Maximum Likelihood Perform Better Than Maximum Allele Count? J Forensic Sci. 56 (2011) 23-28. DOI: 10.1111/j.1556-4029.2010.01550.x.

27. T. Egeland, I. Dalen, P.F. Mostad, Estimating the number of contributors to a DNA profile, Int J Legal Med. 117 (2003) 271-275. DOI: 10.1007/s00414-003-0382-7.

28. D.R. Paoletti, D.E. Krane, M.L. Raymer, T.E. Doom, Inferring the Number of Contributors to Mixed DNA Profiles, IEEE/ACM Transactions on Computational Biology and Bioinformatics. 9 (2012) 113-122. DOI: 10.1109/TCBB.2011.76.

29. Dimitri P. Bertsekas and John N. Tsitsiklis, Introduction to Probability, first ed., Athena Scientific, 2002.

30. National Research Council (NRCII) Committee on DNA Forensic Science, The evaluation of Forensic DNA Evidence. Washington, D.C: National Academy Press, 1996.

31. D.J.C. Mackay, Introduction to Monte Carlo Methods, Springer Netherlands, 1998.

32. ENFSI guideline for evaluative reporting in forensic science: Strengthening the Evaluation of Forensic Results across Europe (STEOFRAE) 2015.

33. H. Kelly, J. Bright, J. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Sci Int Genet. 6 (2012) 191-197. DOI: 10.1016/j.fsigen.2011.04.013.

34. D. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci Int Genet. 4 (2009) 1-10. DOI: 10.1016/j.fsigen.2009.03.003.

35. P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci Int Genet. 7 (2013) 251-263. DOI: 10.1016/j.fsigen.2012.11.002.

36. M.W. Perlin, M.W. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating TrueAllele DNA Mixture Interpretation. J Forensic Sci. 56:6 (2011) 1430-1447. DOI: 10.1111/j.1556-4029.2011.01859.x.

37. M. W. Perlin, K. Dormer, J. Hornyak, L. Schiermer-Wood, S. Greenspoon, TrueAllele Casework on Virginia DNA Mixture EvidenceL Computer and Manual Interpretation in 72 Reported Criminal Cases. PLoS ONE 9 (2014). DOI: 10.1371/journal.pone.0092837.

38. R.G. Cowell, S.L. Lauritzen, J. Mortera, Probabilistic expert systems for handling artifacts in complex DNA mixtures, Forensic Sci Int Genet. 5 (2011) 202-209. DOI: 10.1016/j.fsigen.2010.03.008.

39. R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, D. Balding, Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci Int Genet. 7 (2013) 555-563. DOI: 10.1016/j.fsigen.2013.05.009.

40. D. Taylor, J. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci Int Genet. 7 (2013) 516-528. DOI: 10.1016/j.fsigen.2013.05.011.

41. P. Gill, J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, J. Lambert, Interpretation of complex DNA profiles using empirical models and a method to measure their robustness, Forensic Sci Int Genet. 2 (2008) 91-103. DOI: 10.1016/j.fsigen.2007.

42. G. Dørum, ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, T. Egeland, Exact computation of the distribution of likelihood ratios with forensic applications, Forensic Sci Int Genet. 9 (2014) 93-101. DOI: 10.1016/j.fsigen.2013.11.008.

43. M. Kruijver, Efficient computations with the likelihood ratio distribution, Forensic Sci Int Genet. 14 (2015) 116-124. DOI: 10.1016/j.fsigen.2014.09.018.

44. D. Taylor, J. Buckleton, I. Evett, Testing likelihood ratios produced from complex DNA profiles. Forensic Sci Int Genet. 16 (2015) 165-171. DOI: 10.1016/j.fsigen.2015.01.008.

45. M. Kruijver, R. Meester, K. Slooten, *p*-values should not be used for evaluating the strength of DNA evidence. Forensic Sci Int Genet. 16 (2015) 226-231. DOI: 10.1016/j.fsigen.2015.01.005.

46. Ø. Bleka, G. Dørum, H. Haned, P. Gill, Database extraction strategies for low-template evidence. Forensic Sci Int Genet. 9 (2014) 134-141. DOI: 10.1016/j.fsigen.2013.11.006.

47. D.J.C. Mackay, Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003.

CURRICULUM VITAE

Harish Swaminathan

Center for Computational and Integrative Biology
Rutgers University - Camden, NJ 08102

(240) 460-6518 • harish.3689@gmail.com

EDUCATION AND TRAINING

Rutgers, the State University of New Jersey   Camden, New Jersey
Doctor of Philosophy in Computational and Integrative Biology, expected October 2015
Thesis: Computational Methods for the Interpretation of Forensic DNA Samples
Advisor: Desmond S. Lun

Anna University   Chennai, India
Bachelor of Technology in Biotechnology, 2011

POSITIONS

2014   **Intern**, National Institute of Science and Technology   Gaithersburg, Maryland
Development of computational methods to analyze population STR and NGS data

PUBLICATIONS

K. E. Wiens, **H. Swaminathan**, R. Copin, D. S. Lun, and J. D. Ernst. Equivalent T Cell Epitope Promiscuity in Ecologically Diverse Human Pathogens. *PLoS ONE* 8(8), 2013.

H. Swaminathan, C. M. Grgicak, M. Medard, and D. S. Lun. NOC*It*: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. *Forensic Sci. Int.: Genet.,* 16:172-180, 2015.

H. Swaminathan, A. Garg, C. M. Grgicak, M. Medard, and D. S. Lun. MatchIt: A Computational Tool for the Interpretation of STR Profiles. Submitted - *Forensic Sci. Int.: Genet.*

TALKS

Annual Meeting of North Eastern Association of Forensic Scientists. *NOCIt: A computational tool to infer the number of contributors to a forensic DNA sample.* Cromwell, Connecticut, September 2013.

AWARDS

NIJ Graduate Research Fellowship Program in Science, Technology, Engineering and Mathematics   January 2015 – June 2016
 Computational Methods for the Interpretation of Forensic DNA Samples
 Award amount: $34, 999

PATENTS

C. M. Grgicak, D. S. Lun, M. Medard, and **H. Swaminathan**. Method for determining the number of contributors to a forensic DNA sample. WO Application PCT/US2014/59503, October 7, 2014.

A. Garg, C. M. Grgicak, D. S. Lun, M. Medard, and **H. Swaminathan**. Systems and methods for computational interpretation of STR profiles, US provisional patent application.

TEACHING ASSISTANTSHIPS

50:198:213 *Data Structures*, Rutgers – Camden, Fall 2014
50:198:491/56:198:691 *Applied Probability*, Rutgers – Camden, Spring 2015