# Abstract

BEECHAM Jr., GARY WAYNE. Statistical methods for the Analysis of Forensic DNA Mixtures. (Under the direction of Bruce S. Weir.)

As forensic DNA typing technologies become more sensitive, there is increased chance of evidence being in the form of a mixture of two or more persons' DNA. Though there are methods established to calculate the strength of the evidence, such as the likelihood ratio used here, several statistical issues remain.

The first chapter introduces forensic DNA typing and some of the statistical concepts that are later expanded. The second chapter discusses a confidence interval (CI) for the forensic likelihood ratio (LR) in the mixture context. The LR, since it is based on the allele frequency estimates, is itself an estimate and consequently has variation. To report this uncertainty in the LR, we propose the use of a CI. A formula for the CI based on normal theory is herein explained and a computer program has been made available.

A method to include peak intensity information in the LR is discussed in the third chapter. Intensities, which are reflections of the amount of evidence material with a particular allele, give some information about which alleles are more likely to be from the same person in a DNA mixture. We propose the use of the maximum likelihood method to weight particular genotypes based on the observed intensities. A Normal model and a Dirichlet model are given and compared. In addition, this method can be applied to cases of allelic dropout.

In the final chapter, several different situations are explored. The standard cases considered are single and two-contributor evidence, the paternity index, and the consideration of relationship by pedigree. These four standard cases are used as an introduction to basic concepts, which are in turn used to discuss more complicated cases later in the chapter. The more complicated cases discussed include analysis of a paternity index from a mixture, relatives and mixtures, consideration of relatives in the presence of population substructure, and a case of canine parentage under varying degrees of relatedness.

# STATISTICAL METHODS FOR THE ANALYSIS OF FORENSIC DNA MIXTURES


BY

Gary Wayne Beecham Jr.


A DISSERTATION SUBMITTED TO THE GRADUATE FACULTY OF
NORTH CAROLINA STATE UNIVERSITY
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BIOINFORMATICS

RALEIGH

July 6, 2006

APPROVED BY:

| | |
|---|---|
| Dr. Bruce Weir (Chair) | Dr. Eugene Eisen |
| Dr. Sujit Ghosh | Dr. Dahlia Nielsen |

# Dedication

*Soli Deo Gloria*

Glory to God Alone

# Biography

Gary Wayne Beecham Jr was born on May 27, 1981, in Jacksonville, North Carolina, USA. He is the son of Gary Wayne and Susan Lynne Beecham. He has an older sister, Erin Celeste. While growing up he enjoyed exploring the forests surrounding his home, playing clarinet in the school band, judo, reading, and various church-related activities. In high school he got his first job at a local grocery store where he worked for three years.

In 1999 Gary graduated from Swansboro High School, in Swansboro, North Carolina. At the time, he intended to join the Marine Corps to play in the Marine Corps band. However, when that fell through he attended Coastal Carolina Community College in Jacksonville, NC and worked as a direct care aid at a home for profoundly handicapped individuals. While at the community college, he discovered a knack for mathematics and after two years of general education courses he transfered to North Carolina State University in Raleigh NC where he completed a degree in Statistics.

While studying statistics Gary became interested in genetics. As a senior, that interest led to a job working with Dr. Weir on forensic DNA problems. After graduating in 2003 with a BS in Statistics, the choice for graduate school was obvious, and the next fall Gary entered the Bioinformatics program at NCSU. There he continued working on forensic DNA problems with Dr. Weir as a research assistant. During this time he was nominated to the Sigma Xi honor society, and has provided expert advice to forensic scientists numerous times. Upon graduation, Gary will be accepting a postdoctoral position at Duke University, working in their Center for Human Genetics. He will be participating in genetic research in Alzheimer's disease.

# Acknowledgements

This work was funded through the Bioinformatics Research Center's training grant, through the National Institute of Health, and National Institute for Environmental Health Sciences. Office space, computers, and technical support were provided by the Bioinformatics Research Center and staff. Thank you to my committee, Dr. Weir, Dr. Ghosh, Dr. Nielsen, and Dr. Eisen for your helpful comments on this dissertation.

I am especially grateful to Dr. Weir, who took me under his wing when I was still an undergraduate, and continued to provide guidance and insight throughout my time here. His patience and wisdom as a mentor are unparalleled. I am honored to have been his student, and I look forward to working with him in the future.

Much thanks goes to the BRC crowd who put up with my questions on a regular basis. Thank you Errol Strain and Frank Mannino for help with R. I'm indebted to Oliver Serang for his Java and algorithm assistance – without him there would still be coding left to do. Thank you Amanda Hepler for letting me impose on your time with often lengthy discussions and brainstorming. Much thanks also to Raymond, Rachel, Christine, Hermonta, and many others I've consulted with over the years. A special thanks goes to JB for helping me get all the paperwork and admin red-tape sorted out, and for supplying comic relief and the occasional chocolate milkshake.

I am eternally grateful to my parents, Gary and Susan Beecham. I thank you for raising me "in the fear and admonition of the Lord," and especially for modeling that lifestyle to me. I thank you for being supportive, and encouraging me to be faithful with the gifts God has given me. Failure was always an option, but giving up and not trying never was. This dissertation is more a testament to your guidance and expectations as parents than any work I ever did. Thank you.

Thank you Ashley for more than could ever be put into words. Thank you for your patience and understanding when I had to work and thank you for your encouragement and goading when I didn't want to work. Thank you for your love and friendship. Thank you for the laughs and thank you especially for your prayers. God has used you to sustain

me in countless ways, countless times.

Lastly and most importantly, I thank my God and Lord, Christ Jesus for his grace to me. His blessings to me – food, clothing, family, friends, education, a degree – are all totally undeserved and I owe him a debt that can never be paid. To him be all glory and honor, forever and ever amen.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

The analysis of DNA evidence has proven useful in a number of forensic situations. It has been used to associate a suspect with the crime scene, weapons, and other evidence such as articles of clothing. It has been used to associate the victim with items in the possession of the suspect [1]. In sexual assault cases, DNA evidence can be used to associate a suspect directly with the victim [2]. DNA evidence has been used to help resolve parentage questions in paternity disputes, estate disputes, and alleged incest cases [3, 4, 5, 6, 7, 8, 9]. It has been used to associate remains with missing persons [10] and is increasingly being used to identify remains in mass-disaster situations, such as the World Trade Center attacks, or the 2004 Asian tsunami [11, 12, 13, 14, 15].

For our purposes, the analysis of forensic DNA evidence can be divided into two parts: the genotypic analysis and the statistical analysis. In the genotypic analysis DNA is extracted from some source (blood, semen, saliva, skin cells, suspect, victim, potential relative, etc), amplified, and then genotyped. The statistical analysis takes the results of the genotypic analysis (genotypes of the evidence, suspect, victim, etc) and attempts to assign a numerical weight to the strength of the evidence. This work is concerned only with the second, statistical analysis – specifically the statistical analysis of DNA mixtures. However, an understanding of the genotyping process will be helpful to understand later concepts.

## 1.1 Genotypic Analysis of Forensic DNA Evidence

The genotyping process begins by extracting the DNA from physical evidence or from contributions given by persons involved in the case. The physical evidence comes in many forms. It could be blood from a murder, semen from a sexual assault, perpetrator skin

cells under the fingernails of an assault victim, skin cells on a cigarette butt, skin cells in a stolen car, skin cells on the toothbrush of a now missing person. DNA material has also been extracted from burnt bones, insect guts, decomposed tissue, hair follicles, feces, shoe insoles, finger prints, teeth, cosmetics, fingernail clippings, dandruff, and nearly anything else that has had close contact with those involved in a crime. In criminal trials, the person typed is usually the suspect, but could also include the victim, a relative of the suspect, or other non-suspects placed at the scene of the crime. In the non-criminal context, the persons typed could be a mother, child, and alleged father in a paternity dispute, a son and alleged relative in an estate dispute, or the relative of a missing person. Contributions from those involved in the case are usually in the form of a cheek swab. Extraction processes differ depending on the nature of the source, and are not discussed here.

Though work has been done with Y-chromosome markers and mitochondrial sequences, most forensic analyses involve short-tandem repeat markers (STRs, also known as microsatellites). STRs are regions of DNA that have a short sequence that repeats many times, and whose different alleles at a locus differ in the number of these repeats. Most STRs are dinucleotide repeats (e.g., AG repeated many times). However, forensic markers are usually tetranucleotide repeats (e.g., the FGA locus has the repeat motif CTTT). Since different alleles have differing numbers of repeats, it follows that the length of the DNA is different for differing alleles. For example, if one allele has fifteen copies of the repeat motif and another has seventeen copies, the second allele will be eight basepairs longer than the first.

Once the DNA is extracted from the source material (physical evidence, cheek swab, etc), it is PCR amplified. In the forensic context, primers are used to amplify specific marker regions of the genome. These primers are also labeled using fluorescent dyes. This both increases the quantity of the DNA material and attaches labels to markers. These two factors allow the sample to be genotyped.

Capillary electrophoresis is an ideal method for genotyping the samples since the alleles differ in size. In capillary electrophoresis, the negatively charged DNA is pulled by

a positive charge through a buffer solution that is housed in a capillary. In this solution, the smaller fragments (i.e., the smaller alleles) will move through the solution quicker than the larger fragments. This separates the DNA by size along the capillary. The DNA is pulled past a laser, which excites the fluorescent dye labels that were bonded to the DNA in the PCR process. The excited dyes fluoresce at differing wavelengths depending on the dye, and the fluorescence is detected by a camera.

The end result is an electropherogram, which is a plot of the amount of fluorescence by time. In this plot, there will be peaks in fluorescence that indicate DNA material of different sizes. The peaks that appear earlier in time indicate the smaller alleles, while the later peaks are the larger alleles. In addition to the alleles, a size standard (material of known size) is electrophoresed in the same run. This standard enables the determination of the exact size in basepairs of the markers in the evidence, which in turn enables the determination of the genotype of the evidence.

For an excellent review of capillary electrophoresis in the forensic STR context, see Butler et al. [16].

## 1.2 Statistical Analysis of Forensic DNA Evidence

DNA evidence is often considered a sort of genetic photograph of a perpetrator. However, because the scientist only observes a limited number of markers, the evidence is much closer to a description than a photograph. Saying the evidence was left by someone with alleles A and B at locus 1 is analogous to saying that the perpetrator had brown eyes or black hair. It does not present a complete picture, but does give some identifying information about the person involved. As more markers are added, the description gets more informative, just as including information about age, weight, sex, height, clothing, and voice would add to the informativeness of a physical description.

This description analogy is useful to illustrate why a statistical analysis of the forensic DNA evidence is necessary. For example, if a physical description of a perpetrator was given and a suspect matched that description, it may be tempting to just say that the

suspect matches the description. This simple match or no-match analysis is insufficient because it ignores the informativeness of the description. Suppose that a witness reported that the perpetrator "had a head, two legs, and two arms." Certainly this is a description of the perpetrator, and the suspect probably matches the description, but this isn't exactly surprising. On the other hand, the description that the perpetrator "was seven feet tall, had orange hair, and was dressed like a pirate," is much more informative. If a suspect was found that matched that description, we would consider the evidence very strong against the suspect. This extreme example shows why it is necessary to report the weight of the evidence and not just that the genotype of a suspect matches the genotype of the evidence. Some evidence is stronger than others, and the statistical analysis enables a scientist to better report the strength for a particular case.

We advocate the use of the likelihood ratio to report the strength of the evidence. In the likelihood framework, the probability of observing the evidence under the prosecution hypothesis and then the defense hypothesis is calculated. The likelihood ratio is then the ratio of these two values, with the prosecution hypothesis being in the numerator and the defense hypothesis being in the denominator. This leads to the interpretation, "The evidence is LR times more likely when assuming the prosecution hypothesis than when assuming the defense hypothesis."

Some authors propose the use of the "random-man not excluded" approach (RMNE). In RMNE, the strength of the evidence is reported in terms of the probability that a random individual will be excluded as a contributor. There is some validity to this approach (see Laszlo Szabo's comment in Box 7.2 of [2]). However, the RMNE approach does not consider all the information. RMNE is based on the evidence alone and does not include the information in the suspect's genotype [17, 2]. Brenner provides additional comments showing the superiority of the likelihood ratio method [18].

Though the calculations necessary to produce a likelihood ratio will vary depending on the circumstances of the case, most likelihood ratios can be constructed in a similar fashion. First it is necessary to specify the evidence and the two alternative hypotheses. Determining the evidence is typically obvious. The evidence is generally that which the

prosecution is attempting to associate with the defendant. In an assault case it may be blood that the prosecution claims came from the suspect. In a rape case, it may be a mixture of the perpetrator's and victim's DNA that the prosecution claims came from the suspect and the victim. In a paternity case, the evidence may be the child's genotype, which the mother alleges came from her and the alleged father. In an estate dispute, the evidence may be the genotypes of two alleged half-siblings.

In the forensic context, a hypothesis is a statement about the source of a particular piece of evidence, a statement about the relatedness between two (or more) genotypes, or often both of these. For example, if it appears that a piece of evidence was contributed by a single person, and the genotype of the evidence matches the genotype of the suspect, then the prosecution may hypothesize that the evidence was contributed by the suspect. The defense would then typically hypothesizes that the evidence was contributed by a random individual. In the case of a DNA mixture, the prosecution may propose that the suspect and the victim contributed the evidence, while the defense would propose that the victim and an unknown person contributed the evidence.

Paternity disputes involve hypotheses regarding the relatedness of the child and the alleged father. The plaintiff generally proposes that the mother and alleged father are the true parents, while the defendant proposes that the mother and a random man are the true parents. Similarly, in estate disputes the plaintiff may propose that two persons are half-siblings while the defense may propose that the two persons are unrelated. These pairs of alternate hypotheses can potentially get much more complicated. For example, in Chapter 4 of this work, mixtures are discussed when a relative of the suspect is a potential unknown contributor. In this situation, the prosecution proposes that a mixture was contributed by the suspect and the victim, and the defense proposes that the mixture was contributed by the victim and a relative of the suspect.

After the evidence and hypotheses are defined, the probability of the evidence ($E$) given a hypothesis can be calculated. In some situations, such as the estate dispute, this is just the probability of two genotypes. However, in most cases, this probability needs to be rewritten as the probability of observing the evidence given the contributor genotypes

(**G**) and the hypothesis times the probability of the genotypes given the hypothesis. If there are multiple possible genotypes, this quantity is summed across all possible genotypes.

$$Pr(E|H) = Pr(E|\mathbf{G}, H)Pr(\mathbf{G}|H) = \sum_i Pr(E|G_i, H)Pr(G_i|H)$$

Most likelihood ratios can be formed by calculating the above under both hypotheses and then taking the ratio. Note that the values of the terms depend on the nature of the hypothesized relationships. For example, in a single contributor stain, the first term ($Pr(E|G, H)$) is one if the genotype of the alleged contributor matches the genotype of the evidence and zero otherwise. However, in a paternity situation, the first term is the probability of observing the genotype of the child given the genotypes of the mother and an alleged father, which can take the values zero, one-fourth, one-half, or one, depending on the genotypes of those involved. As shown in Chapter 3 of this work, when peak intensities are considered in mixed-stain cases, this same term can be used to weight the proposed genotypes, and can take nearly any non-negative value. The second term also depends on any hypothesized relationship (including background relatedness due to population substructure).

This method of calculating a likelihood ratio by determining the evidence, hypotheses, and then conditioning on the genotypes of those involved is the same employed in the rest of this work. Chapter 4 is particularly helpful in understanding this process, as it illustrates these ideas repeatedly under a variety of situations, and gives numerical examples for many of them.

## 1.3 DNA Mixtures

This work focuses primarily on statistical analysis of forensic DNA mixtures. A DNA mixture is evidence that consists of contributions from two or more persons. DNA mixtures can be found in any number of situations, but they most often occur in sexual

assaults. In sexual assaults the evidence is frequently a mixture of semen and vaginal material, or perpetrator skin cells found under the victim's fingernails, along with her DNA.

The statistical analysis of mixtures, though theoretically similar to single-person contributions, can be more complicated (see [19, 20] for the standard likelihood ratios for DNA mixtures). When considering an unknown contributor in the single-person case, there is only one feasible genotype at each locus – the genotype of the evidence. However, in a mixture, there can be many possible genotypes for a given unknown contributor. For mixtures with many unknown contributors, the number of possible genotypes can get very large very quickly. This added complexity makes mixtures in general difficult to work with.

For example, the variation in single contributor likelihood ratios under certain assumptions has long been known and can be used to calculate a confidence interval for the likelihood ratio [17]. However, the variance of a mixed-stain likelihood ratio is not as obvious. Bootstrap techniques can be used to obtain a confidence interval of the likelihood ratio, but this interval considers only the variation due to sampling error, and not the variation due to genetic sampling.

In this work (Chapter 2) we provide a formula to calculate the confidence interval. The formula is based on the assumption that the likelihood ratio is log-normal as more loci are added, and it uses Taylor's expansion to calculate an approximation of the variance. An example is given, and several different numerical studies are provided to validate the procedure and show the effects of different parameters. Due to the computational complexity of this method, a computer program is provided and briefly discussed.

Another complication of forensic DNA mixtures is that there is additional information in the electropherograms that can be difficult to utilize. The amount of fluorescence (also called the intensity, peak intensity, peak height, or peak area) for a particular allele is related to the number of copies of that allele that are moving through the buffer solution. When there are more copies present, then more copies of the fluorescent dye are excited by the laser, and the intensity is higher. The number of copies moving through the buffer

solution is related to the number of copies present before the PCR cycles began. Thus, if there are more copies of a particular allele in the evidence mixture, that particular allele will have a higher intensity. Since it is unlikely that two persons will contribute the same amount of material, we would expect that alleles from different contributors will have different intensities and alleles from the same contributor will have similar intensities. It follows that when a proposed set of genotypes violates this expectation, that particular set should be weighted less than a proposed set that violates the expectation to a lesser degree.

This information can be difficult to incorporate into the likelihood ratio. Some propose the elimination of highly unlikely sets of genotypes [21, 22]. Others propose a least-squares approach to pick a single best genotype [23]. These methods, and other similar methods do have their merits (a more detailed discussion of the alternative methods is included in Chapter 3). However, in the context of a likelihood ratio, we would ideally want to weight each possible set of genotypes by the likelihood of observing the peak intensities when that genotype set is the set of the true contributors.

This approach is taken here in Chapter 3. Each possible genotype set is weighted by the likelihood of observing the intensities given that set. The likelihood is calculated under one of two models: a model based on the Normal distribution and a model based on the Dirichlet distribution. An MLE of the mixture proportion (the percent of the material contributed by one of the contributors) is also provided and used in the estimates. Both models tend to weight more correct genotypes higher, though the Normal model puts much more emphasis on the best genotypes than the Dirichlet. This method can also be applied to certain cases of allele drop out (another complication that, though not limited to mixtures, occurs more often in mixture situations).

The final chapter shows the calculation of the likelihood ratio for a variety of situations. Four standard cases (single contributor, two-person mixture, paternity index, and relative index) are used as an introduction. In addition to these four standard cases, there are also several more unusual cases explored. These include a paternity index from a mixture, the situation where a relative of the suspect may be an unknown contributor

to a mixture, a canine parentage case, and the consideration of relationships in the presence of population substructure. As mentioned above, this chapter is particularly useful for seeing how likelihood ratios in the forensic context are constructed since the same method is applied across a variety of situations ranging from relatively simple to fairly complex.

It is hoped that this work as a whole will prove useful for forensic scientists. When analyzing forensic evidence, it is important to not overstate the strength of the evidence, and the confidence interval given here provides a method for doing so. At the same time, it is also important to not ignore evidence, and the method presented for the inclusion of peak intensities provides a way to included that information. In addition, several situations are discussed where the likelihood ratio is not as obvious. The treatment of these situations provides forensic scientists formulae to be used in casework.

# Chapter 2

# Confidence Interval for the Mixed Stain Likelihood Ratio

## 2.1 Introduction

With techniques for obtaining and typing DNA evidence becoming more sensitive, the need for interpreting mixed stains is growing. Unfortunately, analysis of DNA mixtures is both genetically and statistically complex [24] and care is needed to present non-prejudicial analyses. Several lines of research have been explored to simplify the mixture analysis process. Peak areas in chromatograms have been considered to help resolve mixtures [25, 26], and denaturing high-performance liquid chromatography has been used to resolve mitochondrial mixtures [27]. The majority of mixture evidence comes from sexual assault cases, and much work has been devoted to analysis of the Y chromosome in order to identify male perpetrators [28].

While much progress has been made in typing methods, the development of statistical methods has been slower. Weir et al. [19] gave a general formulation under the assumption of allelic independence, and this was extended by Curran et al. [20] to allow for population structure. The case of mixed race populations has been examined [29], and Lauritzen and Mortera [30] provided a useful bound for the number of unknown contributors to a mixture.

One statistical topic that has not received much attention is that of the effects of sampling variation on the numbers presented for DNA mixtures. For single-contributor stains, methods to describe the effects of sampling variation have been reviewed [31, 32]. This chapter proposes the reporting of likelihood ratios for DNA mixtures, and presents an analytical method that can be, and has been, incorporated into a software package.

## 2.2 The Confidence Interval

It has become accepted practice to attach numerical weights to DNA evidence in order to show "whether the patterns are as common as pictures with two eyes, or as unique as the Mona Lisa." (*US v Yee*, 134 FRD 161, 181 [ND Ohio, 1991]). Probability assessments should accurately inform the court of the strength of the evidence. However, a simple quantification of probability does not tell the entire story. Calculations for mixtures, as for single-contributor stains, rest on the frequencies of alleles at the typed markers yet these frequencies are not known. Instead, they are estimated using a sample from the population. Since these samples represent only a small portion of the total population, there is uncertainty about the true frequencies and therefore uncertainty about the resulting calculations. The calculated likelihood ratio is then an estimated likelihood ratio and not the true likelihood ratio.

If the forensic scientist wishes to report on the evidence accurately and thoroughly, the level of uncertainty should in some way be reported. Some investigators (e.g., [33, 34, 31]) advocate the use of Bayesian methods that lead to probability distributions of mixture quantities and there is merit to that approach. Here we present the classical approach of confidence intervals, in part because they avoid the need for complex computations, and in part because they are familiar in the context of public opinion surveys ("47% of those polled support the President on this issue, plus or minus 3 percentage points.") It is understood that the "plus or minus" results from the estimated proportion depending on the particular set of people sampled. We do not mean to imply that forensic scientists should adopt statistical procedures only because they have straightforward calculations or because they can be explained easily, and we point out that the widely-used confidence interval is a statistical tool with a rigorous theoretical basis.

Technically, a confidence interval refers to the range in which a specified central proportion (say 95%) of future estimates would fall if further samples were taken from the population and each one used to provide an estimate. Presenting a confidence interval is the appropriate response to the question "How large a sample is necessary to provide an estimate?" The forensic scientist can explain that the sample size used resulted in a

certain width confidence interval. Smaller samples would widen the interval, and larger samples would make it more narrow. It is worth noting that the common "plus or minus 3 percentage points" generally reflects the uncertainty in the proportion of a population responding to one answer to a single question when 1,000 or so people are questioned. For DNA profiles to match, there is a question ("is there a match?") that must be answered correctly for each allele in the evidence profile, and the resulting confidence interval is more likely to be "plus or minus a factor of 3."

The definition of a confidence interval leads naturally to the technique of bootstrapping, whereby a new sample is created by resampling the sample at hand (Hollander and Wolfe §8.4, [35]). If 1,000 new samples are created in this way, and the 1,000 new estimates are put in rank order, then a 95% confidence interval is bounded by the 26th and the 975th estimates. Although bootstrapping makes few assumptions, other than the original sample being appropriately random, it requires access to the original database from which allele frequencies are estimated. More importantly, bootstrapping from a single population does not address the evolutionary sampling implicit in forensic calculations that employ the "theta correction." The approach we employ for both single- and multiple-contributor stains supposes that allele frequencies are not necessarily available from the most relevant population or subpopulation, but that the available frequencies can be used, along with the population structure parameter $\theta$, in a way that recognizes the variation among populations caused by evolutionary processes.

We offer, therefore, an algebraic treatment that employs sample allele frequencies, and a specified value of $\theta$. Access to the original database is not needed, although a computer program is an advantage. Because it incorporates both "statistical" and "genetic" sampling [36], the intervals we present are wider than those that would be obtained by bootstrapping.

## 2.3 The Likelihood Ratio

The likelihood ratio is particularly well suited for the statistical analysis of forensic DNA evidence because, as in a trial, two alternative hypotheses are compared. In a trial, a jury weighs the prosecution hypothesis and the defense hypothesis and determines which is more likely to explain the evidence. The likelihood ratio method compares the probability of finding the DNA evidence given the prosecution hypothesis ($Pr(E|H_p)$) to the probability of finding the evidence assuming the defense hypothesis ($Pr(E|H_d)$). The comparison is expressed in the form of the ratio $LR$:

$$\text{LR} = \frac{Pr(\text{Evidence}|H_p)}{Pr(\text{Evidence}|H_d)}$$

The likelihood ratio method illustrates what Evett and Weir [10] call the "*First principle of evidence interpretation*." This principle states, "To evaluate the uncertainty of any given proposition, it is necessary to consider at least one alternative proposition." Not only is the method well suited to the mixture situation, it is sometimes necessary. "There is no alternative [to the likelihood ratio] when the evidence is less than certain under the proposition $H_p$." [20].

The likelihood method requires both evidence and hypotheses. The evidence is a DNA profile. The profile may be from a DNA sample left by a single person at a crime, the profile of a child in a paternity case, or a pair of profiles from potential relatives in an estate conflict. In this chapter the term evidence refers to the profile of genetic material that is a mixture from two or more persons, unless otherwise stated. The hypotheses consist of alternative propositions concerning the contributor(s) of the stain or concerning the status of relationships. For example, for a single contributor, the hypotheses may be:

$H_p$: The suspect contributed the evidence.

$H_d$: An unknown person contributed the evidence.

However, for a paternity case, the hypotheses may be:

$H_p$: The alleged father is the true father of the child.

$H_d$: An unknown person is the true father of the child.

If the evidence is a DNA mixture, the most common set of hypotheses are:

$H_p$: The suspect and the victim contributed the evidence.
$H_d$: An unknown person and the victim contributed the evidence.

For more on calculating the likelihood ratio for some of these common examples, see §4.2 of this work for a brief introduction, or [10] for a more thorough treatment. For our purposes here, we now turn specifically to the likelihood ratio in the context of DNA mixtures.

Curran et al. [20] gave an expression for the likelihood ratio for DNA mixtures that allowed for population structure effects. Their formulation rests on the concept that every (sub)population has allele frequencies that can differ from other (sub)populations but that the collection of all sets of frequencies follows a known statistical distribution. If it can be assumed that the populations have reached a state of evolutionary equilibrium, then this distribution is the Dirichlet. A consequence of this distribution is that every observed allele, whether in the evidence profile or in samples taken from individuals, affects the probabilities of allelic types for future observations. In particular, if a set of $n$ alleles has been observed, and if $n_i$ of them are of type $A_i$, then the probability that the next allele is of type $A_i$ is

$$\Pr(A_i|\{n_i\} \text{ of type } A_i) \quad = \quad \frac{n_i\theta + (1-\theta)p_i}{n\theta + (1-\theta)} \tag{2.1}$$

where $\theta$ is the population structure parameter, typically assumed to be in the range 0.01 to 0.05. This expression leads to probabilities for the set of observed alleles under either hypothesis about the evidence profile. Curran et al. [20] generalized (2.1) to calculate the probability of a set of observed alleles, and sum across all possible sets when there is one or more unknown contributors.

$$Pr(E_l|H) = \sum_{r_1=0}^{r}\sum_{r_2=0}^{r-r_1}\cdots\sum_{r_{c-1}=0}^{r-r_1\ldots-r_{c-2}} \frac{(2x)!2^{h_T+h_V}}{\prod_{h=1}^{c} u_h!} \times \frac{\prod_{h=1}^{c}\prod_{j=0}^{t_h+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,h}+j\theta]}{\prod_{j=0}^{2x+2n_T+2n_V-1}[(1-\theta)+j\theta]} \tag{2.2}$$

This expression can be broken up into three parts. The first part, the summations, is used to sum across all possible sets of alleles. For a particular set of alleles, the term $(2x)!2^{h_T+h_V}[\prod_{h=1}^{c} u_h!]^{-1}$ is the number of different ways that set can be ordered. The final term is an extension of (2.1) that is the probability of the entire set, as opposed to just the next allele. Equation (2.2) is calculated at each locus for both $H_p$ and $H_d$. Results are multiplied across loci, leading to the estimated LR,

$$\widehat{LR} = \frac{Pr(E_C|H_p)}{Pr(E_C|H_d)} = \prod_l \frac{Pr(E_l|H_p)}{Pr(E_l|H_d)} \tag{2.3}$$

Though the number of different allelic types will be the same for each hypothesis, the number of alleles contributed per type will be different. This is because alleles seen in a suspect who is not excluded from the evidence may be counted once by the prosecution but twice by the defense who claim the suspect is not a contributor to the evidence. The complete expression for the likelihood ratio requires identification of all alleles from people who have been typed (whether or not they are hypothesized to be contributors to the mixture)) as well as all alleles in the mixture. The number of contributors also need to be specified. For a more detailed explanation of (Equation 2.2, see [20].

## 2.4 Calculating the Confidence Interval

As illustrated in Equation 2.3, under the assumption of independent loci, the overall likelihood ratio is the product of the likelihood ratios for each locus, which leads us to the

heart of our approach. The logarithm of the likelihood ratio is the sum of the logarithms of the likelihood ratios for each locus, and if there are several loci (as there are with the 13-locus CODIS set) this logarithm can be assumed to be normally distributed and standard statistical theory can be invoked to calculate a confidence interval. In particular, a 95% confidence interval for the logarithm of the likelihood ratio is calculated as

$$\text{CI} \quad = \quad \ln(\widehat{\text{LR}}) \pm 1.96\sqrt{\text{Var}[\ln(\widehat{\text{LR}})]} \tag{2.4}$$

where *Var* indicates the variance of the calculated likelihood ratio estimate $\widehat{\text{LR}}$. If we take anti-logs, then this provides a confidence interval $(\widehat{\text{LR}}/C, \widehat{\text{LR}}C)$ for the likelihood ratio estimate, where the quantity $C$ is the anti-log (i.e $e$ to the power of) $1.96\sqrt{\text{Var}[\ln(\widehat{\text{LR}})]}$.

## Variance of the Likelihood Ratio

The most difficult part of calculating the confidence interval is obtaining the variance. Under the independence assumption,

$$\begin{aligned}
\widehat{\text{LR}} &= \prod_l \widehat{\text{LR}}_l \\
\ln(\widehat{\text{LR}}) &= \sum_l \ln(\widehat{\text{LR}}_l) \\
\text{Var}[\ln(\widehat{\text{LR}})] &= \sum_l \text{Var}[\ln(\widehat{\text{LR}}_l)]
\end{aligned}$$

To calculate $\text{Var}[\ln(\widehat{\text{LR}}_l)]$, the Taylor expansion ($\delta$-method) can be used as an approximation. If there are $n$ variables $x$ being estimated then

$$\text{Var}\{g(\mathbf{x})\} \quad = \quad \sum_{i=1}^{n}\{g_i^{'}(\mathbf{x})\}^2\text{Var}(x_i) + \sum_{i=1}^{n}\sum_{\substack{j\neq i \\ j=1}}^{n} g_i^{'}(\mathbf{x})g_j^{'}(\mathbf{x})\text{Cov}(x_i, x_j)$$

Where $g_i^{'}(\mathbf{x})$ is the derivative of $g(\mathbf{x})$ with respect to the $i$th variable $x_i$. Now apply this to $\ln(\widehat{\mathrm{LR}}_l)$.

$$\mathrm{Var}(\ln(\widehat{\mathrm{LR}}_l)) = \sum_{i=1}^{n}\left(\frac{\partial \ln(\widehat{\mathrm{LR}}_l)}{\partial \tilde{p}_{l,i}}\right)^2 \mathrm{Var}(\tilde{p}_{l,i}) + \sum_{i=1}^{n}\sum_{\substack{j\neq i\\ j=1}}^{n} \frac{\partial \ln(\widehat{\mathrm{LR}}_l)}{\partial \tilde{p}_{l,i}} \frac{\partial \ln(\widehat{\mathrm{LR}}_l)}{\partial \tilde{p}_{m,j}} \mathrm{Cov}(\tilde{p}_{l,i}, \tilde{p}_{m,j})$$

To obtain the partial derivative of $\ln(\widehat{\mathrm{LR}}_l)$ with respect to $\tilde{p}_{l,i}$, first use the ln rule for derivatives.

$$\frac{\partial \ln(\widehat{\mathrm{LR}}_l)}{\partial \tilde{p}_{l,i}} = \frac{1}{\widehat{\mathrm{LR}}_l} \times \frac{\partial(\widehat{\mathrm{LR}}_l)}{\partial \tilde{p}_{l,i}} = \frac{Pr(E_l|H_d)}{Pr(E_l|H_p)} \times \frac{\partial\left(\frac{Pr(E_l|H_p)}{Pr(E_l|H_d)}\right)}{\partial \tilde{p}_{l,i}}$$

Next apply the division rule for derivatives and then combine terms.

$$\begin{aligned}
\frac{\partial \ln(\widehat{\mathrm{LR}}_l)}{\partial \tilde{p}_{l,i}} &= \frac{Pr(E_l|H_d)}{Pr(E_l|H_p)} \times \frac{Pr(E_l|H_d)\frac{\partial Pr(E_l|H_p)}{\partial \tilde{p}_{l,i}} - Pr(E_l|H_p)\frac{\partial Pr(E_l|H_d)}{\partial \tilde{p}_{l,i}}}{Pr(E_l|H_d)^2} \\
&= \frac{Pr(E_l|H_d)\frac{\partial Pr(E_l|H_p)}{\partial \tilde{p}_{l,i}} - Pr(E_l|H_p)\frac{\partial Pr(E_l|H_d)}{\partial \tilde{p}_{l,i}}}{Pr(E_l|H_d)Pr(E_l|H_p)} \\
&= \frac{1}{\Pr(E_l|H_p)}\frac{\partial \Pr(E_l|H_p)}{\partial \tilde{p}_{l,i}} - \frac{1}{\Pr(E_l|H_d)}\frac{\partial \Pr(E_l|H_d)}{\partial \tilde{p}_{l,i}}
\end{aligned}$$

It is now necessary to compute the derivative of $Pr(E_l|H)$. Once the notation is cleaned up a bit, the calculations are simple to follow. The summation in (2.2) is a way to iterate all possible genotypes, and many of the terms are constant with respect to $\tilde{p}_{l,h}$. The summations can be collapsed into one summation, and the terms that are constant with respect to $\tilde{p}_{l,h}$ can be combined into one term ($A_k$). This makes (2.2) easier to work

with.

$$\begin{aligned} Pr(E_l|H) &= \sum_{r_1=0}^{r}\sum_{r_2=0}^{r-r_1}\cdots\sum_{r_{c-1}=0}^{r-r_1\ldots-r_{c-2}} \frac{(2x)!2^{h_T+h_V}}{\prod_{h=1}^{c} u_h!} \times \frac{\prod_{h=1}^{c}\prod_{j=0}^{t_h+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,h}+j\theta]}{\prod_{j=0}^{2x+2n_T+2n_V-1}[(1-\theta)+j\theta]} \\ &= \sum_{k=1}^{K} A_k \prod_{h=1}^{c}\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,h}+j\theta] \end{aligned}$$

$K$ is the number of possible genotype combinations that the hypothesis supports, and $A_k$ is the terms constant with respect to $\tilde{p}_{l,h}$ associated with that particular combination.

$$A_k = \frac{(2x)!2^{h_T+h_V}}{\prod_{h=1}^{c} u_h!} \times \prod_{j=0}^{2x+2n_T+2n_V-1}[(1-\theta)+j\theta]^{-1}$$

Next, compute the derivative of $Pr(E_l|H)$ with respect to $\tilde{p}_{l,i}$.

$$\begin{aligned} \frac{\partial Pr(E_l|H)}{\partial \tilde{p}_{l,i}} &= Pr'_{l,i}(E_l|H) \\ &= \frac{\partial}{\partial \tilde{p}_{l,i}}\left(\sum_{k=1}^{K} A_k \prod_{h=1}^{c}\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,h}+j\theta]\right) \\ &= \sum_{k=1}^{K} A_k \frac{\partial}{\partial \tilde{p}_{l,i}}\left(\prod_{h=1}^{c}\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,i}+j\theta]\right) \end{aligned}$$

Now, specifically addressing $\prod_{h=1}^{c}\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,i}+j\theta]$, the terms that are constant with respect to $\tilde{p}_{l,i}$ can now be factored.

$$\frac{\partial}{\partial \tilde{p}_{l,i}}\left(\prod_{h=1}^{c}\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,h}+j\theta]\right) = \left(\prod_{\substack{h=1\\h\neq i}}^{c}\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,h}+j\theta]\right) \times\frac{\partial}{\partial \tilde{p}_{l,i}}\left(\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,i}+j\theta]\right)$$

Note that $\partial(\prod_{j=0}[(1-\theta)\tilde{p}_{l,i}+j\theta])/\partial\tilde{p}_{l,i}$ is the product of several functions of $\tilde{p}_{l,i}$. By the product rule for derivatives, the derivative of the product of $N$ functions $(F_n)$ with respect to $x$ is

$$\begin{aligned}\frac{\partial\left(\prod_{n=0}^{N}F_n(x)\right)}{\partial x} &= [(\partial F_1(x))\times F_2(x)\times ...\times F_{N-1}(x)\times F_N(x)]\\ &\quad +[(\partial F_2(x))\times F_1(x)\times F_3(x)\times ...\times F_{N-1}(x)\times F_N(x)]\\ &\quad +... + [(\partial F_N(x))\times F_1(x)\times F_2(x)\times ...\times F_{N-1}(x)]\\ &= \sum_{n=0}^{N}\left[\frac{\partial F_n(x)}{\partial x}\prod_{\substack{m=0\\m\neq n}}^{N}F_m(x)\right]\end{aligned}$$

For our function, all the partial derivatives equal $(1-\theta)$, so:

$$\frac{\partial}{\partial \tilde{p}_{l,i}}\left(\prod_{j=0}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,i}+j\theta]\right) = (1-\theta)\sum_{q=0}^{t+u_h+v_h-1}\prod_{\substack{j=0\\j\neq q}}^{t+u_h+v_h-1}[(1-\theta)\tilde{p}_{l,i}+j\theta]$$

Thus:

$$\begin{aligned} Pr'_{l,i}(E_l|H) &= \sum_{k=1}^{K} A_k \left[ \prod_{\substack{h\neq i \\ h=1}}^{c} \prod_{j=0}^{t+u_h+v_h-1} [(1-\theta)\tilde{p}_{l,h} + j\theta] \right] \\ &\quad \times (1-\theta) \sum_{q=0}^{t+u_h+v_h-1} \prod_{\substack{j\neq q \\ j=0}}^{t+u_h+v_h-1} [(1-\theta)\tilde{p}_{l,i} + j\theta] \end{aligned} \tag{2.5}$$

Consider also that $\mathrm{Var}(\tilde{p}) = \frac{\tilde{p}(1-\tilde{p})}{2n_l}[(2n_l-1)\theta+1]$, and $\mathrm{Cov}(\tilde{p}_{l,h}, \tilde{p}_{l,i}) = \frac{-\tilde{p}_{l,h}\tilde{p}_{l,i}}{2n_l}[(2n_l-1)\theta+1]$ with $h \neq i$. So,

$$\begin{aligned} \mathrm{Var}(\ln(\widehat{\mathrm{LR}}_l)) &= \sum_i \left[ \frac{Pr'_{l,i}(E_l|H_p)}{Pr(E_l|H_p)} - \frac{Pr'_{l,i}(E_l|H_d)}{Pr(E_l|H_d)} \right]^2 \frac{\tilde{p}_i(1-\tilde{p}_i)}{2n_l}[(2n_l-1)\theta+1] \\ &\quad + \sum\sum_{i\neq h} \left[ \frac{Pr'_{l,i}(E_l|H_p)}{Pr(E_l|H_p)} - \frac{Pr'_{l,i}(E_l|H_d)}{Pr(E_l|H_d)} \right] \\ &\quad \times \left[ \frac{Pr'_{l,h}(E_l|H_p)}{Pr(E_l|H_p)} - \frac{Pr'_{l,h}(E_l|H_d)}{Pr(E_l|H_d)} \right] \frac{-\tilde{p}_{l,h}\tilde{p}_{l,i}}{2n_l}[(2n_l-1)\theta+1] \end{aligned} \tag{2.6}$$

This variance is then used in Equation 2.4 to calculate the confidence interval of the log-likelihood ratio. The anti-log is taken for the final confidence interval.

We note that the intervals here are larger than those in the National Research Council report on the evaluation of forensic DNA evidence [17]. The increased size of the interval is because our variance includes variation due to genetic sampling, and the formulae used in the NRC report do not. To understand the concept of genetic sampling, it is helpful to consider what population substructure and the parameter $\theta$ means. The parameter is a measure of the relatedness of alleles within a subpopulation, and a measure of the differentiation of the subpopulation from other subpopulations. Allele frequencies are estimated from a population-wide sample and since this is just a sample, there is uncertainty in

the estimates. However, there is also uncertainty in the subpopulation allele frequencies. This is somewhat due to the uncertainty in the population-wide allele frequencies, but is mostly due to the subpopulation's differentiation from the overall population. No amount of increased sampling from the population can give us more certainty about the subpopulation frequencies. As a consequence, the consideration of genetic sampling can significantly increase the size of the confidence interval as $\theta$ increases. These concepts are further illustrated in the numerical studies.

A computer program has been written to calculate the likelihood ratio and its confidence interval. DNAMIX was originally written by Dr. John Storey and was updated once to include the population structure calculations put forth by Curran et al. [20]. These two earlier versions were command line based, and did not calculate the confidence interval. The current version, DNAMIX-3, does calculate the confidence interval, and includes a point and click interface. Source code is available and may be freely modified for research purposes. It is written in Java, so can be run on any operating system with Java installed. The program is available at `http://bioinformatics.ncsu.edu/` and is free to the public.

DNAMIX offers the user the opportunity to manually input the genotypes, or have the program guide them in the process. The guided input can greatly decrease the time needed to calculate a likelihood ratio. In addition, there is the ability to calculate multiple likelihood ratios for different databases, and multiple likelihood ratios for different combinations of hypotheses. A sample screenshot is given in Figure 2.1. We note that the formula used to calculate the likelihood ratio in DNAMIX-3 is slightly altered from that given by Curran et al. [20]: all allelic probabilities which are less than $\theta$ are replaced by the value of $\theta$. This is to allow the calculation of $LR$ for evidence profiles that contain alleles not seen in the database.

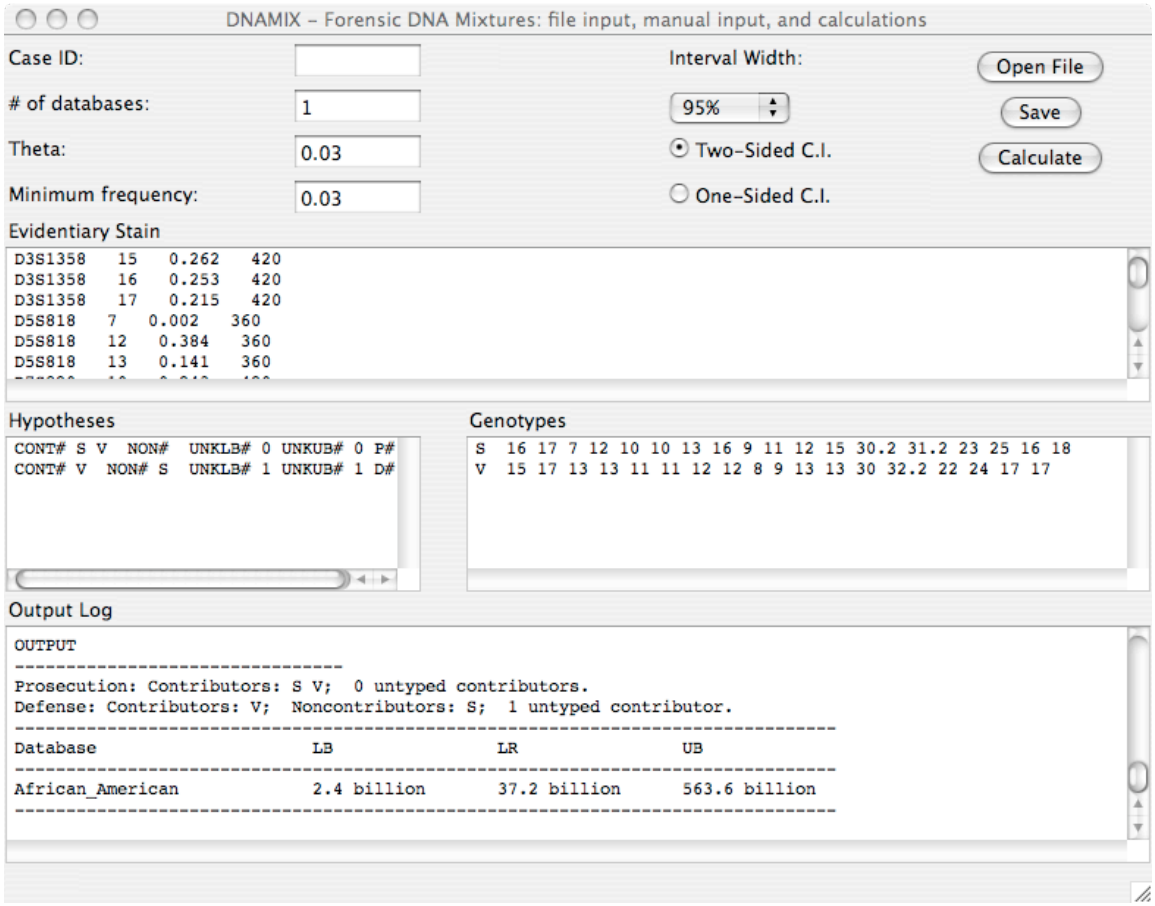



Figure 2.1: DNAMix v3.2 Screenshot

## 2.5 Example

We now present a small, single-locus, single-contributor example to illustrate this approach. Note that in practice this approach is not applicable to a single-locus stain because the normality assumption requires multiple loci. However, it will suffice for illustrative purposes.

Suppose evidence is found and typed, and a suspect is arrested and typed. At a particular locus, both the stain and the suspect are genotyped as being homozygotes for allelic type A. The prosecution proposes that the suspect contributed the stain, and the

defense proposes that an unknown person contributed the stain. Using Equation 2.2, the calculation of $Pr(E|H_p)$, $Pr(E|H_d)$,and $LR$ are straightforward.

$$\begin{aligned} Pr(E|H_p) &= \tilde{p}_A((1-\theta)\tilde{p}_A + \theta) \\ Pr(E|H_d) &= \frac{\tilde{p}_A((1-\theta)\tilde{p}_A + \theta)((1-\theta)p_A + 2\theta)((1-\theta)\tilde{p}_A + 3\theta)}{(1+\theta)(1+2\theta)} \\ \widehat{LR} &= \frac{(1+\theta)(1+2\theta)}{((1-\theta)\tilde{p}_A + 2\theta)((1-\theta)\tilde{p}_A + 3\theta)} \end{aligned}$$

Employing Equation 2.5,

$$\begin{aligned} Pr^{'}_A(E|H_p) &= (1-\theta)\tilde{p}_A + ((1-\theta)\tilde{p}_A + \theta) \\ Pr^{'}_A(E|H_d) &= [((1-\theta)\tilde{p}_A + \theta)((1-\theta)\tilde{p}_A + 2\theta)((1-\theta)p_A + 3\theta) \\ &\quad +(1-\theta)\tilde{p}_A((1-\theta)\tilde{p}_A + 2\theta)((1-\theta)\tilde{p}_A + 3\theta) \\ &\quad +(1-\theta)\tilde{p}_A((1-\theta)\tilde{p}_A + \theta)((1-\theta)\tilde{p}_A + 3\theta) \\ &\quad +(1-\theta)\tilde{p}_A((1-\theta\tilde{p}_A + \theta)((1-\theta)\tilde{p}_A + 2\theta)]\frac{1}{(1+\theta)(1+2\theta)} \end{aligned}$$

These formulae can be used in equation (2.6) to calculate the total variance. For example, if $\tilde{p}_A = 0.10$ and $\theta = 0.03$ and $n = 200$, then $Pr(E|H_p) = 0.0127$ and $Pr(E|H_d) = 0.000342$. This leads to a $LR$ of about 37. Using the above formulae, the variance can also be calculated. Under these same values of $\tilde{p}_A$, $\theta$, and $n$, $Pr^{'}_A(E|H_p) = 0.224$ and $Pr^{'}_A(E|H_d) = 0.009905$. The variance of $\ln(\widehat{\mathrm{LR}})$ is then calculated using (2.6). $\mathrm{Var}(\ln(\widehat{\mathrm{LR}})) = 0.377$ which leads to a 95% confidence interval of (11, 124). Other likelihood ratios and confidence intervals are shown in Table 2.5 for a variety of values of $\theta$ and $\tilde{p}_A$. This shows that even under very controlled circumstances — one locus, one homozygote — the size of the interval varies significantly.

Table 2.1: Example one-locus confidence intervals.

| $\tilde{p}$ | $\theta = 0$ | $\theta = 0.001$ | $\theta = 0.01$ | $\theta = 0.03$ |
|---|---|---|---|---|
| $\tilde{p}_A = 0.01$ | 10000 | 6439.8 | 863.5 | 157.1 |
| | (1422, 70297) | (1014.7, 40869.4) | (244.5, 3049.3) | (68.5, 360.4) |
| $\tilde{p}_A = 0.05$ | 400 | 364.6 | 186.5 | 72.7 |
| | (170, 940) | (139.3, 954.6) | (52.2, 666.4) | (21.3, 247.7) |
| $\tilde{p}_A = 0.10$ | 100 | 95.6 | 67.1 | 37.2 |
| | (56, 180) | (48.5, 188.5) | (23.5,191.8) | (11.2, 123.9) |
| $\tilde{p}_A = 0.20$ | 25 | 24.5 | 20.7 | 15.1 |
| | (17, 37) | (15.5, 38.7) | (9.5, 45.1) | (5.5, 42.0) |

## 2.6 Numerical Studies

### 2.6.1 Interval Range

The Caucasian database for the CODIS loci published by Budowle et al. [37] was used to illustrate the size of confidence intervals for both single-contributor and mixed stains. To indicate the likely range of sizes, two situations were considered: those with the most common alleles at all 13 loci, and those with the least common alleles. Confidence intervals for a two-allele single-contributor profile are shown in Table 2.2, for which the two hypotheses are:

$H_p$: The suspect contributed the evidence.

$H_d$: An unknown person contributed the evidence.

The suspect is assumed to have the same profile as the evidence. "Most" refers to the evidence stain with the most common alleles and "least" refers to the evidence stain with the least common alleles. $\widehat{LR}$, $\widehat{LB}$, and $\widehat{UB}$ are the estimates of the likelihood ratio and the confidence interval. $C$ is the factor defined by $C = e^{z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}(ln(\widehat{\mathrm{LR}}))}}$.

Table 2.2: Range of the CI and bounds for a stain with a single contributor.

| Stain | C.I. | $\widehat{LR}$ | $\widehat{LB}$ | $\widehat{UB}$ | $C$ |
|---|---|---|---|---|---|
| Most | 95.0% | 1.13E11 | 1.97E10 | 6.52E11 | 5.8 |
| | 99.0% | 1.13E11 | 1.14E10 | 1.13E12 | 10.0 |
| | 99.9% | 1.13E11 | 6.01E09 | 2.14E12 | 18.9 |
| Least | 95.0% | 1.41E35 | 7.84E32 | 2.54E37 | 180.0 |
| | 99.0% | 1.41E35 | 1.53E32 | 1.30E38 | 919.8 |
| | 99.9% | 1.41E35 | 2.31E31 | 8.62E38 | 6108.5 |

$\theta = 0.015$

Sets of intervals for a mixed stain, where there are three alleles at every locus in the evidence profile are shown in Tables 2.3 and 2.4. The suspect has two of the alleles at each locus. Hypotheses for "Case 1" (Table 2.3) are:

$H_p$: The suspect and the victim contributed the evidence.

$H_d$: An unknown and the victim contributed the evidence.

Hypotheses for "Case 2" (Table 2.4) are:

$H_p$: The suspect and an unknown contributed the evidence.

$H_d$: Two unknowns contributed the evidence.

As with the single contributor stain, "most" refers to the evidence stain with the most common alleles and "least" refers to the evidence stain with the least common alleles. $\widehat{LR}$, $\widehat{LB}$, and $\widehat{UB}$ are the estimates of the likelihood ratio and the confidence interval. $C$ is the factor defined by $C = \exp z_{\frac{\alpha}{2}} \sqrt{\mathrm{V}(ln(\widehat{\mathrm{LR}}))}$.

### 2.6.2 Theta and the Confidence Interval

What effect does the population parameter $\theta$ have on the confidence interval? To examine this question, the likelihood ratio, upper bounds, and lower bounds were calculated for

Table 2.3: Case 1: Range of the CI and bounds for a stain with two contributors.

| | C.I. | $\widehat{LR}$ | $\widehat{LB}$ | $\widehat{UB}$ | $C$ |
|---|---|---|---|---|---|
| Most | 95.0% | 3.70E08 | 7.15E07 | 1.92E09 | 5.2 |
| | 99.0% | 3.70E08 | 4.26E07 | 3.21E09 | 8.7 |
| | 99.9% | 3.70E08 | 2.34E07 | 5.85E09 | 15.8 |
| Least | 95.0% | 5.86E27 | 1.39E26 | 2.48E28 | 42.2 |
| | 99.0% | 5.86E27 | 4.27E25 | 8.03E29 | 137.1 |
| | 99.9% | 5.86E27 | 1.09E25 | 3.15E30 | 537.1 |

$\theta = 0.015$

Table 2.4: Case 2: Ranges of the CI and bounds for a stain with two contributors.

| | C.I. | $\widehat{LR}$ | $\widehat{LB}$ | $\widehat{UB}$ | $C$ |
|---|---|---|---|---|---|
| Most | 95.0% | 7.76E02 | 1.59E02 | 3.78E03 | 4.9 |
| | 99.0% | 7.76E02 | 9.68E01 | 6.21E03 | 8.0 |
| | 99.9% | 7.76E02 | 5.44E01 | 1.11E04 | 14.3 |
| Least | 95.0% | 3.20E22 | 1.13E21 | 9.06E23 | 28.3 |
| | 99.0% | 3.20E22 | 2.96E20 | 2.59E24 | 80.9 |
| | 99.9% | 3.20E22 | 1.17E20 | 8.76E24 | 273.8 |

$\theta = 0.015$

a range of values for $\theta$. Three different mixtures were used. The first is an artificial thirteen-locus mixture composed of the most common alleles (Figure 2.2). The second is an artificial thirteen-locus mixture composed of the least common alleles (Figure 2.3). For the last (Figure 2.4), real data from a six-locus mixture was used [26]. For each mixture the value of $\theta$ ranged from 0 to 0.0625 at increments of 0.0005. Each figure shows the $\ln(\widehat{\mathrm{LR}})$ (middle line), the log of the upper bound (the upper "+" line), and the lower bound (the lower "-" line).

Several things are immediately observable. These figures show, as noted in the example above, the values of $\widehat{\mathrm{LR}}$ and the bounds depend greatly on the frequency of the alleles and the values of $\theta$. Also of note is that as $\theta$ increases the likelihood ratio decreases and the interval bounds tend to increase. An increase in $\theta$ indicates that there is less variation within the subpopulation. This decreases the $\widehat{\mathrm{LR}}$ because the unknown contributor is more likely to be similar to the known contributor. However, an increase in $\theta$ also indicates increased differentiation between subpopulations. This means there is more uncertainty about the true subpopulation frequencies, so the variation of the likelihood ratio estimate increases.

### 2.6.3 Validation

Although our assumption that the logarithm of the $\widehat{\mathrm{LR}}$ is normally distributed is convenient, we wanted to be sure that it gave valid results. To test this, we used the bootstrap method. Since bootstrapping does not consider between population variation, the analytical method was modified to account for the differences in variation. Instead of using the standard variance and covariance formulae for the Dirichlet distribution ($\mathrm{Var}(\tilde{p}_{l,i}) = p_{l,i}(1-p_{l,i})[(2n_l-1)\theta+1]/2n_l$, $\mathrm{Cov}(\tilde{p}_{l,i},\tilde{p}_{l,j}) = -p_{l,i}p_{l,j}[(2n_l-1)\theta+1]/2n_l$), formulae were used that only consider within population variation ($\mathrm{Var}(\tilde{p}_{l,i}) = p_{l,i}(1-p_{l,i})(1-\theta)/2n_l$, $\mathrm{Cov}(\tilde{p}_{l,i},\tilde{p}_{l,j}) = -p_{l,i}p_{l,j}(1-\theta)/2n_l$).

We computed confidence intervals by bootstrapping, using 10,000 resamplings, and compared them to the modified analytical method. These values are shown in Table 2.5. In this table, "Most" refers to the evidence stain with the most common alleles and

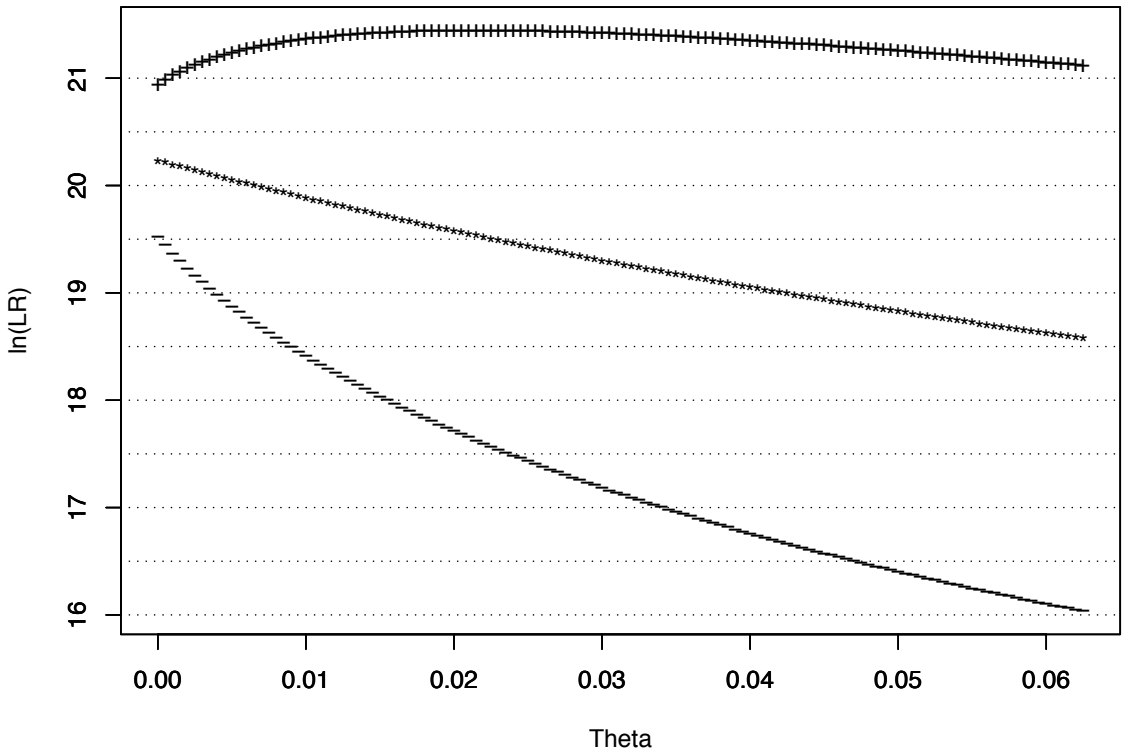



Figure 2.2: $\ln(\widehat{\mathrm{LR}})$ and bounds by $\theta$: Most common alleles

"Least" refers to the evidence stain with the least common alleles. $\widehat{LR}$, $\widehat{LB}$, and $\widehat{UB}$ are the estimates of likelihood ratio and the confidence interval. For the analytical method, $C$ is the factor defined by $C = e^{z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}(ln(LR))}}$. For the bootstrap method, $C$ is defined as $(\frac{\widehat{UB}}{\widehat{LR}} + \frac{\widehat{LR}}{\widehat{LB}})/2$. The figures for the analytical method in this table do not consider variation between populations.

Although the two sets of intervals are very similar, the normal-approximation method intervals were slightly wider and so are more conservative. The differences are greater for the profiles constructed with the least common alleles. The normal-approximation seems to be a reasonable assumption.

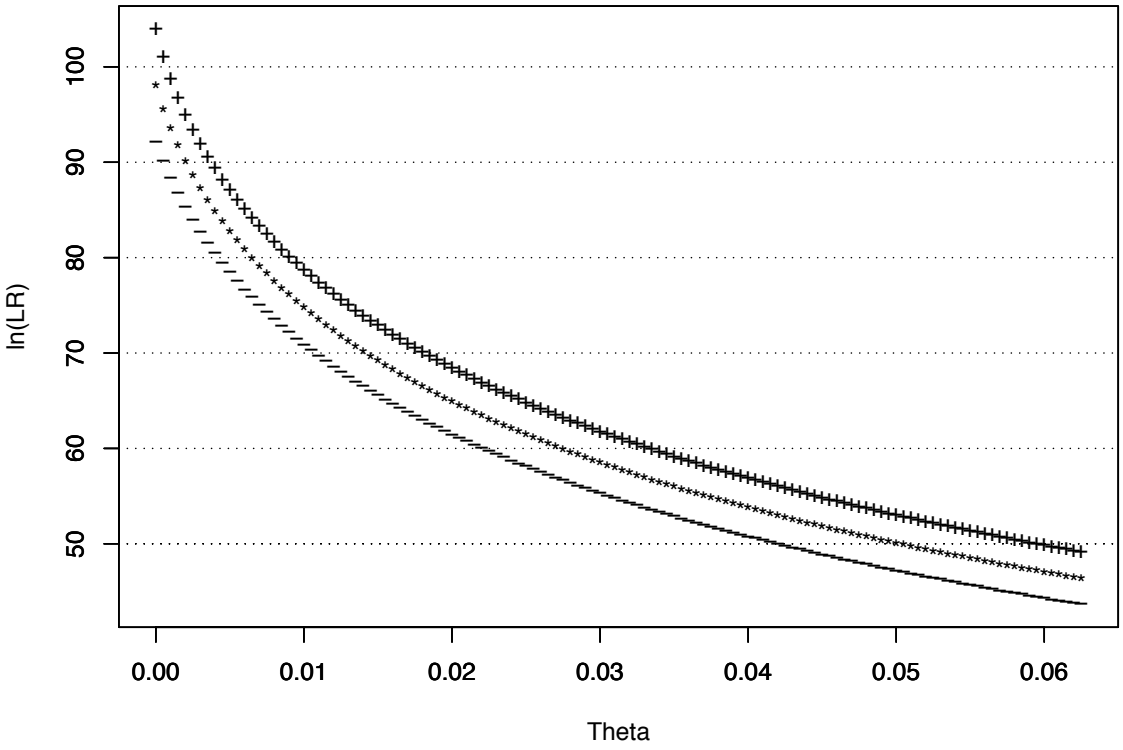



Figure 2.3: $\ln(\widehat{\mathrm{LR}})$ and bounds by $\theta$: Least common alleles

### 2.6.4 Analytical vs. Bootstrap

We also compared the bootstrapping method to the analytical method using the proper Dirichlet variance and covariance formulae. Figure 2.5 shows the importance of the analytical method. The central line in the graph is the likelihood ratio as a function of $\theta$. The LR is that of a mixed DNA stain with two contributors, twelve loci, and the hypotheses:

$H_p$: The suspect and victim contributed the evidence.

$H_d$: The victim and an unknown contributed the evidence.

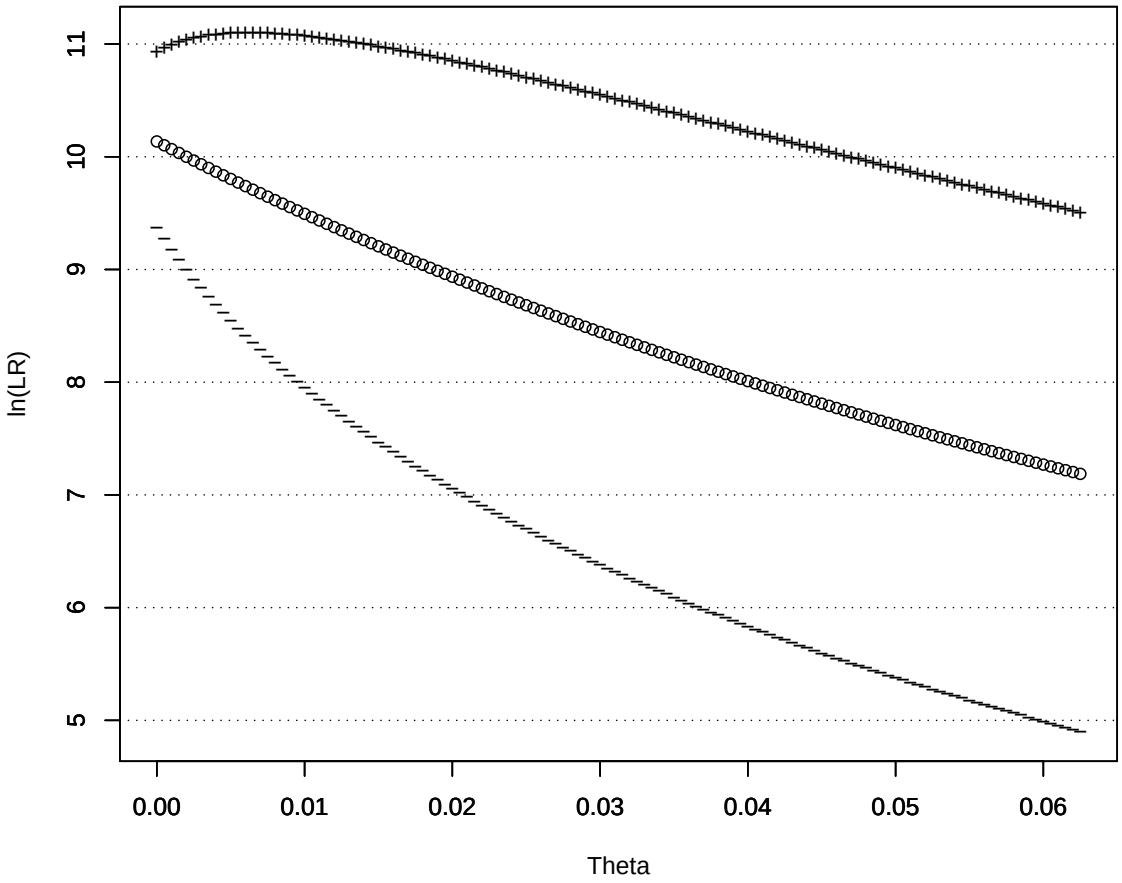



Figure 2.4: $\ln(\widehat{\mathrm{LR}})$ and bounds by $\theta$: Real data

Table 2.5: Results of the bootstrap method compared to the analytical method.

| Stain/CI | Method | $\widehat{LR}$ | $\widehat{LB}$ | $\widehat{UB}$ | $C$ |
|---|---|---|---|---|---|
| Most | Analytical | 3.70E08 | 1.98E08 | 6.90E08 | 1.87 |
| 95.0% | Bootstrap | 3.97E08 | 2.35E08 | 6.72E08 | 1.69 |
| Most | Analytical | 3.70E08 | 1.64E08 | 8.35E08 | 2.27 |
| 99.0% | Bootstrap | 3.97E08 | 1.89E08 | 8.30E08 | 2.10 |
| Least | Analytical | 5.86E27 | 1.42E27 | 2.43E28 | 4.14 |
| 95.0% | Bootstrap | 5.21E27 | 2.63E27 | 1.03E28 | 1.98 |
| Least | Analytical | 5.86E27 | 9.06E26 | 3.79E28 | 6.47 |
| 99.0% | Bootstrap | 5.21E27 | 1.89E27 | 1.34E28 | 2.66 |

$n = 10,000$

$\theta = 0.015$

The pair of lines closest to the $\widehat{LR}$ line is the 95% confidence interval that takes into account within population variation. This is the confidence interval that a bootstrap method would produce. The outer pair of lines indicate a 95% confidence interval that includes within population variation and between population variation. This is the confidence interval that our analytical method produces. It is evident that the confidence interval determined by the bootstrap does not include a significant portion of the variance included in the analytical method. This can lead to a difference of several orders of magnitude between the bounds of the confidence intervals.

### 2.6.5 Sample Size

Using the same data from [26], the effects of sample size were examined. Figure 2.6 displays the likelihood ratio and bounds as sample size increases when $\theta = 0$. Figure 2.7 displays the same thing when $\theta = 0.03$. When $\theta = 0$, the confidence interval continues to decrease as sample size increases. Eventually the variance due to sampling will be

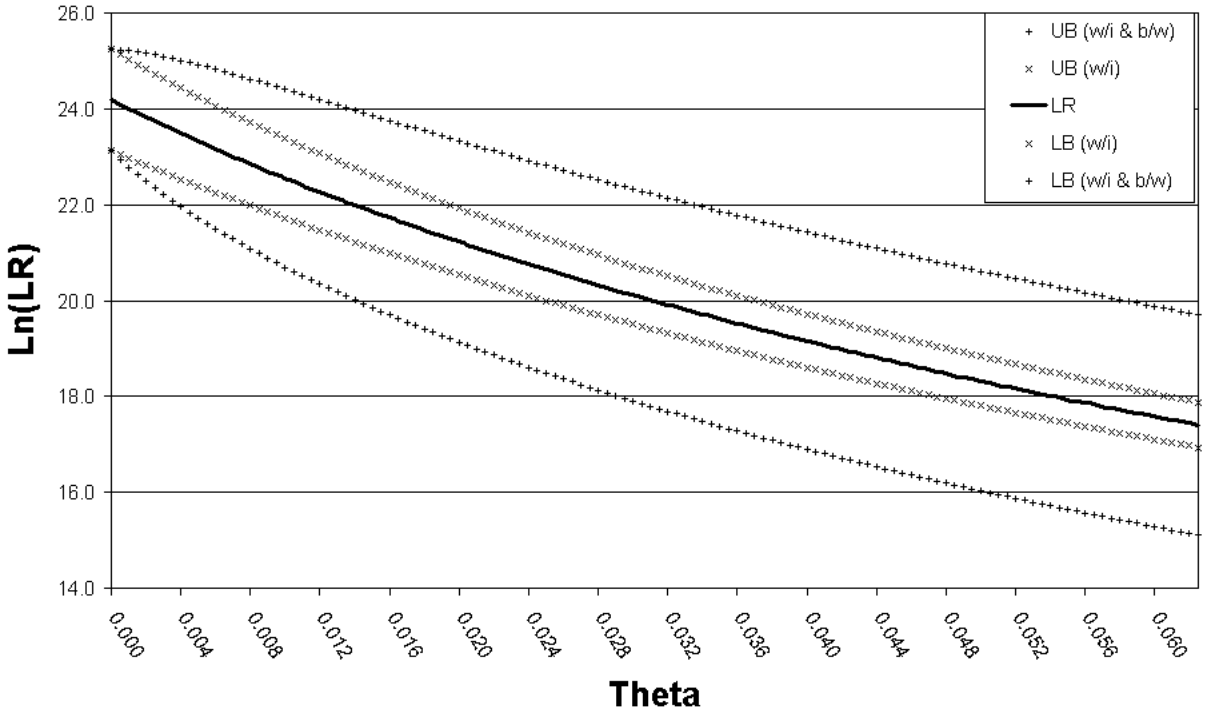



Figure 2.5: Comparison of bootstrap and analytical confidence intervals

eliminated and the variance will go to zero. However, when $\theta = 0.03$ the between subpopulation variance introduced by $\theta$ cannot be eliminated through increased sampling. As a result, the interval decreases only until the sampling error is eliminated.

## 2.7 Discussion

This formula is presented to provide a method for evaluating the confidence of a likelihood ratio in the context of forensic DNA stains. The previous method, nonparametric bootstrapping, does not capture all the variation, can be computationally intensive and is often impractical to implement for forensic scientists. Thus, this method, and DNAMIX v.3, provide another step forward in the interpretation of DNA evidence.

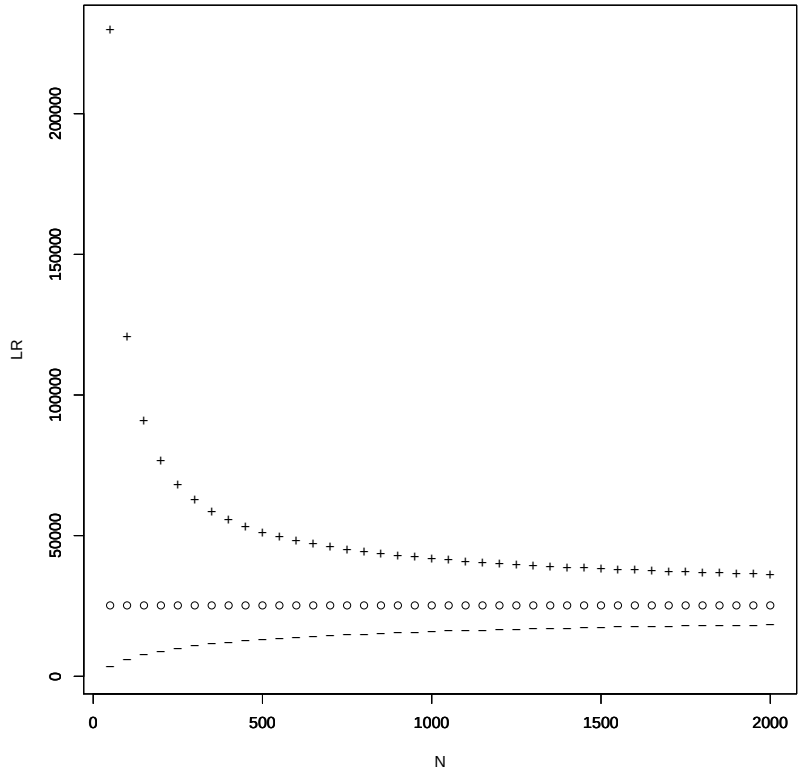



Figure 2.6: $\widehat{LR}$ and bounds by sample size, $\theta = 0.00$

The confidence interval results from a straightforward application of the $\delta$-method to the formula put forth in [20]. Numerical studies have shown that the size of the confidence interval depends largely upon the $\theta$ parameter, and the rarity of alleles observed. The sample size, though exhibiting some effect on the confidence interval, is not a major factor if the usual conservative values of $\theta$ are used. A comparison of a modified analytical confidence interval and the bootstrap confidence interval has shown that the normal-approximation is a valid assumption. A comparison of the non-modified analytical confidence interval with the traditional bootstrap confidence interval shows that considering between population variance is important because it considerable increases

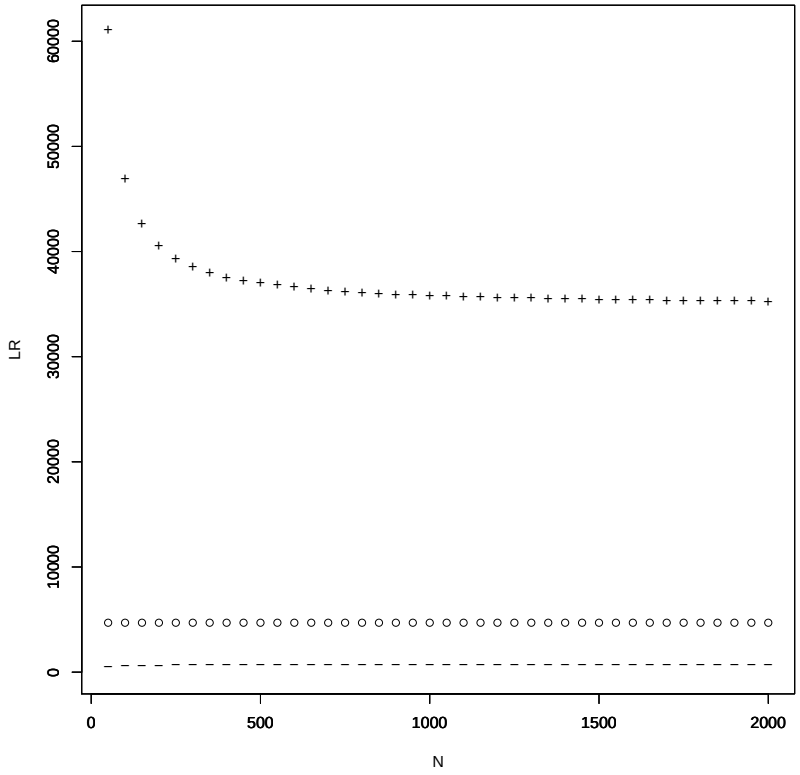



Figure 2.7: $\widehat{LR}$ and bounds by sample size, $\theta = 0.03$

the uncertainty in our estimate. An analysis of the theta parameter has shown it can have a major effect on the confidence interval and the likelihood ratio, especially for stains with a large number of rare alleles.

It is noteworthy that Curran, et al's treatment of mixed stains is applicable even when there is only one contributor to the "mixture." Therefore, this confidence interval can also be applied to single contributor stains.

# Chapter 3

# Using Peak Intensity Information with Maximum Likelihood

## 3.1 Introduction

It is not uncommon for forensic scientists to find DNA evidence that is a mixture from two or more contributors. Mixtures can be the result of a rape, where the evidence is a combination of the victim's DNA and the perpetrator's DNA (and possibly a consensual partner). A blood mixture could result from a fight in which two people are injured. Mixtures can also be found on confiscated material (i.e., a bag of narcotics that has been passed from one person to another). Weir et al. [19] laid the groundwork for analyzing forensic mixtures. Their method relies on the likelihood ratio approach to consider a prosecution hypothesis and a defense hypothesis. This method was later extended to include population substructure, and software was written to compute the likelihood ratio [20].

However, the methods of [20] do not use all available information. When evidence is typed for forensic markers, capillary electrophoresis is usually used, the output of which is a plot of time (related to marker fragment size) by amount of fluorescence (related to the amount of genetic material). The fluorescence intensity, measured by the height of the curve or area under the curve, at a particular size can be used as a measure of the amount of genetic material of that particular size. These intensities provide more information about the genotypes of contributors.

In a forensic mixture, each person contributes an amount of biological material that contains genetic information (i.e., blood, semen, skin cells). The amount each person contributes can be expressed as a proportion of the total amount contributed, and these

proportions are generally not equal. For example, in a two person mixture, a suspect may contribute proportion $m$ ($0 < m < 1$), and the victim would contribute the remaining proportion $1-m$. It is expected that this same proportion holds across all loci. That is, if the suspect contributes proportion $m$ at the first locus, then they contribute proportion $m$ at each of the other loci. It is also expected that the DNA material is evenly distributed between maternal and paternal alleles at each locus. So, if a heterozygous individual contributes proportion $m$, then they will contribute $m/2$ to the first allele, and $m/2$ to the second allele.

Intuitively it is easy to see how these assumptions should affect the statistical calculations. Alleles that have similar peak intensities are more likely to be from the same individual than peaks that have vastly different intensities. Therefore, genotypes that drastically violate the above expectations should be considered less likely than genotypes that violate the expectations to a lesser degree. If a suspect's genotype suggests that a pair of alleles should have similar peak intensities, but the evidentiary stain has intensities that are vastly different, it seems that the suspect may not be the contributor. If that is the case, the suspect's genotype should be weighted less than other genotypes. But, how exactly should the genotypes be weighted?

Several authors have sought to use peak intensities in forensic DNA analyses. Evett, Gill, and Lambert [26] developed much of the probabilistic theory of peak intensities, and fit it within the context of the likelihood ratio. Perlin, Lancia, and Ng [38] and Perlin and Szabady [39] developed the linear model equations used in our approach. Wang, Xue, and Wickenheiser [23] and [39] used similar least-squares approaches to pick a single "best" set of contributor genotypes.

Clayton et al. [25] use the assumption that the mixture proportion is the same across loci. The proportion is estimated from loci with no shared alleles contributed (i.e., four alleles at a locus with two people contributing) and genotypes that appear inconsistent in light of this estimate and the observed intensities are excluded. Gill et al. [22] use the assumption that a heterozygous person's contribution is split evenly between the maternal and paternal alleles. If the assumption holds, the ratio of two peaks from

the same person at the same locus should be approximately equal to one. This ratio, also called heterozygous balance ($Hb$), is a statistic of interest to [22]. When $Hb$ for a proposed heterozygous genotype is below or above certain threshold (Gill et al. use a lower threshold of 0.6) then that proposed genotype is considered highly unlikely, and the genotype is excluded.

All of these works represent significant contributions in the area of forensic mixture analysis, but none of them fully get at the problem at hand. The methods of [39] and [23] are able to pick a "best" genotype or potentially a group of "best" genotypes. Their methods, however, ignore any other potential genotypes. Genotypes that are close could be ignored, which may be prejudicial against the suspect. Clayton et al. [25] and Gill et al. [22] eliminate highly unlikely genotypes, but fail to weight those included. Evett et al. [26] attempt to weight genotypes by their likelihood, but require large amounts of data from tests cases to obtain the distributions of the various test statistics they use. While their methodology is correct, typing hundreds of test cases to validate distributions may be infeasible for many labs and using distributions of another laboratory would not be advisable since protocols may differ.

Our method is based on classical maximum likelihood theory. The method addresses the above issues by weighting each possible genotype, without the need for information foreign to the current evidence mixture at hand. The method fits into the current likelihood ratio approach to DNA evidence.

## 3.2 The Likelihood Ratio

We advocate the use of the likelihood ratio (LR) to analyze forensic DNA mixtures. In a LR, two hypotheses are compared by taking the ratio of the likelihoods given the hypotheses. The LR is ideal because typically in forensic casework, there are two hypotheses (prosecution hypothesis, $H_p$; and defense hypothesis, $H_d$) that need to be compared. Also, if the evidence under ($H_p$) is less than certain, then it is necessary to use a likelihood ratio. To write the LR we define $E$ as the evidence (peak intensities) and $G$ as

a list of genotype sets, with $G_i$ indicating the $i^{th}$ genotype set in the list. There will be different lists of genotype sets under each hypothesis. A genotype set consists of a single genotype for each contributor in the hypothesis. The likelihood ratio then takes the form

$$\begin{aligned} LR &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(E|G,H_p)\Pr(G|H_p)}{\Pr(E|G,H_d)\Pr(G|H_d)} \\ &= \frac{\sum_i \Pr(E|G_i,H_p)\Pr(G_i|H_p)}{\sum_i \Pr(E|G_i,H_d)\Pr(G_i|H_d)} \end{aligned} \tag{3.1}$$

Note that numerator and denominator both involve pairs of terms for each genotype. When peak intensities are not used, the only evidence is the set of alleles seen in the mixture. The first term then, $\Pr(E|G_i,H)$, is 1 if the union of the $K$ genotypes (one for each of $K$ contributors) in $G_i$ equals the alleles seen in $E$, and zero otherwise. $G_{ik}$ represents the genotype of the $k^{th}$ contributor in the $i^{th}$ proposed genotype set.

$$\Pr(E|G_i,H) = \begin{cases} 1 & \text{if } E = G_{i1} \cup G_{i2} \cup \cdots \cup G_{iK} \\ 0 & \text{otherwise.} \end{cases}$$

The second term, $\Pr(G_i|H)$, is calculated as in [20]. When population structure is accounted for, the distribution of a set of alleles follows the Dirichlet distribution, assuming populations have reached evolutionary equilibrium. Under this distribution, the probability of seeing an allele depends on how many of that allele type have been seen before. Thus, the probability of observing $A_j$ given that $n_j$ alleles of the $j^{th}$ type have already been observed is

$$\Pr(A_j|\{n_j\} \text{ of type } A_j) = \frac{n_j\theta + (1-\theta)p_j}{n\theta + (1-\theta)}$$

Typically, the list of genotype sets $G$ is selectively composed such that $\Pr(E|G_i,H) = 1$ for each set $i$. Using the above formula and the selectively composed list of genotype

sets, the LR calculation is straightforward.

## 3.3 Methods

We propose using the maximum likelihood approach to weight each genotype set using peak intensities. This method utilizes the peak information without completely eliminating possible genotypes, and easily fits within the likelihood ratio context. To begin we define the likelihood function. If $f(\mathbf{y}|\Theta)$ is the joint density of the observed $\mathbf{y}$, then the likelihood function of parameters $\Theta$ is defined by

$$L(\Theta|\mathbf{y}) = f(\mathbf{y}|\Theta)$$

The maximum likelihood estimators of parameters ($\tilde{\Theta}$) are then the values of $\Theta$ that maximize the likelihood function $L(\Theta|\mathbf{y})$. The likelihood of a particular estimate $\hat{\Theta}$ is given by $L(\hat{\Theta}|\mathbf{y})$ and for two different estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$, if $L(\hat{\Theta}_1|\mathbf{y}) > L(\hat{\Theta}_2|\mathbf{y})$ then $\hat{\Theta}_1$ is said to be more likely or more plausible than $\hat{\Theta}_2$.

In our work, the joint pdf is given by one of two models, the observations $\mathbf{y}$ are the peak intensities $E$, and many sets of parameters are considered. Each set of parameters will include a proposed genotype set $G_i$ and the mixture proportion $m$. For each $G_i$, the maximum likelihood estimate of $m$ is calculated ($\tilde{m}_i$) and the corresponding genotype is weighted by $L(\tilde{m}_i, G_i, H_p|E)$. The likelihood ratio, including the information in the peak intensities, then becomes

$$\begin{aligned} LR &= \frac{\sum_i \Pr(E|G_i, H_p)\Pr(G_i|H_p)}{\sum_i \Pr(E|G_i, H_d)\Pr(G_i|H_d)} \\ &= \frac{\sum_i L(\tilde{m}_i, G_i, H_p|E)\Pr(G_i|H_p)}{\sum_i L(\tilde{m}_i, G_i, H_d|E)\Pr(G_i|H_d)} \end{aligned} \tag{3.2}$$

We examined two different models: a Normal model and a Dirichlet model. Under the Normal model, each peak intensity is the result of a linear model with errors that are

Normally distributed around zero. Under a Dirichlet model, the set of peaks at one locus follow a Dirichlet distribution, with the parameters given by the proposed genotype set.

### 3.3.1 The Linear Expectation

As said previously, it is expected that a person contributes the same amount of genetic material at each locus, and the contribution to a locus is split evenly between the maternal and paternal alleles. A heterozygous locus will have half of the contribution at one allele and half at the other. A homozygous locus will have all of the contribution at the one allele. However, this does not directly translate into raw intensities that are exactly equal to or twice as large as each other across all loci and alleles. Before the markers are typed, the genetic material is PCR amplified. Since different loci have different rates of amplification, the raw intensities are not directly comparable across loci. To remedy this, the intensities should be scaled such that all the intensities at that locus sum to one.

When the intensities are written as proportions of the total intensities at a locus, they can be expressed mathematically using a linear model, as in [39]. If a mixture of $K$ contributors yields a total of $J$ alleles, then the $i^{th}$ proposed set of genotypes ($G_i$) is a $J \times K$ matrix, composed of $K$ vectors of length $J$. Each vector ($G_{ik}$) is a list of the number of copies of each allele in the genotype of the $k^{th}$ contributor. For example, if alleles 12, 14, and 15 are seen at the D8 locus, and the $i^{th}$ proposed genotype is that contributor one is a homozygote for allele 14 and contributor two is a heterozygote for the other two alleles, then $G_{i1} = [0, 2, 0]^T$ and $G_{i2} = [1, 0, 1]^T$. If $m_k$ is the proportion of the mixture contributed by the $k^{th}$ person ($\sum_{k=1}^{K} m_k = 1$), and $Y$ is the scaled intensities, then the linear model is written as below. The factor of $\frac{1}{2}$ is included because the contribution is split evenly between the maternal and paternal alleles.

$$\begin{aligned}
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_J \end{bmatrix} &= \frac{1}{2}\begin{bmatrix} G_{i1} & G_{i2} & \dots & G_{iK} \end{bmatrix} \times \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_K \end{bmatrix} + e \\
Y &= \frac{1}{2}\left[m_1 G_{i1} + m_2 G_{i2} + \cdots + m_K G_{iK}\right] + e \\
Y &= \frac{1}{2} G_i \times M + e
\end{aligned} \tag{3.3}$$

If errors are expected to be zero, then the expected peak intensities given the $i^{th}$ proposed genotype and proportions $M$ are simply $G_i \times M$.

### 3.3.2 The Normal Model

When using the Normal model, the peak intensities follow a multivariate normal distribution, with mean $G_i \times M$ and error $\sigma^2\mathbf{I}$. This model assumes independence of observations, which is incorrect since the intensities at each locus sum to one. However, we feel that despite this violation, the model is still a useful approximation. If $e_{ij}$ is the error associated with the $j^{th}$ allele given the $i^{th}$ proposed genotype, then under the Normal model

$$\begin{aligned}
L(m, \sigma^2, G_i, H|E) &= \Pr(e|m, \sigma^2, G_i, H) \\
&= \prod_{j=1}^{J} \phi(e_{ij}) \\
&= \prod_{j=1}^{J} \phi\left(y_j - \sum_{k=1}^{K} G_{ijk} m_k\right)
\end{aligned}$$

$$
\begin{aligned}
&= \prod_{j=1}^{J} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[\frac{-\left(y_j - \sum_{k=1}^{K} G_{ijk} m_k\right)^2}{2\sigma^2}\right] \\
&= (2\pi\sigma^2)^{-\frac{J}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\left(y_j - \sum_{k=1}^{K} G_{ijk} m_k\right)^2\right]
\end{aligned}
$$

To simplify, we will turn to the two contributor case. Since the mixture proportions sum to one, there is only one independent mixture proportion when there are two contributors. If $Y' = 2Y - G_{i2}$ and $G'_i = G_{i1} - G_{i2}$, the linear model can be written as

$$
\begin{aligned}
Y &= \frac{1}{2} G_i \times M + e \\
Y' &= G'_i m_1 + e
\end{aligned}
$$

This formulation simplifies the calculations to

$$
L(m, \sigma^2, G_i, H|E) = (2\pi\sigma^2)^{-\frac{J}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\left(y'_j - G'_{ij} m_1\right)^2\right] \tag{3.4}
$$

Using this formula, the maximum likelihood estimates $\tilde{m}$ and $\tilde{\sigma}^2$ can easily be calculated. These estimates are the values of $m$ and $\sigma^2$ that maximize the above likelihood.

$$
\tilde{m}_{i1} = \frac{\sum_j^J G'_{ij} y'_j}{\sum_j^J (G'_{ij})^2} \tag{3.5}
$$

$$
\tilde{\sigma}_i^2 = \frac{1}{J}\sum_{j}^{J}\left(y'_j - G'_{ij}\tilde{m}_{i1}\right)^2 \tag{3.6}
$$

Derivations are given in a comment at the end of this chapter. It is of note that these estimators for $\tilde{m}$ and $\tilde{\sigma}^2$ are equivalent to the least-squares estimates, and the normality assumption is not necessary to derive the estimates. The likelihood for $G_i$

is found by using genotype $G_i$ and estimators (3.5) and (3.6) in the likelihood function $L(\tilde{m}, \tilde{\sigma}^2, G_i, H|E)$.

### 3.3.3 The Dirichlet Model

We considered a Dirichlet model because the Dirichlet distribution accounts for the dependencies in the data that the Normal model ignores. The Dirichlet is a multivariate distribution that describes the distribution of a set of variable frequencies $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ given a set of parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$. The frequencies are constrained such that $\sum_j^n p_j = 1$.

$$\Pr(\mathbf{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^n \alpha_j)}{\prod_{j=1}^n \Gamma(\alpha_j)} \prod_{ji=1}^n p_j^{\alpha_j - 1}$$

In our model, the observed frequencies $\mathbf{p}$ are the peak intensities at a specific locus $l$ ($\mathbf{Y}_{(l)}$). The parameters $\boldsymbol{\alpha}$ are the expected intensities under the particular genotype set times the number of alleles at the locus. That is, suppose $G_{i(l)}$ is the subset of $G_i$ that corresponds to the $l^{th}$ locus. $n_l$ is the number of alleles at that locus. In the two person case,

$$\begin{aligned} \boldsymbol{\alpha} &= n_l \times G_{i(l)} \times M \\ &= n_l(mG_{i(l)1} + (1-m)G_{i(l)2}) \end{aligned}$$

The likelihood function is then the probability of each set of peak intensities multiplied across loci.

$$L(\tilde{m}, G_i, H|E) = \prod_{l=1}^{L} \left[ \frac{\Gamma(\sum_{j=1}^{n_l} n_l G_{i(l)j} M)}{\prod_{j=1}^{n_l} \Gamma(n_l G_{i(l)j} M)} \prod_{j=1}^{n_l} y_{(l)j}^{n_l G_{i(l)j} M - 1} \right] \tag{3.7}$$

Determining an analytical solution for the MLE of $m$ is extremely difficult under the Dirichlet model. However, there are algorithms that can be used to quickly maximize the likelihood function with respect to $m$ along the range of zero to one. These algorithms can be found in most statistical computing packages. We used the `optimize` function in R. The `optimize` function uses golden section search and parabolic interpolation to find the maximum ([40], §10.1–§10.2). Golden section search first brackets the maximum, and then successively narrows the range of the brackets until they converge to the maximum with the desired precision. The parabolic interpolation simply speeds up the narrowing process. All actual likelihood functions we've observed have been both unimodal and parabolic, so the golden section search with parabolic interpolation algorithm seems to be sufficiently accurate and sufficiently fast for our purposes.

**Example** We now turn to a small example to illustrate this approach. This example uses three loci from the data in [26, Table 13]. We are using loci D8, D21, and vWA. Suppose that the prosecution hypothesizes that there are two contributors: the suspect and the victim. Under this prosecution hypothesis, there is only one genotype set, that of the suspect and victim. A genotype set is constructed by condensing the two individual genotype vectors into a larger genotype set matrix. So, the single genotype set is

$$G_1 = \left[\begin{array}{cc} V & S \end{array}\right] = \left[\begin{array}{ccc|cccc|cccc} 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \end{array}\right]^T$$

The defense will likely say that the mixture came from the victim and one unknown contributor. The genotype sets under this hypothesis are constructed in a similar fashion to the prosecution hypothesis. Under the defense hypothesis, the unknown contributor must have genotype (17,19) at the vWA locus and (59,65) at the D21 locus. However, at the D8 locus, the unknown contributor could be (10,11), (11,11), or (11,14). So, there are three genotype sets under the defense hypothesis (Table 3.1). When peak intensities are not being used

Table 3.1: Example One: Genotypes of V, S, and U.

| Locus | Sample | V | S | U |
|---|---|---|---|---|
| D8 | 10 | $\begin{bmatrix}1\\0\\1\end{bmatrix}$ | $\begin{bmatrix}1\\1\\0\end{bmatrix}$ | $\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}0\\2\\0\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}$ |
| | 11 | | | |
| | 14 | | | |
| vWA | 16 | $\begin{bmatrix}1\\0\\1\\0\end{bmatrix}$ | $\begin{bmatrix}0\\1\\0\\1\end{bmatrix}$ | $\begin{bmatrix}0\\1\\0\\1\end{bmatrix}$ |
| | 17 | | | |
| | 18 | | | |
| | 19 | | | |
| D21 | 59 | $\begin{bmatrix}0\\0\\1\\1\end{bmatrix}$ | $\begin{bmatrix}1\\1\\0\\0\end{bmatrix}$ | $\begin{bmatrix}1\\1\\0\\0\end{bmatrix}$ |
| | 65 | | | |
| | 67 | | | |
| | 70 | | | |

$$\begin{aligned}\text{LR} &= \frac{\Pr(G_{p1}|H_p)}{\Pr(G_{d1}|H_d)+\Pr(G_{d2}|H_d)+\Pr(G_{d3}|H_d)}\\ &= \frac{(3.69\text{E}{-10})}{(1.36\text{E}{-14})+(5.59\text{E}{-15})+(2.10\text{E}{-14})}\\ &= 9181.4\end{aligned}$$

Now compare this to the LR when using the peak intensities. Under the Normal Model, $Y'$ and $G_i'$ are calculated as in equation 5 for each genotype set. The maximum likelihood estimators are then calculated as in Equations 3.5 and 3.6, and used to find the likelihood of the estimates as in Equation 3.4. Under the Dirichlet Model, the likelihood was maximized with respect to $m$ using `optimize` in the R statistical package. Results are in Table 3.2. The values of $L(\tilde{m})$ shown under the two models are uninformative outside of the likelihood context, so to put them in that context

Table 3.2: Example One: Results

| Mixture | | | Pros | Defense | | |
|---|---|---|---|---|---|---|
| Locus | Allele | $Y$ | $G_1$ | $G_1$ | $G_2$ | $G_3$ |
| D8 | 10 | 0.515 | 1 1 | 1 1 | 1 0 | 1 0 |
| | 11 | 0.031 | 0 1 | 0 1 | 0 2 | 0 1 |
| | 14 | 0.454 | 1 0 | 1 0 | 1 0 | 1 1 |
| vWA | 16 | 0.444 | 1 0 | 1 0 | 1 0 | 1 0 |
| | 17 | 0.089 | 0 1 | 0 1 | 0 1 | 0 1 |
| | 18 | 0.449 | 1 0 | 1 0 | 1 0 | 1 0 |
| | 19 | 0.018 | 0 1 | 0 1 | 0 1 | 0 1 |
| D21 | 59 | 0.060 | 0 1 | 0 1 | 0 1 | 0 1 |
| | 65 | 0.070 | 0 1 | 0 1 | 0 1 | 0 1 |
| | 67 | 0.433 | 1 0 | 1 0 | 1 0 | 1 0 |
| | 70 | 0.437 | 1 0 | 1 0 | 1 0 | 1 0 |
| Normal Model | | $\tilde{m} =$ | 0.8899 | 0.8899 | 0.9191 | .9021 |
| | | $\tilde{\sigma}^2 =$ | 0.0014 | 0.0014 | 0.0041 | 0.003 |
| | | $L(\tilde{m}, \tilde{\sigma}^2) =$ | 7.70E8 | 7.70E8 | 2.19E6 | 3.80E6 |
| Dirichlet Model | | $\tilde{m} =$ | 0.7646 | 0.7646 | 0.8017 | 0.7661 |
| | | $L(\tilde{m}) =$ | 2.61E3 | 2.61E3 | 1.99E3 | 2.50E3 |

$$\begin{aligned} \text{LR} &= \frac{\sum_i L(\tilde{m}_i, \tilde{\sigma}_i^2, G_i, H_p|E)\Pr(G_i|H_p)}{\sum_i L(\tilde{m}_i, \tilde{\sigma}_i^2, G_i, H_d|E)\Pr(G_i|H_d)} \\ &= \frac{L(\tilde{m}_{p1}, \tilde{\sigma}_{p1}^2, G_{p1}, H_p|E)\Pr(G_{p1}|H_p)}{\sum_{i=1}^{3} L(\tilde{m}_{di}, \tilde{\sigma}_{di}^2, G_{di}, H_d|E)\Pr(G_{di}|H_d)} \end{aligned}$$

**Normal Model**

$$\begin{aligned} \text{LR}_\text{N} &= \frac{(7.70\text{E}8)\Pr(G_{p1}|H_p)}{(7.70\text{E}8)\Pr(G_{d1}|H_d) + (2.19\text{E}6)\Pr(G_{d2}|H_d) + (3.80\text{E}6)\Pr(G_{d3}|H_d)} \\ &= \frac{(7.70\text{E}8)(3.69\text{E}{-10})}{(7.70\text{E}8)(1.36\text{E}{-14}) + (2.19\text{E}6)(5.59\text{E}{-15}) + (3.80\text{E}6)(2.10\text{E}{-14})} \\ &= 26895.95 \end{aligned}$$

**Dirichlet Model**

$$\begin{aligned} \text{LR}_\text{D} &= \frac{(2.61\text{E}3)(3.69\text{E}{-10})}{(2.61\text{E}3)(1.36\text{E}{-14}) + (1.99\text{E}3)(5.59\text{E}{-15}) + (2.50\text{E}3)(2.10\text{E}{-14})} \\ &= 9716.4 \end{aligned}$$

In this example, there is an increase in the LR of about 190% when using the peak intensities under the Normal model. Under the Dirichlet, there is a modest increase in the LR of 6%. When there are more possible genotypes under the defense hypothesis, the increase in the LR is often greater, since more genotypes are down-weighted.

This method can also be beneficial to the suspect. If the peak intensities imply a genotype different from the suspect's, then the suspects genotype will be weighted against, and the likelihood ratio will decrease. In the example above, the suspect's genotype was $[1, 1, 0]^T$ at the D8 locus. Suppose that the suspect's was instead $[0, 2, 0]^T$ at D8. Then, without using peak intensities, we have

$$\begin{aligned} \text{LR} &= \frac{(1.43\text{E}{-10})}{(5.59\text{E}{-15}) + (3.55\text{E}{-15}) + (1.07\text{E}{-14})} \\ &= 7203.5 \end{aligned}$$

Using peak intensities under the Normal model,

$$\begin{aligned} \mathrm{LR_N} &= \frac{(2.19\mathrm{E}6)(1.43\mathrm{E}{-10})}{(7.70\mathrm{E}8)(5.59\mathrm{E}{-15}) + (2.19\mathrm{E}6)(3.55\mathrm{E}{-15}) + (3.80\mathrm{E}6)(1.07{-14})} \\ &= 71.9 \end{aligned}$$

Under the Dirichlet Model

$$\begin{aligned} \mathrm{LR_D} &= \frac{(1.99\mathrm{E}3)(1.43\mathrm{E}{-10})}{(2.61\mathrm{E}3)(5.59\mathrm{E}{-15}) + (1.99\mathrm{E}3)(3.55\mathrm{E}{-15}) + (2.50\mathrm{E}3)(1.07{-14})} \\ &= 5879.0 \end{aligned}$$

When using the peak intensities, the likelihood ratio decreases by about 99% under the Normal model, and decreases by 20% under the Dirichlet model. This is because the intensities do not imply the suspect's genotype ($[0, 2, 0]^T$).

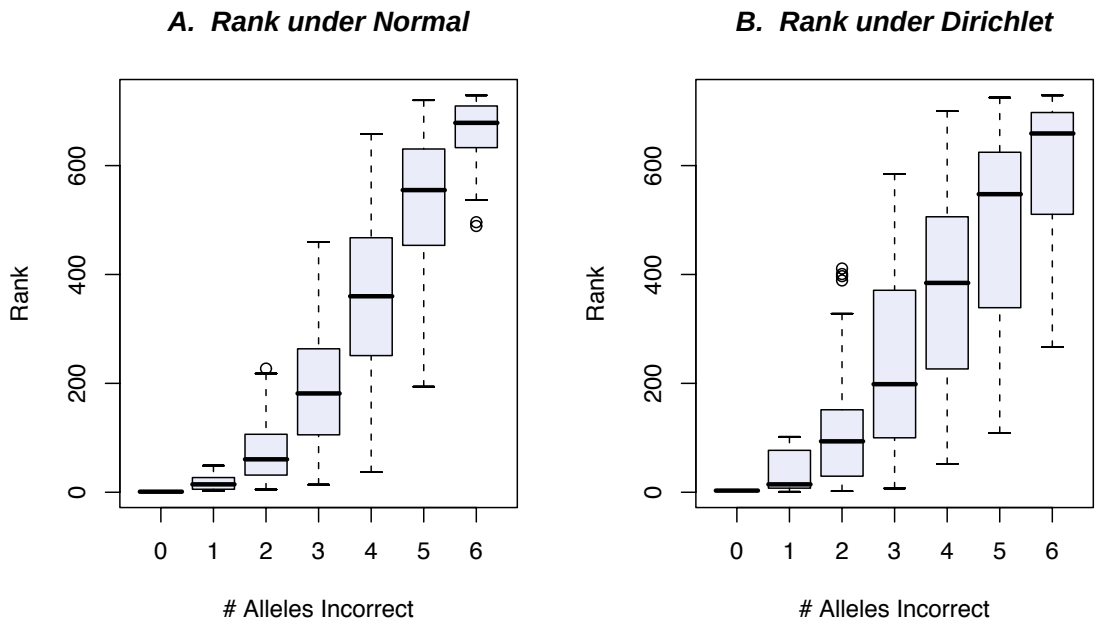
## 3.4 Model Performance

We compared the performance of these two models using several known two-person mixture examples from the literature [26, 39, 23]. For each example, we defined two hypotheses.

> $H_p$: The evidence was supplied by contributors A and B.
> $H_d$: The evidence was supplied by contributor A and one unknown contributor.

Under $H_p$ there is only one genotype set, that of the contributors A and B. However, under $H_d$ there are many possible genotypes for the one unknown contributor. The performance of the models can be evaluated by examining the weights each model gives

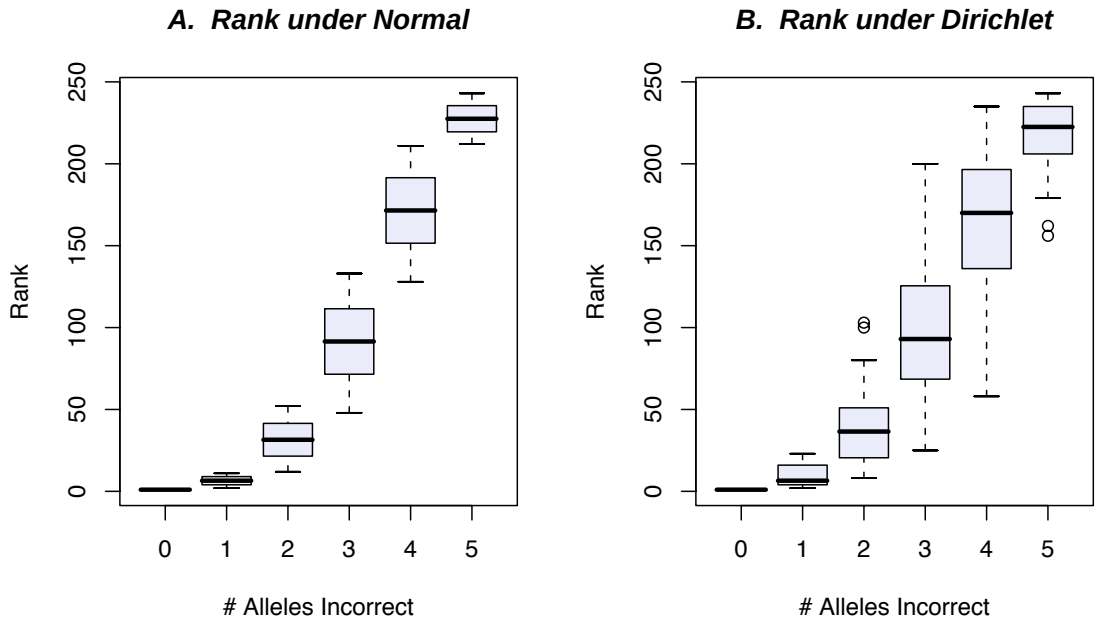
Figure 3.1: Rank of possible genotypes split by number of incorrect alleles (I).




for all possible unknown contributor genotypes, and then noting how correct those genotypes are in relation to the true contributor B. The degree to which a possible unknown genotype is incorrect can be roughly measured by the number alleles in the proposed genotype that differ from the true contributor B's genotype.

We first looked at the rank of the weights under $H_d$ using the Normal model and the Dirichlet model. For each possible genotype, the weight was calculated under both models. Then the weights were sorted, and each genotype was given a rank under each model (the largest weight receives rank 1, next largest receives rank 2, etc). Figure 3.1 and Figure 3.2 are plots of rank, split by the number of incorrect alleles – plot A using the Normal model, plot B using the Dirichlet. The genotypes and peak intensities used are from [23]. These are just two examples, but they are indicative of all the plots we've seen. For both models there is a clear correlation between the number of incorrect alleles and increased rank. Thus, genotypes closer to the true genotype are weighted more.

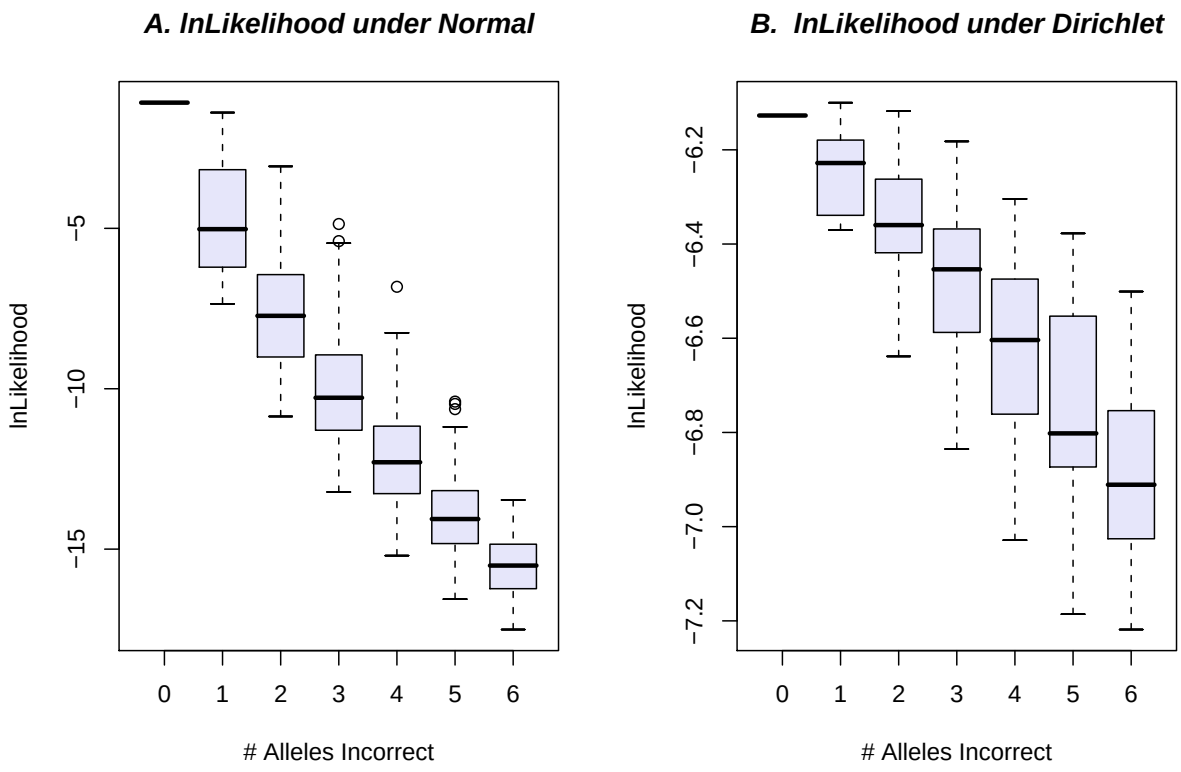
Figure 3.2: Rank of possible genotypes split by number of incorrect alleles (II).




Overlap between two different boxes in the same plot indicates that one group of proposed genotypes (i.e., those genotypes with two alleles incorrect) has genotypes that rank higher and genotypes that rank lower than some of the genotypes in the second group (i.e., those with three alleles incorrect). Since the number of incorrect alleles is a only rough measure of incorrectness, some overlap is expected. However, substantial overlap would mean that the model cannot differentiate between the different groups. Both of these plots show that the Dirichlet has slightly more overlap than the Normal, which is consistent with other plots.

Second we looked at the actual weights of the genotypes under each model. To better compare between models, weights were scaled to be proportions of the sum of the weights under that model. Figure 3.3 and Figure 3.4 are boxplots of the log of these scaled weights, split by the number of incorrect alleles. It is immediately clear that under the Normal model, there is much more difference between the more correct genotypes

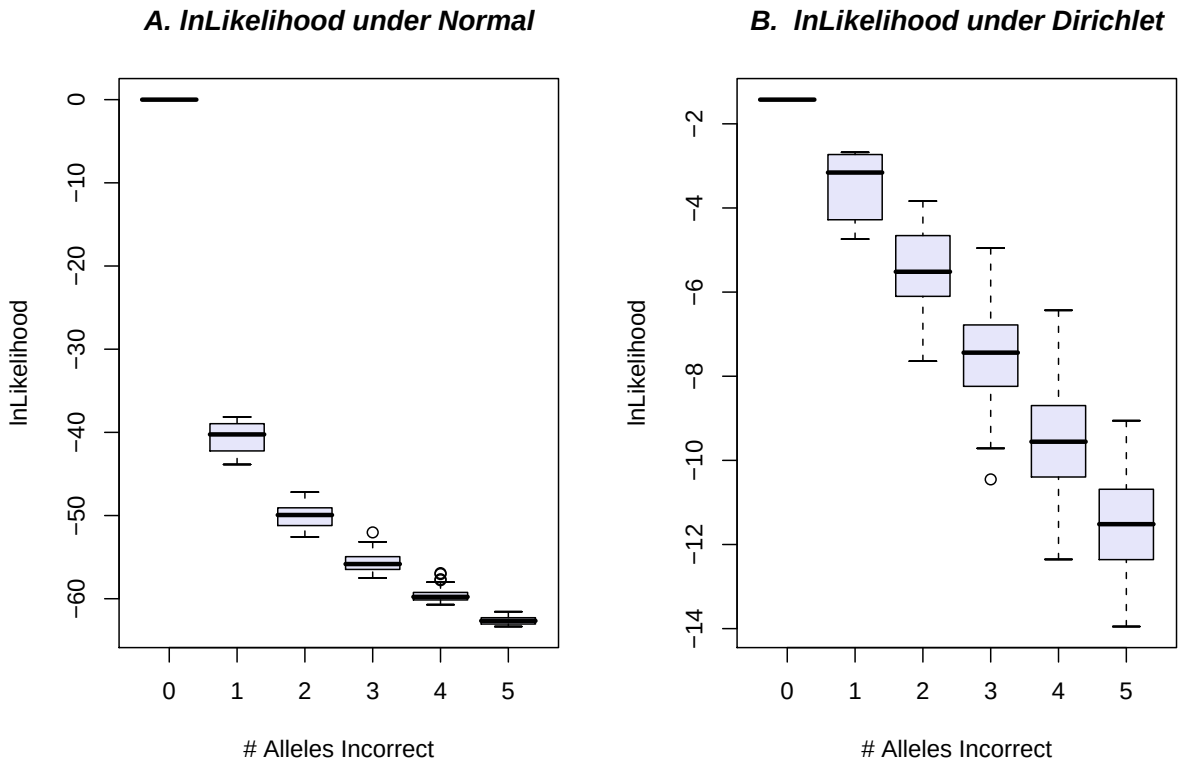
Figure 3.3: Log-likelihood of possible genotypes split by number of incorrect alleles (I).




and the more incorrect genotypes, than under the Dirichlet. This means that, when using the Normal model, the genotypes that better fit the observed peak intensities are weighted heavily, and those that are bad fits are severely down-weighted. When using the Dirichlet, the genotypes that fit the observed intensities are given more moderate weights, and the down-weighting of those with a bad fit is not nearly as severe.

Since both models give better ranks to genotypes closer to the true genotype, the choice of model largely becomes a question of degree rather than of accuracy. To what degree should the observed peak intensities affect the likelihood ratio? If an investigator gives much credence to the intensities, then they may want to use the Normal model. If

Figure 3.4: Log-likelihood of possible genotypes split by number of incorrect alleles (II).




not, then they may want to use the Dirichlet model.

## 3.5 Application to Unseen Alleles

This likelihood approach can be applied to cases of unseen alleles. Suppose that a mixture is typed and two potential contributors are also typed. The first potential contributor (let's say, the victim) is not excluded by any locus and appears to contribute most of the genetic material. The second potential contributor (suspect) is excluded by a single locus. At this particular locus, one of the contributor's alleles is not found in the mixture.

Typically if one of the suspect's alleles is missing they are excluded and therefore are not considered a contributor. However, since the suspect is not excluded at any other locus, and it appears that they only contributed a small proportion of material to the sample, it seems plausible that the allele may actually be in the mixture, but is unseen in the electropherogram because the peak intensity falls below the detection threshold.

The prosecution wants to use the genetic data against the suspect. However, because of the exclusion at one locus, they must either conclude that the suspect did not contribute, or they must claim an error at the troublesome locus and then ignore that locus. Our likelihood method provides a way to use the data without excluding the suspect and without ignoring the locus.

In this situation, the prosecution is claiming that an allele is present but that its peak intensity $y_u$ is less than some detection threshold $t$. Using our method, we can weight each genotype under all the possible values of $y_u$. When considering a single proposed genotype,

$$\begin{aligned}
\Pr(E|H) &= \Pr(E|G_1, H)\Pr(G_1|H) \\
&= \sum_{y_u=0}^{t} \Pr(E, y_u|G_1, H)\Pr(y_u|H)\Pr(G_1|H) \\
&= \sum_{y_u=0}^{t} \mathrm{L}(\tilde{m}, G_1, H|E, y_u)\Pr(y_u|H)\Pr(G_1|H)
\end{aligned} \tag{3.8}$$

The likelihood $\mathrm{L}(\tilde{m}, G_i, H|E, y_u)$ is calculated using our method – the only difference being that we are also including the unseen peak intensity $y_u$ in addition to the seen intensities $E$. The probability of $y_u$ can be calculated from a uniform distribution across the range of $y_u$, or some other appropriate distribution. This likelihood is then used in the LR for the appropriate hypothesis. If multiple genotype sets are proposed, as occurs under defense hypotheses, the above formula is used for each genotype set $G_i$ that has unobserved alleles.

**Unseen Allele Example** Our example is taken from data in [23]. In one of the data sets, the D7 locus has three alleles observed (8, 10, 12), with intensities 1685, 1409, and 144 respectively. From looking at the rest of the data it is obvious that there is probably a mixture proportion of 90% from one contributor and 10% from the second contributor.

Suppose that a suspect is typed, and his genotype seems to match that of the alleles with the smaller peak intensities. However, at the D7 locus the suspect has genotype 11,12 (the victim genotype being 8, 10). To include this locus, we can use our likelihood method to consider alleles whose intensities are below a threshold. We start with the hypotheses

> $H_p$: The evidence was contributed by the victim and the suspect.
> $H_d$: The evidence was contributed by the victim and one unknown contributor.

It is important to remember that when the prosecution says that there is an unseen allele, the defense can easily do the same and claim that any allele can be present but unseen. Thus, when considering only the D7 locus, the hypotheses can be rewritten as

> $H_p$: The evidence was supplied by the suspect contributing genotype (11, 12) and the victim contributing (8, 10).
> $H_d$: The evidence was supplied by the victim contributing (8, 10) and one unknown person contributing either (8, 12), (10, 12), (11, 12), (12, 12), or (12, x), where x is any allele other than 8, 10, 11, and 12.

There are five different genotype sets under the defense hypothesis. The victim's genotype will be in each set, along with one of the five different unknown contributor genotypes mentioned above. In the first {(8,10),(8,12)}, second {(8,10),(10,12)} and fourth set {(8,10),(12,12)}, the defense is proposing some combination of the observed peak intensities. For these three sets, $\Pr(E|G_i, H)$ is calculated using Equation 3.4 for the Normal model, or Equation 3.7 for the Dirichlet model. For the prosecution's hypothesis and the third and fifth genotype set under the defense hypothesis, it is necessary to use

Equation 3.8. The likelihood of each is calculated using $\sum_{y_u=0}^{t} \mathrm{L}(\tilde{m}, G_i, H|E, y_u)\Pr(y_u|H)$ and will be the same for all three. However, the $\Pr(G_i|H)$ for each will differ. A threshold value of 125 was used. This leads to

$$\begin{aligned}\Pr(E|Gi, H) &= \sum_{y_u=1}^{125} \Pr(E, y_u|G_i, H)\Pr(y_u|H) \\ &= \sum_{y_u=1}^{125} \mathrm{L}(\tilde{m}, \{(8,10),(12,x)\}, H|1685, 1409, 144, y_u)\Pr(y_u|H)\end{aligned}$$

The likelihood term is calculated for each $y_u$ using Equations 3.4 or 3.7, depending on the model chosen. For example, when $y_u = 125$, the set of intensities is (1685, 1409, 144, 125) and Y = (0.5010, 0.4189, 0.0428, 0.0372). When $y_u = 1$, the set of intensities is (1685, 1409, 144, 1) and Y = (0.5202, 0.4350, 0.0445, 0.0003). For a proposed genotype where the first two alleles are contributed by the same person, and the second two are contributed by a different person, under the Normal model,

$$\begin{aligned}\mathrm{L}(\tilde{m}, G_i, H|1685, 1409, 144, 125) &= 299.42 \\ \mathrm{L}(\tilde{m}, G_i, H|1685, 1409, 144, 1) &= 161.65\end{aligned}$$

Each likelihood is multiplied by $\Pr(y_u|H)$ and then summed across all $y_u$, with $\Pr(y_u|H)$ taken from a discrete uniform distribution. This leads to $\Pr(E|Gi, H) = 235.44$ under the Normal model. Table 3.3 shows both the probability of intensities given each genotype set and the probability of the genotype set given the hypothesis. Putting these together in the likelihood ratio (3.1) results in

$$LR = \frac{L(\tilde{m}_i, G_1, H_p|E)\Pr(G_1|H_p)}{\sum_{i=1}^{5} L(\tilde{m}_i, G_i, H_d|E)\Pr(G_i|H_d)} = 2.15$$

Table 3.3: Unseen Allele Example

| Hyp | Vic | Sus | Unk | $\Pr(E\|G_i)$ | $\Pr(G_i\|H)$ |
|---|---|---|---|---|---|
| $H_p$ | 8,10 | 11,12 | – | 235.44 | 9.528E-4 |
| $H_{d1}$ | 8,10 | – | 8,12 | 4965.04 | 0.183E-4 |
| $H_{d2}$ | 8,10 | – | 10,12 | 48.44 | 0.316E-4 |
| $H_{d3}$ | 8,10 | – | 11,12 | 235.44 | 0.227E-4 |
| $H_{d4}$ | 8,10 | – | 12,12 | 146.12 | 0.136E-4 |
| $H_{d5}$ | 8,10 | – | 12,a | 235.44 | 0.195E-4 |

The likelihood ratio is only slightly in favor of the prosecution at this locus. This is primarily because, in this example, the peak intensities suggest the defense hypothesis that the contributors were (8,10) and (8,12) much more likely than the prosecution's hypothesis that the 11 allele is unseen.

Typically, an investigator would not be interested in the likelihood ratio of a single locus. However, this example is just for illustrative purposes. The same method can be applied to multiple loci, when only one allele is declared unseen. Extensions could be made for multiple unseen alleles, but these are not explored here.

## 3.6 Discussion

Currently the intensities of allele peaks are not used to weight genotypes in forensic DNA mixture analysis. When intensities are used, it is typically an informal treatment that eliminates highly unlikely genotypes. Any remaining genotypes are weighted equally.

We have outlined two methods that weight proposed genotypes using the observed peak intensities. Our methods include peak intensities in the likelihood ratio by weighting genotypes according to how well the observed intensities agree with expected intensities under the proposed genotype. Both methods are based on likelihood theory; one uses a

Normal model and the other uses a Dirichlet model.

Both models put more weight on the more correct genotypes. Genotypes with fewer incorrect alleles usually get weighted heavier than those genotypes with more incorrect alleles. The primary difference between the models is that under the Normal model the best models are weighted more intensely than under the Dirichlet model. This was clearly seen in our example, where under the Normal model, the most likely genotype had 99.2% of the weight, and the other two genotypes had 0.3% and 0.5% of the weight. Under the Dirichlet model, the most likely genotype had 37% of the weight and the other two had 28% and 35% of the weight.

These methods assume that linearity is maintained in the amplification process. That is, if there is a certain proportion of genetic material in the original sample, that same proportion is maintained throughout the amplification process – across loci and alleles. Differential amplification and stutter artefacts can introduce some nonlinearity, but both are usually minor, predictable and can be corrected prior to using our likelihood methods.

One issue to be resolved is that this method can be computationally intensive. Since we are estimating a common mixture proportion across all loci, the loci are not independent with respect to the calculations. This means that we must consider each possible genotype combination, across all loci. When there is one unknown contributor, there can be at most three possible genotypes at each locus. However, if thirteen loci are considered, there can be as many as $3^{13}$ genotypes. Therefore we may need to consider upwards of one and a half million genotypes. It is important to remember that this is the worst-case scenario for the one-unknown contributor hypothesis using thirteen loci. Typically there will be several thousand possible genotypes instead of over a million. When the number of possible genotypes is in the thousands, a well-written program can easily handle the computation on a desktop computer.

The computational intensity problem is exacerbated when considering more unknowns in the hypothesis. If it is hypothesized that there are two unknown contributors, there can be as many as twelve possible genotypes at each locus. At thirteen loci, that is $12^{13}$ ($1.07E13$) possible genotypes. This is again the worst-case scenario, but even in the

best cases under the two-unknown hypothesis, the number of genotypes to consider can reach into the billions. For these cases, it may be necessary to use high-end computing facilities, or to develop a heuristic to eliminate highly unlikely genotype combinations without evaluating their likelihoods.

Despite this potential for computational intensity, we feel these methods are a valuable contribution to the analysis of forensic DNA mixtures. When these methods are implemented, more information is used and the resulting likelihood ratio is a more accurate measure of the strength of the evidence. The method is not inherently for or against the defendant, and can even benefit the defendant even when they are not excluded. These methods can be applied to cases of allele drop out. These methods could also be applied, with some modifications, to hypotheses involving three or more contributors though computational intensity would be even more of a problem.

**Comment**

To calculate the maximum likelihood estimator for $m$, it is easiest to set the derivative of $L(m, \sigma^2|E, G_i, H)$ with respect to $m$ equal to zero. We show the simpler two contributor case here, but the method can extend to other cases.

$$
\begin{aligned}
L(m_1, \sigma^2|E, G_i, H) &= (2\pi\sigma^2)^{-\frac{J}{2}} \exp\left[-\frac{1}{2\tilde{\sigma}^2}\sum_j^J \left(y'_j - G'_{ij}m_1\right)^2\right] \\
ln(L(m_1, \sigma^2|E, G_i, H)) &= ln\left[(2\pi\sigma^2)^{-\frac{J}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_j^J \left(y'_j - G'_{ij}m_1\right)^2\right]\right] \\
&= -\frac{J}{2}ln(2\pi) - \frac{J}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_j^J \left(y'_j - G'_{ij}m_1\right)^2 \\
\frac{\partial}{\partial m_1}ln(L(m_1, \sigma^2|E, G_i, H)) &= \frac{\partial}{\partial m_1}\left[-\frac{1}{2\sigma^2}\sum_j^J (y'_j - G'_{ij}m_1)^2\right] \\
&= -\frac{1}{2\sigma^2}\sum_j^J \frac{\delta}{\delta m_1}(y'_j - G'_{ij}m_1)^2 \\
&= -\frac{1}{2\sigma^2}\sum_j^J 2(y'_j - G'_{ij}m_1)(-G'_{ij}) \\
&= \frac{1}{\sigma^2}\sum_j^J (G'_{ij}y'_j - (G'_{ij})^2 m_1) \\
&= \frac{1}{\sigma^2}\left[\sum_j^J G'_{ij}y'_j - \sum_j^J (G'_{ij})^2 m_1\right]
\end{aligned}
$$

$$
\begin{aligned}
0 &\Leftarrow \frac{\partial}{\partial m_1} ln(ln(L(m_1, \sigma^2 | E, G_i, H))) \\
&= \frac{1}{\sigma^2} \left[ \sum_j^J G'_{ij} y'_j - \sum_j^J (G'_{ij})^2 \tilde{m}_i \right] \\
\sum_j^J G'_{ij} y'_j &= \sum_j^J (G'_{ij})^2 \tilde{m}_i \\
\tilde{m}_i &= \frac{\sum_j^J G'_{ij} y'_j}{\sum_j^J (G'_{ij})^2}
\end{aligned}
$$

To calculate the estimator for $\sigma^2$, set the derivative of $L(m, \sigma^2 | E, G_i, H)$ with respect to $\sigma^2$ equal to zero.

$$
\begin{aligned}
ln(L(m_1, \sigma^2 | E, G_i, H)) &= -\frac{J}{2} ln(2\pi) - \frac{J}{2} ln(\sigma^2) - \ 12\sigma^2 \sum_j^J \left( y'_j - G'_{ij} m_1 \right)^2 \\
\frac{\partial}{\partial \sigma^2} ln(L(m_1, \sigma^2 | E, G_i, H)) &= \frac{\partial}{\partial \sigma^2} \left[ -\frac{J}{2} ln\sigma^2 - \frac{1}{2\sigma^2} \sum_j^J (y'_j - G'_{ij} m_1)^2 \right] \\
&= -\frac{J}{2\sigma^2} + 2 \left( \frac{1}{2\sigma^2} \right)^2 \sum_j^J (y'_j - G'_{ij} m_1)^2 \\
0 &\Leftarrow -\frac{J}{2\tilde{\sigma}_i^2} + 2 \left( \frac{1}{2\tilde{\sigma}_i^2} \right)^2 \sum_j^J (y'_j - G'_{ij} \tilde{m}_{i1})^2 \\
\frac{J}{2\tilde{\sigma}_i^2} &= 2 \left( \frac{1}{2\tilde{\sigma}_i^2} \right)^2 \sum_j^J (y'_j - G'_{ij} \tilde{m}_{i1})^2 \\
\frac{(2\tilde{\sigma}_i^2)^2}{2\tilde{\sigma}_i^2} &= \frac{2}{J} \sum_j^J (y'_j - G'_{ij} \tilde{m}_{i1})^2 \\
\tilde{\sigma}_i^2 &= \frac{1}{J} \sum_j^J (y'_j - G'_{ij} \tilde{m}_{i1})^2
\end{aligned}
$$

# Chapter 4

# Unique Applications in Forensic Science

## 4.1 Introduction

The analysis of forensic data is by its very nature a messy endeavor. The data are from real people and real situations, and each case may have its own unique twists. For each case, many different factors must be brought into consideration. What exactly is the evidence? How many people contributed to the sample? Who is related to who? What is the effect of population subpopulation? Could the unknown contributor be related to the alleged contributor? What is the prosecution proposing? What is the defense proposing? Usually the answers to these questions are obvious and a case will fit nicely into a certain well-described scenario. However, occasionally a situation will arise that does not fit the well-described scenarios.

During the course of this research, I have come across several of these unusual situations. In these situations, the usual approaches either fail to accurately account for all information, or are just completely not applicable. Many of these are actual cases that were relayed to me by the forensic scientists working on them. Others were suggested in literature and are sufficiently complicated that the peer-reviewed literature on them is unclear or even incorrect.

What follows is an overview of some of the unusual circumstances encountered during this research. First is an introduction to several of the more standard cases: single contributor stains, the two person mixture, paternity index, and the relative index. The following section details some of the more interesting and complex cases I've encountered. They include a paternity index from a mixture, several situations involving relatives, and a canine parentage situation under a variety of assumptions.

## 4.2 Standard Cases

### 4.2.1 Single Contributor

The simplest case is that of a single contributor. Evidence is left at the scene of a crime and it appears there is only one contributor. A suspect is arrested and the genotype of the suspect ($G_S$) matches the genotype of the evidence ($E$). Typically the prosecution will propose that the suspect contributed the evidence, and the defense will propose that the suspect did not contribute the evidence.

$H_p$: The suspect contributed the evidence.

$H_d$: An unknown person contributed the evidence.

The associated likelihood ratio is then

$$\begin{aligned} \text{LR} &= \frac{Pr(E|H_p)}{Pr(E|H_d)} \\ &= \frac{Pr(E|G_S, H_p)Pr(G_S|H_p)}{Pr(E|G_U, G_S, H_d)Pr(G_U, G_S|H_d)} \\ &= \frac{Pr(E|G_S, H_p)Pr(G_S|H_p)}{Pr(E|G_U, G_S, H_d)Pr(G_U|G_S, H_d)Pr(G_S|H_d)} \end{aligned}$$

For the defense hypothesis, the suspect is not a contributor so $Pr(E|G_U, G_S, H_d) = Pr(E|G_U, H_d)$. When the genotype of the suspect or unknown contributor (depending on the hypothesis) matches that of the evidence ($E = G$), then $Pr(E|G, H) = 1$. When $E \neq G$ then $Pr(E|G, H) = 0$. The suspect is known to match the evidence, and typically only those unknown contributor genotypes that match the evidence are considered. So, the $Pr(E|G, H)$ terms conveniently go to one. The suspect's genotype is known, so neither the prosecution's hypothesis nor the defense's hypothesis has any bearing on the known genotype of the suspect, and $Pr(G_S|H_p) = Pr(G_S|H_d) = 1$ and those terms are no longer needed. This leaves

$$\text{LR} = \frac{1}{Pr(G_U|G_S, H_d)} \tag{4.1}$$

Table 4.1: One-locus, one-contributor likelihood ratios, $\theta = 0$

| $E$ | $q$ | $p = 0.01$ | $p = 0.05$ | $p = 0.10$ | $p = 0.20$ |
|---|---|---|---|---|---|
| PP | N/A | 10000 | 400 | 100 | 25 |
| PQ | 0.01 | 5000 | 1000 | 500 | 250 |
| PQ | 0.05 | 1000 | 200 | 100 | 50 |
| PQ | 0.10 | 500 | 100 | 50 | 25 |

If population substructure is not being considered ($\theta = 0$), then the probability of $G_U$ is independent of $G_S$ and, assuming Hardy-Weinberg Equilibrium, the standard Hardy-Weinberg genotype frequencies hold.

$$\begin{aligned} \text{LR} &= \frac{Pr(\text{PP}|H_p)}{Pr(\text{PP}|H_d)} = \frac{1}{p^2} \\ \text{LR} &= \frac{Pr(\text{PQ}|H_p)}{Pr(\text{PQ}|H_d)} = \frac{1}{2pq} \end{aligned} \tag{4.2}$$

These are the formulae for the likelihood ratio at a single locus when there is no population substructure. Table 4.1 shows their values for various allele frequencies. Loci are generally assumed to be independent, so the overall likelihood ratio would be the product of these formulae across all loci.

When considering population substructure, Equation 4.1 still holds. However, the denominator term $Pr(G_U|G_S, H_d)$ is calculated differently. If the suspect and the unknown contributor are from the same subpopulation, $G_U$ is no longer independent of $G_S$. Under the population structure model, alleles in one subpopulation may have different frequencies than the same alleles in a differing subpopulation. The set of frequencies however, are known to follow a Dirichlet distribution if the populations are in evolutionary equilibrium [41]. Under this distribution, when a particular allele is observed, it increases the chance of seeing that allele again. If a set of $n$ alleles has been observed, and if $n_i$ of

them are of type $A_i$, then the probability that the next allele is of type $A_j$ is

$$\Pr(A_i|\{n_i\} \text{ of type } A_i) = \frac{n_i\theta + (1-\theta)p_i}{n\theta + (1-\theta)} \tag{4.3}$$

The population parameter $\theta$ is a measure of the relatedness of the individuals in the subpopulation, and typically ranges between 0.01 and 0.05. Using this formula,

$$\begin{aligned} Pr(E = \text{PP}|G_S = \text{PP}, H_d) &= \frac{((1-\theta)p + 2\theta)((1-\theta)p + 3\theta)}{(1+\theta)(1+2\theta)} \\ Pr(E = \text{PQ}|G_S = \text{PQ}, H_d) &= \frac{2((1-\theta)p + \theta)((1-\theta)q + \theta)}{(1+\theta)(1+2\theta)} \end{aligned} \tag{4.4}$$

Note that the probability of the heterozygote genotype is multiplied by two since the order in which the different alleles are sampled does not matter. The single contributor likelihood ratio is then the reciprocal of these formulae. When $\theta = 0$, Equations 4.4 reduce to Equations 4.2. Table 4.2 shows the value of the single-locus likelihood ratio for different allele frequencies and values of $\theta$. For multi-locus likelihood ratios, the single-locus likelihood ratios are multiplied together.

### 4.2.2 Two Person Mixture

The two-person mixture case is very similar to the one-person case. Evidence is left at the scene of a crime, the evidence is a mixture of DNA from two people. This frequently occurs during rape investigations where a vaginal swab picks up DNA from the victim and the perpetrator. A DNA mixture could also occur as the result of a fight in which two persons are injured, and their blood is mixed together at the scene.

#### Case 1

Usually two-person mixtures have one of two sets of hypotheses. In the first set of hypotheses, the prosecution claims that the suspect and some other known person (usually

Table 4.2: One-locus, one-contributor likelihood ratios with subpopulation.

| $\theta$ | $E$ | $q$ | $p = 0.01$ | $p = 0.05$ | $p = 0.10$ | $p = 0.20$ |
|---|---|---|---|---|---|---|
| 0.001 | PP | N/A | 6439.8 | 364.6 | 95.7 | 24.5 |
| | PQ | 0.01 | 4152.2 | 895.6 | 452.3 | 227.3 |
| | PQ | 0.05 | 895.6 | 193.1 | 97.6 | 49.0 |
| | PQ | 0.10 | 452.3 | 97.6 | 49.3 | 24.8 |
| 0.01 | PP | N/A | 863.5 | 186.5 | 67.1 | 20.7 |
| | PQ | 0.01 | 1300.7 | 435.0 | 237.5 | 124.4 |
| | PQ | 0.05 | 435.0 | 145.5 | 79.4 | 41.6 |
| | PQ | 0.10 | 237.5 | 79.4 | 43.4 | 22.7 |
| 0.03 | PP | N/A | 157.1 | 72.7 | 37.2 | 15.1 |
| | PQ | 0.01 | 347.4 | 175.2 | 108.3 | 61.4 |
| | PQ | 0.05 | 175.2 | 88.6 | 54.8 | 31.0 |
| | PQ | 0.10 | 108.3 | 54.8 | 33.8 | 19.2 |

the victim) contributed the evidence, and the defense claims that the known person and an unknown person contributed the evidence.

Case 1:

$H_p$: The victim and suspect contributed the evidence.

$H_d$: The victim and an unknown person contributed the evidence.

The likelihood ratio is constructed in a similar manner to the one-person case, but this time many different genotypes must be considered. There is the genotype of the victim ($G_V$), the suspect ($G_S$), and the unknown contributor ($G_U$). Under the Case 1

hypotheses,

$$\begin{aligned} \text{LR} &= \frac{Pr(E|H_p)}{Pr(E|H_d)} \\ &= \frac{Pr(E|G_S, G_V, H_p)Pr(G_S, G_V|H_p)}{Pr(E|G_U, G_S, G_V, H_d)Pr(G_U, G_S, G_V|H_d)} \\ &= \frac{Pr(E|G_S, G_V, H_p)Pr(G_S, G_V|H_p)}{Pr(E|G_U, G_S, G_V, H_d)Pr(G_U|G_S, G_V, H_d)Pr(G_S, G_V|H_d)} \end{aligned}$$

This can be cleaned up considerably by noting several things. As in the single-contributor case, the hypotheses have no bearing on the known genotypes of the suspect and the victim. So, $Pr(G_S, G_V|H_p) = Pr(G_S, G_V|H_d) = 1$ and the terms are no longer needed. Under the defense hypothesis, the suspect is not a contributor, so $Pr(E|G_U, G_S, G_V, H_d) = Pr(E|G_U, G_V, H_d)$. For two genotypes $G_i, G_j$, if peak intensities are not being considered,

$$Pr(E|G_i, G_j, H) = \begin{cases} 1 & \text{if } E = G_i \cup G_j \\ 0 & \text{otherwise.} \end{cases}$$

If unknown genotypes for which $Pr(E|G_U, G_V, H_d) = 1$, then the likelihood ratio reduces to

$$\text{LR} = \frac{1}{Pr(G_U|G_S, G_V, H_d)}$$

The probability of the evidence given the two unknown contributor genotypes will often equal one for many different unknown-contributor genotypes ($Pr(E|G_U, G_V, H_d) = 1$). The denominator is the sum over all such genotypes.

$$\begin{aligned} \text{LR} &= \frac{1}{Pr(G_U|G_S, G_V, H_d)} \\ &= \frac{1}{\sum_i Pr(G_{Ui}|G_S, G_V, H_d)} \end{aligned} \tag{4.5}$$

If population substructure is not being considered, then the Hardy-Weinberg frequencies for each $G_{Ui}$ apply. If substructure is being considered, the genotype probabilities are calculated using the sampling formula (Equation 4.3).

**Example** If the mixture is of allelic type P and the suspect and victim are both of type PP, then four copies of allele P have been observed. The probability of seeing a fifth copy of P is then $\frac{(1-\theta)p+4\theta}{1+3\theta}$. Given that this fifth copy, along with the previous four have been seen, the probability of the sixth copy being seen is now $\frac{(1-\theta)p+5\theta}{1+4\theta}$. This leads to the following, which reduces to $p^2$ when $\theta = 0$ and is the reciprocal of the likelihood ratio.

$$Pr(\text{PP}|\text{PPPP}) = \frac{((1-\theta)p+4\theta)((1-\theta)p+5\theta)}{(1+3\theta)(1+4\theta)}$$

**Example** For an example when there are multiple possible genotypes under the defense hypothesis, suppose that $E = \text{PQ}$, $G_S = \text{PP}$, and $G_V = \text{QQ}$. In this situation, the unknown contributor must contribute at least one copy of allele P, and may contribute an addition P or an addition Q. So, $G_{U1} = \text{PP}$ and $G_{U2} = \text{PQ}$. This leads to the following, which reduces to $p^2 + 2pq$ when $\theta = 0$ and is the reciprocal of the likelihood ratio.

$$\begin{aligned}
Pr(G_U|G_S, G_V, H_d) &= Pr(G_{U1}|G_S, G_V, H_d) + Pr(G_{U2}|G_S, G_V, H_d) \\
&= Pr(\text{PP}|\text{PPQQ}) + Pr(\text{PQ}|\text{PPQQ}) \\
&= \frac{((1-\theta)p+2\theta)((1-\theta)p+3\theta)}{(1+3\theta)(1+4\theta)} + \frac{2((1-\theta)p+2\theta)((1-\theta)q+2\theta)}{(1+3\theta)(1+4\theta)} \\
&= \frac{((1-\theta)p+2\theta)\left[((1-\theta)p+3\theta)+2((1-\theta)q+2\theta)\right]}{(1+3\theta)(1+4\theta)}
\end{aligned}$$

Table 4.3 contains $Pr(G_U|G_S, G_V, H_d)$ for all possible genotypes of $G_S$ and $G_V$, with $\theta = 0$ and $\theta \neq 0$ for two-person, single-locus DNA mixtures.

Table 4.3: Genotype probabilities given two observed genotypes, for a one-locus, two-person mixture (Case 1)

| $E$ | $G_S$ | $G_V$ | $G_U$ | $Pr(G_U\|G_V,G_S)$[a] | $Pr(G_U\|G_V,G_S)$[b] |
|---|---|---|---|---|---|
| P | PP | PP | PP | $p^2$ | $((1-\theta)p+4\theta)((1-\theta)p+5\theta)$ |
| PQ | PP | PQ | PP,PQ,QQ | $p^2+2pq+q^2$ | $\big[((1-\theta)p+3\theta)((1-\theta)p+4\theta)+$ $2((1-\theta)p+3\theta)((1-\theta)q+\theta)+$ $((1-\theta)q+\theta)((1-\theta)q+2\theta)\big]$ |
| | | QQ | PP,PQ | $p^2+2pq$ | $\big[((1-\theta)p+2\theta)((1-\theta)p+3\theta)+$ $2((1-\theta)p+2\theta)((1-\theta)q+2\theta)\big]$ |
| | PQ | PP | PQ,QQ | $2pq+q^2$ | $\big[2((1-\theta)p+3\theta)((1-\theta)q+\theta)+$ $((1-\theta)q+\theta)((1-\theta)q+2\theta)\big]$ |
| | | PQ | PP,PQ,QQ | $p^2+2pq+q^2$ | $\big[((1-\theta)p+2\theta)((1-\theta)p+3\theta)+$ $2((1-\theta)p+2\theta)((1-\theta)q+2\theta)+$ $((1-\theta)q+2\theta)((1-\theta)q+3\theta)\big]$ |
| PQR | PP | QR | PP,PQ,PR | $p^2+2pq+2pr$ | $\big[((1-\theta)p+2\theta)((1-\theta)p+3\theta)+$ $2((1-\theta)p+2\theta)((1-\theta)q+\theta)+$ $2((1-\theta)p+2\theta)((1-\theta)r+\theta)\big]$ |
| | PQ | QR | PP,PQ,PR | $p^2+2pq+2pr$ | $\big[((1-\theta)p+\theta)((1-\theta)p+2\theta)+$ $2((1-\theta)p+\theta)((1-\theta)q+2\theta)+$ $2((1-\theta)p+\theta)((1-\theta)r+\theta)\big]$ |
| | | RR | PQ | $2pq$ | $2((1-\theta)p+\theta)((1-\theta)q+\theta)$ |
| PQRS | PQ | RS | PQ | $2pq$ | $2((1-\theta)p+\theta)((1-\theta)q+\theta)$ |

[a] $\theta = 0$

[b] $\theta \neq 0$; Divide each entry in this column by $(1+3\theta)(1+4\theta)$

**Case 2**

Another situation occurs when the prosecution claims the suspect and an unknown person contributed the evidence while the defense says that two unknown persons contributed the evidence.

$H_p$: The suspect and a unknown person contributed the evidence.

$H_d$: Two unknown persons contributed the evidence.

To calculate the likelihood ratio associated with these hypotheses,

$$\begin{aligned}
\text{LR} &= \frac{Pr(E|H_p)}{Pr(E|H_d)} \\
&= \frac{Pr(E|G_S, G_U, H_p)Pr(G_S, G_U|H_p)}{Pr(E|G_S, G_{U1}, G_{U2}, H_d)Pr(G_S, G_{U1}, G_{U2}|H_d)} \\
&= \frac{Pr(E|G_S, G_U, H_p)Pr(G_U|G_S, H_p)Pr(G_S|H_p)}{Pr(E|G_S, G_{U1}, G_{U2}, H_d)Pr(G_{U1}, G_{U2}|G_S, H_d)Pr(G_S|H_d)}
\end{aligned}$$

When considering only those combinations of unknown genotypes that would yield the evidence mixture, this reduces to

$$\text{LR} = \frac{\sum_i Pr(G_{U_i}|G_S, H_p)}{\sum_j Pr(G_{U1_j}, G_{U2_j}|G_S, H_d)} \tag{4.6}$$

Both numerator and denominator are calculated using Equation 4.3 for each possible set of unknown contributor genotypes, and summing across all possible sets. Curran et al. [20] developed a general formula to calculate $Pr(E|H)$ for an arbitrary number of known contributors, known non-contributors, and unknown contributors.

$$Pr(E_l|H) = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1 \ldots - r_{c-2}} \frac{(2x)!2^{h_T+h_V}}{\prod_{h=1}^{c} u_h!} \times \frac{\prod_{h=1}^{c} \prod_{j=0}^{t_h+u_h+v_h-1} [(1-\theta)\tilde{p}_{l,h} + j\theta]}{\prod_{j=0}^{2x+2n_T+2n_V-1} [(1-\theta) + j\theta]} \tag{4.7}$$

Instead of dealing directly with the probability of genotypes, Equation 4.7 centers around the probability of sets of alleles and can be broken up into three parts. The first part, the summations, is used to sum across all possible sets of alleles. For a particular set, the term $(2x)!2^{h_T+h_V}[\prod_{h=1}^{c} u_h!]^{-1}$ is the number of different ways that set can occur. This is the number of different genotype sets that form the same allele set. For example, the allele set (AABB) could be composed of the genotype sets (AA, BB) and (AB, AB). This term considers all possible genotypes for the allele set, and all possible orders of alleles within a genotype. The final term is an extension of (4.3) that is the probability of the allele set, as opposed to just the next allele. Table 4.4 explains the notation used by [20].

In practice, Equation 4.7 is used for each locus under both hypotheses. The likelihood ratio at each locus is calculated, and the product of these likelihood ratios is the overall likelihood ratio. This approach requires the number of unknown contributors to be known. It also assumes that the evidence and all known contributors are typed exactly; stutter and allelic dropout are not factors.

### 4.2.3 Paternity Index

Another common forensic DNA analysis is the paternity index, which is of use in paternity disputes. In a civil case, a plaintiff may claim that her child was fathered by the defendant. The defendant claims that he is not the father, and all three – mother, alleged father, and child – are genotyped with the hope that the DNA will help resolve the case. The set of hypotheses are

$H_p$: The mother and the alleged father are the parents of the child.
$H_d$: The mother and an unknown person are the parents of the child.

The paternity situation is similar in some ways to the mixture situation. The evidence, the child's genotype, is a partial mixture between the father and the mother (Fig. 4.1). The father contributes half of his genotype and the mother contributes half of hers. This is different from the mixture scenario in that each person involved in a mixture

Table 4.4: Notation for Equation 4.7

| | |
|---|---|
| Alleles in the profile of the evidence sample. | |
| $C$ | The set of alleles in the evidence profile. |
| $C_g$ | The set of distinct alleles in the evidence profile. |
| $n_C$ | The known number of contributors to $C$ |
| $h_C$ | The unknown number of heterozygous contributors. |
| $c$ | The known number of distinct allelels in $C_g$. |
| $c_i$ | The unknown number of copies of allele $A_i$ in $C$. $1 \leq c_i \leq 2n_c$, $\sum_{i=1}^{c} c_i = 2n_C$ |
| Alleles from typed people that $H$ declares to be contributors. | |
| $T$ | The set of alleles carried by the declared contributors to $C$. |
| $T_g$ | The set of distinct alleles carried by the declared contributors. |
| $n_T$ | The known number of declared contributors to $C$. |
| $h_T$ | The known number of heterozygous declared contributors. |
| $t$ | The known number of distinct allelels in $T_g$ carried by $n - T$ declared contributors. |
| $t_i$ | The known number of copies of allele $A_i$ in $T$. $1 \leq t_i \leq 2n_T$, $\sum_{i=1}^{c} t_i = 2n_T$. |
| Alleles from unknown people that $H$ declares to be contributors. | |
| $U$ | The set of alleles carried by the unknown contributors to $C$. |
| $x$ | The specified number of unknown contributors to $C : n_C = n_T + x$. |
| $c - t$ | The known number of alleles that are required to be in $U$. |
| $r$ | The known number of alleles in $U$ that can be any allele in $C_g$, $r = 2x - (c - t)$. |
| $n_x$ | The number of different sets of alleles $U$, $n_x = (c + r - 1)!/[(c - 1)!r!]$. |
| $r_i$ | The unknown number of copies of allele $A_i$ in $U$. $0 \leq r_i \leq r$, $\sum_{i=1}^{c} r_i = r$. |
| $u_i$ | The unknown number of copies of $A_i$ in $U : c_i = t_i + u_i$, $\sum_{i=1}^{c} u_i = 2x$. |
| | If $A_i$ is in $C_g$ bu not in $T_g : u_i = r_i + 1$. If $A_i$ is in $C_g$ and also in $T_g : u_i = r_i$. |
| Alleles from typed people that $H$ declares to be non-contributors. | |
| $V$ | The set of alleles carried by typed people declared not to be contributors to $C$. |
| $n_V$ | The known number of people declared not to be contributors to $C$. |
| $h_V$ | The known number of heterozygous declared non-contributors. |
| $v_i$ | The known number of copies of $A_i$ in $V : \sum_i v_i = 2n_V$. |

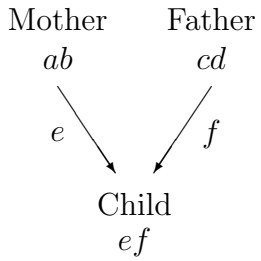



Figure 4.1: Mother, father, and child trio.

contributes all of their genotypes instead of half. Another difference is that in mixtures, if peak intensities are not being considered, the number of copies of an allele is usually unknown. In a paternity case, however, it is known that homozygotes have two copies of the allele while heterozygotes have one copy of two different alleles.

The likelihood ratio is formed in a fashion similar to that of the mixtures and single-contributor case. It is the ratio of the probability of observing the evidence (the genotype of the child, $G_C$) assuming that the mother ($G_M$) and the alleged father ($G_{AF}$) are the true parents of $G_C$ and the probability of observing $G_C$ assuming that the $G_M$ and one unknown contributor $G_U$ are the true parents. This likelihood ratio is known as the Paternity Index ($PI$).

$$\begin{aligned}
PI &= \frac{Pr(E|H_p)}{Pr(E|H_d)} \\
&= \frac{Pr(G_C|G_M, G_{AF}, H_p)Pr(G_M, G_{AF}|H_p)}{Pr(G_C|G_M, G_U, G_{AF}, H_d)Pr(G_M, G_U, G_{AF}|H_d)} \\
&= \frac{Pr(G_C|G_M, G_{AF}, H_p)Pr(G_M, G_{AF}|H_p)}{Pr(G_C|G_M, G_U, G_{AF}, H_d)Pr(G_U|G_M, G_{AF}, H_d)Pr(G_M, G_{AF}|H_d)} \\
&= \frac{Pr(G_C|G_M, G_{AF}, H_p)}{Pr(G_C|G_M, G_U, H_d)Pr(G_U|G_M, G_{AF}, H_d)} \\
&= \frac{Pr(G_C|G_M, G_{AF}, H_p)}{\sum_i Pr(G_C|G_M, G_{U_i}, H_d)Pr(G_{U_i}|G_M, G_{AF}, H_d)}
\end{aligned}$$

When calculating the likelihood ratio for mixtures, the first term, $Pr(E|G,H)$ is either zero or one. For the paternity index, assuming there is no inbreeding, or population substructure, the first term can be either zero, one-quarter, one-half, or one.

For example, if the child is genotype PP and the parents are each genotype PP, then both parents necessarily will give a P allele to the offspring, resulting in a PP genotype. Therefore, $Pr(G_C = \text{PP}|\text{PP},\text{PP},H) = 1$ (Figure 4.2A). If the child is genotype PP and mother is PQ and the father is PP, then mother will pass the P allele half the time, the father will always pass the P allele, and $Pr(G_C = \text{PP}|\text{PQ},\text{PP},H) = \frac{1}{2}$ (Figure 4.2B). If the child is genotype PP and the mother and father are both PQ, then the mother will pass the P allele half the time, the father will pass the P allele half the time, and $Pr(G_C = \text{PP}|\text{PQ},\text{PQ},H) = \frac{1}{4}$ (Figure 4.2C). If the child is genotype PP, the mother is PP, and the father is QR, then it is impossible for the father to pass a P allele. Therefore, $Pr(G_C = \text{PP}|\text{PP},\text{QR},H) = 0$ (Figure 4.2D).

As with DNA mixtures, when an unknown contributor is considered, the probability is summed over all possible genotypes of the unknown person. For example, suppose $G_C = \text{PP}$ and $G_M = \text{PP}$. If there is no substructure, then

$$\begin{aligned}
Pr(E|H_d) &= \sum_i Pr(G_C|G_M,G_{U_i},H_d)Pr(G_{U_i}|G_M,G_{AF},H_d) \\
&= \sum_i Pr(\text{PP}|\text{PP},G_{U_i},H_d)Pr(G_{U_i}) \\
&- Pr(\text{PP}|\text{PP},\text{PP},H_d)Pr(\text{PP}) + Pr(\text{PP}|\text{PP},\text{P}\bar{\text{P}},H_d)Pr(\text{P}\bar{\text{P}}) \\
&= (1)Pr(\text{PP}) + (\frac{1}{2})Pr(\text{P}\bar{\text{P}}) \\
&= p^2 + (\frac{1}{2})2p(1-p) \\
&= p(p+(1-p)) \\
&= p
\end{aligned}$$

In this example, if the alleged father is genotype PP, then $Pr(E|H_p) = 1$ and the paternity index is $PI = \frac{1}{p}$. Table 4.5 contains $Pr(E|H_p)$, $Pr(E|H_d)$, and the $PI$ for all

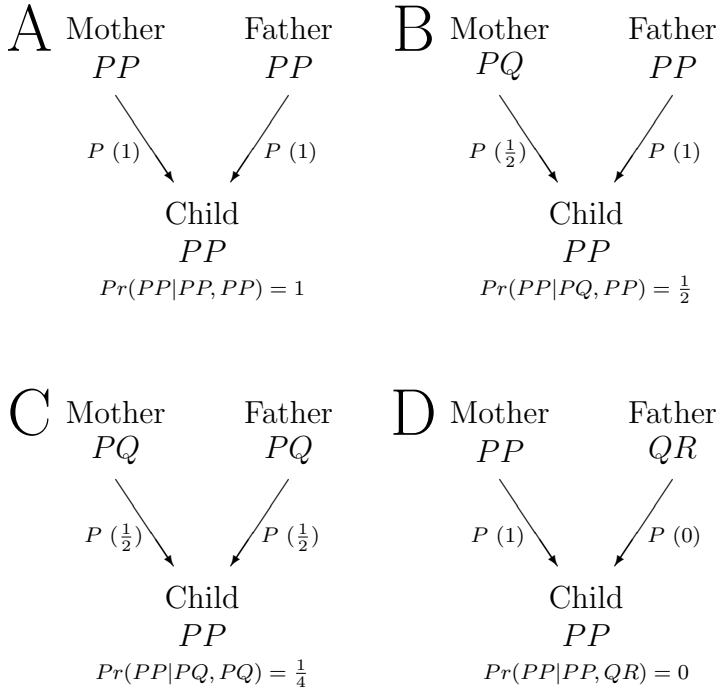



Figure 4.2: Mother, father, and child trios with genotype probabilities of the child.

possible combinations of child, maternal, and (alleged) paternal genotypes. This table is identical to Table 6.2 in [10]. These formulae are single locus measures. For multi-locus profiles, the LR is calculated at each locus, and then multiplied across all loci.

### 4.2.4 Relation by Pedigree

Of particular interest to forensic scientists is the effect of relatedness on the likelihood ratio calculations. Relatedness can come up in a number of different situations. In missing persons cases the DNA of the missing person may not be available, but the DNA of a known relative may be available. The suspect in a criminal case may claim that DNA at the scene came from a sibling or parent instead of some random person. In paternity disputes, the alleged father may claim that a relative is the true father. In all these situations, the claim of relatedness will alter the likelihood ratio calculations. To

Table 4.5: Paternity Index for various combinations of $G_C$, $G_M$, and $G_{AF}$

| $G_C$ | $G_M$ | $G_{AF}$ | Num.[a] | Den.[b] | PI |
|---|---|---|---|---|---|
| PP | PP | PP | 1 | $p$ | $\frac{1}{p}$ |
| | | PQ | $\frac{1}{2}$ | $p$ | $\frac{1}{2p}$ |
| | | QR | 0 | $p$ | 0 |
| | PQ | PP | $\frac{1}{2}$ | $\frac{p}{2}$ | $\frac{1}{p}$ |
| | | PQ | $\frac{1}{4}$ | $\frac{p}{2}$ | $\frac{1}{2p}$ |
| | | QR | 0 | $\frac{p}{2}$ | 0 |
| PQ | PP | QQ | 1 | $q$ | $\frac{1}{q}$ |
| | | QR | $\frac{1}{2}$ | $q$ | $\frac{1}{2q}$ |
| | | RS | 0 | $q$ | 0 |
| | PQ | PP | $\frac{1}{2}$ | $\frac{p+q}{2}$ | $\frac{1}{p+q}$ |
| | | PQ | $\frac{1}{2}$ | $\frac{p+q}{2}$ | $\frac{1}{p+q}$ |
| | | QR | $\frac{1}{4}$ | $\frac{p+q}{2}$ | $\frac{1}{2(p+q)}$ |
| | | RS | 0 | $\frac{p+q}{2}$ | 0 |
| | PR | QQ | $\frac{1}{2}$ | $\frac{q}{2}$ | $\frac{1}{q}$ |
| | | QR | $\frac{1}{4}$ | $\frac{q}{2}$ | $\frac{1}{2q}$ |
| | | RS | 0 | $\frac{q}{2}$ | 0 |

[a]Num $= Pr(G_C|G_M, G_{AF}, H_p)$

[b]Den $= \sum_i Pr(G_C|G_M, G_{U_i}, H_d)Pr(G_{U_i})$

account for the relatedness, we must first consider what it means to be related.

### Relatedness

Two people are said to be "related" if they share one or more common ancestors [10]. The closer the common ancestor, the more related the two people. This commonality of ancestor(s) genetically translates into the two persons having more alleles in common than would be expected randomly. If two persons are unrelated, we would expect the frequency of common alleles to be determined by the frequency of the alleles in the

population. However, if two persons are related, then the frequency of common alleles depends on the closeness of the relationship, and the frequency of the alleles.

One important concept used to describe the relatedness between two people is the concept of *identity by descent* (ibd). Two alleles are said to be ibd if they are copies of the same allele from a common ancestor. Note that *identity by descent* is not the same as *identity by state* (ibs). Alleles that are identical by state are alleles that are of the same allelic type. However, that does not mean they are necessarily from a common ancestor. All alleles that are ibd are ibs, but not all alleles that are ibs are ibd. Identity by descent status is often written in terms of equivalence. That is, if alleles $a$ and $b$ are ibd, then $a \equiv b$. If alleles $a$ and $b$ are not ibd, then $a \not\equiv b$.

The relatedness of two non-inbred relatives can be described by the probability that they will have 0, 1, or 2 pairs of alleles that are identical by descent. These probabilities, $P_0$, $P_1$, and $P_2$ respectively, are from [42] and are derived in Evett and Weir [10].

$$\begin{aligned} P_0 &= 1 - 4\theta_{XY} + 2\Delta_{\ddot{X}+\ddot{Y}} \\ P_1 &= 4(\theta_{XY} - \Delta_{\ddot{X}+\ddot{Y}}) \\ P_2 &= 2\Delta_{\ddot{X}+\ddot{Y}} \end{aligned}$$

The value $\theta_{XY}$ is the probability that two alleles (one taken at random from $X$ and one taken from $Y$) are identical by decent. The value $\Delta_{\ddot{X}+\ddot{Y}}$ refers to the identity of two pairs of alleles – one allele in each pair from each individual. Using Table 4.8 from Evett and Weir, the probabilities can be simplified into a single table, based on varying degrees of relation (Table 4.6). These probabilities were first developed by Malécot in [43].

**Half-sib Example** Figure 4.3 shows the pedigree for a pair of half-sibs. The letters in the pedigree are labels for alleles, and do not indicate the state or type of the allele. If there is no population substructure or inbreeding then none of the parental genotypes are identical by descent ($a \not\equiv b \not\equiv c \not\equiv d \not\equiv e \not\equiv f$). This assumption necessitates that $g \not\equiv h \not\equiv j$ and $g \not\equiv i \not\equiv j$. However, since $h$ and $i$ come from the same person, the father, it could be that $h$ and $i$ are copies of the same allele — that they are ibd ($h \equiv i$). On

Table 4.6: Probability of 0, 1, or 2 alleles being ibd, under varying degrees of relationship.

| Relationship | $P_0$ | $P_1$ | $P_2$ |
|---|---|---|---|
| Not related | 1 | 0 | 0 |
| Parent–child | 0 | 1 | 0 |
| Grandparent–grandchild | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 |
| Full sibs | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |
| Half sibs | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 |
| Uncle–nephew | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 |
| First cousins | $\frac{3}{4}$ | $\frac{1}{4}$ | 0 |
| Double first cousins | $\frac{9}{16}$ | $\frac{3}{8}$ | $\frac{1}{16}$ |
| Identity | 0 | 0 | 1 |

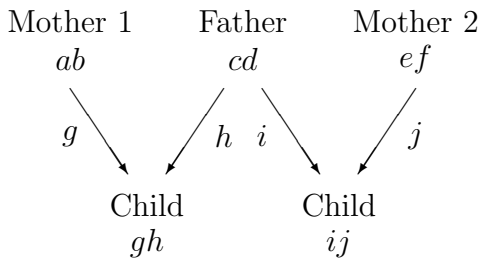



Figure 4.3: Half-sib pedigree

average, half the time the father will pass the same alleles to the two children. So, for half-sibs, the probability that they have one allele ibd allele is one-half ($P_1 = \frac{1}{2}$). If the father does not pass the same allele to each child, they have no alleles ibd, so $P_0{=}\frac{1}{2}$. In the absence of inbreeding and substructure, $P_2 = 0$, since the half-sibs cannot have two pairs of ibd alleles.

### Joint Genotype Probabilities

The values of $P_0$, $P_1$, and $P_2$ and the allele frequencies can be used to calculate the joint probability of two related genotypes. If no alleles are ibd there are four distinct alleles (even if some are identical by state), and the joint probability is the product of the allele frequencies. If one pair of alleles is ibd, there are only three distinct alleles because two of them are copies of the distinct ancestral allele. The frequency is then the product of the frequencies of these three distinct alleles. If two pairs are ibd, there are only two distinct alleles, and the frequency is the product of the frequencies of these two distinct alleles.

The total probability for the genotype is then the sum across all three possibilities (0, 1, or 2 alleles ibd) of the frequency of the distinct alleles, weighted by the frequency of the alleles being distinct. Tables 4.6 and 4.7 have the values necessary for this calculation. The values of $P_0$, $P_1$, and $P_2$ are in Table 4.6. Each row in Table 4.7 represents a different pair of genotypes, and the three columns on the right are the frequencies of the distinct alleles.

**Half-sib Example Continued** Suppose the half-sibs above are typed and both are homozygous with the same allele at a particular marker. By Table 4.7 we see that for any relationship the joint probability of two homozygotes is as follows.

$$Pr_R(\mathrm{PP}, \mathrm{PP}) = P_0 p^4 + P_1 p^3 + P_2 p^2$$

Since these two are half-sibs, $P_0 = P_1 = \frac{1}{2}$ and $P_2 = 0$. Therefore, the joint probability reduces to

$$Pr_{\mathrm{HS}}(\mathrm{PP}, \mathrm{PP}) = \frac{p^3(1+p)}{2}$$

The joint genotype probabilities under various other relationships are easily found using Tables 4.6 and 4.7. NR, PC, FS, and FC stand for Not Related, Parent-Child, Full

Table 4.7: Joint probabilities of two non-inbred related genotypes.

| Genotypes | $P_0$ | $P_1$ | $P_2$ |
|---|---|---|---|
| PP,PP | $p^4$ | $p^3$ | $p^2$ |
| PP,QQ | $p^2q^2$ | | |
| PP,PQ | $2p^3q$ | $p^2q$ | |
| PP,QR | $2p^2qr$ | | |
| PQ,PQ | $4p^2q^2$ | $pq(p+q)$ | $2pq$ |
| PQ,PR | $4p^2qr$ | $pqr$ | |
| PQ,RS | $4pqrs$ | | |

Sibs, and First Cousins respectively.

$$\begin{aligned} Pr_R(\mathrm{PP},\mathrm{PP}) &= P_0p^4 + P_1p^3 + P_2p^2 \\ Pr_{\mathrm{NR}}(\mathrm{PP},\mathrm{PP}) &= p^4 \\ Pr_{\mathrm{PC}}(\mathrm{PP},\mathrm{PP}) &= p^3 \\ Pr_{\mathrm{FS}}(\mathrm{PP},\mathrm{PP}) &= \frac{1}{4}p^4 + \frac{1}{2}p^3 + \frac{1}{4}p^2 \\ Pr_{\mathrm{FC}}(\mathrm{PP},\mathrm{PP}) &= \frac{3}{4}p^4 + \frac{1}{4}p^3 \end{aligned}$$

**Unknown Relative**

Another common situation involving relatives is that of the unknown relative. Suppose DNA evidence is found at a scene and a suspect is arrested. The suspect claims that the real contributor is a relative. The relative's genotype is not available, so the relevant hypotheses would be

$H_p$: The suspect contributed the evidence.

$H_d$: A relative of the suspect, with unknown genotype, contributed the evi-

dence.

In the associated likelihood ratio, the denominator will be the probability of seeing an unknown contributor $G_U$ given we've seen the related genotype $G_S$ ($Pr_R(G_U|G_S, H_d)$). This conditional probability can be derived from the joint genotype probabilities. The joint probability $Pr_R(G_U, G_S|H_d)$ can be calculated from Tables 4.6 and 4.7. Dividing these joint probabilities by the probability of the suspect's genotype is the probability of the evidence given that a relative ($G_S$) has been observed.

$$Pr_R(G_U|G_S, H_d) = \frac{Pr_R(G_U, G_S|H_d)}{Pr(G_S|H_d)} \tag{4.8}$$

**Half-sib Example Continued** If both evidence and suspect are homozygotes for allele P, then the defense will propose that an unknown relative of the suspect will have the same genotype. The joint probability of the unknown and the suspect is $P_0p^4 + P_1p^3 + P_2p^2$. Dividing this by the suspect's genotype yields the conditional probability $P_0p^2 + P_1p + P_2$. This can be used to form the likelihood ratio.

$$\begin{aligned} \text{LR} &= \frac{Pr(E|G_S, H_p)}{Pr_R(E|G_S, H_d)} \\ &= \frac{1}{Pr(E|G_U, H_d)Pr_R(G_U|G_S, H_d)} \\ &= \frac{1}{Pr_R(PP|PP, H_d)} \\ &= \frac{1}{P_0p^2 + P_1p + P_2} \end{aligned}$$

If the suspect and the unknown are half-sibs as in our previous examples, then $P_0 = P_1 = \frac{1}{2}$ and $P_2 = 0$. This leads to the following likelihood ratio. Formulae for $Pr_R(G_1|G_2, H_d)$ under various combinations of genotypes are given in Table 4.8.

$$\text{LR} = \quad = \frac{2}{p(1+p)} \tag{4.9}$$

Table 4.8: Conditional probabilities of non-inbred related genotypes, $Pr_R(G_1|G_2)$.

| $G_1$ | $G_2$ | $P_0$ | $P_1$ | $P_2$ |
|---|---|---|---|---|
| PP | PP | $p^2$ | $p$ | 1 |
| PP | QQ | $p^2$ | | |
| PP | PQ | $p^2$ | $p/2$ | |
| PP | QR | $p^2$ | | |
| PQ | PQ | $2pq$ | $(p+q)/2$ | 1 |
| PQ | PR | $2pq$ | $q/2$ | |
| PQ | RS | $2pq$ | | |

**Relative Index**

Similar to the paternity index is the relative index. The relative index is used in situations where there is a question about the relatedness of two people. For example, in an estate dispute, someone may claim to be an illegitimate child of the deceased. If the deceased's DNA is not available, often times a known child's DNA is used. This leads to the question of whether or not the two individuals are half-siblings. Another common use of the relative index is that of a missing person and remains situation. Suppose a person goes missing, and some time later remains of some unknown person are found. No DNA that is known to be from the missing person is available, but the DNA of a relative is available. Is the deceased related to the known relative? The relative index is also of use in mass casualty situations, such as the World Trade Center tragedy [13]. It may be known that a particular person has died in the disaster, but remains are beyond recognition. Again, if DNA known to be from the deceased is unavailable, then DNA from a relative may be used. There is then the question of whether or not the unknown remains are related to the known relative of the victim. A typical set of hypotheses for these situations might be

$H_p$: The two persons are siblings.

$H_d$: The two persons are unrelated.

Note that any relationship could be substituted for "siblings." Also, in these situations there may not be a prosecution or defense as there is with a criminal case. However, the $H_p$, $H_d$ notation is still used to maintain the convention of $H_p$ specifying the numerator hypothesis, and $H_d$ specifying the denominator hypothesis.

The joint genotype probabilities from Table 4.7 can be used to construct the relative index. If we consider the two genotypes $(G_1, G_2)$ as the "evidence," and the hypotheses as the presence of a specific relationship versus the absence of a relationship, then the logical likelihood ratio would be as follows, where under $H_p$, the values of $P_0$, $P_1$, and $P_2$ are determined by the hypothesized relationship, from Table 4.6.

$$\mathrm{LR}_R \;=\; \frac{Pr_R(E|H_p)}{Pr_{NR}(E|H_d)} = \frac{Pr(G_1, G_2|P_0 \neq 1)}{Pr(G_1, G_2|P_0 = 1)} \tag{4.10}$$

This is the relative index. It is the likelihood of the genotypes given the relationship against no relationship. An alternative but equal formulation of the relative index was given by Brenner and Weir [13].

They define $\mathcal{U}$ as the average of the $u_i$ variables ( $\frac{1}{4}(u_1 + u_2 + u_3 + u_4)$ ), where the $u_i$ variables are defined for the four allele pairs 1: *ac*; 2: *ad*; 3:*bc*; 4: *bd*. If the pair $i$ are the same allele, then $u_i$ is the reciprocal of the allele frequency. If the pair are not the same, then $u_i$ equals zero. Define $\mathcal{W}$ as $\frac{1}{2}(u_1u_4 + u_2u_3)$. Thus, $I = P_0 + \mathcal{U}P_1 + \mathcal{W}P_2$ is the likelihood of the suspected relationship vs the individuals being unrelated.

**Half-sib Example Continued.** In the previous half-sib example, the values of $P_0$, $P_1$, and $P_2$ are $\frac{1}{2}$, $\frac{1}{2}$, and 0 respectively. Using these values in Equation 4.10 gives the likelihood ratio

$$\mathrm{LR}_{\mathrm{HS}} \;=\; \frac{Pr(G_1, G_2|P_0 = \frac{1}{2}, P_1 = \frac{1}{2})}{Pr(G_1, G_2|P_0 = 1)}$$

Table 4.9: Relative Index Formulae.

| Genotypes | $P_0$ | $P_1$ | $P_2$ |
|---|---|---|---|
| PP,PP | 1 | $4(1/4p)$ | $2(1/2p^2)$ |
| PP,QQ | 1 | | |
| PP,PQ | 1 | $2(1/4p)$ | |
| PP,QR | 1 | | |
| PQ,PQ | 1 | $(1/4p + 1/4q)$ | $1/2pq$ |
| PQ,PR | 1 | $1/4p$ | |
| PQ,RS | 1 | | |

If $G_1$ and $G_2$ are both homozygotes for allele $P$, then

$$\begin{aligned} \mathrm{LR}_{HS} &= \frac{Pr(\mathrm{PP},\mathrm{PP}|P_0 = \frac{1}{2}, P_1 = \frac{1}{2})}{Pr(\mathrm{PP},\mathrm{PP}|P_0 = 1)} \\ &= \frac{(P_0p^4 + P_1p^3 + P_2p^2|P_0 = \frac{1}{2}, P_1 = \frac{1}{2})}{(P_0p^4 + P_1p^3 + P_2p^2|P_0 = 1)} \\ &= \frac{\frac{1}{2}p^4 + \frac{1}{2}p^3}{p^4} \\ &= \frac{1}{2} + \frac{1}{2p} \end{aligned}$$

This would be interpreted as, "The pair of genotypes $G_1$ =PP, $G_2$ =PP are $\mathrm{LR}_{HS}$ times more likely if the two persons are half-sibs than if they are unrelated." Table 4.9 shows the relative index for all possible pairs of genotypes, for arbitrary allele frequencies, and arbitrary values of $P_0$, $P_1$, and $P_2$. These index formulae are single locus measures. Forensic markers are generally considered independent, so the overall index is the product of the single-locus measures.

## 4.3 Non-Standard Cases

### 4.3.1 Paternity Index from Mixtures

The article by Liao et al [6], discusses an interesting problem: determining a paternity index (PI) from a DNA mixture. The authors start with the usual PI hypotheses.

$H_p$: The alleged father is the true father of the child.
$H_d$: The alleged father is not the true father of the child.

The difference between this problem and a standard PI is that in this case the genotype of the alleged father is not known; it is inferred from a DNA mixture stain that is contributed by the alleged father and mother. The authors incorrectly apply formulae from Evett and Weir [26]. Their mistake is best seen by examining each hypothesis separately.

A more precise definition of $H_p$ would be: Mother (M) and an unknown co-contributor to mixture (S) are the parents of child (C). Let genotypes $G_M$, $G_S$, and $G_C$ be the genotypes of the mother, the mixture stain, and the child respectively. Using these genotypes as the evidence, the correct $\Pr(E|H_p)$ can be formed. Expansion of the probability is carried out by the third law of probability and the law of total probability is invoked to handle the mixture [26]. Every possible pair of alleles that is a subset of $G_S$ must be considered. These pairs represent the possible genotypes of the alleged father ($G_{AF}$)

$$\begin{aligned}
Pr(E|H_p) &= Pr(G_C, G_M, G_S|H_p) \\
&= P(G_C|G_S, G_M, H_p)Pr(G_S, G_M|H_p) \\
&= Pr(G_C|G_S, G_M, H_p)Pr(G_S|G_M, H_p)Pr(G_M|H_p) \\
&= \sum_i Pr(G_C|G_{AF_i}, G_S, G_M, H_p)Pr(G_{AF_i}, G_S|G_M, H_p)Pr(G_M|H_p) \\
&= \sum_i Pr(G_C|G_{AF_i}, G_S, G_M, H_p)Pr(G_{AF_i}|G_S, G_M, H_p)Pr(G_S|G_M, H_p)Pr(G_M|H_p)
\end{aligned}$$

$G_M$ and $G_S$ are both known under $H_p$, so $\Pr(G_M|H_p) = \Pr(G_S|G_M, H_p) = 1$. Since we are iterating over possible $G_{AF}$, the $G_S$ in the first term is unnecessary. This leaves

$$Pr(E|H_p) = \sum_i Pr(G_C|G_{AF_i}, G_M, H_p)Pr(G_{AF_i}|G_S, G_M, H_p)$$

It is in the calculation of this probability that Liao et al, were mistaken. The authors considered only those $G_{AF_i}$ that could produce $G_C$. While it is true that the hypothesis states that A must be the father of C, the expansion of the probability formulae dictates that A is conditional on S and M, not C. It appears that Liao et al have made A conditional on C and M, instead of S and M.

The first term, $Pr(G_C|G_{AF_i}, G_M, H_p)$, is the same as Evett and Weir, Table 6.2, numerator column [10]. Under $H_p$, one of the genotypes that are iterated through must be the correct genotype. Thus, $\sum_i Pr(G_{AF_i}|G_S, G_M, H_p) = 1$, and $\Pr(G_{AF_i}|G_S, G_M, H_p)$ is defined as

$$Pr(G_{AF_i}|G_S, G_M, H_p) = \frac{Pr(G_{AF_i})}{\sum_i Pr(G_{AF_i})}$$

The defense hypothesis ($H_d$) states that the alleged (A) is not a contributor. To restate, the mother (M) and an unknown contributor (U) are the parents. Under this hypothesis, the mixed stain is irrelevant since the father is an unknown contributor (U) and is not the co-contributor to S. Thus, the denominator is the same as in the standard PI.

$$\begin{aligned} Pr(E|H_d) &= Pr(G_C, G_M|H_d) \\ &= Pr(G_C|G_M, H_d)Pr(G_M|H_d) \\ &= \sum_i Pr(G_C|G_{U_i}, G_M, H_d)Pr(G_{U_i}|G_M, H_d) \\ &= \sum_i Pr(G_C|G_{U_i}, G_M, H_d)Pr(G_{U_i}|H_d) \end{aligned}$$

Though the notation is slightly different, this yields the same results found in Evett and Weir (1998, Table 6.2, denominator column). Using P($E|H_p$) from above and the standard P($E|H_d$), the PI can be calculated.

$$PI = \frac{Pr(E|H_p)}{Pr(E|H_d)} = \frac{\sum_i Pr(G_C|G_{AF_i}, G_M, H_p)Pr(G_{AF_i}|G_S, H_p)}{\sum_i Pr(G_C|G_{U_i}, G_M, H_d)Pr(G_{U_i}|H_d)} \tag{4.11}$$

Now we look at the examples used in [6].

**Example 1** $G_S = \{P\}$, $G_M = \{PP\}$, $G_C = \{PP\}$. It follows that under $H_p$ the alleged must have genotype $G_{AF} = \{PP\}$ and under $H_d$ the unknown contributor has genotype $G_U = \{Px\}$, where $x$ is any allele (including $P$). This result is the same as in Liao et al. Using Equation 4.11,

$$\begin{aligned} PI &= \frac{Pr(PP|PP, PP, H_p)Pr(PP|P, PP, H_p)}{Pr(PP|PP, PP, H_d)Pr(Px|H_d)} \\ &= \frac{(1)Pr(PP|P, PP, H_p)}{(1)Pr(Px|H_d)} = \frac{1}{p} \end{aligned}$$

**Example 2** $G_S = \{PQ\}$, $G_M = \{PQ\}$, $G_C = \{PP\}$. It follows that under $H_p$ the alleged must have one of $G_{AF} = \{PP, PQ, QQ\}$ and under $H_d$ the unknown contributor has genotype $G_U = \{Px\}$, where $x$ is any allele (including $P$).

$$\begin{aligned} PI &= \Big[Pr(PP|PP, PQ, H_p)Pr(PP|PQ, PQ, H_p) \\ &\quad + Pr(PP|PQ, PQ, H_p)Pr(PQ|PQ, PQ, H_p) \\ &\quad + Pr(PP|QQ, PQ, H_p)Pr(QQ|PQ, PQ, H_p)\Big] \\ &\quad \times \frac{1}{Pr(PP|Px, PQ, H_d)Pr(Px|H_d)} \\ &= \frac{1}{(p/2)}\left[\frac{1}{2}\Big(\frac{p^2}{p^2+2pq+q^2}\Big) + \frac{1}{4}\Big(\frac{2pq}{p^2+2pq+q^2}\Big) + 0\Big(\frac{q^2}{p^2+2pq+q^2}\Big)\right] \\ &= \frac{2}{p}\left[\frac{p^2+pq}{2(p+q)^2}\right] = \frac{1}{p+q} \end{aligned}$$

**Example 3** $G_S = \{PQR\}$, $G_M = \{PQ\}$, $G_C = \{PR\}$. It follows that under $H_p$ the alleged must have one of $G_{AF} = \{PR, QR, RR\}$ and under $H_d$ the unknown contributor has genotype $G_U = \{Rx\}$, where $x$ is any allele (including $R$).

$$\begin{aligned} PI &= \Big[Pr(PR|PR, PQ, H_p)Pr(PR|PQR, PQ, H_p) \\ &\quad +Pr(PR|QR, PQ, H_p)Pr(QR|PQR, PQ, H_p) \\ &\quad +Pr(PR|RR, PQ, H_p)Pr(RR|PQR, PQ, H_p)\Big] \\ &\quad \times \frac{1}{Pr(PR|Rx, PQ, H_d)Pr(Rx|H_d)} \\ &= \frac{1}{r/2}\left[\frac{1}{4}\Big(\frac{2pr}{2pr+2qr+r^2}\Big) + \frac{1}{4}\Big(\frac{2qr}{2pr+2qr+r^2}\Big) + \frac{1}{2}\Big(\frac{r^2}{2pr+2qr+r^2}\Big)\right] \\ &= \frac{p+q+r}{r(2p+2q+r)} \end{aligned}$$

Using notation as in Liao et al., the PI for all two contributor cases can be found in Table 4.10. A comparison between our results and the results of Liao et al. reveals that the results are the same in six of the eleven cases. However in all five cases where Liao et al are incorrect, their errors can be shown to be prejudicial against the suspected father.

### 4.3.2 Relatives and Mixtures

When faced with DNA evidence against his client, a defense attorney in Tasmania, Australia said that the defendant would claim in court "his son did it." The son had not been tested, so the prosecutor needed to determine the probability of the evidence if the son, rather than the father, was the perpetrator [Laszlo Szabo, personal correspondence]. Since father and son share half their DNA, the probability of the son having the evidence depends on the profile of the father. It is not uncommon that the accused, when confronted with DNA evidence, will say a relative committed the crime. As more attorneys become familiar with the way relatedness affects probability estimates, the more this issue will be brought up.

Table 4.10: Paternity index when paternal genotype is inferred from a mixture.

| $G_S$ | $G_M$ | $G_C$ | $G_{AF}$ | $\Pr(E\|H_p)$ | $\Pr(E\|H_d)$ | PI |
|---|---|---|---|---|---|---|
| P | PP | PP | PP | 1 | $p$ | $\frac{1}{p}$ |
| PQ | PQ | PQ | PP, PQ, QQ | $\frac{1}{2}$ | $\frac{p+q}{2}$ | $\frac{1}{p+q}$ |
| | | PP | PP, PQ, QQ | $\frac{p}{2(p+q)}$ | $\frac{p}{2}$ | $\frac{1}{p+q}$ |
| | PP | PQ | PQ, QQ | $\frac{p+q}{2p+q}$ | $q$ | $\frac{p+q}{q(2p+q)}$ |
| | | PP | PQ, QQ | $\frac{p}{2p+q}$ | $p$ | $\frac{1}{2p+q}$ |
| PQR | PP | PQ | QR | $\frac{1}{2}$ | $q$ | $\frac{1}{2q}$ |
| | PQ | PP | PR, QR, RR | $\frac{p}{2(2p+2q+r)}$ | $\frac{p}{2}$ | $\frac{1}{2p+2q+r}$ |
| | | QQ | PR, QR, RR | $\frac{q}{2(2p+2q+r)}$ | $\frac{q}{2}$ | $\frac{1}{2p+2q+r}$ |
| | | PQ | PR, QR, RR | $\frac{p+q}{2(2p+2q+r)}$ | $\frac{p+q}{2}$ | $\frac{1}{2p+2q+r}$ |
| | | PR | PR, QR, RR | $\frac{p+q+r}{2(2p+2q+r)}$ | $\frac{r}{2}$ | $\frac{p+q+r}{r(2p+2q+r)}$ |
| PQRS | PS | PR | RS | $\frac{1}{4}$ | $\frac{r}{2}$ | $\frac{1}{2r}$ |

If the DNA evidence in the aforementioned case came from a single contributor, the analysis would not be difficult and has been discussed by Evett and Weir [10] and here in §4.2.4. Evidence stains are not always this simple. If it is from a rape case, the stain could contain material from the victim, consensual partners, and the perpetrator — sometimes multiple perpetrators. The interpretation of mixed stains has been discussed in §4.2.2. However, these formulae do not consider situations where the contributor may be a relative of a typed individual.

The approach in this section takes into account the relatedness between an unknown contributor and a typed individual — usually the suspect. We show two simple formulae used in likelihood ratios when relatedness is a factor.

Our method assumes there is a known number of unknown contributors. It also assumes that one of the unknowns is related to a typed individual who is not a contributor under a given hypothesis. Any other relatedness is not considered. It is assumed that

known contributors are typed at the loci considered. Population structure is ignored.

**Case 1** There are two different situations discussed here. In the first case the prosecution proposes that two typed individuals contributed the evidence (usually the victim and suspect). Also, the defense proposes that one typed individual contributed to the evidence, in addition to one unknown person who is related to a typed person. Typically, though not necessarily, in the defense hypothesis the victim will be the typed individual and the unknown contributor will be some relative of the suspect. This leads to the hypotheses:

> $H_p$: The victim and the suspect are the contributors of the mixture.
> $H_d$: The victim and one unknown person are the contributors to the evidence mixture. The unknown contributor is related to the suspect by relationship $R$.

The appropriate likelihood ratio for these hypotheses is

$$\begin{aligned} \text{LR} &= \frac{Pr(E|H_p)}{Pr(E|H_d)} \\ &= \frac{Pr(E|G_S, G_V, H_p)Pr(G_S, G_V|H_p)}{Pr(E|G_U, G_S, G_V, H_d)Pr_R(G_U|G_S, G_V, H_d)Pr(G_S, G_V|H_d)} \\ &= \frac{Pr(E|G_S, G_V, H_p)Pr(G_S, G_V|H_p)}{\sum_i Pr(E|G_{Ui}, G_S, G_V, H_d)Pr_R(G_{Ui}|G_S, G_V, H_d)Pr(G_S, G_V|H_d)} \end{aligned}$$

The probability of the suspect and victim genotypes are independent of the hypotheses, so those terms cancel out. If the genotypes of suspect and victim are consistent with the evidence, then under the prosecution's hypothesis the probability of the evidence given the two genotypes is one. Under the defense hypothesis, the probability of the evidence is independent of the suspect genotype, and the probability of the unknown genotype is independent of the victim genotype. This reduces the likelihood ratio to

$$\text{LR} = \frac{1}{\sum_i Pr(E|G_{Ui}, G_V, H_d)Pr_R(G_{Ui}|G_S, H_d)}$$

The second term in the denominator is the probability of the unknown contributor's genotype assuming the related genotype $G_S$ has been seen. This conditional probability $Pr(G_U|G_S)$, is the same as in §4.2.4, Table 4.8. This table can be used to obtain the conditional probabilities of the unknown contributor. The denominator is then the sum of the conditional probabilities across all $G_{Ui}$ for which $Pr(E|G_{Ui}, G_V, H_d) = 1$. The final likelihood ratio is then,

$$\text{LR} = \frac{1}{\sum_i Pr_R(G_{Ui}|G_S, H_d)} \tag{4.12}$$

**Example** Suppose an evidence stain has alleles P, Q, and R. A suspect and victim are found. The victim has genotype PQ and the suspect has genotype PR. Under the defense hypothesis, the unknown contributor's genotype could be PR, QR, or RR. Also, the defense claims that an untyped relative of the suspect is the true contributor. Using Equation 4.12,

$$\begin{aligned}
\text{LR} &= \frac{1}{\sum_i Pr_R(G_{Ui}|G_S, H_d)} \\
&= \left[\sum_i Pr_R(G_{Ui}|\text{PR}, H_d)\right]^{-1} \\
&= [Pr_R(\text{PR}|\text{PR}, H_d) + Pr_R(\text{QR}|\text{PR}, H_d) + Pr_R(\text{RR}|\text{PR}, H_d)]^{-1} \\
&= \left[(P_0 2pr + P_1\frac{p+r}{2} + P_2) + (P_0 2qr + P_1\frac{r}{2}) + (P_0 r^2 + P_1\frac{r}{2})\right]^{-1} \\
&= \left[P_0(2pr + 2qr + r^2) + P_1\frac{p+3r}{2} + P_2\right]^{-1}
\end{aligned}$$

**Case 2** In the second case, the prosecution proposes that one typed individual (the suspect) and one person with unknown genotype contributed the evidence. The defense proposes that two unknown persons contributed the evidence, one of which is related to a typed individual (usually the suspect). The associated hypotheses are as follows.

$H_p$: The suspect and one unknown person are the contributors of the mixture.

$H_d$: Two unknown persons contributed the evidence, one of which is related to the suspect by relationship $R$.

The appropriate likelihood ratio would be

$$\begin{aligned}
\text{LR} &= \frac{Pr(E|H_p)}{Pr(E|H_d)} \\
&= \frac{Pr(E|G_S, G_U, H_p)Pr(G_U|G_S, H_p)Pr(G_S|H_p)}{Pr(E|G_{U1}, G_{U2}, G_S, H_d)Pr(G_{U2}|G_{U1}, G_S, H_d)Pr_R(G_{U1}|G_S, H_d)Pr(G_S|H_d)} \\
&= \frac{\sum_i Pr(E|G_S, G_{Ui}, H_p)Pr(G_{Ui}|G_S, H_p)Pr(G_S|H_p)}{\sum_{i,j} Pr(E|G_{U1i}, G_{U2j}, G_S, H_d)Pr(G_{U2}|G_{U1}, G_S, H_d)Pr_R(G_{U1}|G_S, H_d)Pr(G_S|H_d)}
\end{aligned}$$

In the likelihood ratio above, the probability of the suspect's genotype cancels out, as before. Under the prosecution's hypothesis, the genotype of the unknown contributor is independent of the suspect's genotype. Under the defense hypothesis, the probability of the evidence is independent of the suspect's genotype. Under the defense hypothesis, one of the unknown genotypes (say, $G_{U2}$) is independent of the suspect and the other unknown contributor. The other unknown contributor is related to the suspect. If considering only the genotypes of $G_U$ for which $Pr(E|G_S, G_U, H_p) = 1$, then the $Pr(E|G_S, G_U, H_p)$ term can be ignored. Similarly, if considering only those combinations of $G_{U1}$ and $G_{U2}$ for which $Pr(E|G_{U1}, G_{U2}, H_d) = 1$, then this term can also be ignored. This leaves the final likelihood ratio for Case 2.

$$\text{LR} = \frac{\sum_i Pr(G_{Ui}|H_p)}{\sum_{i,j} Pr(G_{U2}|H_d)Pr_R(G_{U1}|G_S, H_d)}$$

The probability of a genotype given the hypothesis is $p^2$ for homozygotes and $2pq$ for heterozygotes. The probability of a genotype given the observation of a relative is the same as in §4.2.4, Table 4.8. Using these probabilities and tables, we can construct the likelihood ratio for a mixed stain when a relative of the suspect may be a contributor.

Table 4.11: Case 2 example calculations.

| | |
|---|---|
| $Pr(\text{RS}\|H_p)$ | $= 2rs$ |
| $Pr(\text{PQ}\|H_d)Pr_R(\text{RS}\|\text{PQ}, H_d)$ | $= 2pq(P_0 2rs)$ |
| $Pr(\text{PR}\|H_d)Pr_R(\text{QS}\|\text{PQ}, H_d)$ | $= 2pr(P_0 2qs + P_1 \frac{s}{2})$ |
| $Pr(\text{PS}\|H_d)Pr_R(\text{QR}\|\text{PQ}, H_d)$ | $= 2ps(P_0 2qr + P_1 \frac{r}{2})$ |
| $Pr(\text{RS}\|H_d)Pr_R(\text{PQ}\|\text{PQ}, H_d)$ | $= 2rs(P_0 2pq + P_1 \frac{p+q}{2} + P_2)$ |
| $Pr(\text{QS}\|H_d)Pr_R(\text{PR}\|\text{PQ}, H_d)$ | $= 2qs(P_0 2pr + P_1 \frac{r}{2})$ |
| $Pr(\text{QR}\|H_d)Pr_R(\text{PS}\|\text{PQ}, H_d)$ | $= 2qr(P_0 2ps + P_1 \frac{s}{2})$ |

**Example** Suppose a DNA mixture is found and a suspect is arrested. The mixture genotype is PQRS, and the suspect genotype is PQ. Under the prosecution's hypothesis, the unknown contributor must be genotype RS. Under the defense hypothesis, there are two unknown contributors and are therefore six possible sets (PQ,RS; PR,QS; PS,QR; RS,PQ; QS,PR; QR,PS). It is worth noting that typically when there are two unknown contributors the order of the genotypes doesn't matter. That is, typically the genotype set (PQ,RS) is equivalent to (RS,PQ). However, in our case since one unknown contributor is related to the suspect, these two sets are not equivalent. One says that the genotype PQ is related to the suspect's genotype while the other say that the genotype RS is related to the suspect's genotype.

In this example, the likelihood ratio is as below. The calculations for each possible unknown contributor genotype are listed in Table 4.11.

$$\text{LR} = \frac{Pr(\text{RS}|H_p)}{\sum_{i,j} Pr(G_{U2}|H_d)Pr_R(G_{U1}|G_S, H_d)}$$

As can be seen, the probability of a genotype conditional of the observed relative (Table 4.8) can be used to address these two cases of DNA mixtures when a relative is a potential contributor. Other situations, such as related unknown contributors, or a relative index from a mixture could similarly be addressed in like manner. The particular

problem here has also been addressed by [44].

### 4.3.3 Relatives and Population Substructure

As discussed previously, the joint genotype frequencies of related individuals are of interest to forensic scientists. They are useful in estate disputes, missing persons cases, and when a suspect claims that a relative committed a crime. Up to this point, when calculating joint genotype probabilities for related individuals, it has been assumed that there is no population substructure. We now look to drop this assumption and calculate joint genotype frequencies of related individuals in the presence of population substructure.

The assumption of no population substructure is largely one of convenience. In the absence of substructure and inbreeding, alleles can be identical by descent between two relatives, but alleles within an individual cannot be ibd. Suppose two persons have alleles $ab$ and $cd$ respectively, where $ab$ and $cd$ are labels and do not indicate allelic type. If there is no inbreeding then there are only seven different patterns of ibd ($\{a \equiv c, b \equiv d\}$; $\{a \equiv d, b \equiv c\}$; $\{a \equiv c\}$; $\{a \equiv d\}$; $\{b \equiv c\}$; $\{b \equiv d\}$; {None ibd}). The probability of each pattern is $\delta_{ac \cdot bd}$, $\delta_{ad \cdot bc}$, $\delta_{ac}$, $\delta_{ad}$, $\delta_{bc}$, $\delta_{bd}$, and $\delta_0$ respectively. Note that these different patterns can be put into three different categories – zero, one, or two alleles in common between the different individuals. These categories are visualized in Figure 4.4. The probabilities of the ibd patterns are used to calculate the $P_0$, $P_1$, and $P_2$ summary probabilities that we've already seen.

$$\begin{aligned} P_0 &= \delta_{ac \cdot bd} + \delta_{ad \cdot bc} \\ P_1 &= \delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd} \\ P_2 &= \delta_0 \end{aligned} \tag{4.13}$$

The difficulty with population substructure is that in addition to identity by descent between two persons, there can also be identity by descent within a person. This introduction of inbreeding further increases the complexity of the calculations. Instead

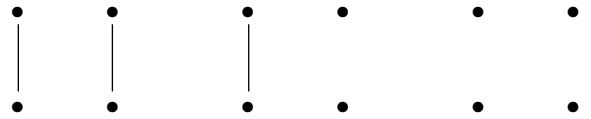



Figure 4.4: Three non-inbred summary ibd categories

Table 4.12: Fifteen ibd patterns and probabilities

| ibd status | Probability | ibd status | Probability |
|---|---|---|---|
| $a \equiv b \equiv c \equiv d$ | $\delta_{abcd}$ | $a \equiv b$ | $\delta_{ab}$ |
| $a \equiv b \equiv c$ | $\delta_{abc}$ | $a \equiv c$ | $\delta_{ac}$ |
| $a \equiv b \equiv d$ | $\delta_{abd}$ | $a \equiv d$ | $\delta_{ad}$ |
| $a \equiv c \equiv d$ | $\delta_{acd}$ | $b \equiv c$ | $\delta_{bc}$ |
| $b \equiv c \equiv d$ | $\delta_{bcd}$ | $b \equiv d$ | $\delta_{bd}$ |
| $a \equiv b, c \equiv d$ | $\delta_{ab \cdot cd}$ | $c \equiv d$ | $\delta_{cd}$ |
| $a \equiv c, b \equiv d$ | $\delta_{ac \cdot bd}$ | None | $\delta_0$ |
| $a \equiv d, b \equiv c$ | $\delta_{ad \cdot bc}$ | | |

of seven measures of ibd that are easily condensed into the three summary measures, there are now fifteen measures that express all possible combinations of identity – within individuals and between. Table 4.12 lists these different patterns and their associated probabilities. Note that any alleles not listed are not ibd to those listed, nor each other.

The no substructure assumption is convenient because these fifteen measures can only be reduced to nine categories (Figure 4.3.3). In addition, the $P_0$, $P_1$, and $P_2$ can be easily calculated following simple mendelian inheritance laws, while the fifteen measures are not so readily calculated. However, though convenient, the no substructure assumption is inappropriate in many situations. In some cases the individuals involved belong to a small, closely related population (such as an island population, or the population of a Native-American reservation). In others, there may not be any obvious population

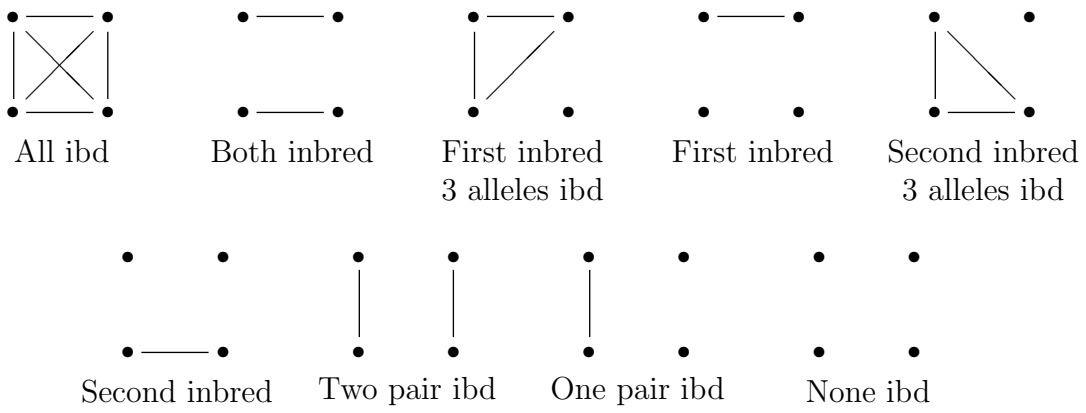



Figure 4.5: Nine inbred summary ibd categories

substructure but the scientist may want to correct for substructure effects to ensure that they are not being prejudicial against a defendant. As such, we are interested in the joint genotype probabilities of relatives in the presence of population substructure, despite the added complexity

**Joint Probabilities: General Formulae**

The joint genotype probabilities can be expressed in terms of the fifteen ibd measures and the allele frequencies [10, 45]. This is the most general form of the joint genotype probabilities and can account for all possible degrees and types of pairwise relatedness.

$$
\begin{aligned}
Pr(\mathrm{PP},\mathrm{PP}) = \quad & \delta_{abcd}p + (\delta_{abc} + \delta_{abd} + \delta_{acd} + \delta_{bcd})p^2 + (\delta_{ab\cdot cd} + \delta_{ac\cdot bd} + \delta_{ad\cdot bc})p^2 \\
& +(\delta_{ab} + \delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd} + \delta_{cd})p^3 + \delta_0 p^4 \\
Pr(\mathrm{PP},\mathrm{QQ}) = \quad & \delta_{ab\cdot cd}pq + \delta_{ab}pq^2 + \delta_{cd}p^2q + \delta_0 p^2q^2 \\
Pr(\mathrm{PP},\mathrm{PQ}) = \quad & (\delta_{abc} + \delta_{abd})pq + (2\delta_{ab} + \delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd})p^2q + 2\delta_0 p^3 q \\
Pr(\mathrm{PP},\mathrm{QR}) = \quad & 2\delta_{ab}pqr + 2\delta_0 p^2 qr
\end{aligned}
$$

$$
\begin{aligned}
Pr(\mathrm{PQ},\mathrm{PQ}) = &\quad 2(\delta_{ac\cdot bd}+\delta_{ad\cdot bc})pq+(\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd})pq(p+q)+4\delta_0 p^2q^2\\
Pr(\mathrm{PQ},\mathrm{PR}) = &\quad (\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd})pqr+4\delta_0 p^2qr\\
Pr(\mathrm{PQ},\mathrm{RS}) = &\quad 4\delta_0 pqrs
\end{aligned}
$$

These probabilities depend on both the allele frequencies and the fifteen ibd measures. The allele frequencies are usually determined by sampling. The ibd probabilities change depending on the amount of substructure, and the relationship (by pedigree) of the two individuals.

### IBD Probabilities

As described in [45], there are many different alternative formulations of the descent measures. Some of them are technically equivalent, but differ in the way the measures are parameterized. Others make different assumptions about the distribution of ibd alleles. Here we assume that the underlying population is mating at random, and has reached evolutionary equilibrium. In this situation, the descent measures can be calculated by determining the probability of selecting two, three, two-pairs, or four ibd alleles at random in the population. These probabilities ($\theta$, $\gamma$, $\Delta$, and $\delta$ respectively) do not change over time under the equilibrium assumption.

To illustrate the use of $\theta$, $\gamma$, $\Delta$, and $\delta$, consider two individuals who are in the same subpopulation, but are not immediately related due to pedigree. In this situation the probability of all four alleles between the two persons being ibd is the probability of four random alleles from the population being ibd ($\delta_{abcd}=\delta$). The probability that only three alleles are ibd is the probability of any three random alleles from the population being ibd, minus the probability of all four being ibd ($\delta_{abc}=\delta_{abd}=\delta_{acd}=\delta_{bcd}=\gamma-\delta$).

Similarly, in Table 4.13 all fifteen ibd measures are rewritten in terms of $\theta$, $\gamma$, $\Delta$, and $\delta$ for several different relationships. In Table 4.13, for the parent-child relationship, the child ($cd$) recieves allele $c$ from the parent ($ab$). In the grandparent-grandchild relationship, the grandparent ($ab$) passes allele $c$ to grandchild $cd$. One parent passes alleles $a$ and $c$ to full sibs $ab$ and $cd$, while the other parent passes alleles $b$ and $d$. In the half sib

relationship, the common parent passes alleles $b$ and $c$ to half sibs $ab$ and $cd$. Cousins $ab$ and $cd$ recieve alleles $b$ and $c$ from their parents, who are full siblings.

Under the assumption of evolutionary equilibrium, the allele frequencies follow a Dirichlet distribution [33], and the $\gamma$, $\Delta$, and $\delta$ probabilities can be written as functions of the population parameter $\theta$. These functions for $\gamma$, $\Delta$, and $\delta$ (Equations 4.14), in conjunction with the ibd formulae from Table 4.13 and the joint genotype equations (Equations 4.14), can be used to calculate the joint probability of two related genotypes in the presence of population substructure. The only necessary parameter values are the allele frequencies and the population parameter $\theta$.

$$\begin{aligned} \gamma &= \frac{2\theta^2}{1+\theta} \\ \Delta &= \frac{\theta^2(1+5\theta)}{(1+\theta)(1+2\theta)} \\ \delta &= \frac{6\theta^3}{(1+\theta)(1+2\theta)} \end{aligned} \tag{4.14}$$

**Example** What is the probability that a father and son would both have the same homozygous genotype? Using the appropriate equation from Equations 4.14 and formulae from Table 4.13,

$$\begin{aligned} Pr_{\mathrm{PC}}(\mathrm{PP},\mathrm{PP}) = \quad & \delta_{abcd}p + (\delta_{abc}+\delta_{abd}+\delta_{acd}+\delta_{bcd})p^2 + (\delta_{ab\cdot cd}+\delta_{ac\cdot bd}+\delta_{ad\cdot bc})p^2 \\ & +(\delta_{ab}+\delta_{ac}+\delta_{ad}+\delta_{bc}+\delta_{bd}+\delta_{cd})p^3 + \delta_0 p^4 \\ = \quad & \gamma p + \left((\theta-\gamma)+0+\frac{\theta-\gamma}{2}+\frac{\theta-\gamma}{2}\right)p^2 \\ & + \left(0+\frac{\theta-\gamma}{2}+\frac{\theta-\gamma}{2}\right)p^2 \\ & + \left(0+\frac{(1-3\theta+2\gamma)}{2}+0+\frac{(1-3\theta+2\gamma)}{2}+0+0\right)p^3 + (0)p^4 \\ = \quad & \gamma p + 3(\theta-\gamma)p^2 + (1-3\theta+2\gamma)p^3 \end{aligned}$$

Table 4.13: Fifteen ibd measures for various relationships, in terms of $\theta, \gamma, \Delta$, and $\delta$.

| $Pr$(ibd) | Unrelated | Parent–Child | Grandparent–Grandchild |
|---|---|---|---|
| $\delta_{abcd}$ | $\delta$ | $\gamma$ | $(\gamma+\delta)/2$ |
| $\delta_{abc}$ | $\gamma-\delta$ | $\theta-\gamma$ | $(\theta-\delta)/2$ |
| $\delta_{abd}$ | $\gamma-\delta$ | 0 | $(\gamma-\delta)/2$ |
| $\delta_{acd}$ | $\gamma-\delta$ | $(\theta-\gamma)/2$ | $(\theta+\gamma-2\delta)/4$ |
| $\delta_{bcd}$ | $\gamma-\delta$ | $(\theta-\gamma)/2$ | $(\theta+\gamma-2\delta)/4$ |
| $\delta_{ab\cdot cd}$ | $\Delta-\delta$ | 0 | $(\Delta-\delta)/2$ |
| $\delta_{ac\cdot bd}$ | $\Delta-\delta$ | $(\theta-\gamma)/2$ | $(\theta-\gamma+2\Delta-2\delta)/4$ |
| $\delta_{ad\cdot bc}$ | $\Delta-\delta$ | $(\theta-\gamma)/2$ | $(\theta-\gamma+2\Delta-2\delta)/4$ |
| $\delta_{ab}$ | $\theta-2\gamma-\Delta+2\delta$ | 0 | $(\theta-2\gamma-\Delta+2\delta)/2$ |
| $\delta_{ac}$ | $\theta-2\gamma-\Delta+2\delta$ | $(1-3\theta+2\gamma)/2$ | $(1-\theta-2\gamma-2\Delta+4\delta)/4$ |
| $\delta_{ad}$ | $\theta-2\gamma-\Delta+2\delta$ | 0 | $(\theta-2\gamma-\Delta+2\delta)/2$ |
| $\delta_{bc}$ | $\theta-2\gamma-\Delta+2\delta$ | $(1-3\theta+2\gamma)/2$ | $(1-\theta-2\gamma-2\Delta+4\delta)/4$ |
| $\delta_{bd}$ | $\theta-2\gamma-\Delta+2\delta$ | 0 | $(\theta-2\gamma-\Delta+2\delta)/2$ |
| $\delta_{cd}$ | $\theta-2\gamma-\Delta+2\delta$ | 0 | $(\theta-2\gamma-\Delta+2\delta)/2$ |
| $\delta_0$ | $1-6\theta+8\gamma+3\Delta-6\delta$ | 0 | $(1-6\theta+8\gamma+3\Delta-6\delta)/2$ |
| $Pr$(ibd) | Full-sib | Half-Sib | First Cousins |
| $\delta_{abcd}$ | $(\theta+2\gamma+\delta)/4$ | $(\gamma+\delta)/2$ | $(\gamma+3\delta)/4$ |
| $\delta_{abc}$ | $(\theta-\delta)/4$ | $(\theta-\delta)/2$ | $(\theta+2\gamma-3\delta)/4$ |
| $\delta_{abd}$ | $(\theta-\delta)/4$ | $(\gamma-\delta)/2$ | $3(\gamma-\delta)/4$ |
| $\delta_{acd}$ | $(\theta-\delta)/4$ | $(\gamma-\delta)/2$ | $3(\gamma-\delta)/4$ |
| $\delta_{bcd}$ | $(\theta-\delta)/4$ | $(\theta-\delta)/2$ | $(\theta+2\gamma-3\delta)/4$ |
| $\delta_{ab\cdot cd}$ | $(\Delta-\delta)/4$ | $(\Delta-\delta)/2$ | $3(\Delta-\delta)/4$ |
| $\delta_{ac\cdot bd}$ | $(1+\theta-2\gamma+\Delta-\delta)/4$ | $(\Delta-\delta)/2$ | $3(\Delta-\delta)/4$ |
| $\delta_{ad\cdot bc}$ | $(\Delta-\delta)/4$ | $(\theta-\gamma+\Delta-\delta)/2$ | $(\theta-\gamma+3\Delta-3\delta)/4$ |
| $\delta_{ab}$ | $(\theta-2\gamma-\Delta+2\delta)/4$ | $(\theta-2\gamma-\Delta+2\delta)/2$ | $3(\theta-2\gamma-\Delta+2\delta)/4$ |
| $\delta_{ac}$ | $(1-2\theta-\Delta-2\delta)/4$ | $(\theta-2\gamma-\Delta+2\delta)/2$ | $3(\theta-2\gamma-\Delta+2\delta)/4$ |
| $\delta_{ad}$ | $(\theta-2\gamma-\Delta+2\delta)/4$ | $(\theta-2\gamma-\Delta+2\delta)/2$ | $3(\theta-2\gamma-\Delta+2\delta)/4$ |
| $\delta_{bc}$ | $(\theta-2\gamma-\Delta+2\delta)/4$ | $(1-2\theta-\Delta+2\delta)/2$ | $(1-4\gamma-3\Delta+6\delta)/4$ |
| $\delta_{bd}$ | $(1-2\theta-\Delta-2\delta)/4$ | $(\theta-2\gamma-\Delta+2\delta)/2$ | $3(\theta-2\gamma-\Delta+2\delta)/4$ |
| $\delta_{cd}$ | $(\theta-2\gamma-\Delta+2\delta)/4$ | $(\theta-2\gamma-\Delta+2\delta)/2$ | $3(\theta-2\gamma-\Delta+2\delta)/4$ |
| $\delta_0$ | $(1-6\theta+8\gamma+3\Delta-6\delta)/4$ | $(1-6\theta+8\gamma+3\Delta-6\delta)/2$ | $3(1-6\theta+8\gamma+3\Delta-6\delta)/4$ |

Using Equations 4.14 in the above leads to the probability of a father and son having the same homozygous genotype in the presence of population substructure, as a function of $\theta$ and the allele frequency $p$.

$$\begin{aligned} Pr_{\mathrm{PC}}(\mathrm{PP},\mathrm{PP}) &= \left(\frac{2\theta^2}{1+\theta}\right)p + 3\left(\theta - \frac{2\theta^2}{1+\theta}\right)p^2 + \left(1 - 3\theta + 2\frac{2\theta^2}{1+\theta}\right)p^3 \\ &= \frac{p}{1+\theta}(2\theta^2 + 3\theta(1-\theta)p + (1-\theta)^2p^2) \\ &= \frac{p}{1+\theta}(\theta + (1-\theta)p)(2\theta + (1-\theta)p) \end{aligned} \tag{4.15}$$

**Ayres, Fung, Carracedo, and Hu**

Ayres [46], Fung et al [47], and Fung and Hu [48] all propose an alternative method for the calculation of joint genotype probabilities in the presence of population substructure. Ayres noted that the non-inbred joint genotype frequencies from Evett and Weir (listed here in Table 4.7) are of a particular form. The form noted is that they are the non-inbred ibd probabilities ($P_0$, $P_1$, and $P_2$) multiplied by the probabilities of non-inbred sets of alleles. Table 4.14 shows this more general form (the multiples of 2 and 4 account for the different possible orderings within heterozygous genotypes).

Ayres proposes that, to account for substructure, instead of multiplying the non-inbred ibd probabilities by the probability of non-inbred sets of alleles, they should be multiplied by the probability of the sets using Ewen's sampling formula (Equation 4.3). This approach is also used in [47] and [48]. For example, if interested in the probability of relatives having the genotypes PP and PQ in the presence of substructure, Ayres proposes

$$\begin{aligned} Pr(\mathrm{PP},\mathrm{PQ}) = \quad & P_0 2Pr(\mathrm{PPPQ}) + P_1 Pr(\mathrm{PPQ}) \\ = \quad & P_0 \frac{2pq(1-\theta)(\theta + (1-\theta)p)(2\theta + (1-\theta)p)}{(1+\theta)(1+2\theta)} \\ & + P_1 \frac{pq(1-\theta)(\theta + (1-\theta)p)}{(1+\theta)} \end{aligned} \tag{4.16}$$

Table 4.14: Joint probabilities of two related genotypes (Ayres, 2000).

| Genotypes | $P_0$ | $P_1$ | $P_2$ |
|---|---|---|---|
| PP,PP | $Pr(\text{PPPP})$ | $Pr(\text{PPP})$ | $Pr(\text{PP})$ |
| PP,QQ | $Pr(\text{PPQQ})$ | | |
| PP,PQ | $2Pr(\text{PPPQ})$ | $Pr(\text{PPQ})$ | |
| PP,QR | $2Pr(\text{PPQR})$ | | |
| PQ,PQ | $4Pr(\text{PQPQ})$ | $Pr(\text{PPQ}) + Pr(\text{PQQ})$ | $2Pr(\text{PQ})$ |
| PQ,PR | $4Pr(\text{PQPR})$ | $Pr(\text{PQR})$ | |
| PQ,RS | $4Pr(\text{PQRS})$ | | |

This method is much more convenient than the calculations involving the fifteen ibd probabilities. The method uses the three summary measures $P_0$, $P_1$, and $P_2$, in addition to the familiar Ewen's sampling formula. However, the question remains, is the method correct? Ayres claims that in Evett and Weir's treatment of joint genotype probabilities for non-inbred relatives, the no-substructure assumption only affects the allele probabilities and not the $P_0$, $P_1$, $P_2$. This claim is incorrect, because $P_0$, $P_1$, and $P_2$ are calculated with Equations 4.13 and Table 4.13 when $\theta = \gamma = \Delta = \delta = 0$. If these parameters are not equal to zero, then $P_0$, $P_1$, and $P_2$ by definition no longer equal the values that Ayres is using.

Though Ayres's claim is incorrect, the method may still be correct – not because $P_0$, $P_1$, and $P_2$ are unaffected by the substructure, but because they account for the relatedness due to pedigree while the use of Ewen's sampling formula accounts for the relatedness due to population substructure. If this is the case, $P_0$, $P_1$, and $P_2$ do not reflect the true ibd probabilities of the pair of relatives in the subpopulation, but instead reflect ibd status that can be attributed to pedigree relatedness.

### Testing Ayres's Method

To test Ayres's method, we attempted to show that it is mathematically equivalent to the standard method using the fifteen ibd measures. We start by noting that our formulation of the fifteen ibd measures relies on the same assumptions as the Ewen's sampling formula used in Ayres's method. Also, when considering unrelated individuals, the fifteen measure method is equivalent to results using Ewen's.

**Unrelated Example** For example, in the presence of population substructure, the probability of two unrelated heterozygotes PQ,RS is $\frac{4pqrs(1-\theta)^3}{(1+\theta)(1+2\theta)}$, from the sampling formula (Equation 4.3). Using the fifteen measure method,

$$\begin{aligned}
Pr(PQ,RS) = \quad & 4\delta_0 pqrs \\
= \quad & 4pqrs(1-6\theta+8\gamma+3\Delta-6\delta) \\
= \quad & 4pqrs\left(1-6\theta+\frac{16\theta^2}{(1+\theta)}+\frac{3\theta^2(1+5\theta)}{(1+\theta)(1+2\theta)}-\frac{36\theta^3}{(1+\theta)(1+2\theta)}\right) \\
= \quad & \frac{4pqrs}{(1+\theta)(1+2\theta)}\Big[(1-6\theta)(1+\theta)(1+2\theta) \\
& \qquad + (16\theta^2)(1+2\theta)+3\theta^2(1+5\theta)-36\theta^3\Big] \\
= \quad & \frac{4pqrs}{(1+\theta)(1+2\theta)}\Big[(1-3\theta-16\theta^2-12\theta^3) \\
& \qquad + (16\theta^2+32\theta^3)+(3\theta^2+15\theta^3)-36\theta^3\Big] \\
= \quad & \frac{4pqrs}{(1+\theta)(1+2\theta)}\Big[1-3\theta+3\theta^2-\theta^3\Big] \\
= \quad & \frac{4pqrs(1-\theta)^3}{(1+\theta)(1+2\theta)}
\end{aligned}$$

This equivalence means that it is possible to rewrite the results of the sampling formula in terms of $\theta, \gamma, \Delta$, and $\delta$. The joint probabilities using Ayres's method can then be rewritten in these same terms and directly compared to the joint probabilities using fifteen measure method. We will do this here with one of the joint genotype formulae to

illustrate the method. The other formulae will not be compared here, but are included in Table 4.15.

### Pr(PP,PP)

The probability of two genotypes being homozygous for the same allele is given below using Ayres's method. The probabilities of the allele sets are then rewritten in terms of $\theta, \gamma, \Delta$, and $\delta$. The formula is then rearranged such that it is in the same form as Equations 4.14

$$\begin{aligned}
Pr(\text{PP},\text{PP}) &= P_0 Pr(\text{PPPP}) + P_1 Pr(\text{PPP}) + P_2 Pr(\text{PP}) \\
&= P_0 \frac{p(\theta+(1-\theta)p)(2\theta+(1-\theta)p)(3\theta+(1-\theta)p)}{(1+\theta)(1+2\theta)} \\
&\quad + P_1 \frac{p(\theta+(1-\theta)p)(2\theta+(1-\theta)p)}{(1+\theta)} + P_2 p(\theta+(1-\theta)p) \\
&= P_0 \Big[\delta p + 4(\gamma-\delta)p^2 + 3(\Delta-\delta)^2 + 6(\theta-2\gamma-\Delta+2\delta)p^3 \\
&\quad + (1-6\theta+8\gamma+3\Delta-6\delta)p^4\Big] + P_1\Big[\gamma p + 3(\theta-\gamma)p^2 + (1-3\theta+2\gamma)p^3\Big] \\
&\quad + P_2\Big[\theta p + (1-\theta)p^2\Big] \\
&= p(P_0\delta + P_1\gamma + P_2\theta) + p^2(P_0[4(\gamma-\delta)+3(\Delta-\delta)] + P_1 3(\theta-\gamma) + P_2(1-\theta)) \\
&\quad + p^3(P_0 6(\theta-2\gamma-\Delta+2\delta) + P_1(1-3\theta+2\gamma)) \\
&\quad + p^4(P_0(1-6\theta+8\gamma+3\Delta-6\delta))
\end{aligned}$$

If Ayres's method is correct then the above must be equal to $Pr$(PP,PP) using the fifteen measure method. Specifically, the following equalities must be true for for all relationships. To see if the equalities are true, the $P_0$, $P_1$, and $P_2$ in the equalities are replaced with the appropriate values from Table 4.6, and the fifteen ibd parameters are replaced with their appropriate values from Table 4.13.

$$\delta_{abcd} = P_0\delta + P_1\gamma + P_2\theta$$

$$\delta_{abc}+\delta_{abd}+\delta_{acd}+\delta_{bcd}+\delta_{ab\cdot cd}+\delta_{ac\cdot bd}+\delta_{ad\cdot bc} = P_0[4(\gamma-\delta)+3(\Delta-\delta)]+P_1 3(\theta-\gamma)+P_2(1-\theta)$$

$$\delta_{ab} + \delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd} + \delta_{cd} = P_0 6(\theta - 2\gamma - \Delta + 2\delta) + P_1(1 - 3\theta + 2\gamma)$$

$$\delta_0 = P_0(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)$$

**Parent-Child** In the parent-child relationship, $P_0$, $P_1$, and $P_2$ are 0, 1, and 0 respectively. As seen in Table 4.13, for the parent-child relationship, $\delta_{abcd} = \gamma$. It is clear that $(0)\delta + (1)\gamma + (0)\theta = \gamma$, so the first equality holds for the parent-child relationship. Since $P_0 = P_2 = 0$, the right side of the second equality reduces to $3(\theta - \gamma)$ which is equal to the sum of the ibd measures on the left side. So, the second equality also holds. Similarly, the right side of the third equality reduces to $(1 - 3\theta + 2\gamma)$, which is the sum of the ibd measures on the left side. Finally, both $\delta_0$ and $P_0$ equal zero, showing that all four equalities hold for the parent-child relationship.

**Full-Sibs** In a full sib relationship, $P_0$, $P_1$, and $P_2$ are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ respectively. The right side of the first equality is then $\frac{1}{4}\delta + \frac{1}{2}\gamma + \frac{1}{4}\theta$ which is equal to $\delta_{abcd}$. The right side of the second equality is

$$\begin{aligned} P_0[4(\gamma - \delta) + 3(\Delta - \delta)] &+ P_1 3(\theta - \gamma) + P_2(1 - \theta) \\ &= \frac{1}{4}[4(\gamma - \delta) + 3(\Delta - \delta)] + \frac{1}{2}3(\theta - \gamma) + \frac{1}{4}(1 - \theta) \\ &= \frac{1}{4}[1 + 5\theta - 2\gamma + 3\Delta - 7\delta] \end{aligned}$$

This is equal to the sum of the ibd measures on the left side of the equation which shows that the second equality holds for the full-sib relationships. Similarly the third and fourth equalities hold for full-sibs.

**Remarks** We have shown that Ayres's method is equivalent to the 15 ibd measure method when calculating the joint probability of two homozygous genotypes when the two genotypes are related by the parent-child relationship, or the full-sib relationship. Though not shown here, all other relationships considered were also equivalent (Unre-

Table 4.15: Joint probabilitiy of related genotypes, with population substructure. Standard and Ayres's method.

| $Pr(G_1, G_2)$ | 15 Measure Method | Ayres′s Method |
|---|---|---|
| $Pr(\mathrm{PP}, \mathrm{PP})$ | $\Big[\delta_{abcd}p + (\delta_{ab\cdot cd} + \delta_{ac\cdot bd} + \delta_{ad\cdot bc})p^2$ $+(\delta_{abc} + \delta_{abd} + \delta_{acd} + \delta_{bcd})p^2$ $+(\delta_{ab} + \delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd} + \delta_{cd})p^3 + \delta_0 p^4\Big]$ | $\Big[p(P_0\delta + P_1\gamma + P_2\theta) + p^2P_0[4(\gamma-\delta) + 3(\Delta-\delta)]$ $+p^2(3P_1(\theta-\gamma) + P_2(1-\theta))$ $+p^3(6P_0(\theta - 2\gamma - \Delta + 2\delta) + P_1(1 - 3\theta + 2\gamma))$ $+p^4P_0(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)\Big]$ |
| $Pr(\mathrm{PP}, \mathrm{QQ})$ | $\delta_{ab\cdot cd}pq + \delta_{ab}pq^2 + \delta_{cd}p^2q + \delta_0p^2q^2$ | $\Big[pqP_0(\Delta-\delta) + (pq^2 + p^2q)P_0(\theta - 2\gamma - \Delta + 2\delta)$ $+p^2q^2P_0(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)\Big]$ |
| $Pr(\mathrm{PP}, \mathrm{PQ})$ | $\Big[(\delta_{abc} + \delta_{abd})pq + (2\delta_{ab} + \delta_{ac}$ $+\delta_{ad} + \delta_{bc} + \delta_{bd})p^2q + 2\delta_0p^3q\Big]$ | $\Big[pq(2P_0(\gamma-\delta) + P_1(\theta-\gamma))$ $+p^2q(6P_0(\theta - 2\gamma - \Delta + 2\delta) + P_1(1 - 3\theta + 2\gamma))$ $+2p^3qP_0(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)\Big]$ |
| $Pr(\mathrm{PP}, \mathrm{QR})$ | $2\delta_{ab}pqr + 2\delta_0p^2qr$ | $\Big[2pqrP_0(\theta - 2\gamma - \Delta + 2\delta)$ $+2p^2qrP_0(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)\Big]$ |
| $Pr(\mathrm{PQ}, \mathrm{PQ})$ | $\Big[2(\delta_{ac\cdot bd} + \delta_{ad\cdot bc})pq + 4\delta_0p^2q^2$ $+(\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd})pq(p+q)\Big]$ | $\Big[pq(4P_0(\Delta-\delta) + 2P_1(\theta-\gamma) + P_2(1-\theta))$ $+4pq(p+q)P_0(\theta - 2\gamma - \Delta + 2\delta)$ $+pq(p+q)P_1(1 - 3\theta + 2\gamma))$ $+4p^2q^2P_0(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)\Big]$ |
| $Pr(\mathrm{PQ}, \mathrm{PR})$ | $(\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd})pqr + 4\delta_0p^2qr$ | $\Big[pqr(4P_0(\theta - 2\gamma - \Delta + 2\delta) + P_1(1 - 3\theta + 2\gamma))$ $+4p^2qr(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)\Big]$ |
| $Pr(\mathrm{PQ}, \mathrm{RS})$ | $4\delta_0pqrs$ | $4P_0(1 - 6\theta + 8\gamma + 3\Delta - 6\delta)$ |

lated, Grandparent-Grandchild, Half-Sibs, First Cousins). In addition, Ayres's method is equivalent for all relationships tested for each of the seven genotype combinations. The results using both the 15 ibd measures and Ayres's method are in Table 4.15. To test equivalence, use the appropriate values from Tables 4.6 and 4.13.

This result has several significant implications. Three are listed here. (1) The first implication is that it appears the complicated fifteen measure method is not necessary for the standard relationships in the presence of population substructure. Ayres's method can be used to avoid the fifteen different probability measures, in favor of the three summary measures and the application of the sampling formula.

(2) This result also implies that the three summary measures can be used to describe relatedness attributed to pedigree in the presence of substructure, not just the realized relatedness in the absence of substructure. This is not to say that the non-inbred summary probabilities reflect the actual probability of zero, one, or two pairs ibd between the two genotypes. However, they do reflect the probability of zero, one, or two pairs ibd that can be attributed to the pedigree relationship between the two persons.

(3) The result also implies that more complicated joint probabilities could be considered. We may want, for example, to consider the probability that an unknown contributor and the suspect are related, given that an unrelated victim is in the same subpopulation. Since the $P_0, P_1$, and $P_2$ account for the pedigree relatedness, the sampling formula can be used to account for the subpopulation relatedness, by simply conditioning on already observed genotypes. This last implication in particular is of great convenience, because calculating three or four genotype descent probabilities for all possible patterns of ibd when a pedigree relationship is involved would be beyond most scientists.

### 4.3.4 Parentage Index

Rob Ogden of Wildlife DNA Services relayed the following situation [personal correspondence]. One of a breeder's puppies goes missing, and he accuses a local of stealing the dog. A specific dog is found in the possession of the accused that the breeder claims is his puppy. The puppy is typed, as are the dogs that the breeder alleges are the parents.

The alleged parents are not excluded as possible parents. The situation brings up the following pair of hypotheses.

$H_p$: The puppy is the offspring of the two alleged parents.
$H_d$: The puppy is the offspring of two other dogs.

These hypotheses can be used to form the following likelihood ratio.

$$\begin{aligned} \text{LR} &= \frac{Pr(E|H_p)}{Pr(E|H_d)} \\ &= \frac{Pr(G_C|G_{AF}, G_{AM}, H_p)Pr(G_{AF}, G_{AM}|H_p)}{Pr(G_C|G_{AF}, G_{AM}, H_d)Pr(G_{AF}, G_{AM}|H_d)} \end{aligned}$$

In this likelihood ratio, $Pr(G_{AF}, G_{AM}|H_p)$ and $Pr(G_{AF}, G_{AM}|H_d)$ are both independent of the hypothesis, and so cancel out. This leaves

$$\text{LR} = \frac{Pr(G_C|G_{AF}, G_{AM}, H_p)}{Pr(G_C|G_{AF}, G_{AM}, H_d)}$$

**No Population Substructure**

When there is no population substructure, under the defense hypothesis, the genotype of the child/puppy is independent of the alleged parents, so $Pr(G_C|G_{AF}, G_{AM}, H_d) = Pr(G_C|H_d)$. The LR then reduces to

$$\text{LR} = \frac{Pr(G_C|G_{AF}, G_{AM}, H_p)}{Pr(G_C|H_d)}$$

Notice that the numerator for this LR is the same form as the numerator for the paternity index. For the numerator, then, we can use the same values from the numerator column of Table 6.2 from Evett and Weir [10]. This table is included as Table 4.5 in this work. These values are also included below in the numerator column of Table 4.16 for a variety of parental genotype combinations.

Also notice that, when it is assumed there is no inbreeding, the denominator is the same as in a single contributor case. This is because if both parents are unknown (as in

the defense hypothesis), then there is no information about the child's genotype, and the genotype frequency is just the child's genotype frequency (i.e., $p^2$ and $2pq$). Table 4.16 shows the denominator formulae for a variety of parental genotypes ($\text{Dem}^{\text{b}}$).

### Inbreeding, No Population Substructure

In Mr. Ogden's case, the suspect happens to be an amateur dog breeder, and he claims that the puppy is the offspring of two of his dogs. He does not know which of his dogs are the parents, and since he has several dozen dogs, none of the potential parents have been genotyped. Due to the amateur nature of the suspect's dog breeding, it is possible that the puppy is inbred. That is, under the defense hypothesis, the unknown parents may be related.

Inbreeding within an individual $I$ is typically measured by the inbreeding coefficient $F_I$. The inbreeding coefficient is the probability that an individual receives alleles from their parents that are identical by descent; that is, they are copies of the same ancestral allele. $F_I$ is usually calculated using the path counting method (For more on path-counting, see pages 98 and 99 of Evett and Weir [10]). Note that any population substructure in the common ancestors will be accounted for later, so any ancestral inbreeding *due to population substructure* need not be considered (i.e., assume $\theta = 0$ when calculating $F_I$).

In the presence of inbreeding, the probability of a genotype is,

$$\begin{aligned} Pr_{F_I}(\text{PP}) &= F_I p + (1 - F_I)p^2 \\ Pr_{F_I}(\text{PQ}) &= 2pq(1 - F_I) \end{aligned} \tag{4.17}$$

Since assuming no population substructure, the genotype of the child is independent of the alleged parents under the defense hypotheses. Therefore, the above two formulae are used for the denominator of the likelihood ratio, and the numerator column from Table 4.16 is used for the numerator of the likelihood ratio. For example, again suppose that the puppy is genotype PP, and each parent is PQ. The numerator is again $\frac{1}{4}$. The

denominator is now $F_I p + (1 - F_I)p^2$. This gives a likelihood ratio of:

$$\begin{aligned} \text{LR} &= \frac{Pr(G_C|G_{AM}G_{AF}, H_p)}{Pr_{F_I}(G_C|H_d)} \\ &= \frac{1}{4(F_I p + (1 - F_I)p^2)} \end{aligned}$$

**Population substructure**

It may be that the defense proposes that the alleged parents are not the true parents, but are in the same subpopulation as the unknown, true parents. Under this defense hypothesis the probability of observing the puppy's genotype is no longer independent of the alleged parents. That is, $Pr(G_C|G_{AF}, G_{AM}, H_d) \neq Pr(G_C|H_d)$. To calculate $Pr(G_C|G_{AF}, G_{AM}, H_d)$ Ewen's sampling formula (Equation 4.3) is used.

In this case, we are interested in the probability of a genotype (the puppy/child genotype, $G_C$) given that we have already seen two other genotypes ($G_{AM}$ and $G_{AF}$). For example, if the child genotype is PP and both parents are genotype PP, then

$$\begin{aligned} Pr(G_C|G_{AM}, G_{AF}, H_d) &= Pr(\text{PP}|\text{PP}, \text{PP}, H_d) \\ &= Pr(\text{P}|\text{PPPP}, H_d)Pr(\text{P}|\text{PPPPP}, H_d) \\ &= \frac{(4\theta + (1 - \theta)p)(5\theta + (1 - \theta)p)}{(1 + 3\theta)(1 + 4\theta)} \end{aligned} \tag{4.18}$$

Table 4.16 lists the denominator formulae when including population substructure, under a variety of parental genotypes (Dem[c]).

Table 4.16: Parentage Index for various combinations of $G_C$, $G_{AM}$, and $G_{AF}$

| $G_C$ | $G_{AM}$ | $G_{AF}$ | Num.[a] | Den.[b] | Den.[c] | Den.[d,e] |
|---|---|---|---|---|---|---|
| PP | PP | PP | 1 | $p^2$ | $F_I p + (1-F_I)p^2$ | $(4\theta + (1-\theta)p)(5\theta + (1-\theta)p)$ |
| | | PQ | $\frac{1}{2}$ | $p^2$ | $F_I p + (1-F_I)p^2$ | $(3\theta + (1-\theta)p)(4\theta + (1-\theta)p)$ |
| | | QR | 0 | $p^2$ | $F_I p + (1-F_I)p^2$ | $(2\theta + (1-\theta)p)(3\theta + (1-\theta)p)$ |
| | PQ | PP | $\frac{1}{2}$ | $p^2$ | $F_I p + (1-F_I)p^2$ | $(3\theta + (1-\theta)p)(4\theta + (1-\theta)p)$ |
| | | PQ | $\frac{1}{4}$ | $p^2$ | $F_I p + (1-F_I)p^2$ | $(2\theta + (1-\theta)p)(3\theta + (1-\theta)p)$ |
| | | QR | 0 | $p^2$ | $F_I p + (1-F_I)p^2$ | $(1\theta + (1-\theta)p)(2\theta + (1-\theta)p)$ |
| PQ | PP | QQ | 1 | $2pq$ | $2pq(1-F_I)$ | $2(2\theta + (1-\theta)p)(2\theta + (1-\theta)q)$ |
| | | QR | $\frac{1}{2}$ | $2pq$ | $2pq(1-F_I)$ | $2(2\theta + (1-\theta)p)(\theta + (1-\theta)q)$ |
| | | RS | 0 | $2pq$ | $2pq(1-F_I)$ | $2(2\theta + (1-\theta)p)(1-\theta)q$ |
| | PQ | PP | $\frac{1}{2}$ | $2pq$ | $2pq(1-F_I)$ | $2(3\theta + (1-\theta)p)(\theta + (1-\theta)q)$ |
| | | PQ | $\frac{1}{2}$ | $2pq$ | $2pq(1-F_I)$ | $2(2\theta + (1-\theta)p)(2\theta + (1-\theta)q)$ |
| | | QR | $\frac{1}{4}$ | $2pq$ | $2pq(1-F_I)$ | $2(\theta + (1-\theta)p)(2\theta + (1-\theta)q)$ |
| | | RS | 0 | $2pq$ | $2pq(1-F_I)$ | $2(\theta + (1-\theta)p)(\theta + (1-\theta)q)$ |
| | PR | QQ | $\frac{1}{2}$ | $2pq$ | $2pq(1-F_I)$ | $2(\theta + (1-\theta)p)(2\theta + (1-\theta)q)$ |
| | | QR | $\frac{1}{4}$ | $2pq$ | $2pq(1-F_I)$ | $2(\theta + (1-\theta)p)(\theta + (1-\theta)q)$ |
| | | RS | 0 | $2pq$ | $2pq(1-F_I)$ | $2(\theta + (1-\theta)p)(1-\theta)q$ |

[a]Num $= Pr(G_C|G_{AM}, G_{AF}, H_p)$

[b]Den $= Pr(G_C|H_d), \theta = 0, F_I = 0$

[c]Den $= Pr(G_C|H_d), \theta = 0, F_I \neq 0$

[d]Den $= Pr(G_C|H_d), \theta \neq 0, F_I = 0$

[e]Each column entry is divided by $(1+3\theta)(1+4\theta)$.

**Inbreeding with population substructure**

To account for population substructure and inbreeding, Equations 4.17 must be revisted. Note that they could be rewritten as

$$\begin{aligned} Pr_{F_I}(\mathrm{PP}) &= F_I Pr(\mathrm{P}) + (1 - F_I) Pr(\mathrm{PP}) \\ Pr_{F_I}(\mathrm{PQ}) &= Pr(\mathrm{PQ})(1 - F_I) \end{aligned}$$

When considering population substructure, any other observed genotypes affect the frequencies of subsequent alleles. In this case, the probability of seeing the puppy genotype is affected by the observation of the alleged parents, even though the alleged parents are not considered the true parents under the defense hypothesis. This alters the calculations such that

$$\begin{aligned} Pr_{F_I}(\mathrm{PP}|G_{AM}G_{AF}) &= F_I Pr(\mathrm{P}|G_{AM}G_{AF}) + (1 - F_I) Pr(\mathrm{PP}|G_{AM}G_{AF}) \\ Pr_{F_I}(\mathrm{PQ}|G_{AM}G_{AF}) &= Pr(\mathrm{PQ}|G_{AM}G_{AF})(1 - F_I) \end{aligned} \tag{4.19}$$

As in Ayres's method, each $Pr(\cdots|G_{AM}G_{AF})$ is calculated using Ewen's sampling formula. This accounts for the relatedness due to population substructure, while the $F_I$ and $(1 - F_I)$ accounts for the inbreeding within the puppy.

Returning to the example where the puppy is genotype PP and both parents are genotype PQ, the numerator is again the numerator column from Table 4.16. The denominator is the first formula from Equations 4.19.

$$\begin{aligned} Pr(\mathrm{P}|\mathrm{PQ},\mathrm{PQ}) &= \frac{(2\theta + (1-\theta)p)}{1 + 3\theta} \\ Pr(\mathrm{PP}|\mathrm{PQ},\mathrm{PQ}) &= \frac{(2\theta + (1-\theta)p)(3\theta + (1-\theta)p)}{(1 + 3\theta)(1 + 4\theta)} \\ Pr_{F_I}(\mathrm{PP}|\mathrm{PQ},\mathrm{PQ}) &= F_I Pr(\mathrm{P}|\mathrm{PQ},\mathrm{PQ}) + (1 - F_I) Pr(\mathrm{PP}|\mathrm{PQ},\mathrm{PQ}) \\ &= F_I \frac{(2\theta + (1-\theta)p)}{1 + 3\theta} + (1 - F_I) \frac{(2\theta + (1-\theta)p)(3\theta + (1-\theta)p)}{(1 + 3\theta)(1 + 4\theta)} \end{aligned}$$

This leads to the following likelihood ratio

$$\mathrm{LR} = \frac{1}{4}\left[F_I \frac{(2\theta + (1-\theta)p)}{1+3\theta} + (1-F_I)\frac{(2\theta + (1-\theta)p)(3\theta + (1-\theta)p)}{(1+3\theta)(1+4\theta)}\right]^{-1}$$

**Conclusion** We have presented here different ways to calculate a "parentage index." That is, the probability of observing the child's genotype given a particular set of parents, versus the probability of seeing the child's genotype at random. Since this situation would most often come up in animal populations, several different scenarios were considered involving inbreeding and population substructure.

# Literature Cited


[1] B S Weir. DNA statistiics in the simpson matter. *Nature Genetics*, 11, 1995.

[2] J Buckleton. *Forensic DNA Evidence Interpretation*. CRC, 2004.

[3] J W Morris, R A Garber, J d'Autremont, and C H Brenner. The avuncular index and the incest index. *Advances in Forensic Haemogenetics 1*, pages 607–611, 1988.

[4] D J Balding and R A Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3–12, 1995.

[5] B S Weir. Quantifying the genetic structure of populations with application to paternity calculations. In M E Halloran and S Geisser, editors, *Statistics in Genetics Volume 112 of IMA Volumes in Mathematics and its Applications*, pages 31–44. Springer-Verlag, New York, 1998.

[6] X H Liao, T S Lau, K F N Ngan, and J Wang. Deduction of paternity index from DNA mixture. *Forensic Science International*, 128:105–107, 2002.

[7] M Macan, P Uvodić, and V Botica. Paternity testing in case of brother-sister incest. *Croatian Medical Journal*, pages 347–349, 2003.

[8] C H Brenner. Multiple mutations, covert mutations, and false exclusions in paternity casework. In *Progress in Forensic Genetics*, volume 10, pages 112–114, 2003.

[9] K L Ayres and D J Balding. Letter to the editor: Paternity index calculations when some individuals share common ancestry. *Forensic Science International*, 151:101–103, 2005.

[10] I W Evett and B S Weir. *Interpreting DNA Evidence*. Sinauer, Sunderland, MA, 1998.

[11] B Olaisen, M Stenersen, and B Mevag. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics*, 15(4):402–405, 1997.

[12] C M Hsu, N E Huang, L C Tsai, L G Kao, C H Chao, A Linacre, and J C Lee. Identification of victims of the 1998 Taoyuan Airbus crash accident using DNA analysis. *International Journal of Legal Medicine*, 113(1):43–46, 1999.

[13] C H Brenner and B S Weir. Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical Population Biology*, 63(3):176–178, 2003.

[14] B Leclair, C J Fregeau, K L Bowen, and R M Fourney. Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: The Swissair Flight 111 disaster. *Journal of Forensic Sciences*, 49(5):939–953, 2004.

[15] C H Brenner. Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities. *Forensic Science International*, 157:172–180, 2006. Unpubished.

[16] J M Butler, E Buel, F Crivellente, and B R McCord. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for (str) analysis. *Electrophoresis*, 25, 2004.

[17] National Research Council. *The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, DC, 1996.

[18] C H Brenner. What's wrong with the "exclusion probability", November 1997.

[19] B S Weir, C M Triggs, L Starling, L I Stowell, K A Walsh, and J Buckleton. Interpreting DNA mixtures. *Journal of Forensic Sciences*, 44(2):213–222, 1997.

[20] J M Curran, C M Triggs, J Buckleton, and B S Weir. Interpreting DNA mixtures in structured populations. *Journal of Forensic Sciences*, 44(5):987–995, 1999.

[21] P Gill, R Sparkes, R Pinchin, T Clayton, J Whitaker, and J Buckleton. Interpreting simple STR mixtures using allele peak areas. *Forensic Science International*, 91(1):55–70, 1998.

[22] P Gill, C H Brenner, J C Buckleton, A Carracedo, M Krawczak, W R Mayr, N Morling, M Prinz, P M Schneider, and B S Weir. DNA commission of the international society of forensic genetics: recommendations on the interpretation of mixtures. *International Journal of Legal Medicine*, 2006. Submitted for publication.

[23] T W Wang, N Xue, and R Wickenheiser. Least-square deconvolution (LSD): A new way of resolving STR/DNA mixture samples. Presented at: 13th International Symposium Human Identification, October 7-10, 2002.

[24] Y Torres, I Flores, V Prieto, M Lopez-Soto, M J Farfan, A Carracedo, and P Sanz. DNA mixtures in forensic casework: a 4-year retrospective study. *Forensic Science International*, 134(2-3):180–186, 2003.

[25] T M Clayton, J P Whitaker, R Sparkes, and P Gill. Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, 91(1):55–70, 1998.

[26] I W Evett, P D Gill, and J A Lambert. Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Sciences*, 43(1):62–69, 1998.

[27] G S LaBerge, R J Shelton, and P B Danielson. Forensic utility of mitochondrial DNA analysis based on denaturing high-performance liquid chromatography. *Croatian Medical Journal*, 44(3):281–288, 2003.

[28] N Cerri, U Ricci, I Sani, A Verzeletti, and F De Ferrari. Mixed stains from sexual assault cases: autosomal or Y-chromosome short tandem repeats? *Croatian Medical Journal*, 44(3):289–292, 2003.

[29] C Triggs, S A Harbison, and J Buckleton. The calculation of DNA match probabilities in mixed race populations. *Science and Justice*, 40(1):33–38, 2000.

[30] S L Lauritzen and J Mortera. Bounding the number of contributors to mixed DNA stains. *Forensic Science International*, 130(2-3):125–126, 2002.

[31] J M Curran, J S Buckleton, C M Triggs, and B S Weir. Assessing uncertainty in DNA evidence caused by sampling effects. *Science and Justice*, 42:29–37, 2002.

[32] P Gill, L Foreman, J S Buckleton, C M Triggs, and H Allen. A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations. *Forensic Science International*, 131(2-3):184–196, 2003.

[33] D J Balding and R A Nichols. DNA profile match probability calculation - How to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*, 64(2–3):125–140, 1994.

[34] D J Balding. Estimating products in forensic identification using DNA profiles. *Journal of the American Statistical Association*, 90:839–844, 1995.

[35] M Hollander and D Wolfe. *Nonparametric Statistical Methods*. Wiley-Interscience, New York, 1999.

[36] B S Weir. *Genetic data analysis II*. Sinauer, Sunderland, MA, 1996.

[37] B Budowle, T R Morett, A L Baumstark, D A Defenbaugh, and K M Keys. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U S Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *Journal of Forensic Sciences*, 1999.

[38] M W Perlin, G Lancia, and S K Ng. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. *American Journal of Human Genetics*, 57(5):1199–1210, 1995.

[39] M W Perlin and B Szabady. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*, 46(6):1372–1378, 2001.

[40] W H Press, S A Teukolsky, W T Vetterling, and B P Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 2002.

[41] S Wright. The genetical structure of populations. *Annuals of Eugenics*, 15:332–354, 1951.

[42] C Ws Cotterman. *A Calculus for Statistico-Genetics*. PhD thesis, Ohio State University, 1940.

[43] G Malécot. Étude mathématique des populations "mendeliennes". *Ann Univ Lyon Sci Sec A*, 4:45–60, 1941.

[44] Y Hu and W K Fung. Interpreting DNA mixtures with the presence of relatives. *International Journal of Legal Medicine*, 117:39–45, 2003.

[45] W Liu and B S Weir. Genotypic probabilities for pairs of inbred relatives. *Philosophical transaction of the Royal Society - B*, 360:1379–1385, 2005.

[46] K L Ayres. Relatedness testing in subdivided populations. *Forensic Science International*, 114:107–115, 2000.

[47] W K Fung, A Carracedo, and Y Hu. Testing for kinship in a subdivided population. *Forensic Science International*, 135:105–109, 2003.

[48] W K Fung and Y Hu. Interpreting DNA mixtures with related contributors in subdivided populations. *Scandinavian Journal of Statistics*, 31:115–130, 2004.