Evaluation of statistical methods for the analysis of forensic DNA mixtures

Hinda Haned

Numéro d'ordre: 231-2010 Année 2010

THÈSE

Présentée

devant L'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

soutenue le 29 Octobre 2010

par

Hinda HANED

Évaluation de méthodes statistiques pour l'interprétation des mélanges d'ADN en science forensique

Volume 1: texte principal

Directeurs de thèse: Dominique Pontier
Laurent Pène
Frank Sauvage

Jury: Chritisan Biémont (Rapporteur)
Peter Gill (Rapporteur)
Denis Laloë (Examinateur)
Eric Petit (Rapporteur)
Laurent Pène (Co-directeur)
Dominique Pontier (Directrice)
Frank Sauvage (Co-directeur)

Évaluation de méthodes statistiques pour l'interprétation des mélanges d'ADN en science forensique

Résumé de la thèse:

L'analyse et l'interprétation d'échantillons constitués de mélanges d'ADN de plusieurs individus est un défi majeur en science forensique. Lorsqu'un expert de la police scientifique a affaire à un mélange d'ADN il doit répondre à deux questions: d'abord, "combien de contributeurs y a-t-il dans ce mélange ?"et puis, "quels sont les génotypes des individus impliqués ?"

Le typage seul de cet ADN ne permet pas toujours de répondre à ces questions. En effet le problème est posé dès lors que plus de deux allèles sont observés à un locus donné, plusieurs combinaisons génotypiques sont alors à envisager et il est impossible de déterminer avec certitude le nombre d'individus qui ont contribué au mélange. De plus, la présence d'anomalies liées à l'analyse de marqueurs génétiques, comme la contamination ou la perte d'allèles ("drop-out"), peut davantage compliquer l'analyse.

Les nombreux développements statistiques dédiés à ces problématiques n'ont pas eu le succès escompté dans la communauté forensique, essentiellement, parce que ces méthodes n'ont pas été validées. Or sans cette validation, les experts de la police scientifique ne peuvent exploiter ces méthodes sur des mélanges issus d'affaires en cours d'investigation.

Avant d'être validées, ces méthodes doivent passer par une rigoureuse étape d'évaluation. Cette dernière soulève deux questions: d'abord, la question de la méthodologie à adopter, puis, celle des outils à déployer. Dans cette thèse, nous tentons de répondre aux deux questions. D'abord, nous menons des études d'évaluation sur des méthodes dédiées à deux questions clés: i) l'estimation du nombre de contributeurs à un mélange d'ADN et ii) l'estimation des probabilités de "drop-out". En second lieu, nous proposons un logiciel "open-source" qui offre un certain nombre de fonctionnalités permettant de faciliter l'évaluation de méthodes statistiques dédiées aux mélanges d'ADN.

Cette thèse a pour but d'apporter une réponse concrète aux experts de la police scientifique en leur fournissant à la fois une démarche méthodologique pour l'évaluation de méthodes, et la possibilité d'analyser la sensibilité de leurs résultats au travers d'un outil informatique en libre accès.

Evaluation of statistical methods for the analysis of forensic DNA mixtures

Abstract:

Analysis of forensic DNA mixtures recovered from crime scenes is one of the most challenging tasks in forensic science. DNA mixture raise two main questions: "how many contributors are there" and "what are the genotypes of the contributing individuals?" The genetic characterization alone of such samples does not always answer these questions. In fact, whenever more than two alleles are observed at a given locus, several distinct genotypic combinations are plausible for the unknown contributors to the sample, and it is not possible to determine the number of these contributors with absolute certainty. Besides, the presence of anomalies related to DNA typing techniques, such as contamination or allele loss (drop-out), can further complicate the analysis.

Numerous statistical developments facilitating DNA mixtures interpretation were proposed, but they did not receive the expected success in the forensic community. The main explanation for this is that these methods are not validated for forensic casework.

In order to achieve this validation criterion, the methods must undergo a rigorous evaluation step. The latter raises two questions: i) how methods should be evaluated? and ii) what tools can be used to conduct evaluation studies? In this thesis we attempt to answer both questions. First, we evaluate methods dedicated to two key issues, the estimation of the number of contributors to DNA mixtures and the estimation of drop-out probabilities. Second, we propose an "open-source" software that offers a number of functionalities dedicated to facilitating method evaluation through the simulation of data commonly encountered in forensic settings.

This thesis aims to provide a concrete answer to the issues raised by forensic DNA mixtures, by providing a methodology for method evaluation and by offering necessary tools to enable method evaluation.

Contents

Chapter 1

Introduction

Contents

1.1 Background

The scene takes place in a house where a victim of a suspicious drug overdose is found dead in her living room. A homicide police detective (Jim Brass) and a forensic scientist (Warrick Brown) are in the victim's bedroom seeking clues that may help them understand the suspicious circumstances of the death. Detective Brass turns his attention to balloons that have probably been used to handle drugs.

> Brass: *"Can you get a print off these balloons?"*
> Warrick (obviously annoyed by the question): *"I can get a print off the air."*

This exchange, extracted from CBS's hit television series *CSI: Crime Scene Investigation*[1], is typical of the numerous television shows featuring forensic scientists solving murders, rapes, and other violent crimes using a number of advanced technologies, including DNA typing. The techniques presented in these series are, of course, extremely fast and always capable of producing sound scientific evidence, but what is most striking is the rapidity with which new technologies or methods are introduced into laboratory practice. During a single episode, a forensic scientist may develop a method to address a particular issue (for example, air fingerprinting) and apply it to a piece of evidence.

By exaggerating the abilities of the techniques employed in forensic science, these popular shows have contributed to raising the expectations of viewers, among whom are eligible jurors, regarding forensic science and, in particular, DNA evidence.

Obviously, the reality of forensic science is quite different from what is described in these TV shows. For instance, before new techniques can even be applied to real cases, they have to i) be adapted to the reality of casework, in which forensic scientists are often confronted with samples that are limited in quantity and quality, and ii) overcome a series of validation studies to ensure that the yielded results are reliable and admissible as scientific evidence. Once these criteria are fulfilled, forensic laboratories establish standardized protocols that ensure that the technology will be used under appropriate conditions that will enable the generation of reliable results (Rudin and Inman, 2002; Jobling and Gill, 2004).

Rigorous validation of novel methods and technologies are currently watchwords in forensic science, but this has not always been the case.

Forensic science crosses the boundaries of a broad spectrum of sciences: biology, chemistry, mathematics and physics provide a wide variety of tools to analyze evidence as varied as shoe prints, bite marks, fingerprints, blood stains, glass shards, bullets, ear-prints, recorded voice, etc (Siegel et al., 2000). Despite this wide variety of investigative tools, it is the DNA evidence that revolutionized forensic investigations. Immediately after its introduction into forensic practice in the early 1980's, DNA was perceived by the general public, justice officials and a number

[1] *CSI: Crime Scene Investigation*, "Burked" (CBS television broadcast September 27, 2001). Transcript available at `http://www.twiztv.com/scripts/csi/season2/csi-201.txt`.

of scientists, as the most powerful means of identification available (Aronson, 2007). As a consequence, DNA evidence was rapidly introduced into courtrooms, although standards of use and interpretation had not yet been discussed or published.

The "DNA revolution" began in 1984 when Alec Jeffreys discovered a particular sequence of nucleotides in varying repetitive patterns in mammalian genomes. Jeffreys discovered that the number of repetitions of these sequences, as well as the length of the repeat units, varied from one individual to another, constituting length polymorphisms (Jeffreys et al., 1985b). The potential of these regions, which became known as variable number of tandem repeats (VNTR), was quickly realized as being tremendous for forensic identification purposes:

> "*The implications for individual identification and kinship analysis were obvious... It was clear that these hypervariable DNA patterns offered the promise of a truly individual-specific identification system. We therefore coined the term "DNA fingerprinting" as a deliberate move to emphasize the new forensic paradigm that we could foresee if these probes could be used in criminal and civil investigations.*" (Jeffreys, 1993).

The first time DNA evidence was introduced in a criminal case was at the end of the 1980s. The case involved a double homicide in Leicestershire, England. The DNA evidence, consisting of the semen recovered from both victims, was characterized using restriction length polymorphisms, which had only just been discovered by Alec Jeffreys. Interestingly, before the DNA evidence could help identify the perpetrator of the crimes, it first helped demonstrate the innocence of a man who had confessed to the killing of one of the victims.

Immediately after this "baptism by fire" for forensic DNA typing, DNA evidence started pouring into forensic laboratories, which found themselves, at least in the U.S., forced to outsource a part of their casework to private companies. The rush of DNA evidence to courts had undesirable effects. Though the technologies and procedures used in forensic DNA typing were rooted in molecular biology, there was a considerable lack of (peer-reviewed) scientific papers describing and justifying the specific DNA typing procedures being used. The desire to introduce forensic DNA typing as rapidly as possible led to a lack of rigor in the practice of analyzing and interpreting DNA evidence (Lander, 1989).

Some of these poor laboratory practices had been revealed during several U.S. trials by the end of the 1980s, leading to the dismissal of DNA as a form of evidence because of poor-quality laboratory work. Eric Lander was one of the first scientists to attack the "poorly defined procedures and interpretation" of many of the companies selling forensic DNA typing services to the police (Lander, 1991).

These first failures marked an important step in the perception of DNA evidence, at least in the scientific community: DNA evidence was no longer considered flawless, and as a result, forensic DNA typing was more than ever subject to debate and criticism (Lander and Budowle, 1994). While there was no disagreement in the community regarding the need for standards

and interpretation guidelines, it was another aspect of forensic DNA typing that allowed the controversy to be settled at greater length: the model of population genetics underlying the calculation of the weight of DNA evidence.

When DNA evidence reveals a match between a crime scene profile and the profile of a tested suspect, the next step is to determine the probability that the match at the tested loci is coincidental. This probability is usually calculated as a product of allele frequencies across all available loci, thereby assuming independence both between and within loci. The first points in this controversy were raised by critics claiming that these assumptions of independence are violated in human populations (Lander, 1989; Cohen, 1990). It was claimed that human populations might be structured into subpopulations, causing variations of allele proportions among these subpopulations. The existence of population structure would invalidate independence assumptions and thus the model used to calculate the profile frequencies.

Many studies using the data that were available at the time followed to refute these assertions (Devlin et al., 1990; Chakraborty and Jin, 1992; Weir, 1992a,b). The dispute escalated further with the publication of a paper by two pioneers of population genetics: Richard Lewontin and Daniel Hartl (Lewontin and Hartl, 1991). The numerous papers and letters to the editor in response to this paper and other, related articles demonstrated the heated exchanges around this subject (Chakraborty and Kidd, 1991; Risch and Devlin, 1992; Morton, 1992; Devlin and Risch, 1992, to cite a few).

Two factors helped end this dispute, which was referred to as the "DNA wars" by the press[2]. First, a series of population studies carried out by the U.S. Federal Bureau of Investigation showed that subpopulation effects were not as substantial as originally claimed (U.S. Department of Justice, Federal Bureau of Investigation, 1993). Second, a series of publications, including a report from the U.S. National Research Council[3], proposed statistical methods to account for population substructure through the correction of allele frequencies (Balding and Nichols, 1994; Evett and Weir, 1998).

The controversy related to both the statistical and biological aspects of forensic DNA typing has contributed to shaping the science of forensic genetics as it is known today. This has greatly contributed to establishing good practices and has made quality assurance and quality control watchwords in forensic laboratories. The methods involved in forensic DNA typing have co-evolved with dynamic validation processes, permitting the progressive introduction of tools at the cutting edge of developments in molecular biology, as shown by today's powerful genotyping system based on short tandem repeat loci multiplexes. This system has undergone extremely rigorous validation studies across the globe (Jobling and Gill, 2004). Similarly, methods for interpreting a match in single-source profiles, i.e., profiles where the DNA of only one individual

[2]Humes, E. "DNA War" *L.A. Times Magazine*, Novembre, 29, 1992.

[3]NRC II - National Research Council Committee on DNA Forensic Science, National Academy Press: Washington, D. The Evaluation of Forensic DNA Evidence, 1996.

is involved, are rooted in a rigorous population genetics framework, which was notably initiated by the DNA wars.

While there is agreement on the interpretation of single-source stains, the interpretation of DNA profiles consisting of mixtures of DNA from several individuals has not received the same attention from the scientific community, at least not at a level comparable to the DNA wars, and to date, there is no consensus on how DNA mixtures should be interpreted (Perlin, 2006).

The statistical analysis of DNA mixtures is challenging: While in single-source stains, only one genotype is possible for each locus, several genotypes are possible in DNA mixtures. Therefore, whenever a stain is suspected to be a mixture, reporting officers are challenged with at least two questions: i) how many contributors are there, and ii) what are the genotypes of the contributors to the mixture?

It is often not easy to answer these questions. In fact, the individuals who contributed to the mixture may carry the same alleles at one or more loci, which is a phenomenon known as allele sharing, reducing the available information and making it hard or even impossible to either determine the number of individuals involved in the stain or to resolve the mixture into individual components. In addition, DNA profiles can be affected by a number of anomalies that render their interpretation even more challenging. Indeed, forensic DNA samples are often recovered in less than ideal situations: biological material at the crime scene can be subject to extreme environmental conditions, leading to the degradation of the DNA. This degradation, in addition to the limited quantities of DNA available from the evidence, tends to increase the probability of allele loss and the introduction of spurious alleles due to contamination.

It is within this context that our collaboration with the forensic laboratory in Lyon[4] was initiated: as reporting officers are confronted daily with DNA mixtures, they require statistical tools that may help them throughout the interpretation process. A review of the relevant literature revealed that the field of statistical interpretation of DNA mixtures has, in fact, been one of the most active fields in forensic science in the past decade in terms of methodological developments. However, these developments have not produced the expected level of success in the forensic community, and there are two main reasons for this: first, courtroom resistance to probabilistic reasoning, and second, the issue of the validation of methods for forensic science.

Courtroom resistance to probabilistic reasoning

In criminal cases, forensic scientists are too often asked by the court to report DNA evidence in a binary fashion: "Is the suspect guilty, yes or no?" Indeed, it is generally thought that a DNA match proves that the suspect is guilty, which is not true. A number of situations can lead to a match between two samples, including coincidence; for example, the suspect could have left the DNA trace during an occasion unrelated to the crime, or a coincidental match may occur between the perpetrator and the suspect profiles. In any case, background evidence must be

[4]Laboratoire de Police Scientifique de Lyon, Institut National de Police Scientifique (INPS).

used by the judge and the jury to reach a conclusion about the guilt or innocence of the suspect. Relying on probabilistic reasoning, a scientist can only give a measure of the weight of DNA evidence, which indicates the strength of the association between the compared samples.

However, courtrooms have a history of reluctance towards "complex" probabilistic reasoning. This is understandable because judges, juries and lawyers have little experience in this field (Taroni et al., 2002). Therefore, simple approaches are preferred to more complicated ones. As a consequence, introducing methods based on probabilistic reasoning is difficult. In particular, a number of methods have been proposed to facilitate mixture interpretation through the decomposition of DNA profiles into individual components. However, to our knowledge, none of them is currently in use in any forensic laboratory.

Method validation issues

The methods used for genotypic or statistical analysis in a forensic setting must undergo rigorous evaluation through studies that eventually lead to the validation of the method for use in forensic casework. Validation studies usually consist of comparing a new method to a reference method or checking the consistency of the tested method with respect to criteria predefined by laboratory practice.

Statistical methods dedicated to DNA evidence interpretation and, in particular, to analysis of mixtures are usually complex and require computer software. However, there is a clear lack of tools enabling the implementation and testing of these methods. This lack of tools has seriously limited the introduction of statistical improvements into forensic practice. Indeed, it is easier for a forensic laboratory to test tools and procedures that are related to DNA analysis than to evaluate statistical methods, simply because tools for these tests are not available (whether in free or in commercial software).

Given these observations, we conclude that there are at least two strategies to improve forensic interpretation of DNA mixtures. The first consists of providing education and training in statistics for forensic scientists, judges and lawyers to enhance the admissibility of statistical methods in the courtroom. The second, which is the one that concerns us here, is the development of accessible evaluation tools that would facilitate the validation of statistical methods and, hence, their introduction into forensic casework. Ideally, both strategies should be pursued simultaneously. Our thesis aims to answer this need for evaluation tools both methodologically and technically. Throughout this manuscript, we will try to demonstrate how our contribution, on both methodological and technological levels, can help introduce statistical reasoning into courtrooms.

To help frame the motivation behind the work presented in this thesis, we describe a number of aspects of DNA evidence interpretation and address common problems encountered in forensic casework. For our purposes, the analytical process associated with the analysis of evidentiary

DNA stains can be described in two stages: first, the genetic analysis, and second, the statistical analysis. The first stage involves all of the procedures that lead to the generation of a DNA profile. The second stage most often consists of assigning a weight to the association observed between a crime scene profile and a suspect profile. In what follows, we provide key information on methods and procedures related to these two stages but without actually describing them in detail.

1.2 Overview of forensic DNA typing

DNA testing is a relatively recent advent in forensic science, with its first applications in the early 1980s. The desire to achieve a gold standard, i.e., an "ideal" in terms of reliability and speed, has prompted forensic scientists and private companies to introduce genotyping tools for use as soon as they are developed. The current methodology employed for DNA evidence stains is based on the simultaneous analysis of 10 to 15 short tandem repeat (STR) loci. Before the advent of this powerful system, investigation of VNTR loci was the preferred method of forensic DNA typing used in most U.S. and European forensic labs.

As a matter of fact, the first case of human identification in a forensic setting, which came during the investigation of the aforementioned Leicestershire double murder, involved the analysis of recently discovered VNTR loci. The technique used to examine VNTR loci at that time was based on restriction fragment length polymorphism (RFLP) analysis (Jeffreys et al., 1985a; Gill et al., 1985).

This method involves the use of restriction enzymes to cut the DNA surrounding VNTR regions. Once the DNA is digested at specific restriction sites, the resulting DNA fragments are separated according to their size via electrophoresis. The results of the RFLP typing process consist of a multi-banded two-dimensional picture in which each band corresponds to a probe matching a DNA fragment with a desired sequence (Figures 1.1 & 1.2).

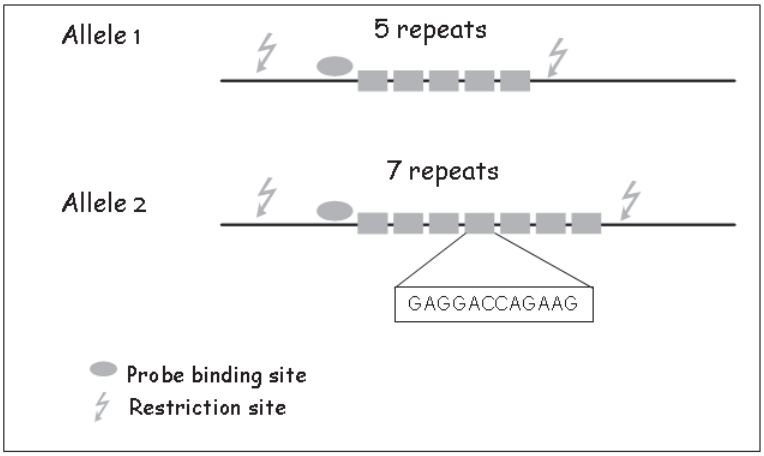


FIGURE 1.1: Illustration of the analysis of VNTR markers through RFLP analysis. The alleles of a heterozygote individual, differing in the number of repeats, are represented.

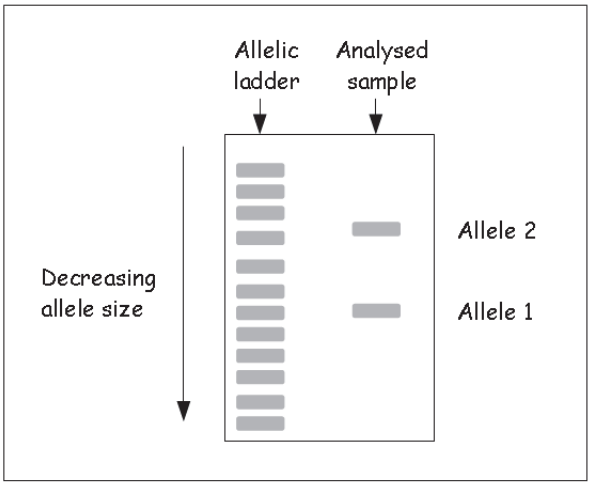


FIGURE 1.2: Representation of an autoradiogram showing the alleles bands for a single-locus VNTR analysis (adapted from Butler, 2001).

The original technique involved in the analysis of VNTR loci was particularly time-consuming, as the entire process involved up to several months to characterize a single DNA stain. Another concern regarding RFLP-based methods is the high sensitivity to degraded stains with limited amounts of DNA. Degradation might be caused by extreme environmental conditions (e.g., water, fire, body decomposition) and is commonly encountered in forensic contexts. As a consequence of these limitations, alternative procedures based on the polymerase chain reaction (PCR) emerged. This technique permits the amplification of numerous DNA copies through the use of the enzyme DNA polymerase.

The first attempts to use PCR to amplify VNTR loci were confronted with the limitation imposed on fragment lengths. The method was not adapted to large DNA fragments (more than 2,000 bp), and therefore, it was not adapted to VNTR loci. The search for shorter VNTR loci has led to the discovery of STR loci, which have shorter sequences (less than 500 bp) (Weber and May, 1989). PCR can target specific areas of the genome (STR) and generate billions of identical copies of these regions. Once the DNA amplification step is completed, PCR-based analysis of STR follows the same principle as described above for VNTR and the results are visualized through an autoradiogram.

The introduction of STR loci analyzed by PCR increased the sensitivity of this analysis by allowing small amounts of DNA to be analyzed. Additionally, the introduction of DNA sequencing technologies led to reductions in the time and the cost needed to carry out the analyses. Another major advance that led to the techniques currently used in forensic typing was the advent of fluorescent detection of STR loci (Edwards et al., 1991). Fluorescent labeling permits the simultaneous detection of multiple STR loci, referred to as STR multiplexing.

The principle of DNA detection remains the same; PCR is performed to amplify specific regions of the genome followed by capillary electrophoresis to separate the DNA fragments according to size. The originality of this method resides in the fluorescent labeling of the PCR

probes. The DNA fragments separated by size during the capillary electrophoresis are drawn past a laser that excites the florescent dye labels that are linked to the DNA during PCR. The amounts of fluorescence are then reported in relative fluorescent units (RFUs). The end result of the process is an electropherogram plotting the amount of fluorescence over time, with fluorescent peaks indicating DNA material of different sizes (Figure 1.3)

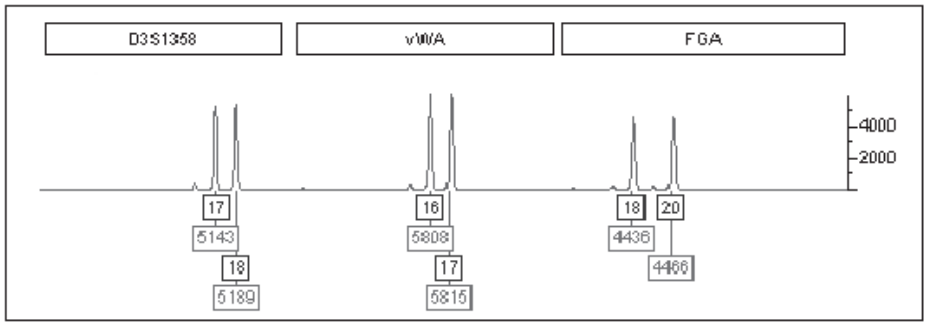


FIGURE 1.3: Extract of an electropherogram of the analysis of a single-source stain using STR markers. Three markers are shown: "D3S1358", "vWA" and "FGA". Allele peaks are labelled with boxes containing two information: the first one indicates the allele call, determined from the number of tandem repeats, the second one is the peak height in RFUs. For example, at locus "FGA", the typed individual carries alleles "18" and "20" (figure adapted from Doom and Raymer, 2004).

"DNA profile by itself is fairly useless" (Butler, 2001). Although this statement can be nuanced, DNA evidence is usually considered in the context of a comparison between two samples: a "questioned sample", usually recovered from a crime scene or from a victim, and a "known sample", or reference sample, usually consisting of the DNA of an individual suspected of committing the investigated crime. If no match is observed between the two samples, then the analysis does not go further. On the contrary, if there is a match between the sample profiles, then a *weight* is assigned to the observed match.

Figure 1.4 illustrates such a situation, where a comparison between a blood stain and a suspect profile is carried out. In this example, the blood stain is characterized with only three loci, but forensic DNA typing is usually carried out using at least 10 or more STR loci.

Observing a match does not, in and of itself, prove that the suspect is the culprit. It is the province of the judge and the juries to determine this in the light of other (non-genetic) evidence. A match only indicates that the culprit and the suspect have the same profile. To quantify the strength of the observed match, a statistical evaluation must be carried out. Indeed, the size of the human genome makes it unrealistic to entirely sequence the DNA samples involved in a crime scene investigation. Instead, only a limited number of markers are typed to genetically characterize the suspect or the crime scene profile genotype. As we previously mentioned, STR markers have several desirable properties for use in forensic DNA typing, but their most important feature resides in their high power of discrimination between individuals.

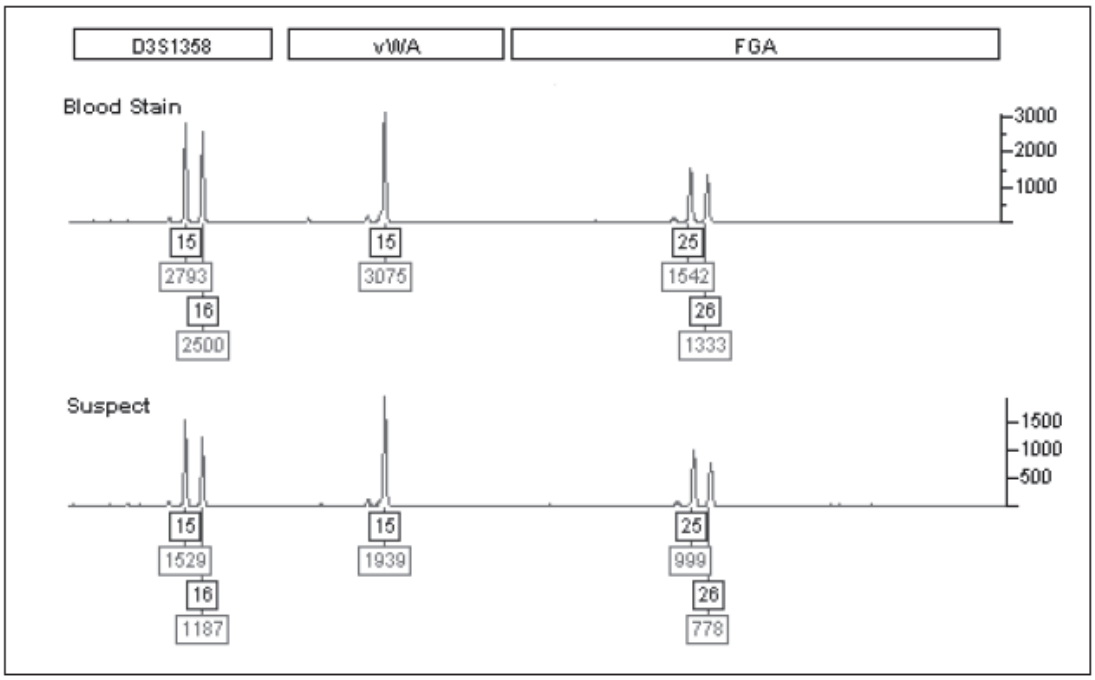


FIGURE 1.4: Illustration of a comparison between a blood stain recovered from a crime scene, and a matching DNA profile of a suspect (adapted from Gilder, 2007).

For each of the STR loci used in forensic DNA typing, at least 10 alleles have been reported in human populations. As a consequence of this considerable diversity, the probability of any two unrelated individuals having exactly the same allele at each STR locus is extremely low. If we consider a single STR locus with 15 alleles, then there are 120 possible genotypes at this locus, and if several loci are considered simultaneously, for example, the 13 loci from the Combined DNA Index System in use in the U.S., then at least 10^{27} profiles are possible.

Because of the different process shaping the genetic pool of populations (selection, migration, genetic drift), some alleles are more frequent than others, making it unlikely that allele frequencies are uniformly distributed in the population. Therefore, only a small fraction of the possible profiles is actually observed in the population. Still, because a complete description of the suspect's genome is not obtained, a central remaining question is whether it is plausible that another (unknown) person with the same profile as the suspect could be the actual contributor to the DNA evidence. Weighting DNA matching evidence generally consists of assigning a probability to this hypothesis. This probability varies depending on the frequency of the profile in the target population: the rarer the profile in the population, the higher the value of the DNA evidence.

1.3 Statistical evaluation of DNA evidence

Attaching a statistical weight to an observed match avoids the binary process of the match/no-match analysis, which overstates the informativeness of the DNA evidence. In this regard, it is

interesting to note that the term "genetic fingerprint" was quickly forsaken in favor of "genetic profile". The old terminology wrongly implied that DNA evidence is comparable to fingerprints, which were long portrayed as perfectly individual (Coquoz and Taroni, 2006). The term "profile" introduces a necessary nuance with respect to the informativeness of DNA evidence. The genotype of an individual described with 10 STR markers is not an exhaustive or full description of his person but rather a description of some of his features, in this case, a description of a part of his genome.

Several statistical approaches can be employed to attach a statistical weight to DNA evidence. The methods differ in their underlying philosophical approaches and in the type of data used to report the strength of the DNA evidence. Buckleton et al. (2005) proposed a classification of these methods, distinguishing the frequentist approach from the Bayesian approach.

To illustrate these approaches, we will use a fictitious scenario. Suppose that a crime has been committed and that a blood stain has been recovered from a crime scene. The crime scene investigators believe that this stain was left by the offender. Let us now suppose that as a result of the police investigation, a suspect is detained and he provides a sample of his blood. A forensic laboratory provides the profiles for both the blood stain and the suspect, and the two samples match at all of the STR loci used for the DNA analysis (Table 1.1).

	Crime scene profile		Suspect profile
Locus	Alleles	Frequencies	
D8S1179	14	0.16556	14
CSF1PO	11	0.30132	11
	14	0.00828	14
VWA	18	0.20033	18
FGA	20	0.12748	20
	25	0.07119	25

TABLE 1.1: A fictitious case of match between a blood stain recovered from a crime scene and a suspect profile. The blood stain was characterized with four loci, the observed alleles and their corresponding frequencies in the U.S. Caucasian popoulation (Butler et al., 2003) are given in the table.

1.3.1 The frequentist approach to DNA evidence interpretation

In the frequentist approach, the strength of the DNA evidence is reported in terms of probabilities. Two common methods are used in this approach: the random man not excluded (RMNE) probabilities and the coincidence probabilities.

Random man not excluded (RMNE)

In the RMNE approach, the strength of the evidence is reported in terms of the probability that a random individual will not be excluded as a contributor to the DNA evidence. In the example shown in Table 1.1, a match is observed between the suspect's profile and the crime

scene profile. The table displays the alleles observed at four loci, along with the corresponding allele frequencies in the U.S. Caucasian population (Butler et al., 2003). If the DNA stain has alleles $A_1, ..., A_n$ at a given locus l, with frequencies $p_1, ..., p_n$, the RMNE for locus l is given by: $RMNE_l = 1 - \left(\sum_{i=1}^{n} p_i\right)^2$, and the RMNE across multiple loci is given by: $1 - \prod_l (1 - RMNE_l)$.

Based on the data from Table 1.1, the RMNE statistic is approximately equal to 0.99, and using this approach, the DNA evidence will be presented this way: "Approximately 99% of unrelated individuals would be excluded as the source of this DNA."

Using the probability of exclusion is primarily justified by its simplicity and the fact that no assumption is issued with respect to the number of potential contributors to the mixture. However, it should be noted that this method does not use the information available on the genetic profiles of known contributors to the mixture (e.g., victim or suspect).

Coincidence probabilities

Coincidence probabilities (or random match probabilities) give the probabilities of a coincidental match to the trace evidence sample. Based on the suspect's genotypic frequencies in the target population, the random match probabilities assess whether the match between the suspect and the crime scene profiles could occur by coincidence. The probabilities of the multi-locus genotypes are usually calculated by multiplying together the frequencies of the per-locus genotype, which are in turn, calculated by multiplying the frequencies of the alleles observed in the profile (and including a factor of two for each heterozygous genotype). This implies that the population being considered (in our example the Caucasian population) is under Hardy-Weinberg equilibrium at the considered STR loci and that linkage equilibrium is verified across loci. These simplifying assumptions fall within a simple model of population genetics, termed the "product rule", that has been in use since the advent of the first DNA typing cases (Evett and Weir, 1998).

Returning to the former example, the match probability is obtained by multiplying the allele frequencies appearing in Table 1.1, and including a factor of 2 whenever a heterozygous locus is observed. Thus, the probability that the four-locus match occurred by chance is approximately equal to 9×10^{-8}. A reporting officer will generally report this probability in a manner such as, "the probability of observing the DNA profile if the blood stain came from someone else than the suspect is approximately 9×10^{-8}".

Frequentist methods are used in many forensic laboratories, and in many countries, RMNE probabilities constitute the only admissible statistical method for reporting the weight of DNA evidence (Buckleton and Curran, 2008). This is understandable, as frequentist methods seem more natural for the court (judges, juries, lawyers) and are easily understandable by non-statisticians. However, this approach has serious limitations, mainly because the probability of the evidence is considered under a unique hypothesis, which is usually that "someone other than the suspect left the DNA stain recovered from the crime scene". Critics of this approach

argue that DNA evidence cannot be properly assessed without considering at least a second alternative explaining the origin of DNA evidence. This issue is even more noticeable when the DNA evidence is complex. Consider the previous example in which the suspect's profile matched the crime scene profile. We now consider a case in which the crime scene profile is a mixed stain of two individual profiles (Table 1.2).

	Crime scene profile	Suspect profile
Locus	Alleles	
D8S1179	14	14
	16	
CSF1PO	11	11
	13	
	14	14
VWA	14	
	18	18
FGA	20	20
	25	25

TABLE 1.2: Profile of a DNA mixture recovered from a crime scene. The suspect's profile matches only a part of the alleles from the crime scene profile.

In such a scenario, frequentist methods can still be used, provided that the suspect is included in the DNA evidence. Extra information is not used in the evaluation of the DNA evidence, and in this case, the conclusions regarding the RMNE or the match probabilities would be the same as in the example in Table 1.1.

It is obvious from this example that DNA evidence cannot be correctly assessed under only one particular hypothesis. Ideally, the evidence should be evaluated under a flexible framework that allows taking into account all of the information provided by genetic data available and that includes the extra alleles observed in the sample that are not explained by the suspect's profile. Therefore, the frequentist approach is limited whenever any complication arises; for instance, a mixed stain cannot be evaluated properly because the methods do not make use of all the available information that might be relevant to the case.

These limitations, described as "frustrations" by Buckleton et al. (2005), led to the development of alternative methods for assessing the strength of DNA evidence. The Bayesian approach has naturally emerged as a possible alternative, especially because it has been implemented in paternity testing since the 1930s (Essen-Möller., 1938)

1.3.2 The Bayesian approach to DNA evidence interpretation

The Bayesian approach is based on three principles of interpretation that were first formulated by Evett and Weir (1998):

1. "*To evaluate the uncertainty of any given proposition, it is necessary to consider at least one alternative proposition.*"

2. "*Scientific interpretation is based on questions of the kind "What is the probability of the evidence given the proposition?"*"

3. "*Scientific interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated.*"

The Bayesian approach is thereby based on the calculation of conditional probabilities, which usually consist of the probabilities of observing the genetic evidence given alternative hypotheses. A typical Bayesian analysis of a crime scene sample would consist in weighting two alternative hypotheses formulated by two adverse camps: the prosecution who wants to convict the suspect, and the defence, who wants to exonerate the suspect. We introduce the following notations:

E:	the genetic evidence, generally consists of the crime scene profile and the suspect profile
Hp:	the prosecution hypothesis, in general the prosecution will seek to associate the suspect profile with the profile of the crime scene
Hd:	the defence hypothesis, in general the defence camp will seek to exonerate the suspect
I:	non-genetic evidence; all the background information relevant to the case that have led the investigators to designate the suspect as the possible offender

Note that more than two alternative hypotheses can be considered under the Bayesian approach, here we only detail the case for two competing hypotheses.
The principles of interpretation are summarized in the odds from Bayes' Theroem:

$$\frac{Pr(H_p|E,I)}{Pr(H_d|E,I)} = \frac{Pr(E|H_p,I)}{Pr(E|H_d,I)} \times \frac{Pr(H_p|I)}{Pr(H_d|I)} \tag{1.1}$$

This formula is also often written as: Posterior odds = Likelihood ratio $\times$ Prior odds.

The formulation in equation (1.1) is particularly attractive because it provides a mathematical representation, of the legal process (Curran, 2009): Prior odds are the odds of the two competing hypothesis (H_p and H_d), given background evidence (I) and before any DNA evidence is received. It can be seen as the prior belief about a suspect's guilt (or innocence). Posterior odds are the odds of the two hypotheses once the DNA evidence is considered, the likelihood ratios relates the two quantities.

It is unlikely that the prior odds are determined numerically by the judges and the jury and it is generally assumed to be 1, thus, the quantity cancels out from equation (1.1). Usually, forensic scientists report only the likelihood ratio. As a consequence, the ratio above is more akin to a "logical approach" than to a Bayesian approach because the prior odds is generally assumed to be 1 (Buckleton et al., 2005).

Using likelihood ratios, the strength of DNA evidence is generally expressed into the following terms: "The evidence is x times more likely if the prosecution hypothesis is true than if the defence hypothesis is true" (Gill, 2009). Unlike in the frequentist approach where the suspect is said to be "excluded" or "included", the LR framework is very flexible: its allows the

simultaneous testing of the prosecution and the defence hypotheses, thereby, any scenario can be considered under each tested hypothesis.

This flexibility led to a number of statistical developments proposing general formulations of the LR accounting for a wide variety of situations, for example population substructure, the existence of multiple contributors to the evidence (Curran et al., 1999), relatedness between contributors to the evidence (Hu and Fung, 2005), etc.

The LR framework also allows accounting for uncertainty when reporting the weight of DNA evidence. Indeed, the analysis of DNA evidence can sometimes be complicated due to stochastic effects (some of which will be detailed later) that are exacerbated when the DNA is in low quantity or quality. For example, allele drop-out, i.e., the non-detection of an allele that is present in the sample, can occur, especially in degraded DNA samples. If an allele is suspected to have dropped-out, relying on samples comparisons, then a probability can be assigned to this hypothesis, instead of removing the problematic locus, or not exploiting the DNA evidence at all. Anchored in a LR framework, Gill et al. (2000) proposed a generalised statistical model that takes into account the stochastic effects that may complicate the interpretation of DNA evidence. To our knowledge, the use of this model is not widespread in forensic laboratories.

Although the LR framework has received general acceptance in the forensic community (Gill et al., 2006), courtrooms reluctance to statistical reasoning has significantly slowed down the introduction of advanced statistical solutions, such as Gill et al.'s model (2000). One of the most emblematic manifestation of this reluctance is reflected in the ruling of the Court of Appeal in the Adams case[5] (1996, UK):

> "*Jurors evaluate evidence and reach a conclusion not by means of a formula, mathematical or otherwise, but by the joint application of their individual common sense and knowledge of the world to the evidence before them.*"

As a consequence of courtroom reluctance to probabilistic reasoning, methods used to report weight of DNA evidence are mostly based on the frequentist approach, thereby limiting the possibilities. However, some countries have played the role of pioneers in adopting improved methodologies. This is the case of the Forensic Science Service in the UK, who employs a LR approach in reporting DNA evidence.

The increased sensitivity of DNA typing methods makes the use of a LR framework more relevant then ever. Indeed, the theoretical sensitivity of current DNA typing technologies based on STR loci is a single (diploid) cell. Still, the stochastic effects mentioned above are particularly exacerbated in samples with low quantities of DNA. As a consequence, forensic scientists are more and more faced to increasingly complex cases, requiring advanced statistical solutions. These observations constitute the starting point of our thesis. Hence, throughout this manuscript, we directly or indirectly advocate the use of statistical models anchored in a Bayesian logic, whether

[5]In this rape case, DNA was the only incriminating evidence heard by the jury, as all the other evidence pointed towards innocence. After three trials, Adams was found guilty.

for reporting the weight of DNA evidence or in a more general objective of assisting reporting officers throughout the interpretation process.

In what follows, we give a description of some of the most confounding issues encountered in forensic DNA typing. These will further highlight the need for statistical tools for DNA evidence interpretation.

1.4 Issues with (forensic) DNA typing of STR loci

The analysis of STR loci is subject to anomalies that can affect the interpretation of DNA evidence. These anomalies can arise either from the stochastic fluctuations that are inherent to the analytical process, the type of the genetic markers employed, or the nature of the sample itself. Hereafter we briefly review two main issues: stochastic effects and DNA mixtures.

1.4.1 Stochastic effects related to the PCR process

The use of the polymerase chain reaction in DNA typing permits the analysis of minute amounts of DNA. However, when only a small starting number of molecules is available, stochastic fluctuations related to sampling effects are exacerbated. These effects lead either to the detection of spurious alleles unrelated to the analyzed profile or to the non-detection of alleles that are actually present in the sample.

Stutter products Stutters are artifactual PCR products specific to STR markers. They are small peaks that have the wrong number of repeats, usually one fewer repeat. They are caused by the slippage of the DNA polymerase during the PCR. Distinguishing a stutter peak from a real allele is not trivial, especially when the evidence is a mixture. However, the characteristics of stutters have been studied relative to the main peak height. Stutters are reported as percentages of the main (or parent) peak height, and for STR markers commonly used in DNA typing, this percentage is generally less than 15% (Whitaker et al., 2001; Frégeau et al., 2003, Figure 1.5).

Allele drop-in Allele drop-in refers to the contamination of evidentiary samples by a source unrelated to the investigated crime. The contaminant alleles, usually consisting of one or two additional alleles in the DNA profile, are DNA fragments that are pervasive in the atmosphere, for example, in plasticware. Allele drop-in must be distinguished from "gross contamination" from an individual profile unassociated with the crime scene profile, which is characterized by the presence of more than two spurious alleles (Gill et al., 2006).

Peak height imbalance and allele drop-out Allele peak heights are good indicators of the post-PCR DNA quantity. Thus, alleles from a given heterozygote are expected to have similar peak heights because the heterozygote contributes the same amount of DNA to each allele. Still, even in pristine DNA samples, it is rare to observe identical peak heights. The variability

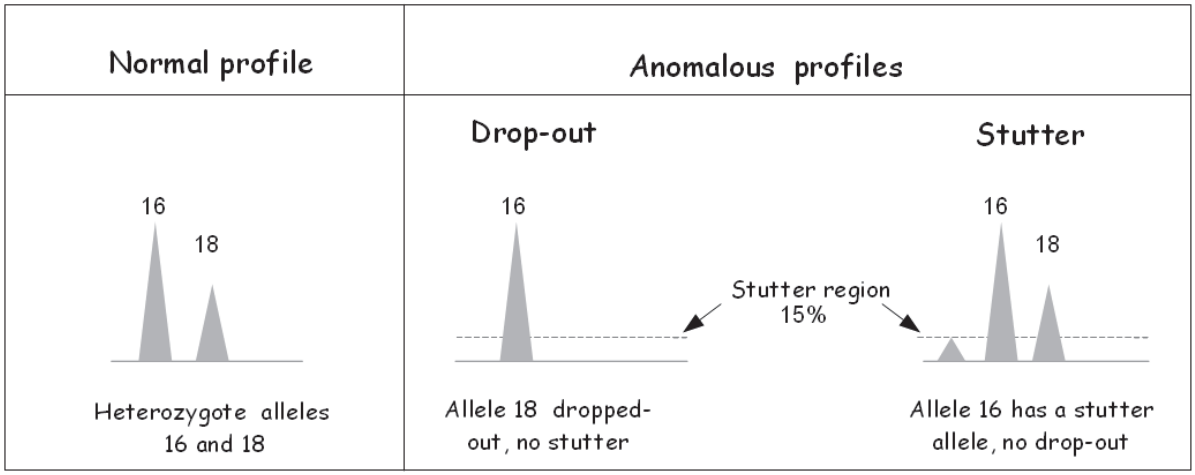


FIGURE 1.5: llustration of two phenomena that can complicate the interpretation of DNA profiles: allelic drop-out and stutters.

around this expectation is explained by the fact that primers do not necessarily bind with the same efficiency to each copy of the target alleles during the first PCR cycle. Hence, there is always a certain amount of imbalance (Gill et al., 2007). Of course, when starting from a small number of DNA molecules, the imbalance is exacerbated.

Allele drop-out can be seen as an extreme form of peak height imbalance (Buckleton et al., 2005); it results from the non-amplification of an allele during PCR (Gagneux et al., 1997). Drop-outs are generally deemed possible when two (or more) samples are compared; typically, an allele is observed at a given locus in the reference sample, but not in the crime scene profile.

1.4.2 DNA mixtures

DNA mixtures involving two or more contributors are often encountered in forensic casework. The high sensitivity of DNA typing systems based on STR multiplexes makes it possible to recover minute amounts of DNA from all types of contacted objects and surfaces. These samples often consist of a mixture of DNAs from several individuals.

Provided that the observed peaks are not artifactual, the presence of a mixture is plausible whenever a locus exhibits more than two alleles. However, allele counting alone is not a reliable indicator of the number of contributors because of allele sharing, i.e., contributors to the mixture may carry the same alleles at one or several loci (Figure 1.6).

Given the heterozygosity of the STR markers used in forensic DNA typing, the typed loci are expected to exhibit enough distinct alleles to permit the detection of a mixture. However, as the number of contributors increases, the number of distinct alleles decreases, and therefore, allele counting has limited efficiency with complex DNA mixtures. This may be problematic for as few as three contributors (Buckleton et al., 2007). The amount of allele sharing is accentuated when contributors are related or when they belong to the same subpopulation, in which the amount of co-ancestry is significant.

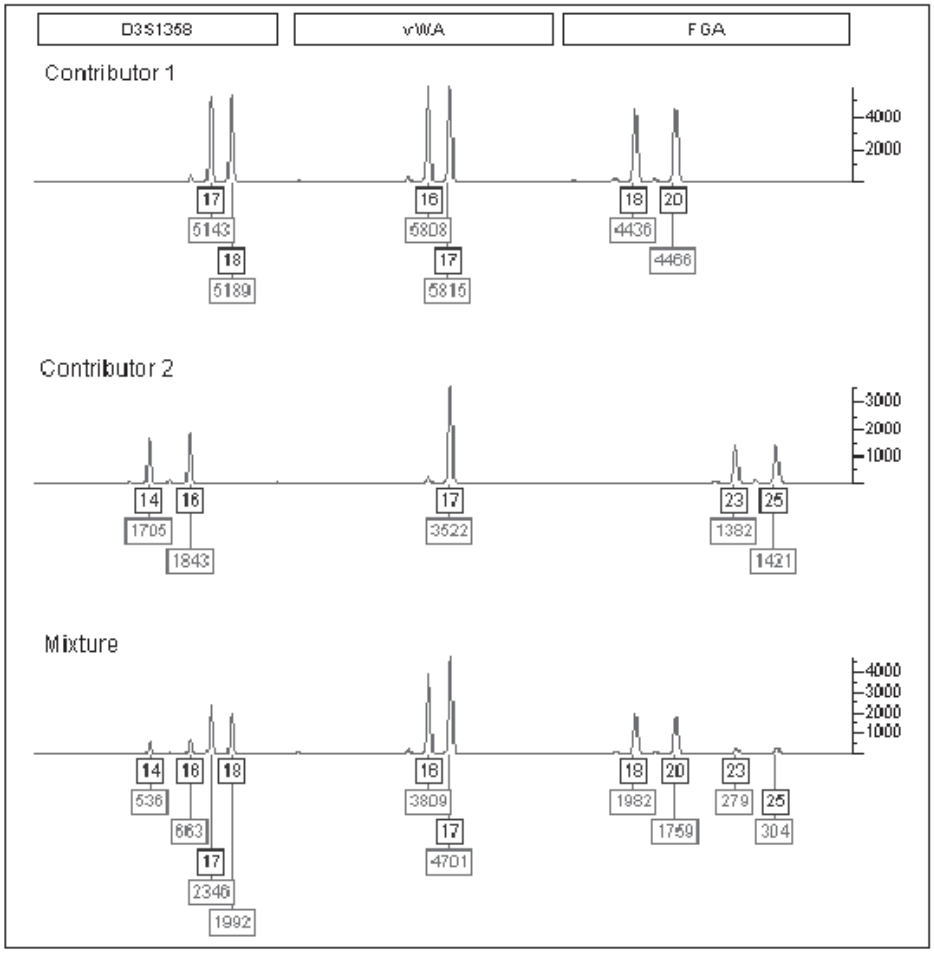


FIGURE 1.6: Electropherogram of a DNA mixture. Profiles of the contributing individuals (contributors 1 and 2) are also provided. The profiles are characterized with three STR loci "D3S1358", "vWA" and "FGA". Contributors 1 and 2 share one allele ("17") at locus "vWA", thereby only two alleles are present at this locus in the electropherogram of the mixture (adapted from Doom and Raymer, 2004).

In addition to allele sharing, the presence of anomalies in a DNA profile can further complicate its interpretation. For example, the presence of a contaminant allele can lead to an over-estimation of the number of contributors to the sample, or it may yield false genotypes for the potential contributors.

We previously explained that the issues raised by the interpretation of forensic DNA mixtures can be summarized in two questions: i) how many contributors are there, and ii) what are the genotypes of the contributors to the mixture?

Several methods have been proposed for the interpretation of forensic DNA mixtures. Some of them focus on putting a weight on the evidence when a suspect is available (Balding and Nichols, 1994; Weir et al., 1997; Curran et al., 1999; Fukshansky and Bär, 1999; Fung and Hu, 2000; Hu and Fung, 2005). More advanced statistical methods deal with mixture deconvolution, i.e., the enumeration of the possible genotypic combinations forming the mixture (Evett et al.,

1998; Clayton et al., 1998; Gill et al., 1998a,b; Perlin and Szabady, 2001; Mortera et al., 2003; Wang et al., 2006; Cowell et al., 2007a,b,a; Curran, 2008; Perlin and Sinelnikov, 2009).

These methods meet the needs that may emerge in different stages of DNA mixture interpretation. Methods dedicated to weighting evidence are typically used once the alleles present in a DNA profile have been called by a reporting officer, and the yielded results are recorded in a report, which is usually transmitted to the police or justice officials. Mixture deconvolution methods can be thought of as particularly helpful when no reference sample, such as a profile from a victim or a suspect, is available. The plausible genotypes can thus be used, for example, in a database search.

However, not all of these methods have been introduced into forensic casework. In particular, methods dedicated to mixture deconvolution are not currently being used. As we previously explained, there is a reluctance to statistical reasoning that may slow down the introduction of new methods, but another important factor is the lack of evaluation of the proposed methods because there are no tools or methods for evaluating these methods, it is not easy, or even possible for many forensic laboratories, to introduce these methods in their casework.

1.5 Thesis outline

The methodological aspects of the issue of method validation have been discussed within a context of a collaboration with the national forensic laboratory in Lyon (Institut National de Police Scientifique, INPS). As in many other forensic laboratories, reporting officers at the INPS are challenged daily by the complexity of DNA mixtures. Our interlocutors were, of course, aware of the multitude of methods that were proposed in the literature, but they were frustrated by not being able to implement them in their routine work because these methods have not been validated for casework.

Our contribution to this problem therefore revolves around the issue of method evaluation. Given the number and the diversity of existing methods for DNA mixture interpretation, it was not realistic to undertake a systematic review of existing tools for the interpretation of mixtures for the purpose of their validation, nor would the results be accurate. Instead, we focused on identifying key issues that affect various aspects of mixture interpretation and, in parallel, provide the necessary tools to enable method evaluation.

The second chapter of this thesis presents the early results of this work. Within the framework of a close collaboration with the national forensic laboratory in Lyon, we developed a method to estimate the number of contributors to a DNA mixture. With the aim of assessing and thus improving existing methods, we evaluated this method and compared its efficiency to the method currently used in most forensic laboratories, which is based on allele counting, and we demonstrate that our method constitutes an accurate alternative to the current practice.

In the third chapter, we focus on the interpretation of anomalous DNA profiles, focusing in particular on recent statistical developments in forensic literature related to the estimation of

allelic drop-out (Gill et al., 2009; Tvedebrink et al., 2009). While these methods appeared to be promising for the forensic community, we were concerned about whether they could be used in practical cases. This led us to propose a framework for evaluating these models.

Our evaluation approach naturally led us to develop software to automate the process. In the fourth chapter of this thesis, we introduce the forensim package for the R statistical program (R Development Core Team., 2006), which provides a set of tools dedicated to the evaluation of statistical methods involved in mixture interpretation. In conclusion, we present a review of our contribution to the interpretation of forensic DNA mixtures, and we detail critical points and perspectives that we intend to explore in the future.

Chapter 2

Determining the number of contributors to forensic DNA mixtures

Contents

2.1 General background

The increase in sensitivity offered by the PCR has led forensic laboratories to generate increasing numbers of profiles from "touch DNA" which is also referred to as DNA traces, collected from all kinds of touched or handled objects. The interpretation of these DNA samples, which are often generated from as little as a single diploid cell (Findlay et al., 1997), is rather problematic because the artifacts related to PCR are exacerbated in DNA samples with a low template amount. Another major issue with these samples is that they generally consist of mixtures. This is not surprising because these traces are collected from objects (e.g., a doorknob) that may have been touched or handled by several individuals.

Despite these difficulties, the great sensitivity of typing methods has greatly facilitated the exploitation of DNA evidence in criminal investigations. As a consequence, many investigators consider DNA the "gold standard" in terms of investigative tools, particularly when no background information, such as from an eyewitness or a victim's testimony, are available. For example, among the 14,000 DNA traces analyzed per year by the national forensic laboratory in Lyon, 90% of the samples consist of traces with no known contributors or identified suspects. Moreover, more than half of these samples are classified as mixtures of more than two contributors[1].

In this context, reporting officers of the forensic laboratory in Lyon are regularly asked by police investigators to provide indications of the possible number of contributors. Because this number can never be known with certainty, it has become a common laboratory practice to set the lower limit of this number to the minimum required to explain the observed profiles (Paoletti et al., 2005; Clayton and Buckleton, 2005). In practice, if a locus shows n distinct alleles, the (minimum) number of contributors is set to $\lceil \frac{n}{2} \rceil$ because each contributor carries at most two different alleles at each locus.

DNA evidence is usually analyzed using 10 or more STR loci, and thus, the locus showing the maximum number of contributors sets the limit. This method is often referred to as the maximum allele count method (Paoletti et al., 2005). However, setting a lower limit is different from attempting to estimate the most supported number of contributors from the data alone. Moreover, potential allele sharing between the mixture contributors limits the ability of allele counting to infer the number of contributors. The limitations of this method have left many practitioners frustrated. It is in this context that the methodological aspect of the question of the estimation of the number of contributors was initially discussed with our collaborators from the forensic laboratory.

Several methods relying on different statistical concepts were developed to facilitate mixture resolution (or deconvolution) into individual components (Evett et al., 1998; Clayton et al., 1998; Gill et al., 1998a,b; Perlin and Szabady, 2001; Mortera et al., 2003; Wang et al., 2006; Cowell et al., 2007a,b; Curran, 2008; Perlin and Sinelnikov, 2009). Additionally, many statistical

[1]Laurent Pène, personal communication.

developments anchored to a Bayesian framework were proposed for assigning the weight of DNA evidence consisting of mixed DNAs (Balding and Nichols, 1994; Weir et al., 1997; Curran et al., 1999; Fukshansky and Bär, 1999; Fung and Hu, 2000; Egeland et al., 2003; Hu and Fung, 2005) [2]. It is notable that, except for Egeland et al. (2003), none of these developments have considered inferring the number of contributors from the available data. This is not surprising because two-person mixtures are believed to account for the majority of DNA mixtures in casework (Torres et al., 2003), and in most cases, one of the contributors, for instance the victim, is known (Aitken and Taroni, 2004).

Given these observations, we were naturally concerned with current practices for determining the number of contributors, and we therefore evaluated the recent developments relevant to this topic. We briefly introduce two commonly used approaches for the interpretation of DNA mixtures. This will help seize the contribution of our first article and will lead us to discuss the issue of making use of quantitative data in the statistical methods dedicated to DNA mixtures.

2.2 DNA mixtures interpretation methods

The simplest method for mixtures resolution consists in excluding or including the suspect profile as a potential contributor to the mixture (Butler, 2001). However, most forensic laboratories rely on both qualitative and quantitative data for the interpretation of mixed stains. Hereafter we briefly introduce two approaches dedicated to mixture interpretation, which differ according to the type of data employed: qualitative data or quantitative data.

2.2.1 Qualitative assessment of DNA mixtures

The question of the number of contributors to DNA mixtures has primarily been addressed by taking into account the uncertainty about this number into the likelihood ratio calculations. This implies that the competing propositions under evaluation, namely the prosecution and the defence propositions, state who the known contributors to the stain were and how many unknown contributors were involved. Plus, the number of unknown contributors might differ under each hypothesis.

Weir et al. (1997) were the first to propose a general formula for the likelihood ratios (LR) allowing for an unknown number of contributors. Stating the number of unknown contributors to the stain, x, the set of alleles in the crime stain, E, and the alleles carried by the unknown contributors, U, the LR calculations are based on the ratio of two probabilities,

$$L = \frac{P(E|H_p)}{P(E|H_d)} = \frac{P_x(U|E, H_p)}{P_x(U|E, H_d)} \tag{2.1}$$

[2] These are not exhaustive lists of existing methods, but rather enumerations of the most significant developments.

where $P_x(U|E, H_p)$ is the probability that x unknown contributors carry the alleles in U, not carried by known contributors under the H_p hypothesis. Subscript x indicates the number of unknown contributors to the stain under the considered hypothesis. Note that x can differ between hypothesis H_d and H_p.

Weir et al. (1997) gave a generic formula for computing this probability:

$$P_x(U|E, H_d) = T_0^{2x} - \sum_j T_{1;j}^{2x} + \sum_{j,k} T_{2;j,k}^{2x} - \sum_{j,k,l} T_{3;j,k,l}^{2x} + ... \tag{2.2}$$

with:

T_0: the sum of frequencies of all alleles in E

$T_{1:j}$: sum of frequencies of all alleles in E except the jth allele in U

$T_{2;j,k}$: sum of frequencies of all alleles in E, except the jth and kth alleles in U, and so on

Curran et al. (1999) elegantly generalized this formula by allowing populations subdivision to be accounted for. This formula is now refereed to as the "general formula for likelihood ratios" (Buckleton et al., 2005).

This approach is usually refereed to as the "non restricted combinatorial approach" because all genotypic combinations of the unknown contributors are considered, without any restrictions (Gill et al., 2006). Significant improvements were proposed for this method, they allow several factors, such as relatedness between contributors or population substructure, to be accounted for (see for example Fung and Hu, 2000).

Even if the above liklelihood ratio formulation explicitly introduces the uncertainty about the number of contributors to the stain, it does not help decomposing the mixture into individual components. Hence, in situations where no known or suspected contributors are available, another strategy is needed. Currently, most forensic laboratories rely on the *Binary model* (Clayton and Buckleton, 2005) that make use of quantitative data, i.e., the alleles peak heights, to infer genotypes from two-person mixture. Note that deconvolution methods can also be used when reference samples are available.

2.2.2 Quantitative assessment of DNA mixtures

While the qualitative assessment of mixtures considers all possible genotypic combinations, the use of quantitative data allows removing the combinations that are not consistent with the data. Although interpretation guidelines vary between laboratories (Perlin, 2006), the rationale behind these practices relies on the Binary model, derived from the work of Evett et al. (1998). This model is a manual method dedicated to the deconvolution of two-person mixtures. The rationale behind it is that quantitative data, i.e., peak heights or areas, are directly related to the post-PCR amount of DNA contributed to the mixture by each individual. Since different contributors contribute different, or at least slightly different amounts of DNA, alleles from different contributors have different intensities, and hence, alleles from the same contributors will have

similar intensities. Thereby, starting from an estimate of the post-PCR mixture proportion, i.e., the relative amount of DNA contributed by each individual participating to the mixture (after the PCR is carried out), genotypic combinations are considered plausible if the corresponding mixture proportion is consistent with the initial estimate. Clayton et al. (1998) formalized the Binary model for the resolution of two-person mixtures, by recommending a series of five steps:

1. Identify the presence of a mixture: once artefactual peaks, resulting for example from stutters, are eliminated, a mixture can be identified by the presence of three or more peaks at one or several loci.

2. Assume the number of contributors: the identification of the number of contributors is based on the maximum number of distinct alleles observed across loci (maximum allele count method). This is straightforward for two-person mixtures. The circumstances of the crime scene can also give indications about the potential number of contributors.

3. Determine the mixture ratio: the mixture ratio, denoted M_x, gives the relative contribution of each individual to the analysed stain. Mixtures can range from equal proportions of each contributing genotype, to one component being greatly in excess (Butler, 2001). Using loci where there are no shared alleles between contributors, for instance, loci with four distinct peaks, it is possible to determine M_x as the ratio between peak heights belonging to each contributor.

4. Enumerate all possible mixture contributor combinations: once the mixture proportion is determined from a given locus, several combinations of genotypes are possible for the remaining loci. A given genotype combination is eliminated if it violates the expected peak profiles for an estimated mixture ratio of $\hat{M}_x$ (hat is for estimation).

5. Compare to reference sample: in order to avoid a suspect-driven approach, where only genotypic combinations matching the suspect's profile are selected, the above steps are taken without considering the available reference samples. The Binary model does not yield a best single genotype, but a list of genotypes that are well supported by the data. If the suspect genotype matches one of the plausible combinations, then it is plausible that he has contributed to the DNA mixture.

Given that currently used methods in mixture interpretation all involve some knowledge about the potential number of contributors, and given the limitations of the maximum allele count method, currently used in most of forensic laboratories, we strove to evaluate existing methods dedicated to estimating the number of contributors. To our knowledge, only one study addressed this question under the same perspective discussed here: Egeland et al. (2003) suggested a likelihood-based estimator of the number of contributors when genetic data alone is used, and irrespective of background information that may be available. As we shall explain in the following article, we modified this estimator to allow taking into account two cases relevant

to forensic casework: multiallelic loci and population substructure. This development was also an opportunity to evaluate the efficiency of this method against the method currently used in many forensic laboratories: the maximum allele count.

2.3 Article 1: "Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count?"

Accepted for publication in the *Journal of Forensic Sciences*, to appear in 2011.

Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count?

H. Haned[1,*], L. Pène[1], J.R. Lobry[1], A.B. Dufour[1], D. Pontier[1]

[a] *Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 69622 Villeurbanne, France*
[b] *Institut National de Police Scientifique, Laboratoire de Police Scientifique de Lyon, France*

Abstract

Determining the number of contributors to a forensic DNA mixture using maximum allele count is a common practice in many forensic laboratories. In this paper, we compare this method to a maximum likelihood estimator, previously proposed by Egeland et al. *Int J Legal Med* 2003;*117*(5):271-5, that we extend to the cases of multiallelic loci and population subdivision. We compared both methods' efficiency for identifying mixtures of two to five individuals in the case of uncertainty about the population allele frequencies and partial profiles. The proportion of correctly resolved mixtures was greater than 90% for both estimators for two and three-person mixtures, while likelihood maximization yielded success rates two to fifteen-fold higher for four- and five-person mixtures. Comparable results were obtained in the cases of uncertain allele frequencies and partial profiles. Our results support the use of the maximum likelihood estimator to report the number of

*Corresponding author
Email address: hinda.haned@univ-lyon1.fr (H. Haned)

Accepted for publication in Journal of Forensic Sciences *To appear (2011)*

contributors when dealing with complex DNA mixtures.

Keywords: Forensic Science, DNA typing, likelihood estimator, STR loci, DNA mixtures, population subdivision, allele count, partial profiles

1. Introduction

Interpretation of forensic DNA mixtures is a challenging task in forensic casework. Mixtures arise when more than one individual contribute to the DNA stain. This is common in cases of sexual assault where the source of DNA evidence can include the victim, the perpetrator(s) and the consensual partner(s) of the victim.

The interpretation of DNA evidence is even more challenging when competing hypotheses are weighted using likelihood ratios because it is implicitly assumed that the number of contributors is known. As misclassified DNA mixtures can lead to dramatic effects on the result of a police investigation, several attempts have been made to assess this problem. Weir (1), Brenner et al. (2), Buckleton et al. (3), and Lauritzen and Mortera (4) have all suggested bounds on likelihood ratios. None of these authors considered the matter of inferring the number of contributors from the data although this is a prevalent line of questioning in court.

It is common laboratory practice to set the lower bound on the number of contributors to the minimum required to explain the observed set of alleles. This bound is based on the maximum allele count throughout the analyzed loci, i.e., the locus showing the maximum number of alleles determines the bound. This method is believed to be an unreliable predictor because of the effect of allele sharing between contributors to the mixture known as

the masking effect (5, 6). Setting a lower bound is obviously different from attempting to estimate the most supported number of contributors from the data alone. Egeland et al (7) proposed to overcome this issue by making explicit use of the available allele frequencies of the target population. They suggested a likelihood-based estimator of the number of contributors using diallelic markers when conditions for Hardy-Weinberg equilibrium are met in the population. This method was shown to perform rather well for at least 200 diallelic markers and for mixtures of two and three contributors.

DNA stains from crime scenes are usually characterized through multiallelic Short Tandem Repeat (STR) loci, so there is a need to investigate which approach is the most efficient in determining the number of individuals involved in a mixture. Moreover, several studies have shown that longer DNA fragment lengths carry a greater probability of lost information from allelic drop out (8) leading the forensic expert to conclude that the DNA evidence has partial profiles.

In this paper, we aim to: i) extend the work of Egeland et al. (2003) to an arbitrary number of alleles per locus and to dependencies between alleles due to population subdivision and ii) investigate through simulations the performance of two methods for estimating the number of contributors to a DNA mixture from the genetic data alone and irrespective of background information that may affect this estimation: The maximum allele count and the maximum likelihood estimator.

We investigate the methods properties in three distinct situations: In the first situation, all contributors to the mixture belong to the same population with known allele frequencies; in the second situation, we take into account

the effect of not knowing with certainty the allele frequencies of the contributors' population, a situation that may arise from population subdivision, in the third situation, we seek to identify the effects of partial profiles on the estimation accuracy for both the maximum allele count and the likelihood-based estimators.

In order to facilitate reproducibility of our results and extension to other situations, our method is freely available in the package forensim for the R statistical software (9).

2. Methods

2.1. Extending the likelihood estimator to the cases of multiallelic loci and population subdivision

Let A be a specific locus with alleles $A_1, ..., A_k$ with frequencies $p_1, ..., p_k$ in a given population. Let m be the set of observed alleles in a DNA stain from a crime scene. We are interested in estimating the probability of observing m knowing that there are x individuals contributing to the mixture. This is the likelihood of the data m conditional on x, denoted: $L_A(x)$.

Example.

Suppose that a crime scene stain shows alleles A_1 and A_2 at locus A and the forensic expert wants to determine the likelihood that two contributors supply these alleles. Combining the observed alleles into two individual genotypes yields 7 distinct pairs of possible genotypes for the two contributors: (A_1A_1, A_2A_2), (A_2A_2, A_1A_1), (A_1A_1, A_1A_2), (A_2A_2, A_1A_2), (A_1A_2, A_1A_1), (A_1A_2, A_1A_2), and (A_1A_2, A_2A_2).

Under the hypothesis that the contributors to the DNA stain are not related, the estimation of each genotype proportion can be obtained as a product of the allele frequencies using the Hardy-Weinberg formula. This assumes the independence of alleles between and within individuals. This simplifying hypothesis as a means to determine the genotype proportions from allele frequencies is termed the "product rule" (10).

The probability of observing the pair of genotypes (A_1A_1, A_1A_2), denoted $Pr(A_1A_1, A_1A_2)$, corresponds to the probability of observing one homozygote for A_1 and one heterozygote A_1A_2, which is $p_1^2 2p_1p_2$. By adding the probabilities for each possible genotype pair, we finally obtain:

$L_A(x = 2) = 4p_1^3p_2 + 6p_1^2p_2^2 + 4p_1p_2^3$

These results could be derived analytically in a simple case (one locus and two hypothetic contributors), but the complexity of the likelihood computation increases dramatically with the numbers of loci and contributors; hence, there is a need for a general formulation of the likelihood function. In order to achieve this generalization, we follow the work of Curran et al. (11) who gave a general framework for interpreting DNA mixtures that can take population subdivision into account. In their paper, a general formula for mixture interpretation evaluation was given in the form: $Pr(E|H)$, where E is the DNA evidence and H is the hypothesis under which the data is being considered, for example, the prosecution hypothesis.

When only genetic data is considered, the evidence E is composed of the set of alleles observed in the mixture, denoted C. This set of alleles is composed of: 1) the set of alleles found in the typed individuals who are known to have contributed to the mixture, denoted T; 2) the set of alleles

found in the typed individuals known to be non-contributors to the mixture, denoted V; and 3) the set of alleles carried by the unknown contributors, denoted U. For instance, in the case of a DNA stain from a rape case, T is the set of alleles carried by the victim, her consensual partner(s), and potentially the suspect(s); V is the set of alleles carried by cleared suspects; and U is the set of alleles carried by the unknown contributors to the mixture.

The general formula of the likelihood can thus be derived from the particular case where all contributors to the mixture are unknown and there are no typed individuals. This corresponds to:
$T = V = \emptyset$ and $C = U$. Note that the equality $C = U$ does not correspond to the degenerate case evoked in (11) where unknown contributors can have any genotypes in C. In our case, the x unknown contributors genotypes must explain all alleles in C; thus, all possible genotypes attributable to the unknown individuals must explain the alleles present in the mixture, and they must all be taken into account in the likelihood calculation.

General formulation of the likelihood function.
Before giving the general formulation of the likelihood function we first specify the notations used in this paper, following Curran et al. (11):
x: The unknown number of contributors to the DNA mixture
c: The distinct number of alleles observed in the DNA stain
r: The number of unconstrained alleles, $r = 2x - c$
r_i: The unknown number of copies of allele A_i among the r unconstrained alleles of the stain
u_i: The unknown number of copies of allele A_i in the stain , with $\sum_{i=1}^{c} u_i = 2x$ and $u_i = r_i + 1$

θ: Wright's F_{ST} coefficient, which gives the probability of identity by descent of two alleles taken at random from a subpopulation, in two distinct individuals.

In our case, all contributors are unknown. Consequently, the DNA evidence, E, is only composed of the alleles present in the stain, C, and all other quantities defined in (11) and related to the typed individuals, whether or not they are known to have contributed to the mixture, are set to zero. The likelihood of having x individuals giving the alleles observed at a locus A in the case of all individuals belonging to the same subpopulation, is given by the general formula:

$$L_A(x) = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1-r_2-\ldots-r_{c-2}} \frac{(2x)!}{\prod_{i=1}^{c} u_i!} \frac{\prod_{i=1}^{c} \prod_{j=0}^{u_i-1} [(1-\theta)p_i + j\theta]}{\prod_{j=0}^{2x-1} [(1-\theta) + j\theta]} \quad (1)$$

Equation (1) takes into account the variation in the subpopulation allele frequencies. When there is no need to consider population subdivision, the likelihood of the data is simply obtained by setting θ to zero.

The likelihood estimator.

The maximum likelihood estimation of x, when a single marker A is considered, satisfies:

$$\max_{j=1,2,3,\ldots} L_A(x=j) \quad (2)$$

When multiple loci are considered simultaneously, the likelihood is calculated as the product of the likelihoods of each locus:

$$\max_{j=1,2,3,\ldots} \prod_{A} L_A(x = j) \quad (3)$$

The result in equation (3) is straightforward for the case of a homogeneous population, that is when $\theta = 0$ in equation (1). When there are allele dependencies in the general population due to subdivision, the overall loci likelihood (in the subpopulation) is still, to a close approximation, the product of the single locus probabilities, because the dependencies between alleles at different loci are corrected through θ (12).

In fact, the likelihood estimator defined by equations (2) and (3) extends the likelihood-based estimator derived by Egeland et al. (7) to the case of multiallelic loci and allows population subdivision to be taken into account through θ. Thus, the value for θ must be chosen according to the level of subdivision of the population. Typically, θ is chosen in the interval [0,0.03] when dealing with human populations (13).

Most forensic DNA mixtures consist of two-person mixtures (14); thus, for the estimator to be biologically meaningful, estimates were searched in the discrete interval [1,6]. This is a sensible upper limit for the number of contributors that can be analyzed in practice.

2.2. Evaluation of the methods' performance

Known allele frequencies case.

We used a published data set of allele frequencies in three US populations

(15): African Americans, Caucasians, and Hispanics. These populations were characterized by 15 STR loci, of which 13 correspond to the core CODIS loci.

Genotypes were simulated by drawing alleles independently at their relative frequencies from each population data base. Mixtures were then simulated by randomly drawing genotypes at each locus. The performances of the likelihood-based estimator and maximum allele count were compared on 1000 simulated mixtures comprising two to five contributors.

Uncertain allele frequencies case.

Generally, in the case of population subdivision, allele frequencies of the subpopulations are not known with certainty. This is due to the difficulty of defining the subpopulation of an individual (16). In this paper, we analyze the effect of uncertainty on allele frequencies by modeling the differences in allele frequencies between the global population and a subpopulation through a Dirichlet model. The term "subpopulation" means that the allele frequencies in the target population are not known with certainty and does not imply allele dependencies between and within and loci.

The allele frequencies for a given locus in a given subpopulation are generated as random deviates from a Dirichlet distribution (17-18). Each allele frequency is a random variable with a parameter $\alpha_i = p_i(1-\theta)/\theta$ where θ is the F_{ST} coefficient. Denoting $p_i^{'}$ the frequency of allele A_i in the subpopulation, the allele frequencies are modeled as:

$$(p_1^{'}, ..., p_k^{'},) \rightarrow Dirichlet(\alpha_1, ..., \alpha_k)$$

The global allele frequencies were taken from the African American pop-

ulation (15). We chose to set $\theta = 0.03$ in the variance parameter α_i. This value corresponds to the correction factor suggested by the National Research Council (19) for dealing with highly-subdivided human populations. Since we were only interested in studying the effect of uncertainty on the subpopulation allele frequencies, all loci were simulated independently within the subpopulation.

We compared the results of the maximum allele count to the likelihood-based estimator on 1000 simulated mixtures of two to five contributors. We investigated the differences between results when the uncorrected form of the likelihood-based estimator is used ($\theta = 0$) and compared them to the results obtained using the corrected form by setting $\theta = 0.03$.

Evaluation of the methods' robustness to partial profiles.
We analyzed the effect of successively removing loci while estimating the number of contributors on 1000 simulated mixtures of two to five individuals. The markers were successively removed according to their alleles expected median length (20). This corresponds to what happens in the case of a degraded DNA sample: Longer DNA fragments drop out first (8).

All programs used for the simulations were implemented in the *forensim* package for the R statistical software, available at `http://forensim.r-forge.r-project.org/`.

3. Results

3.1. Known allele frequencies case

The accuracy of estimations decreased with the number of contributors for both the maximum allele count and the maximum likelihood estimators

(Table 1). The probability of a correct estimation was always greater than 90% for mixtures of two or three individuals. Maximum allele count produced better estimates for three-person mixtures but the efficiency of this method decreased dramatically for complex mixtures of four or five individuals, while maximum likelihood gave a correct classification rate ranging from 64% to 79% in the three populations.

3.2. Uncertain allele frequencies case

The effect of uncertainty on allele frequencies was investigated for the case where the real allele frequencies deviate greatly from those used in the estimator ($F_{ST} = 0.03$, Table 2). Accurate estimates were obtained with the maximum allele count for mixtures with two or three contributors (success rate greater than 90%). The percentage of correctly identified stains was lower when dealing with four or five contributors. For instance, only 21% of five-person mixtures were correctly identified.

The corrected ($\theta = 0.03$) and uncorrected forms ($\theta = 0$) of the likelihood-based estimator produced similar results for mixtures of two or three individuals. The corrected form was more efficient in cases of a greater number of contributors: 60% of five-person mixtures were correctly identified, which was more than twofold the maximum allele count success rate.

3.3. Method robustness to partial profiles

The effects of partial profiles on the estimators' accuracy are shown in Figure 1. Only mixtures simulated from African American allele frequencies are shown here in the known allele frequencies case. Similar results were obtained for the other two populations (Caucasians and Hispanics) as well

as in the uncertain allele frequencies case for all three populations (results not shown). Consistent with previous results (Tables 1 & 2), the accuracy of both methods decreased with the number of contributors. The relative performance of both methods changed with the number of contributors in the mixture. The maximum allele count was revealed to be more efficient for mixtures of two or three persons, while the likelihood-based estimator performed better for mixtures of more than three individuals (see Figure 1). A 90% success rate was reached using the maximum allele count for a two-person mixture when exploring only two loci while five were needed for the maximum likelihood estimator. For three-person mixtures, the loci number increased to 10 and 14 respectively. For complex mixtures of four or five contributors, the success rates fell to 63% for the likelihood-based estimator and to 0.042% for the maximum allele count using all 15 loci.

Finally, to further our understanding of the above results, we looked at the characteristics of the profiles responsible for the biased estimations with the maximum likelihood estimator (Tables 1 & 2, Figure 1). We analyzed the sensitivity of the estimator to allele frequencies. An illustration of our results is shown in Figure 2 for a three-person mixture characterized by one locus. The maximum allele count can only give a lower bound to the real number of people involved in the mixture; thus, it cannot give overestimates. In contrast, maximizing the likelihood can lead to either underestimation or overestimation. Underestimation occurred when there are rare alleles in the mixture, while mixtures with frequent alleles also tended to be misclassified.

4. Discussion

We compared the efficiency of the commonly-used maximum allele count and an estimator based on likelihood maximization in inferring the number of contributors to forensic DNA mixtures.

Globally, maximizing the likelihood did not perform better than the maximum allele count for mixtures of two or three individuals. When all loci were documented and all mixture contributors belonged to the same population with known allele frequencies, the maximum allele count gave lower misclassification rates (varying from 1 to 3%) than the likelihood-based estimator (varying from 6 to 8%). These results corroborate previous findings for the former estimator (5).

Maximum allele count gives correct estimates for mixtures comprising x individuals when there are at least $2x - 1$ alleles at one of the considered loci in the stain. While this condition is often met in two- or three-person mixtures it is unlikely to find as much distinct alleles in mixtures of high order because of allele sharing (6). For instance, five-person mixtures are unlikely to show 9 distinct alleles at any of the considered loci, even if very polymorphic markers are used. Consequently, the maximum allele count method which tends to underestimate the real number of contributors in mixtures of high order ($x > 3$), still gives satisfactory results for two- and three-person mixtures. Maximum likelihood estimator can either over or underestimate the real number of contributors for all mixture types.

As expected, the uncertainty of estimations increased with the number of contributors for both methods while four- and five-person mixtures were more accurately identified by maximizing the likelihood. This is due to allele

sharing between contributors. As maximum allele count relies only on the number of distinct alleles, mixtures with greater numbers of contributors have greater amounts of allele sharing, which leads to the underestimation of the number of contributors.

Previous studies showed that using maximum allele count in the case of substantial allele sharing leads to biased estimates (5). The bias is likely to increase in cases of population subdivision. Here, we were more interested in one of the consequences of subdivision on the likelihood-based estimator, namely, the uncertainty on allele frequencies of the subpopulation, because the estimator explicitly makes use of the allele frequencies. In the case of uncertain allele frequencies we observed that the corrected form of our estimator performed better than the uncorrected one only for mixtures consisting of four or five contributors. Mixtures involving two or three individuals were more accurately classified with the uncorrected form of the estimator. The correction for subdivision was thus efficient in the uncertain allele frequencies case only for complex mixtures but this might not be the case in highly subdivided populations, where the independence of individual genotypes might not be realized.

In the case of partial profiles, both of the estimators showed a similar decrease in precision for two- and three-person mixtures, while the likelihood-based estimator was clearly more robust to partial profiles when dealing with four- and five-person mixtures. The lack of robustness of maximum allele count is explained by the fact that decreasing the number of loci decreases the chance of encountering in the mixture a locus that shows enough distinct alleles to allow a correct estimation using only the maximum allele count.

This effect is likely to be increased when dealing with complex mixtures of more than three contributors.

Overall, it is difficult to specify the minimum number of loci needed to accurately resolve a mixture because this number depends on the tolerated error rate which relies on the forensic experts experience; however, even with all 15 STR loci, five-person mixtures could not be resolved satisfactorily: The maximum allele count yielded an error rate of more than 95% while maximizing the likelihood misclassified more than 30% of the mixtures.

The bias in estimations is due in part to profiles with multiple masked alleles. This problem could be circumvented by using quantitative data given by the mixture profiles peak heights or areas (21). In fact, our estimator only takes into account qualitative information consisting of the allele types present in the stain. We assumed that the forensic expert had already determined the alleles present in the mixture and that there was no ambiguity during this stage of the evidence analysis. Further work could thus include the use of quantitative information to help in revealing masked alleles.

Most forensic laboratories use the maximum allele count method to specify the number of contributors to mixed stains. Complex mixtures comprising multiple masked alleles are likely to be misclassified by this method. This issue could have dramatic consequences especially when the number of contributors is determined solely on genetic data. This might be the case when dealing with DNA casework. Very often no suspect is available in such stains. Consequently, having an estimate of the number of contributors could help investigators when new elements emerge in the case. Therefore, it appeared to us that in case the number of contributors is determined on genetic data,

maximizing the likelihood should be preferred to maximum allele count especially when dealing with stains suspected to be mixtures of three or more individuals.

To conclude, we would like to point out that we do not recommend one method over the other. Our work is intended to provide insight to forensic practitioners on the differences in efficiency between the two estimators with respect to situations frequently encountered in forensic casework, namely uncertainty about the population allele frequencies and partial profiles. Our methodology is freely available in the package *forensim* for the R statistical software to allow investigations in contexts not explored here.

References

1. Weir BS. DNA statistics in the Simpson matter. Nat Genet 1995;11(4):365-8.
2. Brenner CH, Fimmers R, Baur MP. Likelihood ratios for mixed stains when the number of donors cannot be agreed. Int J Legal Med 1996;109(4):218-9.
3. Buckleton JS, Evett IW, Weir BS. Setting bounds for the likelihood ratio when multiple hypotheses are postulated. Sci Justice 1998;38(1):23-6.
4. Lauritzen SL, Mortera J. Bounding the number of contributors to mixed DNA stains. Forensic Sci Int 2002;130(2-3):125-6.
5. Paoletti DR, Doom TE, Krane CM, Raymer ML, Krane DE. Empirical analysis of the STR profiles resulting from conceptual mixtures. J Forensic Sci 2005;50(6):1361-6.
6. Buckleton JS, Curran JM, Gill P. Towards understanding the effect of uncertainty in the number of contributors to DNA stains. Forensic Sci Int Genet 2007;1(1):20-8.

7. Egeland T, Dalen I, Mostad PF. Estimating the number of contributors to a DNA profile. Int J Legal Med 2003;117(5): 271-5.
8. Butler JM. Forensic DNA typing, Elsevier Academic Press 2001.
9. R Development Core Team 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna: Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
10. Evett IW, Weir BS. Interpreting DNA evidence. Sinauer, Sunderland MA 1998.
11. Curran JM, Triggs CM, Buckleton JS, Weir BS. Interpreting DNA mixtures in structured populations. J Forensic Sci 1999;44(5):987-95.
12. Buckleton JS, Triggs CM, Walsh SJ. Forensic DNA evidence interpretation, CRC Press 2005.
13. Curran J, Buckleton JS, Triggs CM. What is the magnitude of the subpopulation effect? Forensic Sci Int 2003;135(1):1-8.
14. Torres Y, Flores I, Prieto V, Lpez-Soto M, Farfn MJ, Carracedo A, Sanz P. DNA mixtures in forensic casework: a 4-year retrospective study. Forensis Sci Int 2003;134(2-3):180-6.
15. Butler JM, Schoske R, Vallone PM, Redman JW, Kline MC. Allele frequencies for 15 Autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. J Forensic Sci 2003;48(4):908-11.
16. Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 1994;64(2-3):125-40.
17. Nicholson G, Smith AV, Jnsson F, Gstafsson O, Stefnsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. J R Stat Soc B 2002;64:69515.

18. Marchini JL, Cardon LR. Discussion of Nicholson et al. J R Stat Soc B 2002;64:7401.

19. NRC II - National Research Council Committee on DNA Forensic Science, The Evaluation of forensic DNA evidence. Washington, DC: National Academy Press 1996.

20. Applied Biosystems (2001) AmpFlSTR Identifiler PCR Amplification Kit Users Manual, Foster City, CA, P/N 4323291.

21. Evett IW, Gill PD, Lambert JA. Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci 1998;43(1):629.

Tables & Figures

x	Maximum allele count (%)	Likelihood estimator (%)
African Americans		
2	100	100
3	99	94
4	45	79
5	5	67
Caucasians		
2	100	99
3	97	92
4	34	77
5	2	64
Hispanics		
2	100	100
3	98	93
4	45	79
5	2	67

Table 1: Percentages of correctly identified mixtures for all three studied populations. The first column gives the true number of contributors, x. The second and third columns give the percentages of mixtures correctly identified by the two methods: The maximum allele count and the maximum likelihood estimator.

		Likelihood estimator (%)	
x	Maximum allele count (%)	Uncorrected form	Corrected form
2	100	99	99
3	94	95	91
4	21	56	76
5	0.7	27	60

Table 2: Percentages of correctly identified mixtures in the uncertain allele frequencies case. The first column gives the true number of contributors, x. The next two columns give the percentages of accurate estimation for the maximum allele count and the maximum likelihood methods. For the latter, two estimates are displayed corresponding to the form used in the estimator: The uncorrected form ($\theta = 0$) and the corrected form ($\theta = 0.03$).

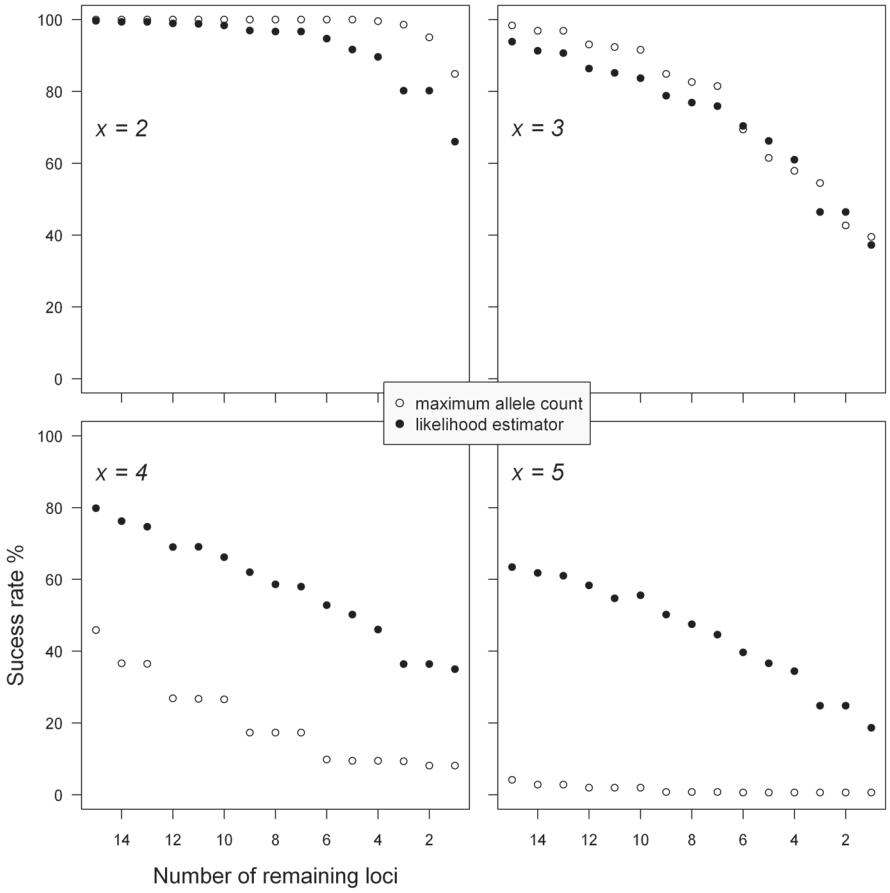


Figure 1: Percentages of correctly identified mixtures for x contributors, where x ranges from 2 to 5 in the case of partial profiles, for the maximum allele count and the maximum likelihood methods.

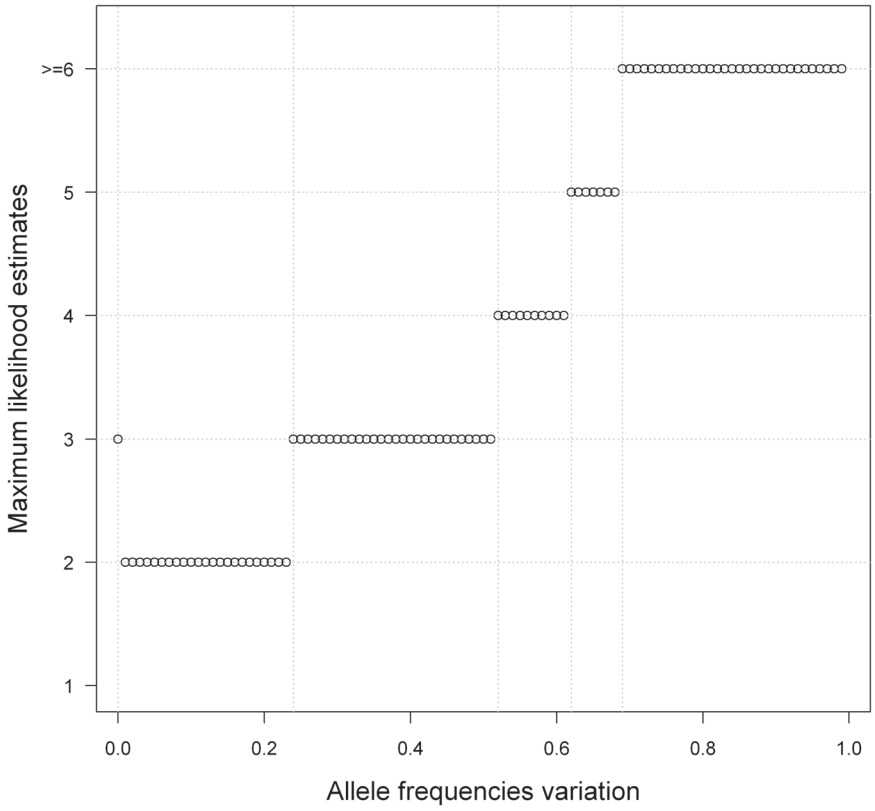


Figure 2: Sensitivity of the maximum likelihood estimations of the number of contributors to variations in allele frequencies for a simulated three-person mixture. A single locus, "vWA", was considered. At this locus, the mixture included alleles "16," "17, " "18,", and "19," with initial allele frequencies taken as 0.25, 0.24, 0.15, and 0.06 from the African American population. We varied the frequency of the less frequent allele "19" from 0 to 1 (x-axis), values of the three other alleles being also varied by keeping their relative frequencies constant. Each point on the plot represents the estimation yielded by the maximum likelihood estimator (y-axis). Correct estimates are obtained with the original allele frequencies (origin of the x-axis), and when the frequency of allele "19" varies between 0.24 and 0.52. Underestimation of the number of contributors occurs when frequency of allele "19" is under 0.24, while overestimations occur when its frequency is greater than 0.52.

2.4 Evaluating the maximum likelihood estimator efficiency

Once we developed a methodological framework to answer the question before us, still being rooted into an evaluation approach, we naturally raised the question of the use of our estimator in practical cases. We completed the work presented in the first article by providing a method to quantify the effectiveness of the estimator. Article 2 below describes this method and illustrates its use in practical cases.

Article 2: "The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture"

Article in press in *Forensic Science International: Genetics*, 2010.

G Model
FSIGEN-598; No. of Pages 4
ARTICLE IN PRESS

Forensic Science International: Genetics xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Original research paper

The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture

H. Haned [a,*], L. Pène [b], F. Sauvage [a], D. Pontier [a]

[a] *Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de biométrie et biologie évolutive, 69622 Villeurbanne, France*
[b] *Institut National de Police Scientifique, Laboratoire de Police Scientifique de Lyon, France*

ARTICLE INFO

Article history:
Received 13 November 2009
Received in revised form 22 February 2010
Accepted 21 April 2010
Available online xxx

Keywords:
DNA mixtures
Likelihood estimator
Traces
Body fluids
Predictive value
Bayes' theorem

ABSTRACT

We propose to quantify the accuracy of a likelihood-based estimator that was recently proposed for the determination of the number of contributors to a DNA mixture, when genetic data alone is considered [H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci., in press]. Using Bayes' theorem, we derive a formula for the calculation of the predictive value (PV) of the likelihood-based estimator. The PV gives the probability that a DNA stain contains the DNAs of *i* people given that the maximum likelihood estimator gave an estimate of *i* contributors for this stain. We illustrate the PV calculations for two different types of DNA evidence: traces and body fluids.

The PV varied according to the number of contributors involved in the DNA stain. Setting the maximum number of possible contributors to five, the lowest predictive values were scored for five-person mixtures with a minimum value of 0.26 for traces, but values were always above 0.94 for stains comprising one, two or three contributors, for both traces and body fluids. Values remained relatively high for four-person mixtures with a minimum value of 0.69. These findings confirm that likelihood-maximization is a powerful approach for the determination of the number of contributors to forensic DNA mixtures.

1. Introduction

As the sensitivity of typing methods is constantly increasing, forensic experts deal with more and more complex cases of evidence containing the DNA of several individuals. Though numerous statistical methods exist to calculate the strength of DNA evidence, the most challenging step in the interpretation of such mixed stains is still the determination of the number of contributors involved [1]. Usually, the circumstances of the investigated crime combined with genetic and non genetic evidence can produce good grounds to the determination of this number. But the task is seriously complicated when scarce data is available about the origin of the stain. This is common in DNA casework where often no suspect or known contributors are available. A common laboratory practice consists on bounding the number of contributors to the minimum required to explain the observed DNA profiles without making any use of the available data except for the number of alleles per locus [2]. Recently, an alternative approach based on the maximum likelihood principle was proposed to overcome this issue [3]. Using qualitative information on which alleles are present in the mixture, this maximum likelihood estimator searches the number of contributors maximizing the likelihood of the observed DNA profiles. Using computer-simulated DNA mixtures, the authors of this study showed that maximizing the likelihood of the data to find the most likely number of contributors gives more accurate estimates than using a lower bound when dealing with mixtures of more than three contributors. However; before considering the use of this estimator in practical cases, it is important to have at disposal a method to quantify the level of confidence that can be given to the yielded results.

In this paper, we propose to globally quantify the accuracy of the maximum likelihood estimator. Relying on Bayes' theorem, we derive a formula for the calculation of the predictive value (PV) of the estimator. The PV aims to give a global appreciation of the confidence that can be given to the estimates meanwhile taking into account prior information about the occurrences of mixed DNA stains in forensic casework. We explain the method and illustrate its potential use in forensic studies.

2. Methods

2.1. Theoretical background

The maximum likelihood estimator takes into account genetic data, namely, the frequencies of the alleles present at each locus

* Corresponding author at.
E-mail address: haned@biomserv.univ-lyon1.fr (H. Haned).

1872-4973/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2010.04.005

G Model
FSIGEN-598; No. of Pages 4
ARTICLE IN PRESS

characterizing the analyzed DNA stain, and searches the number of contributors that maximizes the likelihood of the observed profiles [3]. We define the predictive value of this estimator as the probability of having i contributor(s) to the tested DNA stain, knowing that the likelihood estimator gave an estimate of i contributor(s) for this stain. The PV is data-independent, which means that the observed data, namely the DNA profiles in the stain, are not involved in the calculations. The PV can thus be assimilated to a precision rate of the estimator, specific to each mixture type.

2.2. Formulation of the predictive value of the likelihood estimator

Denoting x the true number of contributors to the mixture and $\hat{x}$ its estimation, the predictive value of the estimator can be written as the conditional probability: $Pr(x = i|\hat{x} = i)$. A simple way to estimate this unknown probability is to rewrite it using its inverse, which is: $Pr(\hat{x} = i|x = i)$. The transformation is simply done using Bayes' formula:

$$Pr(x = i|\hat{x} = i) = \frac{Pr(\hat{x} = i|x = i)Pr(x = i)}{Pr(\hat{x} = i)} \quad (1)$$

The term $Pr(\hat{x} = i|x = i)$ is the probability that the estimator classifies the considered stain as a mixture of i contributor(s), given that there are actually i contributor(s). Haned et al. [3] used a simulation procedure to estimate these conditional probabilities: a thousand mixture comprising two to five contributors were simulated by combining alleles at random, with respect to their allele frequencies. The efficiency of the estimator was estimated as the proportion of correctly identified mixtures. Here, we follow a similar procedure: We simulated 1000 DNA stains containing one to five individuals, using the US African American allele frequencies published in [4]. The conditional probabilities of success of the estimator were then estimated for each simulated number of contributors.

Hereafter, we will refer to the probability $Pr(x = i)$ as the prior probability of encountering a mixture of i contributors. $Pr(\hat{x} = i)$ is the probability of the estimator giving i as an estimate for the number of contributors to the stain, regardless of the concerned mixture type. Using the law of total probabilities we rewrite probability $Pr(\hat{x} = i)$ to a product of conditional and prior probabilities as follows:

$$Pr(x = i|\hat{x} = i) = \frac{Pr(\hat{x} = i|x = i)Pr(x = i)}{\sum_{k=1}^{K} Pr(\hat{x} = i|x = k)Pr(x = k)} \quad (2)$$

$Pr(\hat{x} = i|x = k)$ is the probability that the estimator classifies the considered stain as a mixture of i contributor(s) knowing that there are actually k contributor(s), where k can be equal or differ from i. Values of k range from 1 to K, where K is a biological meaningful threshold for the number of contributors. For illustrative purpose, we set K to 5 and search the maximum likelihood estimates in the discrete interval [1,6]. As we later discuss, this threshold can be extended to $K > 5$.

2.3. Constructing the prior distribution of mixed DNA stains

Thanks to Eq. (2), the only term we have to determine now is the prior probability $Pr(x = i)$. In order to construct this prior distribution we used a survey of the crime scene profiles analyzed at the Institut National de Police Scientifique (INPS), the national forensic laboratory in Lyon, France (data communicated by Laurent Pène). For the year 2008, 8479 crime scene profiles were analyzed at the INPS using the Applied Biosystems AmpFlSTR® Identifiler™ kit [5]. These samples were either classified as traces when they came from contact traces, for instance epithelial cells on a given object or tool, or as body fluids when samples came from biological fluids, namely, blood, saliva and semen. The number of individuals involved in the stain was also indicated. Samples comprising one contributor were classified as "single-source" stains, samples comprising two contributors were classified as "resolvable mixtures" and stains comprising more than two contributors were classified as "unresolvable mixtures". This restricted classification is explained by the difficulty of determining the real number of individuals involved [6].

Two-person mixtures are believed to account for the majority of mixtures encountered in casework [7]. Three-, four- and five-person mixtures are believed to be rarer. But, as a consequence of the restricted classification, very scarce data is available in the literature about the occurrence of these complex mixtures in forensic casework. The construction of a prior distribution of mixtures occurrences in forensic casework was thus necessary for mixtures comprising more than two contributors.

The prior probabilities for stains comprising one or two contributors were set using the available data (survey of the INPS casework for year 2008). We chose to set the remaining probabilities for mixtures comprising more than two contributors using experts' prior beliefs. We asked three experienced forensic experts at the INPS to set the proportions of mixed stains comprising three, four or five contributors. We focused on two key issues in setting up this prior distribution:

(i) the probability of encountering a mixture with i contributors must decrease as i increases,
(ii) the probability of encountering a complex mixture with more than two contributors must be greater in case of traces than in case of body fluids. We justify this by the difficulty in distinguishing single-sources contributors in case of traces [8].

These requirements are meant to help the forensic experts to set the prior distribution but they are not compulsory to the method, and they can of course be modified or dropped.

3. Results and discussion

3.1. Crime scene profiles survey

Among the 8479 casework profiles stains, 5169 were body fluids and 3310 were traces. The majority of stains, 71%, comprised one contributor and was classified as "one contributor stains". Among the remaining 29% stains, 6% were resolvable mixtures classified as two-person mixtures and 23% were classified as unresolvable mixtures. There were more mixed DNA stains among traces than among body fluids (Table 1). This finding agrees with our predictions and can be explained by the fact that in case of body fluids, the major contributor drowns the signal of other contributors to the mixture, whereas in case of traces, the low quantities of DNA contributed by each individual prevent from detecting single-source DNA contributors.

3.2. Predictive value of the likelihood estimator

The conditional probabilities of success were estimated from simulated data (Table 2). We obtained similar results to those of

Table 1
Percentages of crime scene profiles comprising one, two or more than two individuals.

	$x = 1$	$x = 2$	$x > 2$	
Traces	45%	4%	51%	$N = 3310$
Body fluids	87%	7%	6%	$N = 5169$

Please cite this article in press as: H. Haned, et al., The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.04.005

G Model
FSIGEN-598; No. of Pages 4
ARTICLE IN PRESS
H. Haned et al. / Forensic Science International: Genetics xxx (2010) xxx–xxx 3

Table 2
Estimates of the conditional probabilities $Pr(\hat{x} = i|x = k)$. The table is read vertically. For example, the probability of having an estimate of 5, knowing that there are actually 4 people in the DNA stain is 0.127.

	$\hat{x}=1$	$\hat{x}=2$	$\hat{x}=3$	$\hat{x}=4$	$\hat{x}=5$	$\hat{x}=6$
$x=1$	1	0	0	0	0.00	0
$x=2$	0	0.998	0.002	0	0.00	0
$x=3$	0	0.005	0.937	0.058	0.00	0
$x=4$	0	0	0.067	0.805	0.127	0.001
$x=5$	0	0	0	0.131	0.662	0.207

Haned et al. [3]. Different prior values were chosen for traces and body fluids (Table 3).

The predictive values varied according to the prior probabilities used. Where non null priors are used, the predictive values were relatively high, for both traces and body fluids, as values ranged from 0.69 to 1 for stains containing one, two, three or four contributors. The lowest values were scored for five-person mixtures (0.26 for traces). When similar priors are used, the PV slightly differed; in this case, it appeared that the distinction between the types of DNA stains under analysis is not necessary.

The priors used in this study are not arbitrary as they are defined by experts' prior belief. The use of such priors in likelihood ratios is controversial as discussed in Buckleton et al. [9], but in this study, the focus is on methods evaluation and these priors are not related to the prior knowledge about the number of contributors before the DNA evidence is analyzed.

We set the threshold for the number of contributors to five (Tables 3 and 4) which led to searching the maximum likelihood estimates in the discrete interval [1,6]. We believe that this is a biologically meaningful threshold for searching the most plausible number of contributors. However, this threshold can be extended, depending on the crime scene context and the type of evidence being analyzed. For instance, traces are likely to contain more contributors than stains from body fluids. Once the prior distributions of the mixed stains set, the results are straightforward.

Table 3
Prior distribution probabilities, for traces and body fluids, set by three forensic DNA experts: Expert 1, Expert 2 and Expert 3. Values for $x = 1$ and $x = 2$ were set using the data survey shown Table 1. Values for $x = 3, \ldots, 5$ were given by the interviewed forensic experts.

	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$
Expert 1					
Traces	0.45	0.04	0.30	0.15	0.06
Body fluids	0.87	0.07	0.04	0.01	0.01
Expert 2					
Traces	0.45	0.04	0.35	0.15	0.01
Body fluids	0.87	0.07	0.05	0.01	0
Expert 3					
Traces	0.45	0.04	0.25	0.20	0.06
Body fluids	0.87	0.07	0.05	0.01	0

Table 4
Predictive values of the maximum likelihood estimator according to the prior distributions defined by Experts 1–3 and shown Table 3. Predictive values are given for traces and body fluids, according to the number of individuals contributing to the stain ($x = 1, \ldots, 5$).

	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$
Expert 1					
Traces	1	0.96	0.96	0.83	0.67
Body fluids	1	0.99	0.98	0.69	0.84
Expert 2					
Traces	1	0.96	0.97	0.85	0.26
Body fluids	1	0.99	0.98	0.73	0
Expert 3					
Traces	1	0.96	0.94	0.88	0.61
Body fluids	1	0.99	0.98	0.73	0

4. Conclusion

In this paper, we propose the predictive value to be considered as a global measure of the likelihood-based estimator efficiency. It is notable that the PV is not meant to be a measure of the uncertainty related to the estimates.

The values presented in this study depend on the simulated data and the priors we defined. These can be adapted with respect to the context where the DNA evidence is analyzed. PV calculations using priors different from those we propose here can be carried out using the R package *forensim*, available from http://forensim.r-forge.r-project.org/.

The maximum likelihood estimator of the number of contributors to forensic DNA mixtures can be powerful in critical cases, for instance when dealing with DNA casework. Very often in such cases, scarce data is available about the origin of the stain and only genetic data are available. These data consist of qualitative information about which alleles are present in the stain and quantitative information about the alleles' peak heights and areas. The maximum likelihood estimator only considers qualitative data. Quantitative information might not always help to separate the DNA profiles into individual components. Moreover, there is no consensus in the literature about how peak heights or areas should be taken into account, and the developments in the literature dealing with quantitative data [10–15] have not encountered the expected success in the forensic community.

The fact that genetic data support a certain number of contributors to the evidentiary stain can be of significant help for the investigators, before any suspect or comparison between profiles can be processed. When no other information is available, this estimate can guide investigators in their search for potential suspects. To conclude, even if the maximum likelihood approach might seem too complex for presentation in court, it must not be neglected as a valuable tool to determine the number of contributors to DNA stains and forensic experts should be aware that an alternative method to maximum allele count exists.

Acknowledgments

We thank two referees for a thorough review and constructive comments. We are grateful to Anne Viallefont and David Fouchet for their helpful comments.

References

[1] T. Clayton, J. Buckleton, Mixtures, in: J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, 2005, pp. 217–274.
[2] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (2005) 1361–1366.
[3] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci., 2011, in press.
[4] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline, Allele frequencies for 15 Autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations, J. Forensic Sci. 8 (2003) 908–911.
[5] Applied Biosystems (2001) AmpFlSTR® Identifiler™ PCR Amplification Kit User's Manual, Foster City, CA, P/N 4323291.
[6] B. Budowle, J. Onorato, T.F. Callaghan, A.D. Manna, A.M. Gross, R.A. Guerrieri, et al., Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J. Forensic Sci. 54 (2009) 810–821.

Please cite this article in press as: H. Haned, et al., The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.04.005

G Model
FSIGEN-598; No. of Pages 4
ARTICLE IN PRESS

[7] Y. Torres, I. Flores, V. Prieto, M. López-Soto, M.J. Farfán, A. Carracedo, P. Sanz, DNA mixtures in forensic casework: a 4-year retrospective study, Forensic Sci. Int. 134 (2003) 180–186.
[8] J. Buckleton, P. Gill, Low Copy Number, in: J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, 2005, pp. 275–297.
[9] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int. Genet. 1 (2007) 20–28.
[10] P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, J. Buckleton, Interpreting simple STR mixtures using allele peak areas, Forensic Sci. Int. 91 (1998) 41–53.
[11] I.W. Evett, P. Gill, J. Lambert, Taking account of peak areas when interpreting mixed DNA profiles, J. Forensic Sci. 43 (1998) 62–69.
[12] M. Perlin, B. Szabady, Linear mixture analysis: a mathematical approach to resolving mixed DNA samples, J. Forensic Sci. 46 (2001) 1372–1378.
[13] T. Clayton, J. Buckleton, Mixtures, in: J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, 2005, pp. 217–274.
[14] T. Wang, N. Xue, J. Douglas Birdwell, Least-square deconvolution: a framework for interpreting short tandem repeat mixtures, J. Forensic Sci. 51 (2006) 1284–1297.
[15] R. Cowell, S. Lauritzen, J. Mortera, Identification and separation of DNA mixtures using peak area information, Forensic Sci. Int. 166 (2007) 28–34.

Please cite this article in press as: H. Haned, et al., The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.04.005

2.5 Discussion

The work presented here addresses the problem of the estimation of the number of contributors to DNA mixtures, but there is at least one unresolved issue: the use of quantitative data. Indeed, the estimator we presented in Article 1 only considered qualitative data, which leads to considering all possible genotypic combinations inferred from a DNA evidence profile. It has been suggested that the use of allele peak heights could help reduce the number of plausible genotypes by removing those that are not supported by this information (see for example, Evett et al., 1998). Ideally, we would be able to assign a weight to each possible genotype based on both the allele frequencies and the allele peak heights. Unfortunately, in practice, the incorporation of quantitative information is not trivial because there has been little research on peak height distributions.

The rationale behind the use of quantitative data in mixture interpretation resides in the fact that the amount of fluorescence for a given allele is related to the (post-PCR) number of copies of that allele present in the DNA sample analyzed. Allele peak heights are thus expected to reflect the amount of DNA contributed by each individual participating in the mixture.

Following this rationale, the majority of methods proposed for mixture deconvolution, including the Binary model presented earlier, proceed through two major steps: first, the peak heights are used to yield an estimate of the mixture proportions, and second, the genotypes that are inconsistent under the estimate of the mixture proportions are eliminated. Depending on the statistical approach employed, these methods select either the most likely set of genotypes (Clayton et al., 1998; Gill et al., 1998b, 2006) or a single best genotype (Perlin and Szabady, 2001; Wang et al., 2006).

These developments can be seen as a further formalization of the Binary model presented above. As a consequence, their generalization to higher-order mixtures (more than two contributors) is not straightforward. For complex mixtures, another category of methods employing a probabilistic approach seems promising (Mortera et al., 2003; Cowell et al., 2007a,b). In this approach, each of the possible genotypic combinations is assigned a weight that is calculated from a continuous probabilistic distribution, for which the parameters depend on the available qualitative (allele frequencies in the target population) and quantitative data. In this continuous approach, the genotypic combinations are weighted according to how well the observed peak heights agree with the expected intensities under the proposed genotype. This naturally implies that some information, namely the probability distribution functions of the peak heights conditional on the mixture proportions, is available.

Notable attempts to model peak height distributions include the use of a Gaussian model (Evett et al., 1998) and a Gamma model (Cowell et al., 2007a,b). However, none of these methods is currently in use in any forensic laboratory. The foremost reason for this is the complexity of the statistical approaches employed and the lack of their accessibility in the form of (free or proprietary) software. This has limited their use in forensic casework.

The second reason explaining the difficulty of employing quantitative data and, for instance, continuous models, is the lack of (published) data supporting the distributional assumptions for the allele peak heights on an experimental basis. More than a decade ago, Evett et al. (1998) proposed to investigate the distributions of peak heights conditional on the mixture proportion. However, little work has been published on the validation of the continuous probabilistic models, and to our knowledge, only a single recent publication reports an attempt to validate the Gamma model for STR peak heights (Cowell, 2009). Therefore, despite of the developed methods, the incorporation of allele peak heights in DNA mixture interpretation remains to be explored in further detail. The methodological framework proposed here, coupled with an evaluation approach, is an encouraging basis for future methodological developments incorporating quantitative data to determine the number of contributors to DNA mixtures.

Chapter 3

Analysis of low-template DNA samples

Contents

3.1 Introduction

A number of issues can be raised with respect to the analysis of STR loci that might seriously complicate the interpretation of DNA profiles. We discussed some of these in the introductory chapter, and they mainly consist of i) the presence of spurious alleles unassociated with the crime scene profile and ii) the non-detection of alleles originally present in the analyzed stain (allele drop-out). These phenomena occur in all DNA samples, regardless of the quantity or the quality of the DNA sample. However, as in any other random process, these stochastic effects are exaggerated in samples with low levels of template DNA molecules.

The increased sensitivity of PCR-based analysis of STR loci paved the way for studies on DNA typing from minute amounts of DNA, for example, DNA recovered from touched objects, called touch DNA (van Oorschot and Jones, 1997). The analysis of samples originating from as little as a single cell has thus become possible (Findlay et al., 1997).

This increase in sensitivity is generally achieved through modifications of the polymerase chain reaction (PCR), for example, by increasing the number of cycles, or through post-PCR manipulations (Strom et al., 1991; Taberlet et al., 1996; Strom and Rechitsky, 1998; Gill et al., 2000). The technique associated with raising the number of PCR cycles (generally from 28 to 34 cycles) has become known as low copy number (LCN) analysis. It was the Forensic Science Service of the UK (FSS) who pioneered the application of LCN typing to casework in 1999 for samples with low DNA levels, thus widening the scope of application of forensic DNA typing to touch DNA (Gill et al., 2000).

However, increasing the sensitivity of PCR inevitably leads to increased detection of anomalies associated with the analysis of STR loci. As a consequence, standard methods for interpretation cannot be applied to samples with low DNA levels. Still, strict guidelines regarding the collection and processing of samples and the interpretation of results were published closely following the first applications of LCN DNA typing by the FSS (Gill et al., 2000).

Anomalies associated with the typing of minute amounts of DNA samples are not specific to forensic identification, as they are also encountered, for example, in ecology (Pompanon et al., 2005). However, it is in forensic applications that they have raised the most concerns, mainly from a legal perspective, though not from the scientific community. Indeed, apart from the recent exchanges in the forensic literature that we will build on later, there has been no *debate* concerning LCN typing methods, at least nothing comparable to the heated exchanges between scientists at the time of the "DNA wars". However, as some authors refer to these discussions as a debate (Budowle et al., 2010), we will use the same terminology to describe the elements of this discussion.

3.1.1 The low copy number debate

Although LCN techniques have allowed the analysis of DNA stains that weren't exploitable before –nearly 21000 criminal cases were analysed by the FSS alone since 1999 (Gilbert, 2010)–

the challenges accompanying the technique raised concerns about the validity and reliability of the method among the forensic and the legal communities (Budowle et al., 2001, 2009a,d).

The first significant challenge of LCN typing in a courtroom emerged during the Omagh bombings trial (Ireland, UK, 2007), where the prosecution relied on LCN DNA profiles (and also conventional profiles) to link a suspect with a series of bombings in Northern Ireland. During this trial, DNA evidence was dismissed by the judge, who expressed concerns about the reliability of the methods used to generate the DNA profiles. This ruling marked a symbolic turn in forensics, in particular because the UK police suspended the use of LCN DNA, though it was later reinstated following a UK review on the technique, which concluded that LCN typing is a "robust" and "fit for purpose method" (Caddy et al., 2008).

In more recent trials (2009), judges have either rejected the use of DNA evidence analyses based on LCN (Gilbert, 2010) or supported the use of LCN evidence when they considered the method reliable[1]. While a court may not be the most appropriate place to discuss the validity and reliability of these methods, these legal decisions provided an outline of the repercussions of the use of LCN analyses in practical cases.

As we previously mentioned, forensic literature is "discrete" when it comes to discussing LCN typing. The only (non-peer-reviewed) paper discussing the flaws in the technique followed the first applications of LCN techniques (Budowle et al., 2001). The same authors recently advocated limiting the use of LCN typing to investigative purposes or to victim identification and avoiding its use for exculpatory purposes, mainly because of the lack of *reproducibility* of DNA profiling using LCN (Budowle et al., 2009c).

The main argument of the detractors of LCN is thus the lack of reproducibility of the results. This is particularly interesting, since lack of reproducibility may as well be considered as a process inherent in the collection/generation of biological data itself, rather than as "vagaries" (Budowle et al., 2009b) linked to LCN typing. In a discussion on the subject, Gill and Buckleton (2010) argue:

> "*Variability, and indeed uncertainty, is a part of most, if not all, scientific endeavours. It is not the existence of variability but rather the magnitude and potential consequences of any variability that needs to be assessed and reported to the court.*"

Regardless of which side one supports, this debate reveals a need for a methodology that takes into account the uncertainty related to low-template DNA samples. Actually, such a model was proposed a decade ago by Gill et al. (2000). Their methodology, built upon a Bayesian framework, is based on the calculation of the likelihood ratio that accounts for stutter peaks, drop-out and drop-in alleles. Moreover, the model can be applied to cases where replicates of the same sample are available.

This approach offers an important advantage. Because anomalies related to STR typing in critical conditions are accounted for, there is no need to restrict the applications of LCN

[1]People vs. Megnath, Supreme Court, New York, 2010.

samples. Indeed, restrictions inevitably lead to defining what a low-copy number sample is. There have been various definitions in relation to the quantity of DNA, with the most widespread definition corresponding to a limit of less than 100 pg of DNA. Interestingly, manufacturers of STR multiplex analysis systems usually recommend the use of at least 250 pg of DNA (Gill, 2001). A definition based on quantitative thresholds is problematic because it is not compatible with the rather continuous nature of the data. Indeed, one cannot define a threshold for which no artifacts are observed because their random nature makes them more frequent in samples with limited quantities of DNA, though this is not impossible for pristine samples.

The term LCN, which has traditionally been used to designate both a method (for instance, increasing the number of PCR cycles) and samples with a low level of DNA has added confusion to the debate. In the face of this confusion, the developers of the technique felt obliged to redefine samples DNA of low quantity/quality as "low-template DNA" samples instead of low-copy number samples (Gill and Buckleton, 2010). Throughout this manuscript, we use the same terminology.

In the following, we briefly describe Gill et al.'s model and we illustrate its use on a simple case involving a single locus.

3.1.2 The *statistical model* for DNA evidence interpretation: a solution for low copy number samples

Anchored in a Bayesian logic, and more precisely in a likelihood ratio framework, Gill et al. (2000) proposed a model that accounts for the anomalies related to STR loci typing. This model calculates likelihood ratios for a set of replicates of a given DNA profile sample, and simultaneously accounts for drop-outs, drop-ins and stutters phenomena.

This model is termed the *statistical model*, as opposed to the *biological model*. In the latter, a consensus profile is deduced by comparing replicated profiles: only alleles consistent throughout all replicates are reported, and thus a number of subjective decisions are made to reach a consensus. The statistical model is thus much more flexible, and several hypotheses about the origin of the stain can be evaluated simultaneously throughout likelihood ratios.

In the following, we illustrate the use of the statistical model for the case where a unique replicate profile is available.

Illustration

Figure 3.1 gives an example of a situation where a crime scene profile matches only one of the alleles of the suspect profile, the other one being under the limit of detection threshold (set for the purpose of this example to 50 RFUs). A common practice in this case consists in either excluding the suspect as a potential contributor, or, if the profiles match at other loci, to removing this problematic locus from the analysis[2]. Clearly, excluding problematic loci from

[2]Laurent Pène, Personal communication.

the analysis is a sensitive issue. The statistical model offers a framework that deals with these problematic situations: instead of deciding whether a profile is excluded or included into the analysis; probabilities of drop-out are incorporated into the calculations.

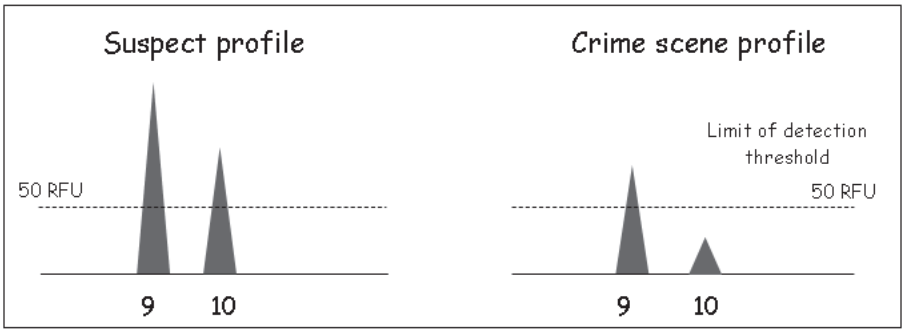


FIGURE 3.1: Illustration of a partial match: a suspect profile at a single locus matches only one of the alleles of the crime scene profile, the second allele is below the limit of detection threshold.

To illustrate Gill et al.'s (2000) approach, we calculate the likelihood ratios for the single-locus profile shown Figure 3.1. If the prosecution evaluates the hypothesis that the profile comes from the suspect, then its is necessary to consider drop-out to explain the evidence. In this example, the competing hypotheses are:

H_p: the crime scene profile comes from the suspect
H_d: the crime scene profile comes from an unknown contributor

Under H_d, the crime scene profile (CSP) comes from an unknown contributor who is either homozygote for allele "9" or heterozygote "9/Q", where allele Q denotes any allele other than "9". Under H_p, the crime scene profile comes from the suspect and a drop-out event is considered for allele "10". These hypotheses are weighted against each other through a likelihood ratio:

$$LR = \frac{Pr(CSP \equiv 9 | Suspect \equiv 9/10, H_p)}{Pr(CSP \equiv 9 | Suspect \equiv 9/10, H_d)} \tag{3.1}$$

Under the H_p hypothesis, we observe one allele and and one drop-out (allele "10"). Assuming that drop-out is independent within a genotype and across loci, the corresponding (conditional) probability is $Pr(D)Pr(\bar{D})$, where $Pr(D)$ is the probability of the event of drop-out for the considered allele and $Pr(\bar{D})$ is the probability of the complementary event (no drop-out). Note that we do not consider the possibility for the present alleles to be drop-ins or stutters.

Under H_d, an unknown contributor, unrelated to the suspect is the source of the evidence. Two situations have to be considered for the unknown contributor:

- the unknown contributor is a homozygote "9/9", and thus there is no drop-out. Denoting p_9 the frequency for allele "9", the corresponding probability is $p_9^2 Pr(\bar{D})^2$,

- the unknown contributor is a heterozygote "$9/Q$", and allele Q has dropped-out, the corresponding probability is $2p_9p_QPr(D)Pr(\bar{D})$, where $p_Q = 1 - p_9$.

The likelihood ratio is finally given by:

$$LR = \frac{Pr(D)Pr(\bar{D})}{p_9^2Pr(\bar{D})^2 + 2p_9p_QPr(D)Pr(\bar{D})} \tag{3.2}$$

From this example, it is obvious that an approach of exclusion/inclusion of the suspect profile would lead to a likelihood ratio of zero, but this not the case when the statistical model is applied.

The reason that a generalized statistical method for DNA evidence interpretation is not currently in use in forensic laboratories is not specific to low-template DNA samples. As we previously explained, courtrooms are resistant to what they see as complicated statistical reasoning. Still, it is notable that the statistical model has been implemented in two commercial "expert systems" (Curran et al., 2005; Gill et al., 2007), and more recently, an open-source implementation of the statistical model was proposed by Balding and Buckleton (2009).

These programs should facilitate the introduction of the model, but an unsolved issue is that of the determination of the probabilities of drop-out, drop-in and stutters, which constitute the model parameters, and have ultimately to be specified by the user.

Usually, these parameters are determined through experimental studies in which conditions favoring the occurrence of drop-outs, drop-ins and other anomalies are created (Gill et al., 2000). However, relying on experimentation alone is contestable because the estimates yielded depend closely on the experimental design, and different experimental runs are likely to yield different results. Parameter estimations should account for data that are available from the sample itself, namely, the allele peak heights that give information about the quality and the quantity of DNA. Ideally, reporting officers should be able to answer the question: "Given the observed peak heights/areas, what is the probability that a given allele has dropped out?"

There has been little work on how these probabilities should be estimated from available data. Therefore, there is clearly a need for establishing a methodology to help determine the model parameters. The drop-out phenomenon has particularly attracted our attention. Indeed, reporting officers are well trained in most forensic laboratories to detect gross contamination, stutters, drop-ins and other STR-related artifacts, but drop-outs seem to be particularly challenging because they raise questions about whether a profile should be included or excluded from the analysis, thereby reducing the strength of the evidence. Hence, our focus in this chapter is on the estimation of drop-out probabilities.

3.2 Estimating drop-out probabilities in forensic DNA samples

3.2.1 Motivation

The importance of taking into account artifacts associated with STR loci is well agreed upon in the forensic community (Gill et al., 2006). Still, a review of the relevant forensic literature

shows that until recently, no significant developments had been proposed to estimate drop-out probabilities in a forensic setting. The drop-out phenomenon is a common problem in degraded DNA samples, which are often encountered in conservation studies of wild populations in ecology or ancient fossils in museums. Typically, small quantities of DNA are recovered from hair or feces, leading to profiles in which drop-outs are very likely to occur (Taberlet et al., 1996; Gagneux et al., 1997; Pompanon et al., 2005).

Unfortunately, statistical developments devoted to estimating drop-out probabilities concern non-forensic applications and mainly suggest the use of replicates of genotypes, along with comparison to known allelic frequencies in the target population to infer the most likely DNA profiles (Miller et al., 2002; Johnson and Haydon, 2007). However, in a forensic setting, systematic replicates are not always possible, and furthermore, the choice of a relevant population database to obtain allele frequencies represents an additional difficulty.

Two recent studies proposed the use of logistic models to estimate the probability of drop-out based on data provided by analyzed DNA profiles in forensic casework (Tvedebrink et al., 2009; Gill et al., 2009). With respect to our evaluation approach, we propose the implementation of a simulation model to evaluate different models based on logistic regression. The article presented below introduces our methodology and illustrates it with an evaluation of some features of the logistic model proposed by Gill et al. (2009).

3.2.2 Article 3: "Estimating drop-out probabilities in forensic DNA samples: a simulation approach to evaluate different models".

Article submitted to *Forensic Science International : Genetics* (2010).

Estimating drop-out probabilities in forensic DNA samples: A simulation approach to evaluate different models

H. Haned[a,*], T. Egeland[b], D. Pontier[a], L. Pène[c], P.Gill[b,d]

[a] *Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 69622 Villeurbanne, France*
[b] *Institute of Forensic Medicine, University of Oslo, 0027 Oslo, Norway*
[c] *Institut National de Police Scientifique, Laboratoire de Police Scientifique de Lyon, France*
[d] *University of Strathclyde, Royal College, 204 George Street, Glasgow G1 1XW, UK*

Abstract

Allele drop-out is a well established phenomenon that is primarily caused by the stochastic effects associated with low quantity or low quality DNA samples. Recently, new interpretation models that employ the use of logistic regression have been utilised in order to estimate the probability of drop-out. The model parameters are estimated using profiles from samples of extracted DNA diluted to low template levels in order to induce drop-out. However, we propose that this approach is over-simplistic, because several sources of variability are not taken into account in this generalised model. For example, in real-life, small (discrete) crime-stains are analysed where cells are (or were) intact. The integrity of the paired chromosomes of the diploid cell is preserved. In *extracted* DNA that is diluted to low template levels, we argue that the paired-chromosome integrity is lost. This directly affects the

*Corresponding author
Email address: hinda.haned@univ-lyon1.fr (H. Haned)

outcome of the logistic model. To date, current experimentation procedures are more akin to *haploid* cells and thus, different logistic models are needed for haploid and diploid cells. In order to simplify the methodology to estimate multiple logistic regressions that may be required, we propose the use of a simulation model of the entire process associated with the analysis of STR loci, as a supplement to the purely experimental approach to support the validation of new methods. We illustrate with an evaluation of some features of the logistic model proposed by Gill et al. [P. Gill, P. Puch-Solis, J. Curran, The low-template-DNA (stochastic) threshold–Its determination relative to risk analysis for national DNA databases, Forensic Sci. Int. Genet. 3 (2009) 104–111] and discuss alternative models.

Keywords: drop-out, logistic regression, simulations, PCR, STR

1. Introduction

A number of phenomena often associated with low-template DNA profiles, especially drop-out, and drop-in, complicate their interpretation [1]. Although it is fairly standard practice to interpret allele drop-out using the "$2p$" rule, Buckleton and Triggs [2] and Balding and Buckleton [3] show that this can be anti-conservative, especially when the probability of drop-out is close to zero. Although allelic drop-out events are encountered more frequently with low quantities of DNA, they cannot be eliminated with certainty even in "gold standard" DNA samples [1, 4].

A consistent approach consists of quantifying and integrating error into the statistical analysis [5, 6]. The likelihood ratio framework is the preferred approach to report the weight of DNA evidence. Likelihood ratios avoid

the binary decision making process of inclusion vs. exclusion, by taking into account all of the potential ambiguities that are incurred by drop-out and drop-in within the context of a continuous model [7]. Although robust interpretation frameworks have been proposed by Gill et al. (2000) [7] and improved by [3, 8, 9, 10, 11, 12], there are few alternative methods to inform the probabilities of drop-out.

In non-forensic applications, maximum likelihood approaches based on replicate genotypes have been proposed in order to estimate the drop-out rates [13, 14]. In forensic applications, different principles have to be followed, because of the limited sample sizes: multiple systematic replicates are usually not available (although it is often the practice to obtain two or three replicates with low level samples). To circumvent this difficulty, simple models based on the logistic regression were originally proposed to estimate the drop-out probability [12, 15]. Gill et al. (2009) [15] suggested that the probability of drop-out could be derived from a measure of the quality of the DNA profiles based on their observed peak heights. Tvedebrink et al. (2009) [15] proposed to condition this probability on an estimate of a proxy for the total DNA

Tvedebrink et al. [15] illustrated how the model parameters can be estimated from experimental data. In their experimental design, the drop-out events were induced using serial dilutions of mixtures. The estimated parameters were then used to predict the probability that a given allele has dropped-out, conditioned on a given DNA profile. The authors noted that the model parameters depend on the process used to generate DNA profiles. For example, parameters estimated from a process with 28 PCR amplification cycles are not valid for a more sensitive process utilising 34 cycles. Additional

variability occurs within the process. For example, the kind of DNA samples used; whether the cells are haploid or diploid; the allelic composition; the STR multiplex used to characterize the sample; the material/machinery used during the process. Moreover, the efficiency of DNA extraction varies between sample-types and probably between laboratories [16].

Given, the potential sources of variation, this raises questions of whether a single logistic regression can be used to model a process, given that there is always variation within any process itself. Experimental data sets are the easiest adapted to study drop-out. Current methodology to determine the drop-out parameter is generally based on serial dilution experiments of disrupted cells. But strictly speaking, these experiments are only valid for haploid cells (i.e. sperm), hence conclusions about the robustness of models cannot necessarily be extrapolated to all kinds of biological material (especially diploid cells).

Ideally, methods evaluation should be carried out using controlled experiments. However, experimental design is often challenging in practice because there are many sources of variability that are difficult to control. Here, we propose a complementary method to the purely experimental approach. Based on the stochastic model of [16], we propose a simulation approach to evaluate the efficiency of the logistic model in describing allele drop-out. The suggested methods have been implemented in the open source software forensim [17].

2. Methods

2.1. Simulation model

The simulation model proposed by [16] described the entire process associated with the generation of a DNA profile using short tandem repeat loci. The model presented here was simplified to incorporate allele drop-out events only. Stutters are not modelled.

Starting from the initial number of cells, n_{cells}, the simulation model applies successive binomial samplings in order to determine the numbers of template molecules that are successfully passed to consecutive steps of the DNA analytical process. This starts with extraction, selection of an aliquot for PCR, the PCR cycling process, and finally visualisation of the alleles. A separate parameter is defined for each of these steps: $\pi_{extraction}$, $\pi_{aliquot}$ and π_{PCReff} respectively.

Two different models are required in order to fully characterise the DNA process. There is a diploid model, where there are $n = 2 \times n_{cells}$ target molecules, and a haploid model with $n = n_{cells}$ target molecules (Appendix A).

The simulation model allows generating data that will be used to estimate the logistic model parameters. Using the simulation model described in the Appendix section, we have generated haploid and diploid data sets in order to test the robustness of the modelling assumptions.

2.2. Definition of the logistic model (heterozygotes)

The event of drop-out at a given locus is defined using a binary variable D. If the allele peak height, denoted h; is below the limit of detection threshold

(LOD) T_{drop}, expressed in RFUs, a drop-out indicator is recorded:

- $D = 1$ if $h \leq T_{drop}$: drop-out has occurred

- $D = 0$ if $h > T_{drop}$: no drop-out observed

In this study, we set the LOD threshold to $T_{drop} = 50$ RFUs. This threshold can be modified to include lower peak heights if necessary.

We further define heterozygote drop-out as the indicator that one of the allele peak heights of the heterozygote is below T_{drop} and the homozygote drop-out as the indicator that the peak height of the homozygote allele is below T_{drop}.

The main difference between Gill et al.'s [12] and Tvedebrink et al.'s [15] model is the ability to directly infer the results obtained from heterozygous profiles to homozygous profiles. Tvedebrink et al. estimate the drop-out probability relative to a proxy for the amount of DNA contributed to the stain, which allows the drop-out probability of homozygotes to be determined, while Gill et al.s model describes heterozygous drop-outs only.

2.3. Heterozygous drop-out

It was originally suggested that the homozygote drop-out probability could be computed from the square of the heterozygote probability [12, 15]. This implied that drop-out events were independent between alleles at a given locus.

Recently, Balding and Buckleton (2009) [3] showed that this method tends to overestimate the homozygous drop-out probability, and suggested a correction factor (α) applied to the square of the heterozygote probability provided

a better estimate. This suggestion has intuitive appeal, but further investigations are required to properly characterise the α parameter relative to DNA quantity and quality. Homozygous drop-out will be evaluated in a separate paper.

In this paper we evaluate Gill et al.'s model [12] only in relation to heterozygous loci. Following Gill et al. we model the probability of drop-out as the conditional probability $P(D = 1|X_{HET})$, where X_{HET} is one of the peak heights (in RFUs) of a heterozygote (whether it's present or not, the peak height can equal zero). Given the peak height of one allele, we predict the probability of *drop-out* of the other allele, that we will refer to as the *companion* allele. Considering a heterozygote with alleles A and B, the peak height of allele B is used as an explanatory variable for the drop-out of allele A (the companion allele). Note that the order in which the alleles are chosen is arbitrary. The rationale behind this model is that the peak height of a given allele can be used to provide information on the presence or absence of its companion allele.

The same logic applies to the simulation model: given the random seed that determines the randomisation procedure, the parameter n_{cells} and the "*process*" parameters: $\pi_{extraction}$, $\pi_{aliquot}$ and π_{PCReff}, uniquely determine the output, i.e., the allelic peak heights. When the "*process*" parameters are constant, the only non-random factor that determines allelic peak heights is the number of cells (equivalent to a known quantity of DNA). The number of cells is generally unknown, but allele peak heights can be used as a proxy estimator [15], as there is an approximately linear relationship between peak heights and DNA quantity.

The logistic model is thus defined as:

$$Pr(D = 1\,|X_{HET}) = \frac{e^{\beta_0+\beta_1 X_{HET}}}{1 + e^{\beta_0+\beta_1 X_{HET}}} \quad (1)$$

An example illustrating the construction of the drop-out variable D and the explanatory variable X_{HET} on simulated heterozygous profiles is given in Appendix B.

The logistic regression (eq. 1) is implemented using the R statistical software [18].

2.4. The traditional experimental design to estimate the drop-out probability

By default of the experimental design used to produce logistic regressions, previous authors [12, 15] have unintentionally described a *haploid* model. In this model, for a given cell, at a given heterozygous locus, only one allele per locus is observed per cell, whereas in the diploid model, alleles from *both* chromosomes are present, hence both alleles are present and associated. The difference between the two models is easily illustrated Figure 1.

Consider the case of a sexual assault, where the evidential DNA sample consists of a sample of sperm cells left by the offender on the crime scene. In the following, a single heterozygous locus is considered, with alleles A and B. We further formalize our fictitious example by considering the offender as an "urn" of n haploid cells (in this case, sperm cells), where n is very large and can be considered as infinite.

In this "urn", the numbers of A vs. B alleles are perfectly balanced (after meiosis): there are $\frac{n}{2}$ alleles of type A and $\frac{n}{2}$ alleles of type B. When DNA is transferred from the offender to the crime scene, this is equivalent to a sample of a certain size being randomly selected from the "urn". Given

that there are equal numbers of alleles of type A and B, the probability of selecting allele A is the same as that of selecting allele B, and is equal to 0.5. However, the actual number of alleles of each type depends on the size of the sample taken from the "urn". For instance, if 10 haploid cells were deposited at the crime scene, then it is unlikely that there are exactly 5 cells carrying allele A and 5 cells carrying allele B. The extent of this imbalance depends of the size of the sample taken: each time a sample of size n_{cells} is taken from the urn, a different estimate of the proportion p, denoted $\hat{p}$ is yielded.

The imbalance between proportions of A vs. B alleles can be further described through the sampling distribution of the proportion estimate $\hat{p}$, which has expectation p and standard error (s.e.) $\sqrt{\frac{p(1-p)}{n_{cells}}}$.

In our case, the samples are taken from an infinite population size, and p is known to be 0.5. The sampling distribution of the proportion can be approximated by a Gaussian distribution with mean 0.5 and standard deviation of $\frac{0.5}{\sqrt{n_{cells}}}$ (Figure 2).

Increasing the size of the sample, increases the probability of having balanced proportions of alleles A and B. It is possible to calculate the sample size needed to achieve a balanced proportion of A vs. B alleles, with a given tolerance defined as a standard error. Recall that the standard error is most easily interpreted in terms of confidence intervals since a 95% confidence interval is obtained as $[\hat{p} - 2 \times s.e., \hat{p} + 2 \times s.e]$. If we need the estimated proportion of A alleles in the sample, denoted $\hat{p}$, to equal 0.5 with a s.e of 1%, then, the number of cells needed for the experiment is $n_{cells} = \left(\frac{\hat{p}(1-\hat{p})}{0.0,1}\right)^2 = \left(\frac{0,5}{0.01}\right)^2 = 2500$

The same calculation is carried out for different values for the standard

error (Table 1).

n_{cells}	Standard error
2500	0.01
625	0.02
278	0.03
156	0.04
100	0.05
69	0.06
51	0.07
39	0.08
30	0.09
25	0.10

Table 1: Number of cells needed to recover a proportion of alleles of type A equal to 0.5, with different values for the standard error.

2.5. Model evaluation

We have evaluated both haploid and diploid models in order to test the robustness of the modelling assumptions.

Diploid case

In order for the simulations to reflect the reality of (low DNA template) forensic casework, we simulated heterozygous profiles from varying numbers of cells n_{cells} sampled from a Gaussian distribution with mean 10 and a standard deviation of 5. A total of 1000 heterozygous profiles were generated. The parameters suggested by Gill et al.'s [16] were used: $\pi_{extraction} = 0.6$,

$\pi_{aliquot} = 20/66$, $T = 28$ and $\pi_{PCReff} = 0.8$. Obviously, parameter values may be changed according to any specific application or process

Haploid case

The diploid cell simulation model can easily be adapted to simulate haploid data, by introducing an additional, preliminary sampling step, where alleles A and B are randomly selected using a binomial distribution (conditioned on n_{cells}). A supplementary parameter, π_A , is used to define the probability of selecting allele A or B at a heterozygous locus. The same simulation procedure is then applied to the resulting numbers of alleles A and B. The usual serial dilution experiments can also be mimicked by varying the probability of selecting alleles A and B, π_A.

Using the same simulation conditions as for the diploid case, we simulated four datasets with varying starting proportions of alleles A vs. B. This was achieved by applying a binomial sampling on the starting number of cells, where the probability of selecting allele A (π_A) is successively taken as 0, 0.3, 0.5 and 0.8, for examples across the range. The probability of observing a sample with a proportion of alleles A of 0.3 for example, is dependent upon the starting number of cells and this is illustrated in Figure 2. The choice $\pi_A = 0$ corresponds to the mono-allelic case, i.e., the probability of selecting allele of type A is null. The simulation model is implemented in the forensim package for the R statistical software [17].

3. Results/ Discussion

3.1. Assessing the fitted model

Performing the Hosmer-Lemeshow goodness of fit test [19] on the different models fitted with either a haploid or a diploid dataset indicated that there was no reason to reject the logistic model, except when $\pi_A = 0$. In other words, our simulation data supports the logistic model except for the case where the data is mono-allelic.

Parameter estimates, along with the confidence intervals, for the diploid model are given in Table 2.

	Estimates– odds ratios	95% confidence interval for odds ratios		P-value
Parameter		Lower	Upper	
e^{β_0}	3.39	2.65	4.35	0
e^{β_1}	0.96	0.96	0.97	0

Table 2: Estimates of the model parameters for the diploid dataset. The table gives the confidence intervals and P-values of the Wald test for the logistic model parameters.

The confidence intervals indicate that there is little uncertainty about the estimation of the model parameters. The number of points in the simulated data sets is 1000, consequently, the P-values are, not surprisingly, very significant. Here, our focus is on parameter interpretation: β_0 , expresses the log of the odds ratio (OR) of the probability of drop-out when the surviving peak height is zero. The parameter β_1 expresses the change in the log of the OR relative to the probability of allele drop-out . We used these parameters to calculate the predicted probabilities of allelic drop-out.

228 *3.2. Expected drop-out probability vs. peak heights*

229 Using the estimated parameters from the simulated data, we investigate
230 the model predictions when the peak height of the companion allele is varied
231 from 0 to 500 RFUs. Figure 3 shows the expected drop-out probabilities of
232 allele A with respect to the peak height of the companion allele, which height
233 is given by X_{HET}. We reiterate the process for different datasets simulated
234 as outlined in section 2.5.

235 Note that the curves do not start at 1 (Figure 3): the probability of allele
236 drop-out in case no surviving allele is observed, $P(D = 1|X_{HET} = 0)$, is
237 estimated as 0.83 for the diploid case; this corresponds to the profiles where
238 only one of the heterozygote's alleles survived. These drop-out events are
239 termed as "extreme drop-out events" in [12] because the size of the companion
240 allele exceeds the stochastic threshold.

241 Except for the case where the data set is mono-allelic ($\pi_A = 0$), the logistic
242 model predicted that the probability of drop-out of allele A decreased as the
243 peak height of allele B increased. Predictions corresponding to different
244 datasets differed in the decrease rate of the drop-out probability, the curve
245 representing the diploid case decreasing more rapidly than the others. This
246 is explained by the imbalance between the initial proportions of alleles A and
247 B in the haploid datasets. If alleles B are under-represented ($\pi_A = 0.8$) then
248 there is not enough quantitative data (peak heights of allele B) to describe the
249 companion peak height. Similarly, when alleles B are over-represented ($\pi_A \in$
250 $\{0, 0.3\}$) the drop-out events are scarcer than in the previous case because
251 fewer alleles of type A are present in average. The imbalance between alleles
252 A and B persists even in the case where they have the same probability to

be selected ($\pi_A = 0.5$); this is due to the variability inherent in the sampling process.

These simulations show that the initial imbalance of the number of alleles A and B is reflected at the prediction level. Hence, the logistic model presented here is not adapted to experimental datasets prepared through dilution experiments. Other models based on the contributed amount of DNA [15] may be better adapted for haploid data.

3.3. Homozygous vs. heterozygous drop-out

Treating homozygote drop-out is not trivial, because both alleles drop-out, hence a model based on the companion peaks of a heterozygote cannot work directly. To circumvent this problem, Balding and Buckleton [3] suggested that the drop-out probability for homozygous profiles, denoted D_2, could be inferred from the heterozygote probability using a scaling factor (the α model). We propose that simulated data can be used to investigate the validity of this model.

3.4. Drop-out threshold

The use of a limit of detection threshold of 50 RFUs is common in forensic practice, however, the underlying data (the peak height intensities) are continuous, hence the application of thresholds can lead to inconsistencies. Following from the drop-out definition in this paper, an allele with a peak height of 49 RFUs is recorded as a drop-out, whereas an allele with a peak of 51 RFUs will not be recorded as a drop-out. The discussion of continuous models is beyond the scope of this paper, however, the use of Bayesian networks to assess the risk associated to thresholds seems to be the way forward

[12].

3.5. Alternative models

There are obvious alternatives and modifications of the logistic model that we have presented and some of these are discussed below:

Consistency requirement

Assume first that we impose the following consistency requirement: If the companion peak height equals the threshold for declaring drop-out, T_{drop}, the model should predict a drop-out with probability 0.5. This is not necessarily the case for models that have been suggested so far - including the approach described in this paper. To derive a model that fulfils the mentioned requirement consider the reparametrisation

$$P(D = 1|X_{HET}) = \frac{e^{\beta_0+\beta_1(X_{HET}-T_{drop})}}{1 + e^{\beta_0+\beta_1(X_{HET}-T_{drop})}} \tag{2}$$

If $X = T_{drop}$ in eq. 2, then we have:

$$P(D = 1|X_{HET} = T_{drop}) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{3}$$

If $\beta_0 = 0$, then $P(D = 1|X_{HET} = T_{drop}) = 0.5$

While this model, i.e., $P(D = 1|X_{HET}) = \frac{e^{\beta_1(X_{HET}-T_{drop})}}{1+e^{\beta_1(X_{HET}-T_{drop})}}$ meets the consistency requirement, only one parameter is estimated and it is therefore less flexible.

Generalisations

The models so far require independent observations (as do other logistic regression models suggested for similar purposes including [15]). The models

(eqs. 1 and 2) therefore do not apply to data with several markers from the same case. In particular, the simulated data of Table 3 cannot be used. From a practical point of view, dependence implies that if we are given information on the peak heights of some markers, then this guides our expectations for the peak heights of remaining markers. Dependence is therefore useful and could be utilised to improve estimates on drop-out probabilities whenever data is available for several markers cases. For the simulated data of Table 3, dependence arises since the number of cells varies between cases but not within cases. We will not pursue the discussion of models accounting for dependence here since this would deviate from the objective of this paper. However, the problem of dependence could be reduced or perhaps removed by conditioning on an accurate estimate of the quantity.

Based on the above observation and analyses that we have carried out (data omitted) we conclude that if the number of n_{cells} is known then it can be used to predict drop-out probabilities. Significantly better predictions will result. In the absence of exact information on n_{cells} we could replace it by an estimate. For instance, n_{cells} could be estimated from quantification. The result is likely to be inaccurate, however. Therefore, it is not obvious that such an approach would improve the model. Alternatively, all peak heights of a case could be used to estimate the DNA quantity or the n_{cells} equivalent. Indeed the model of Tvedebrink et al. [15] can be viewed in this light. Further data, preferably based on experimental data, is needed to resolve the above issues.

Case	Marker	n_{cells}	$\pi_{extraction}$	$\pi_{aliquot}$	π_{PCReff}	H_1	H_2
1	1	10	0.60	0.30	0.80	65	38
1	2	10	0.60	0.30	0.80	0	33
2	1	14	0.60	0.30	0.80	91	49
2	2	14	0.60	0.30	0.80	57	90
3	1	12	0.60	0.30	0.80	66	38
3	2	12	0.60	0.30	0.80	54	67

Table 3: In this table three cases are simulated, each with two heterozygous markers. The number of cells (n_{cells}) are sampled form a Gaussian distribution with the same parameters used for data simulation, whereas the remaining parameters determining the PCR simulation ($\pi_{extraction}$, $\pi_{aliquot}$, π_{PCReff}) have been kept fixed. H_1 and H_2 are the peaks heights of the alleles of the simulated heterozygotes.

4. Conclusion

Accounting for stochastic anomalies linked to STR typing techniques should become a standard in forensic genetics practice [5]. Therefore, there is a need to develop methods to quantify the probabilities associated with the occurrences of events such as drop-out and drop-in. Recent efforts based on logistic regression are promising, but it is necessary to accompany these developments by evaluation and validation procedures based on experimental datasets, in order to enhance their use in forensic casework. Such experimental evaluation is challenging in practice because it is not clear how the experiments should be designed. Here, we do not intend to give a firm direction for experimental protocols, but rather propose the simulation approach as an alternative to investigate the most appropriate variables to take into account in the experimental design. The simulation model presented here

offers a flexible framework to evaluate models based on logistic regression. It should be considered as an aid for methods evaluation that can temporarily replace laboratory experiments. The accessibility of the methodology in free software should also enhance the evaluation procedures on different kinds of data collected in different situations.

338 **Figures**

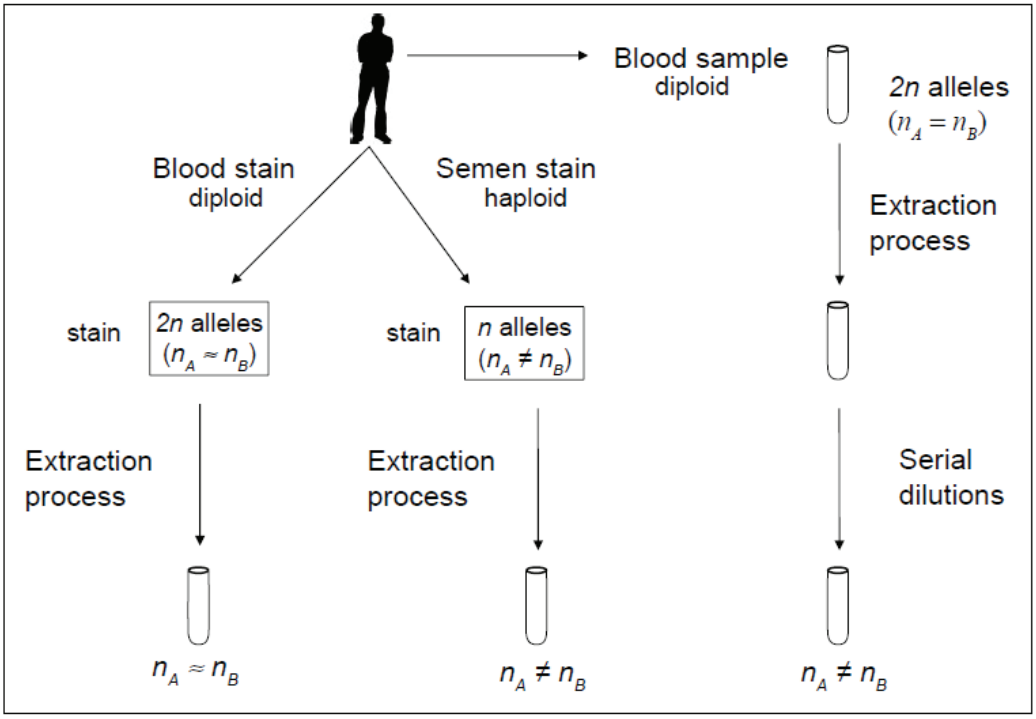


Figure 1: Traditional experiments take a blood sample (or other body fluid of diploid origin) and the DNA is extracted. Then serial dilutions are made, hence the number of of copies of allele A (n_A) does not equal the number of copies of allele B (n_B), and is akin to the haploid model. After extraction in real casework, n_A is approximately the same as n_B.

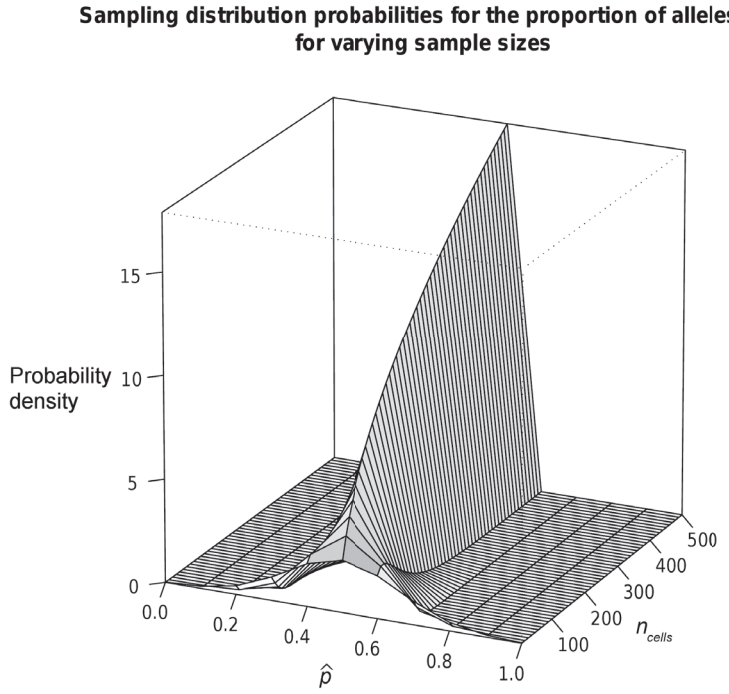


Figure 2: Sampling distribution of the proportion of alleles A conditioned on the sample size n_{cells} (number of cells). n_{cells} was modelled as a continuous variable which values varied from 10 to 500 cells.

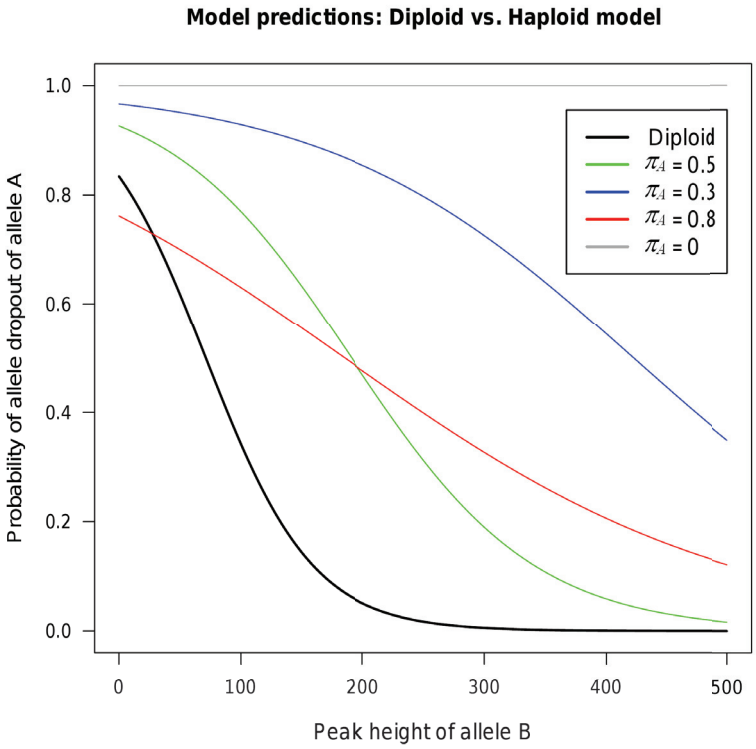


Figure 3: Model predictions, when haploid or diploid data are used. The parameter π_A is the probability parameter used to sample alleles A and B prior to the extraction step. The curves give the probability that allele A has dropped-out giving the peak height intensity of allele B. Peak heights of allele B range from 0 to 500 RFUs.

Appendix A: simulation model

We adapted the simulation model proposed in [16] to only account for allele drop-out, stutter peaks are not modelled in the version presented here. Starting from the initial number of cells, n_{cells}, the simulation model applies successive binomial samplings in order to determine the numbers of template molecules that are successfully passed to consecutive steps of the DNA analytical process. This starts with extraction, selection of an aliquot for PCR, the PCR cycling process, and the final visualisation of the alleles. Two different models are required in order to fully characterise the DNA process. There is a diploid model, where there are $n = 2 \times n_{cells}$ target molecules, and a haploid model with $n = n_{cells}$ target molecules. These models differ from each other in the diploid model, chromosomes are paired and are therefore completely dependent, whereas in the haploid model, chromosomes are unpaired and completely independent. The steps for the diploid cells are subdivided as follows:

i. *DNA extraction*: When DNA is extracted from a crime sample, the process is never 100% efficient. Dependant on the process used, this means that a given template molecule has a probability $\pi_{extraction}$ of being selected. The random variable, $n_A^{survived}$, describes the number of DNA molecules, for a given allele A, surviving this step:
$n_A^{survived} \rightarrow Bin(n_{cells}, \pi_{extraction})$

ii. *Aliquot selection*: Once the DNA has been extracted, an aliquot (e.g. 10 out of 50ul total) is removed in order to be forwarded to the PCR reaction. The decision on the size of aliquot is often informed by a quantitative step that is used to optimise the quantity of DNA. The number

of molecules extracted in the previous step are included in the aliquot (that is forwarded to the PCR reaction) with a probability $\pi_{aliquot}$. The number of molecules succeeding this selection step is modelled as:

$n_A \rightarrow Bin(n_A^{survived}, \pi_{aliquot})$

iii. *PCR amplification*: the n_A molecules in the aliquot suspension are forwarded into a PCR reaction comprising T PCR amplification cycles. At a given locus, a DNA molecule is successfully amplified with a probability π_{PCReff}. In reality, the reaction efficiency decreases with cycle number because the *Taq* polymerase enzyme degrades. In our model, this probability is a single fixed parameter for each cycle t (note that [16] shows that a precise estimate of the PCR efficiency is not critical to the estimation of relative heterozygous balance). The number of amplified molecules for a given allele A, at a given PCR cycle t is given by the recurrence equation:

$n_A(t) = n_A(t-1) + Bin(n_A(t-1), \pi_{PCReff})$

This model is adapted to the haploid case by introducing a supplementary parameter, π_A, which gives the probability of selecting allele A or B at a heterozygous locus, previous to the extraction step. The number of molecules of type A is determined using a binomial sampling of n_{cells} with parameter π_A. Rewriting step i. for the haploid case:

$n_A^{survived} \rightarrow Bin(Bin(n_{cells}, \pi_A), \pi_{extraction})$

The number of surviving molecules of type B is calculated as:

$n_B^{survived} \rightarrow Bin(n_{cells} - Bin(n_{cells}, \pi_A), \pi_{extraction})$

The use of binomial sampling applied to the simulation procedure implies that the template molecules are independent during the different simulation

steps, with the exception of the first (extraction) step where there is dependency of A with B. In subsequent steps we assume independence between alleles A and B.

Peak height determination

In the simulation, the number of molecules of type A and B must achieve a threshold of detection of approximately 2×10^7 molecules in order to trigger the photomultiplier in the automated sequencing machine to generate a signal [16]; thereafter, we use a log-linear relationship in order to determine the amplitude of the allele peak heights. Recall that $n_A(t)$ denotes the number of amplified molecules after t PCR cycles, the peak height is defined by:

$$log\left(\frac{n_A(t)+T^*_{drop}}{T^*_{drop}}\right)K \tag{4}$$

where T^*_{drop} is the threshold of detection expressed in terms of number of amplified molecules that must be achieved in order to generate a signal, and K is a positive constant, (determined by best fit to the observed peak heights) that can be considered as a scaling factor for peak height. In this study we used $K = 55$ and $T^*_{drop} = 2 \times 10^7$.

Appendix B: example dataset

In this Appendix we show how the simulated profiles are manipulated to construct the data sets on which the logistic regression is performed. The simulation model described in this paper was implemented into the forensim package for the R statistical software. Forensim and its documentation can be downloaded from: http://forensim.r-forge.r-project.org/. The following

R code simulates 6 DNA profiles, with the starting numbers of cells varying from 10 ($= n_{cells}$) to 15 cells. Note that the command lines are given as they should be typed in an R console:

```
>library(forensim)
>set.seed(12452)
>sapply(10:15,function(i)simPCR2(ncells=i,probEx=0.6,probAlq=0.30,probPCR=0.8,cyc=28))
```

The simulated data is shown in Table 4.

Height allele A	Drop-out variable allele A	Height allele B	Drop-out variable allele B
19	1	44	1
38	1	84	0
48	1	74	0
73	0	90	0
85	0	70	0
58	0	81	0

Table 4: Simulation results: the table displays the simulated peak heights and their corresponding drop-out variables.

The peak height of the first allele is used to describe the drop-out state of the other allele. To avoid dependencies in the data, only one allele peak height is used. Consequently the simulated data provides two datasets on which the logistic regression can be performed.

Starting with the first column "Height allele A", the drop-out variable D is set to 1 if the peak height is below 50 RFUs, and 0 otherwise. Variable X records the corresponding peak height of allele B. This process is repeated

for allele B. The following tables are obtained:

Dataset 1						
D	1	0	0	0	0	0
X	19	38	48	73	85	58
Dataset 2						
D	1	1	1	0	0	0
X	44	84	74	90	70	81

Table 5: Datasets obtained from Table 4.

The independence of observations is a necessary condition to perform the logistic regression, thus the estimations can be made on each of the yielded datasets above.

References

[1] J. Butler, Forensic DNA typing, Academic Press London, 2001.

[2] J. Buckleton, C. Triggs, Is the 2p rule always conservative?, Forensic Sci. Int. 159 (2-3) (2006) 206–209.

[3] D. J. Balding, J. Buckleton, Interpreting low template DNA profiles. Forensic Sci. Int. Genet. 4 (2009) 1–10.

[4] P. Gill, J. Buckleton, A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number, Forensic Sci. Int. Genet. 4 (2010) 221–227.

[5] P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayer, N. Morling, M. Prinz, P. M. Schneider, B. S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2-3) (2006) 90–101.

[6] F. Van Nieuwerburgh, E. Goetghebeur, M. Vandewoestyne, D. Deforce, Impact of allelic dropout on evidential value of forensic DNA profiles using RMNE, Bioinformatics 25 (2) (2009) 225–229.

[7] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Sci. Int. 112 (1) (2000) 17–40.

[8] J. Mortera, A. P. Dawid, S. L. Lauritzen, Probabilistic expert systems for DNA mixture profiling, Theor. Popul. Biol. 63 (2003) 191–205.

[9] J. M. Curran, P. Gill, M. R. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53.

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

[11] P. Gill, A. Kirkham, J. Curran, LoComatioN: A software tool for the analysis of low copy number DNA profiles, Forensic Sci. Int. 166(2-3) (2007) 128–138.

[12] P. Gill, R. Puch-Solis, J. Curran, The low-template-DNA (stochastic) threshold–Its determination relative to risk analysis for national DNA databases, Forensic Sci. Int. Genet. 3(2) (2009) 104–111.

[13] C. R. Miller, P. Joyce, L. P. Waits, Assessing allelic dropout and genotype reliability using maximum likelihood, Genetics 160 (2002) 357–366.

[14] P. C. Johnson, D. T. Haydon, Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data, Genetics 175 (2007) 827–842.

[15] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of str alleles in forensic genetics, Forensic Sci. Int. Genet. 3(4) (2009) 222–226.

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

474 [17] H. Haned, Forensim: an open source initiative for the evaluation of
475 statistical methos in forensic genetics, Forensic Sci. Int. Genet. in press
476 (2010).

477 [18] R. D. C. Team., R : A Language and Environment for Statistical Com-
478 puting. R Foundation for Statistical Computing, Vienna, Austria. ISBN
479 3-900051-07-0, URL http : //www.Rproject.org/.

480 [19] D. Hosmer, S. Lemeshow, Applied logistic regression, Wiley Interscience,
481 2000.

3.3 Discussion

The statistical model proposed by Gill et al. (2000) is regarded as the best way to move forward in the field of DNA interpretation (Gill and Buckleton, 2009), but much remains to be accomplished before such a model can be introduced in laboratory practice. In this regard, we identify two main stages: first, model validation, and second, implementation through accessible software.

Although Gill et al.'s model is anchored in a Bayesian logic that is well-established in science, validation is necessary to enhance its use in forensic practice. The validation of the statistical model would consist of demonstrating its superior flexibility with respect to currently used methods, derived from biological models. In fact, Gill et al. (2000) already showed how this model can be used to evaluate the consistency achieved by biological models. For this reason, we focused on the most significant issue raised by the model: the estimation of its parameters.

In the article presented here, only heterozygote drop-out is modeled, though the simulation model can also be used to simulate homozygote drop-out. However, the considered logistic models would no longer be valid for such data. Indeed, if no allele is observed at a given locus, then homozygote drop-out must have occurred. However, there are no remaining data at the considered locus that can be used to predict the drop-out probability. Future work should thus include the modification of current models to permit the modeling of homozygote drop-out. In particular, the alpha model proposed by Balding and Buckleton (2009) should be evaluated to determine the most accurate value or set of values for the alpha factor. These modifications will certainly imply the use of data provided by several loci, and the use of models that can account for dependencies in the data.

The second step to be taken to help introduce the use of the statistical model for DNA interpretation is its implementation in a free software platform, as we believe this will facilitate its introduction, at least for validation purposes. Recent work by Balding and Buckleton (2009) constitutes a good starting point for this effort, but a more complete, user-friendly solution is still required to initiate the use of the model in forensic casework. Despite this need, we focused on the "method evaluation" aspect of the problem. However, programming the model of Gill et al. (2000) into the open-source software developed during the course of this thesis work is planned for the near future.

Chapter 4

An open-source initiative for method evaluation in forensic genetics

Contents

4.1 Method evaluation in forensic genetics

Validation of analytical methods is an essential matter in forensic science because it is a prerequisite for the admissibility of evidence in court (Rudin and Inman, 2002). To achieve admissibility, forensic laboratories conduct validation studies on all analytical methods involved with the analysis of evidence.

Most forensic laboratories rely on standards defined by international bodies or organizations, such as the International Organization for Standardization (ISO) [1] or the U.S. Food and Drug Administration (FDA) [2]. For example, the FDA defines validation as "confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use are fulfilled".

Validation studies are generally conducted when a new technology is implemented in the laboratory (Kimpton et al., 1994; Sprecher et al., 1996; Moretti et al., 2001; Junge et al., 2003; Greenspoon et al., 2004). For example, when a new method for detecting blood stains is proposed by a vendor, a series of experiments are conducted to test the effectiveness and the reliability of the new method under conditions appropriate to the operational procedures in use in the laboratory. One of the first tests typically consists of testing the specificity of the method, i.e., its ability "to measure unequivocally and to differentiate the analyte(s) in the presence of components, which may be expected to be present" (Shah et al., 2000). In our example, this would consist of testing the ability of the new method to detect blood from a stain discovered at a crime scene, despite the presence of other components. Relying on the results of the testing procedures, laboratories specify standards that become the established practice for all future casework.

Historically, forensic DNA typing has largely been influenced by the Anglo-American judiciary system and its interactions with scientists over the past several decades. It is precisely in these judiciary systems where definitions of what qualifies as admissible DNA evidence have clearly been stated. In the U.S., it is the Daubert[3] standard that prevails. This standard states four criteria that can be used to assess the reliability of a given method involved in the production of scientific evidence:

1. Is the method testable and has it been tested?
2. Is it possible to quantify the error rate related to the method, and has it been determined?
3. Has the method been submitted to peer review in the relevant scientific community?
4. Has the method reached general acceptance in the relevant scientific community?

[1] `www.iso.org`
[2] www.fda.gov
[3] Daubert standards were articulated during a U.S. Supreme Court case in 1993: Daubert et al. vs. Merrell Dow Pharmaceuticals Inc., 509 US 579 1993.

Methods involved in forensic DNA typing have co-evolved with the dynamics of the validation process, permitting the progressive introduction of tools at the cutting edge of developments in molecular biology, as can be seen in the powerful genotyping system based on STR multiplexes in existence today. However, while validation studies have largely emphasized the reliability of analytical methods, little attention has been paid to validating the methods involved in statistical aspects of forensic DNA profiling (Perlin, 2006). As a consequence, developments dedicated to DNA evidence interpretation and, in particular, to DNA mixtures, have not encountered the expected success in the forensic community. There are two explanations for this: first, courtroom resistance to statistical methods has slowed down validation studies, and second, methods are not always accessible, i.e., they are either not implemented or not released.

Indeed, despite the plethora of tools dedicated to statistical analysis released as either free software, such as R (R Development Core Team., 2006), or commercial software, such as SAS®[4], there are very few software programs dedicated to forensic genetics (we provide a description of available software in the article below) in comparison to what is available in other disciplines, for example, forensic genetics.

In fact, major contributions in DNA mixture interpretation and, for instance, the statistical model we invoked in Chapter 3 have been implemented through the use of expert commercial software (Bill et al., 2005; Gill et al., 2007). For many forensic laboratories, it is not possible to invest in commercial software for evaluation purposes, so there is a significant cost issue. As a consequence, the evaluation of these methods based on real or simulated cases is often not possible unless software development is carried out, which is not necessarily within the scope of most forensic laboratories. Hence, following the Daubert criteria described above, these methods are not testable.

In addition to the inaccessibility of methods, another important issue is the cost associated with the evaluation procedure itself. For example, testing and evaluating new methods for mixture resolution implies that experimental mixtures have to be prepared under particular conditions, for example, with differing ratios of contributions or different marker types. These experiments can be expensive and time-consuming.

An appealing solution to these problems is computer simulations: the generation of thousands of realistic DNA profiles under a varying number of conditions presents no significant time or resource costs. However, as far as we know, there is no (monetarily) free software currently available that provides simulation tools specific to forensic genetics. Therefore, we have identified two needs: first, making available the methods of interest, and second, providing the necessary software tools to conduct evaluation studies. As follows below, we describe the forensim package, which we developed for implementation in the R statistical program to meet these needs.

[4] `www.sas.com`

4.2 Forensim: An open-source tool for method evaluation

The originality of the forensim package is that it provides both statistical methods devoted to the weight of DNA evidence and simulation tools to generate realistic datasets. Relying on object-oriented programming, new classes of objects are defined to represent and to simulate the main data types encountered in forensic genetics: allele frequencies, individual genotypes and DNA mixtures. The package also offers the most commonly used methods in the statistical interpretation of DNA evidence.

The following article describes the structure of the package and its main functionalities. Further information on forensim can also be found in the package documentation, which is provided in the Appendix accompanying this manuscript.

4.2.1 Article 4: "Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics"

Article in press in *Forensic Science International: Genetics*, 2010.

G Model
FSIGEN-593; No. of Pages 4
ARTICLE IN PRESS

Forensic Science International: Genetics xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics

Hinda Haned

Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de biométrie et biologie évolutive, 69622 Villeurbanne, France

ARTICLE INFO

Article history:
Received 13 November 2009
Received in revised form 8 February 2010
Accepted 20 March 2010
Available online xxx

Keywords:
Forensic sciences
DNA evidence
Statistical genetics
Simulation
R software
R-Forge

ABSTRACT

Forensim is a new package for the R statistical software that is dedicated to forensic DNA evidence interpretation. As far as we know, *forensim* is the first open-source tool that allows for the simulation of data encountered in forensic genetics studies. The package also implements common statistical methods used for reporting the weight of DNA evidence. *Forensim* is written in the R language and is freely available from http://forensim.r-forge.r-project.org. This paper presents an overview of the software's functionalities.

1. Introduction

Analysis of DNA samples for the investigation of criminal cases has been widespread since the "DNA revolution" of the early 1980s [1]. As a consequence, a theoretical framework for the statistical interpretation of DNA evidence has been well established. However, one of the most challenging tasks in forensic DNA analysis is the interpretation of DNA mixtures of several genotypes. This type of evidence is difficult to evaluate because individual components may not be separated, giving rise to questions about the origin of the stain [2].

Numerous statistical methods have been proposed to overcome this issue. Bayesian methods, in particular, seem to be very flexible and able to account for a wide variety of situations: population subdivision [3], relatedness between contributors to the evidentiary DNA stain [4], typing errors [5] and laboratory contamination [6]. However, these statistical developments have not been as successful as expected, mainly because of the difficulty in assessing their efficiency and degree of confidence that can be given to the results.

The efficiency of different methods should be tested with experimental DNA stains for which the circumstances of the hypothetical crime are known by the experimenter. The methods can then be evaluated according to their ability to accurately report the weight of the DNA evidence. However, such experimental validation is tedious and expensive in practice because new experiments must be conducted for each tested scenario. Computer simulations appear to be a satisfactory alternative because the generation of thousands of DNA profiles presents no significant time or resource costs. However, as far as we know, there is no free software currently available that provides simulation tools specific to forensic genetics.

The purpose of the *forensim* package for the free R statistical software [7] is to provide these tools. *Forensim* implements new classes of objects and functions devoted to the simulation of genetic data encountered in the evaluation of DNA evidence. The package also provides statistical methods dedicated to DNA evidence interpretation. This paper presents an overview of these functionalities.

2. Software overview

2.1. Simulation tools

Significant attention was devoted to facilitating the simulation of data in *forensim*. For instance, the package relies on object-oriented programming that allows for the definition of three object classes corresponding to the three main types of data commonly encountered in forensic casework: population allele frequencies, individual genotypes and mixed-DNA stains. Data

E-mail address: haned@biomserv.univ-lyon1.fr.

1872-4973/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2010.03.017

G Model
FSIGEN-593; No. of Pages 4
ARTICLE IN PRESS

simulation is achieved using specific functions, called constructors, which are named after the object class to which they are linked. A brief description of the object classes and their constructors follows.

2.1.1. Allele frequencies

Importation of existing allele frequencies is achieved using the `tabfreq` constructor. Input data must be in a particular format that is in widespread use among the forensic community: a matrix in which the first column gives the allele names, and the frequencies for a given allele are read in rows for different loci. The imported allele frequencies are stored in objects of class `tabfreq`. Useful information about the data is stored in different components, or slots, of this object. Each slot can be accessed via the `@` operator. For example, locus names are stored in the `which.loc` slot.

All *forensim* objects are described in the help pages. For instance, typing the command line `class?tabfreq` into an R console gives an exhaustive description of the `tabfreq` class along with examples of its usage.

When no pre-existing data are available, allele frequencies can be simulated for any number and type of loci, either in a homogeneous population, using the `simufreqD` function, or in a subdivided population, using the `simupopD` function. These functions are based on a Dirichlet model, whose parameters are determined by the user, to allow for control of the allele frequencies' means and variances. In subpopulations, the allele frequencies are modeled to deviate from the average values in the global population. The extent of this deviation is specified by the user. The use of the Dirichlet distribution to simulate population subdivision agrees with recent work on the subject [8,9].

2.1.2. Individual genotypes

Individual genotypes are simulated from allele frequencies using the `simugeno` constructor. At a given locus, an individual's genotype is simulated by randomly drawing two alleles (with replacement) at their respective allele frequencies in the target population. The simulated genotypes are independent between and within individuals. Useful information about the individuals, such as their names (stored in the `@indID` slot) and their population of origin (slot `@popind`), is also stored in the `simugeno` objects.

2.1.3. DNA mixtures

Mixtures are simulated from `simugeno` objects using the constructor `simumix`. The resulting mixtures are stored in `simumix` objects that contain two data types: (i) data that are usually available when dealing with real cases of forensic DNA mixtures—the mixture profiles, which are lists of the alleles observed for a given set of loci (slot `@mix.all`), and (ii) data that are usually not available–the number of contributors (slot `@ncontri`) and their genotypes (slot `@mix.prof`), which are known only if the mixture is simulated.

2.2. Statistical tools for DNA evidence interpretation

Forensim supplies the statistical methods used in the most critical steps in the process of evidentiary DNA interpretation: the determination of the number of contributors involved in the stain and the weight of the strength of the DNA evidence [2]. The number of individuals involved in a DNA stain can be determined based on allele counting using the `mincontri` function; this is often referred to as the maximum allele count method [10]. The package also includes an original method based on likelihood maximization (function `likestim` [11]).

Table 1
Summary of main statistical methods implemented in *forensim*.

Method	Function	Reference
Random man exclusion probability	`PE`	Clayton and Buckleton [2]
Random match probability	`RMP`	Balding and Nichols [12]
Likelihood ratios	`LR`	Curran et al. [3]
Conditional profile probability	`Pevid2`	Curran et al. [3]

Basic statistical methods used to report the weight of DNA evidence, such as the exclusion probability, are available in *forensim*. The package also includes more original methods, such as the general formula of likelihood ratios, which accounts for population subdivision [3]. A brief summary is given Table 1.

3. Other available software

Several commercial and open-source software programs have been proposed to deal with the problems raised by DNA evidence interpretation; a non-exhaustive review is given Table 2. It is notable that these programs implement specific methodologies for DNA evidence interpretation and were not originally intended for method evaluation. *Forensim's* contribution is that it offers a global methodological framework for method evaluation. The implemented statistical methods are an important basis for such evaluation. Interested users can implement their own methods in the R language and then use *forensim* object classes and functions to facilitate the evaluation.

4. Conclusion

Forensim aims to provide practical and open-source simulation tools, enabling the evaluation of statistical methods for DNA evidence interpretation. The software sources are available on R-Forge, which offers a central platform for the development of R packages. R-Forge also provides a variety of web-based collaborative tools, allowing several developers around the globe to work on the same project. Contributions from forensic scientists or scientists from other disciplines could be of great help in enhancing this open-source initiative.

The next step in *forensim* development is to increase user-friendliness. Multiple features of the software will be made accessible through a graphical interface that will make *forensim* more accessible to users not familiar with the R software, hopefully encouraging these users to contribute to the package.

Validation of *forensim* is also an important issue. There is no consensus on how software that combines simulation and statistical tools should be validated. The most basic level of validation is verification of the reproducibility of the results obtained by other software. This type of validation was done regularly throughout the programming process. Various examples of DNA evidence interpretation from Fung and Hu [18] were tested, and identical results were obtained. Some of these examples appear in the help pages, and they are also available in the software manual: http://forensim.r-forge.r-project.org/misc/forensim-manual.pdf. The accuracy of the results obtained by the statistical methods implemented in *forensim* was checked against two programs:

- the *DNAMIX* software, available at http://statgen.ncsu.edu/storey/;
- the *forensic* package for the R software, available at http://cran.r-project.org/web/packages/forensic/index.html.

This package implements the main statistical methods used to report the weight of DNA evidence. Similar functions are implemented in *forensim*, but the computational burden was

Please cite this article in press as: H. Haned, *Forensim*: An open-source initiative for the evaluation of statistical methods ?in forensic genetics, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.03.017

G Model
FSIGEN-593; No. of Pages 4
ARTICLE IN PRESS
H. Haned / Forensic Science International: Genetics xxx (2010) xxx–xxx 3

Table 2
Non-exhaustive review of open source and commercial software for forensic genetics. Note that these programs might evolve; thus, information about software features and availability is subject to change.

Software	Features					License	
	Allele calling	Crime case–DNA mixture	Kinship testing	Disaster victim identification	Simulation tools		
DNAMIX		×				Open source	http://statgen.ncsu.edu/storey/.
DNAVIEW		×	×	×		Commercial	http://dna-view.com/dnaview.htm
familias			×			Freeware	Egeland et al. [14]
FEST			×			Open source	Skare et al. [13]
forensic[a]		×				Open source	http://cran.r-project.org/web/packages/forensic
forensim		×			×	Open source	http://forensim.r-forge.r-project.org/
Fss-i3	×	×				Commercial	Bill et al. [15]
GeneMapper ID-X	×	×				Commercial	http://idx.appliedbiosystems.com/
GenoStat		×				Shareware	http://www.bioforensics.com/
Genoproof			×			Commercial	http://qualitype.de/genoproof
Genoproof Mixture	×	×				Commercial	http://qualitype.de/genoproofmixture
Grape		×	×			Commercial	http://dna-soft.com/
M-FISys		×	×	×		Commercial	http://www.genecodesforensics.com/software/
PCRSIM					×	Not released	Gill et al. [16]
TrueAllele	×	×		×		Commercial	Perlin [17]

[a] Similar functions are implemented in *forensim*, but the computational burden was significantly decreased.

significantly decreased thanks to the implementation of certain routines in the C programming language.

Further validation of the software could be achieved by encouraging forensic scientists to test different features of the software and to discuss their conclusions. The *forensim* website offers multiple tools to facilitate such interaction; for instance, a mailing list and a forum are accessible online. Increasing user-friendliness will certainly facilitate this step.

Forensim may become an important tool for method evaluation in forensic genetics. It has already been used in the investigation of the efficiency and robustness of the maximum-likelihood approach for DNA mixture resolution [11]. Because *forensim* is open source, more features can be added to meet the needs of forensic practitioners in situations that are not yet covered by the package.

Acknowledgments

I thank two referees for their thorough reviews and constructive comments. I am also grateful to Laurent Pène, Dominique Pontier and Frank Sauvage for their helpful comments. My thanks to R-Forge for hosting *forensim*.

Appendix A. Example

To illustrate some functionalities of *forensim* for forensic DNA evidence interpretation, we simulated a four-person DNA mixture and then interpreted it as an evidentiary DNA sample.

The package must be installed in the R environment; the procedure is described in the software tutorial available from: http://forensim.r-forge.r-project.org. Once the package has been installed, it must be loaded using the library command. Then, the strusa dataset, containing the allele frequencies for 15 Short Tandem Repeat Loci (STR) in three US populations [19], can be loaded (along with the command data). Hereafter, the command lines are displayed as they should be entered in the R console:

```
> library(forensim)
> data(strusa)
```

To allow other users to reproduce the simulations in this example, the random seed is set:

```
> set.seed(123560)
```

Individual genotypes can easily be simulated using the `simugeno` constructor, which can be entered into the command line as

```
> geno <- simugeno(strusa,n=c(0,1000,0))
```

This command simulates a thousand genotypes from the Caucasian population (the second population in strusa). Then, mixtures of any number of contributors can be simulated; here we simulate a four-person mixture:

```
> mix4 <- simumix(geno,ncontri=c(0,4,0))
```

Now, we assume that this simulated mixture was in fact recovered from a crime scene and that the real number of contributors is unknown. The only information available in this case is the alleles present at each locus. Note that only qualitative data is handled for the moment. For instance, the locus "D21S11" shows the alleles:

```
> mix4@mix.all[ ''D21S11'']
$D21S11
[ 1]   ''28''   ''29''   ''30''   ''31.2''   ''32.2''
```

The forensic expert analyzing this DNA evidence wants to answer the question: "How many people were involved in this stain?" To this end, he chooses to compare two methods: first, the maximum allele count, considering all available STR loci simultaneously:

```
> mincontri(mix4)
[ 1] 3
```

and second, a maximum-likelihood estimation of the number of contributors [11]. For this estimate, the expert assumes that all people involved in the stain are unrelated and come from the American Caucasian population:

```
> likestim(mix4,freq=strusa,refpop=''Cauc'')
                        max                    maxval
[ 1,]                   4                      8.7e-29
```

Note that the allele frequencies to be used in the calculation must be specified (using the argument refpop). Maximizing the likelihood gives the correct estimate of the number of contributors (max, with a likelihood value of maxval), while the maximum allele count underestimates the number of contributors but still gives an informative lower bound.

More complex scenarios can also be considered. For example, one can simulate DNA stains comprising the DNAs of several individuals from

Please cite this article in press as: H. Haned, *Forensim*: An open-source initiative for the evaluation of statistical methods ?in forensic genetics, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.03.017

G Model
FSIGEN-593; No. of Pages 4
ARTICLE IN PRESS

different subpopulations with varying degrees of subdivision (using the function simupopD and the argument alpha1) and evaluate the importance of taking population subdivision into account in the calculations.

References

[1] M.A. Jobling, P. Gill, Encoded evidence: DNA in forensic analysis, Nat. Rev. Genet. 5 (2004) 739–751.
[2] T. Clayton, J. Buckleton, Mixtures, in: J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, 2005, pp. 217–274.
[3] J.M. Curran, C.M. Triggs, J. Buckleton, B.S. Weir, Interpreting DNA mixtures in structured populations, J. Forensic Sci. 44 (1999) 987–995.
[4] W.Q. Fung, Y.Q. Hu, Interpreting DNA mixtures with related contributors in subdivided populations, Scand. J. Statist. 31 (2004) 115–130.
[5] W.C. Thompson, F. Taroni, C.G. Aitken, How the probability of a false positive affects the value of DNA evidence, J. Forensic Sci. 48 (2003) 47–54.
[6] J. Buckleton, P. Gill, Low copy number, in: J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, 2005, pp. 275–297.
[7] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0, available at http://www.R-project.org, 2008.
[8] G. Nicholson, A.V. Smith, F. Jónsson, O. Gústafsson, K. Stefánsson, P. Donnelly, Assessing population differentiation and isolation from single-nucleotide polymorphism data, J. R. Stat. Soc. B 64 (2002) 695–715.
[9] J. Marchini, L.R. Cardon, Discussion on the meeting on "Statistical modelling and analysis of genetic data", J. R. Stat. Soc. B 64 (2002) 740–741.
[10] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (2005) 1361–1366.
[11] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic. Sci. (2011), in press.
[12] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic. Sci. Int. 64 (1994) 125–140.
[13] Ø. Skare, N. Sheehan, T. Egeland, Identification of distant family relationships, Bioinformatics 25 (2009) 2376–2382.
[14] T. Egeland, P.F. Mostad, B. Mevåg, M. Stenersen, Beyond traditional paternity and identification cases. Selecting the most probable pedigree, Forensic Sci. Int. 110 (2000) 47–59.
[15] M. Bill, P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy, J. Buckleton, PENDULUM—a guideline-based approach to the interpretation of STR mixtures, Forensic Sci. Int. 148 (2005) 181–189.
[16] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2005) 632–643.
[17] M.W. Perlin, in: M.I. Okoye, C.H. Wecht (Eds.), Identifying Human Remains Using TrueAllele® Technology. Forensic Investigation and Management of Mass Disasters, Lawyers & Judges Publishing Co., Tucson, AZ, 2007 , pp. 31–38.
[18] W.K. Fung, Y.Q. Hu, Statistical DNA Forensics: Theory, Methods and Computation, John Wiley & Sons, Ltd., 2008.
[19] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline, Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations, J. Forensic Sci. 8 (2003) 908–911.

Please cite this article in press as: H. Haned, *Forensim*: An open-source initiative for the evaluation of statistical methods ?in forensic genetics, Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.03.017

4.2.2 Why the R software?

This past decade, the free R software for statistical computing and graphics (R Development Core Team., 2006) has become an indispensable tool for statistical analysis in areas as diverse as ecology, genetics, pharmacology, genomics and econometrics (to name a few). The software presents several advantages among which several built-in mechanisms for representing information. Thus, creating/importing data, running calculations and representing important information using graphics is straightforward. Another advantage of R is that it is a free software. Free software guarantee to users four essential freedoms[5]:

1. the freedom to run the software for any purpose (for instance, for commercial purposes),
2. the freedom to study how the program works, this implies that the software sources are accessible to everyone (they must be open-source),
3. the freedom to redistribute copies of the programme, for any purpose (even commercial),
4. the freedom to distribute copies of modified versions of the software to other users.

R is thus not only accessible free of charge[6], but it is also "open-source", i.e., its sources are accessible and can be modified by anyone. In addition, R has a community of users that support each other and collaborate on the development of the software. Thanks to this community, the software is continually supplemented with packages containing codes for varying topics.

The R-Forge platform was built to support this community and help developers and users collaborate around the software. The most attractive feature of R-Forge is the ability to work with several programmers on the same project. Forensim source being hosted by R-Forge, contributions from forensic scientists or scientists from other disciplines could be of great help in improving the package.

4.3 Perspectives and future developments

As any other R package, forensim is constantly evolving. Currently available tools aim at facilitating methods evaluation, but hopefully, new features and improvements will be made. For example, an important issue is the handling of quantitative data. As stated in the article above, only qualitative data is handled for the moment. Future developments will include the modifications of the main classes to include the importation and the generation of quantitative data from genotyping software such as GeneMapper IdX (Applied Biosystems).

Another important improvement that can be made to the package consists in making the software more user-friendly. We have already started this effort by making available a number of features in the form of user-friendly graphical interfaces: these include simulation modules

[5]These criteria were defined by the Free software Foundation, who pioneered the free software movement in 1985 , see `www.fsf.org`.

[6]This is not always the case for free software.

(see functions *simPCRTK*, and *Hbsimu*) and a module implementing a mixture deconvolution method (function *mastermix*, Gill et al., 1998b).

Since its release on R-Forge in April 2009, there have been several efforts to introduce the forensim package in the forensic community. A few months month after its release, forensim was presented at the 37th European DNA Profiling Group of the International Society for Forensic Genetics (ISFG). The ISFG members demonstrated an important interest to this initiative, and they now encourage similar projects. As an evidence of this enthusiasm, the ISFG website now includes a section entitled "Free software resources"[7].

In conclusion, forensim is certainly not a perfect solution to the multiple problems raised by methods validation, however, the unique background offered by the R software lets us foresee the possibility of making of forensim a platform centralizing all validation efforts in varying topics related to forensic genetics. Such project already exists in R for other disciplines, for example the "Bioconductor" project (Gentleman et al., 2004) centralizes a very large number of tools dedicated to analysing genomic data. The future platform would include packages dealing with different topics such as relationship testing, disaster victim identification and DNA database search. Hopefully, this will cause a dynamic validation of several methods, and possibly allow their introduction into forensic casework.

[7]http://www.isfg.org/Software

Chapter 5

Conclusions

As we have now presented the questions addressed during the course of this thesis work and the solutions that we proposed for them, we will review our results to identify some of the remaining issues and discuss our research perspectives.

Summary of thesis results

The first issue addressed in this thesis concerned a recurring problem in forensic casework: determining the number of contributors to DNA mixtures. Surprisingly, methods dedicated to DNA evidence interpretation have not often considered this problem, probably because two-person mixtures constitute the main mixture type encountered in casework. Another reason for this is that forensic DNA typing generally involves comparisons between samples, thereby reducing the difficulty of interpretation. However, there is an increasing number of cases in which DNA samples are recovered from crime scenes but no reference sample is available.

Determining the number of contributors to a DNA mixture can be relevant at any stage of the interpretation process. However, there is currently no established method for carrying out the estimation of this number, apart from the rudimentary allele counting method. This led us to propose a method to estimate the number of contributors based on qualitative data. We showed that this method can be efficient, in particular for complex mixtures of more than two contributors, and we also proposed a simple method to quantify the efficiency of the estimator.

The second point addressed in this thesis focused on situations in which the interpretation of DNA evidence is challenging. We were particularly interested in methods dedicated to estimating allelic drop-out. Accounting for allelic drop-out or other anomalies that may affect DNA profiles is essential to avoid biased results or misinterpretations. Gill et al. (2000) have proposed an elegant solution for this problem, but one aspect that has received little attention in the forensic literature is the estimation of the model parameters. Because our focus was on evaluation, we adopted a simulation approach to investigate a number of features of a model proposed by Gill et al. (2009), which relies on the peak heights observed in a DNA profile to estimate drop-out probabilities.

Testing and evaluating models is based on the use of data. Therefore, we were committed to investigating the accuracy of the current experimental approach used to generate "drop-out data". Our investigations showed that these techniques are only valid for haploid data. It was particularly important for us to convey this message, as is demonstrated in Chapter 3 in which we showed through simulations that the tested model tends to give biased results if used with haploid data. However most importantly, we illustrated how model evaluation can be conducted using simulations.

In parallel to our work on methodological aspects, we addressed the question of the implementation of method evaluation. It is difficult to conceive an evaluation effort without software that makes the methods in question accessible. However, methods of interest must not only be made available but also testable. The forensim package was developed to answer both of these conditions by offering key methods involved in the weighting of DNA evidence, along with new object classes adapted to generate and handle data that is commonly encountered in forensic casework. The package has been constantly evolving ever since its first release: existing programs are regularly improved and new features are made accessible through user-friendly graphical interfaces.

Our approach related to method evaluation led us to address key issues related to the interpretation of DNA evidence both in terms of methodology (how to evaluate existing methods?) and in terms of practice (what tools should be used for this evaluation?).

A review of existing practices in terms of the interpretation of DNA profiles in general allowed us to identify several open questions. Still, certain issues appeared to us as more urgent than others, especially in light of our dialogue with forensic scientists from the national forensic laboratory of Lyon.

These issues concerned two stages: the first one is the investigation stage, where investigators search for information and clues to find the perpetrator of a crime; the second one is the trial stage where one or more individuals are suspected of having committed a crime and are judged by a court, at the light of information that were collected during the investigation stage.

Our first concern was that certain practices needed improvement, especially with respect to DNA samples for which no suspects or any reference samples are available. These cases take on increasingly important roles in the volume of casework processed by forensic laboratories around the world, and it is more than legitimate to present the question of their exploitation by forensic scientists and investigators. We believe we have contributed to this question by proposing an estimator of the number of contributors to DNA mixtures. Although this estimate might not necessarily find its place in the courtroom, it might still assist forensic scientists and investigators in their work.

Concerning the methods used for weighting DNA evidence, we were concerned with their introduction into forensic laboratories and, in particular, the introduction of a statistical model for DNA evidence interpretation (Gill et al., 2000). We observed that some issues remain unsolved, such as that the estimation of the model parameters has not been dealt with as well

as it should in the literature. We have thus proposed the use of simulation models to evaluate existing methods to estimate these parameters and to stimulate a discussion about the possible alternatives in this matter. This work, presented in Chapter 3, can be considered as an illustration of an evaluation procedure based on computer simulations, which can be adopted by any forensic laboratory since the underlying methods are implemented in an open-source software.

While the investigation phase is common to all judiciary systems, the trial phase does not necessarily take place. Indeed, a case is not necessarily brought before court. For example in the U.S., a suspect can plead guilty, and his sentence is not debated before a jury[1]. In this case, the weight of scientific evidence is not debated between prosecution and defence parties. In any case, statistical methods are still needed during the investigation phase. Hence, our contribution addresses questions that can be encountered in different judiciary systems.

The last element that we wish to raise in this general assessment is the positive feedback we had from many forensic scientists. Indeed, it is notable that during this thesis work, one of the first educational workshops dedicated to open-source software solutions for DNA mixtures interpretation was held under the auspices of the International Society for Forensic Genetics (ISFG). Feedback received from forensic scientists from different European laboratories show that our methodology and its implementation respond to a real need in forensic genetics.

Research perspectives

The work presented here addresses several open questions, but a number of additional issues remain unsolved. The main unsolved issue that must be addressed in the future is the exploitation of quantitative data. These data can provide important additional information with respect to qualitative data, i.e., the list of alleles in a DNA profile. Indeed, as we explained in the first chapter, there is a relationship between alleles' peak heights and post-PCR DNA quantities, and also between peaks heights and (post-PCR) mixture ratios of contribution. Logically, including such data into statistical methods should improve their efficiency. For instance, in mixture resolution, quantitative data could lead to calculating the probability of observed peak heights conditional on a given genotypic combination. Contributions on the subject either suggested the use of a Gaussian (Evett et al., 1998) or a Gamma model (Cowell et al., 2007b), but somehow failed to justify the use of such distributions on an experimental basis.

We thereby intend to explore the most adequate probabilistic distribution for alleles peak heights. To achieve this work, we will rely on quantitative data, either generated from laboratory experiments or from the numerous data sets that many forensic laboratories collect from control DNA samples used in their routine work.

Of course, a key issue here is the experimental design. We did not consider this question in this thesis beyond proposing a simulation model for helping experimenters during this sensitive

[1]Federal rules of criminal procedure, `http://www.law.cornell.edu/rules/frcrmp/Rule11.htm`.

step, but it is obvious to us that this issue needs to be considered. We believe that simulation models will have an important role to play in this process.

Future results from this research will hopefully lead us to solve several issues in forensic DNA interpretation, among which, DNA mixtures resolution. Of course, the estimator of the number of contributors proposed in Chapter 2 will benefit from such developments, as less plausible genotypes can be eliminated from the likelihood function. Another important step in this work will be the quantification of the impact of incorporating quantitative data in weighting DNA evidence, i.e., what is the amount of additional information provided by such data? We believe this is an important question, especially because it is likely to be discussed in court.

Another issue to address is the definition of a more accurate limit of the detection threshold for declaring allelic drop-out. A commonly employed threshold is 50 RFUs, which indicates whether a signal can be referred to as an allele or as background noise (thus implying that a drop-out can be declared), with a variation of a single RFU. This is clearly problematic. Because the underlying data are continuous, there is no reason why a drop-out should be declared at 49 RFUs and not at 50 RFUs.

Thresholds definitions are essential to analysing DNA evidence, without them, true signals cannot be differentiated from background noise. Still, their definition needs to be improved. We believe that the way forward is to develop a measure of the risk associated with these thresholds. Regardless of the threshold chosen to define drop-out, a measure of the risk associated with such decision is calculated. Obviously, this measure must rely on quantitative data, which contain information about a profile quality. Gill et al. (2009) proposed the use of a graphical model to answer this question, based on the logistic model we studied in Chapter 3.

Combining the results from these investigations with information gained from model evaluation from Chapter 3 will hopefully result in a rigorous statistical model for estimating drop-out probabilities.

Open questions

Throughout the manuscript, we evoked the limits of our work in the discussion sections in the chapters, and in our future research perspectives above. But there are two remaining questions that we wish to extend on in this general conclusion: first, the limits of free software, and the discussion of the concept of validation.

Software limitations

While open-source software offers to the users the freedom "to run, copy, distribute, study, change and improve the software"[2], non open-source software significantly limits the users freedom, by imposing data types, inputs formats, methods, procedures, that do not always correspond nor

[2] `www.fsf.org`

satisfy the users needs. Obviously, it is not realistic to develop a software that entirely satisfies all potential users. Still, non open-source software limits the possibilities offered to the users, who cannot access the software sources to adapt them to their needs. Thus, software that is used to facilitate analysis or research can quickly turn to a straitjacket that inhibits ideas or developments that do not correspond to the possibilities offered by the software.

This seems inevitable in non open-source software, but free software are not completely immune from these constraints. Indeed, some users are not necessarily able to explore all possibilities offered by a free software. For example, if a user wants to change how mixtures are simulated in forensim, he must have good knowledge of R and some programming skills. Hence, even free software have their limitations. It is thus legitimate to wonder whether novice programmers can explore all of forensim possibilities, the answer is a clear "no". Still, we believe that the package provides essential functions and simulation tools, which can easily be used and manipulated, notably thanks to the documentation provided with the package (see Appendix section).

The concept of validation

The discussion about software limitations leads us to question ourselves about the limits of our work. Validation of bioanalytical methods has been widely discussed, and there are several standards established by international bodies such as the FDA or the ISO that give helpful guidelines for testing laboratories. But discussing methods or model validation is a quite different task that raises philosophical questions. The subject is rich and is still debated between those who think validation as essential, and those for whom only invalidation is reasonable (see Rykiel, 1996 for a review of the contradictory opinions on the matter). It would be somewhat inappropriate to discuss here in detail the concept of validation: this is beyond the scope of this thesis. Still, whatever definition is given to validation, it cannot take place unless the method is accessible: only available methods are questionable and testable. Because the implemented procedures only represent a transcription of statistical methods, laboratories that use software like forensim will be testers of the implemented methods. Making our developments available in open-source software is faithful to this principle, and we therefore believe we have provided useful tools enabling method evaluation and thus paving the way to method validation.

From open software to open community

Future improvements in DNA typing will yield new challenges and issues in forensic DNA typing, and the need for freely accessible statistical tools can no longer be neglected. We demonstrated two main factors explaining the difficulty of introducing statistical reasoning in forensic genetics, the first of which is courtroom resistance to probabilistic reasoning and the second the lack of tools allowing for the evaluation of existing methods. Another explanation is the perception of DNA evidence, which has long been considered a flawless form of scientific evidence by the

general public, but also, by a number of scientists (Aronson, 2007). As a consequence, there is a lack of funding to support the improvement of current practices or to introduce advanced statistical tools. For this reason, we believe that open-source software is an appealing solution. In addition to being fully transparent about the content of such software, users can access these programs, modify them and even redistribute them.

The democratization of software tools will allow for constructive discussions on issues raised by DNA mixtures or low template DNA. Indeed, laboratories around the world could pool their data and their results, as in inter-laboratory exercises. Such exercises are already conducted under the auspices of the ISFG (see for example, an inter-laboratory exercise on mitochondrial DNA markers, Prieto et al., 2008). We do not see why the same type of exercises could not be applied to statistical methods, especially as the matter of their accessibility is now answered by an open-source solution.

Another advantage of this democratization, initiated by open-source software, is the opportunity for scientists from diverse communities to discuss questions encountered in their disciplines. For example, the problem of determining the number of contributors to DNA mixtures does not arise only in forensic science. Besides, the forensim package is already visible to other communities. For example, the five first occurrences yielded by a search for R packages related to the key word "DNA" on R website[3] are:

1. package RFLP, dedicated to the analysis of data generated from RFLP analysis
2. package forensic, earliest package in R dedicated to weighting DNA evidence
3. package ape, dedicated to phylogenetic trees
4. package cgh, for micro-arrays analysis
5. package forensim

Although we have not focused on the matter of extrapolating our work outside of a forensic setting, we believe that the issues addressed in this thesis may also be relevant to non-forensic applications. For instance, the matter of estimating drop-out probabilities is encountered in ecological studies in conservation genetics and in behavioral ecology studies in which non-invasive genotyping is performed (Taberlet et al., 1996; Gagneux et al., 1997; Pompanon et al., 2005; Broquet et al., 2007). Another discipline where such difficulties can also be encountered is the analysis of ancient DNA from specimens stored in museums or from bones (Gill et al., 1994; Schmerer et al., 1999; Wandeler et al., 2007).

DNA samples obtained from diverse biological materials such as shed hairs, are often recovered in limited quantity and are thus prone to anomalies such as drop-out, drop-ins and stutters. The simulation framework for testing different models for the estimation of drop-out probabilities,

[3]The Comprehensive R Archive Network (CRAN), `http://cran.r-project.org/`, September 1st, 2010

presented in Chapter 3 can be extended to such situations and may prove helpful in choosing the most accurate variables in the prediction models. Similarly, the need for an estimate of the number of contributors to DNA mixtures can emerge in non-forensic settings. For example, samples of shed hairs, faeces and urine, may contain biological material from different animals, and hence constitute a mixture. Obtaining an estimate of the number of contributors may be helpful for biological studies based on these samples. Moreover, the great flexibility of the simulation tools provided in forensim ensures that simulation models can be applied to genetic markers of different types. For example, single-nucleotide polymorphism markers are commonly used in studies related to ancient DNA. Thus, the simulation framework can easily be extended to model situations that are different from those presented here.

In fact, as we were going alone with the package development, we realized that the scope of the software exceeded our original intentions, since it dealt with a problematic that is commonly encountered in applied sciences. Thus, both the topics addressed and the open-source solution we developed may be relevant to different disciplines and to scientists from different backgrounds. Hopefully, open software will lead to an open community that will be the scene of interesting interactions.

Towards "Bayesian justice"

Although we have addressed somewhat different topics in this work, we have been consistently motivated by the urgent need to introduce advanced statistical tools for the interpretation of DNA evidence. In this regard, we believe that the Bayesian approach based on likelihood ratios is the best way to improve mixture interpretation.

Probabilistic reasoning is already in use in many courtrooms, but not always under their optimal form, i.e., through the calculation of likelihood ratios. We previously explained that this is due to courtrooms reluctance to such logical reasoning.

Still, probabilistic reasoning is essential in forensics because this science deals with complex objects that are subject to a number of uncertainties about their origin. For example, the origin of a DNA evidence cannot be known with certainty even in the case of a match between a suspect profile and a crime scene profile. Typically a suspect could have left some biological material during an occasion unrelated to the crime itself, or another person with the same profile could have left the stain. Hence, the language of probability is meant to translate the uncertainties about the origin of a DNA stain into numerical statements. These statements must be of course comprehensible by judges and juries. Still, statistical methods should not be dismissed because of their complexity, and a compromise should be found between the comprehension level of non-specialists and the need for sound statistical reasoning in court. We are aware that this question is not easy, and would probably require years of collaborative work between forensic scientists, mathematicians, cognitive psychologists, judges, juries, etc.

Any improvement in this field can hardly be achieved if methods of interest are not available. We believe that our work can pave the way to the introduction of Bayesian methods in courtrooms, by allowing many forensic scientists to appropriate these methods, by testing, manipulating, evaluating, and may be even validating them for casework. Hopefully, this will facilitate conveying the message that probabilistic reasoning is essential to assess scientific evidence, and ultimately, that the Bayesian approach is the most adapted for evidence interpretation.

Another key step to achieve resides in changing the perception of DNA evidence itself. DNA is thought of as a gold standard, or as flawless scientific evidence. We believe that this perception cannot be changed as long as DNA is considered as an identification tool, instead of an intelligence tool that can lead to find the perpetrator of a crime, in combination with other scientific evidence. Of course, if forensic scientists have the tools to prove that DNA evidence is neither flawless nor unreliable, they can further contribute to find a compromise between methods complexity and juries understanding.

The problem of method evaluation is not completely resolved, since open questions remain. Nevertheless, given the needs expressed by the forensic community and the positive feedback regarding the availability of an open-source software solution, it is very likely that the tools necessary for "Bayesian justice" will be gradually introduced into forensic practice.

Bibliography

C. Aitken and F. Taroni. *Statistics and the evaluation of evidence for forensics scientists.* John Wiley & Sons Inc, 2004.

J. Aronson. *Genetic witness: science, law, and controversy in the making of DNA profiling.* Rutgers University Press, 2007.

D. Balding and J. Buckleton. Interpreting low template DNA profiles. *Forensic science international. Genetics*, 4(1):1–10, 2009.

D. J. Balding and R. A. Nichols. DNA profile match probability calculation: how to allow for population stratification, relatedness, databse selection and single bands. *Forensic Science International*, 64:125–140, 1994.

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

T. Broquet, N. Menard, and E. Petit. Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates. *Conservation Genetics*, 8(1):249–260, 2007.

J. Buckleton, C. M. Triggs, and S. J. Walsh. *Forensic DNA evidence interpretation.* CRC PRESS, 2005.

J. S. Buckleton and J. Curran. A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Science International: Genetics*, 2:343–348, 2008.

J. S. Buckleton, J. M. Curran, and P. Gill. Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Science International: Genetics*, 1:20–28, 2007.

B. Budowle, D. Hobson, J. Smerick, and J. Smith. Proceedings of the Twelfth International Symposium on Human Identification. In *Available at http://www.promega.com/geneticidproc/ussymp12proc/contents/budowle.pdf.*, 2001.

B. Budowle, M. Bottrell, S. Bunch, R. Fram, D. Harrison, S. Meagher, C. Oien, P. Peterson, D. Seiger, M. Smith, et al. A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *Journal of Forensic Sciences*, 54(4):798–809, 2009a.

B. Budowle, A. Eisenberg, and A. Van Daal. Validity of low copy number typing and applications to forensic science. *Croatian medical journal*, 50(3):207–17, 2009b.

B. Budowle, A. Eisenberg, and A. van Daal. Low copy number typing has yet to achieve "general acceptance". *Forensic Science International: Genetics Supplement Series*, 2(1):551–552, 2009c.

B. Budowle, A. J. Onorato, T. F. Callaghan, A. Della Manna, A. M. Gross, R. A. Guerrieri, J. C. Luttman, and D. L. McClure. Mixture Interpretation: Defining the Relevant Features for Guidelines for the Assessment of Mixed DNA Profiles in Forensic Casework. *Journal of Forensic Sciences*, 54(4):810–821, 2009d.

B. Budowle, A. J. Eisenberg, and A. van Daal. Response to Comment on Low copy number typing has yet to achieve general acceptance (Budowle et al., 2009. Forensic Sci. Int. Genetics: Supplement Series 2, 551-552) by Theresa Caragine, Mechthild Prinz. *Forensic Science International: Genetics*, In Press, 2010.

J. Butler. *Forensic DNA typing*. Academic Press London, 2001.

J. Butler, R. Schoske, M. Vallone, J. W. Redman, and M. C. Kline. Allele frequencies for 15 autosomal str loci on u.s. caucasian, african american, and hispanic populations. *Journal of Forensic Sciences*, 48(8):908–911, 2003.

B. Caddy, G. Taylor, and A. Linacre. A review of the science of low template DNA analysis. Technical report, Office of the Forensic Regulator, 2008.

R. Chakraborty and L. Jin. Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Human genetics*, 88(3):267–272, 1992.

R. Chakraborty and K. Kidd. The utility of DNA typing in forensic work. *Science*, 254(5039): 1735–1735, 1991.

T. Clayton and J. Buckleton. *Forensic DNA evidence interpretation*, chapter Mixtures, pages 217–239. CRS PRESS, 2005.

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

J. Cohen. DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. *American Journal of Human Genetics*, 46(2):358, 1990.

R. Coquoz and F. Taroni. *Preuve par l'ADN: la génétique au service de la justice*. PPUR presses polytechniques, 2006.

R. Cowell. Validation of an STR peak area model. *Forensic Science International: Genetics*, 3 (3):193–199, 2009.

R. Cowell, S. Lauritzen, and J. Mortera. Identification and separation of DNA mixtures using peak area information. *Forensic Science International*, 166(1):28–34, 2007a.

R. Cowell, S. Lauritzen, and J. Mortera. A gamma model for DNA mixture analyses. *Bayesian Analysis*, 2(2):333–48, 2007b.

J. Curran. A MCMC method for resolving two person mixtures. *Science &Justice*, 48(4):168–177, 2008.

J. Curran. Statistics in forensic science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2):141–156, 2009.

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.

J. M. Curran, P. Gill, and M. R. Bill. Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure. *Forensic Science International*, 148:47–53, 2005.

B. Devlin and N. Risch. Ethnic differentiation at VNTR loci, with special reference to forensic applications. *American Journal of Human Genetics*, 51(3):534, 1992.

B. Devlin, N. Risch, and K. Roeder. No excess of homozygosity at loci used for DNA fingerprinting. *Science*, 249:1416, 1990.

E. Doom and M. L. Raymer. Analysis of allele sharing and deconvolution of mixed DNA samples. In *Forensic Bioinformatics, 3rd Annual conference: DNA from crime scene to court room: An expert forum*, 2004.

A. Edwards, A. Civitello, H. Hammond, and C. Caskey. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics*, 49(4):746, 1991.

T. Egeland, I. Dalen, and P. F. Mostad. Estimating the number of contributors to a DNA profile. *International Journal of Legal Medicine*, 117:271–275, 2003.

E. Essen-Möller. Die beweiskraft der hnlichkeit im vatershaftsnachweis teoretische graundlagen. *Mitteilung Antropologishe Graundlagen*, 68:9–53, 1938.

I. Evett and B. Weir. *Interpreting DNA evidence: statistical genetics for forensic scientists*. Sinauer Sunderland, 1998.

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

I. Findlay, A. Taylor, P. Quirke, R. Frazier, and A. Urquhart. DNA fingerprinting from single cells. *Nature*, 389:556–556, 1997.

C. Frégeau, K. Bowen, B. Leclair, I. Trudel, L. Bishop, and R. Fourney. AmpFlSTR Profiler Plus short tandem repeat DNA analysis of casework samples, mixture samples, and nonhuman DNA samples amplified under reduced PCR volume conditions (25 microL). *Journal of Forensic Sciences*, 48(5):1014–1034, 2003.

N. Fukshansky and W. Bär. Biostatistical evaluation of mixed stains with contributors of different ethnic origin. *International Journal of Legal Medicine*, 112(6):383–387, 1999.

W. K. Fung and Y. Q. Hu. Interpreting Forensic DNA Mixtures: Allowing for Uncertainty in Population Substructure and Dependence. *Journal of the Royal Statistical Society*, 163(2): 241–254, 2000.

P. Gagneux, C. Boesch, and D. Woodruff. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Molecular Ecology*, 6(9): 861–868, 1997.

R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

N. Gilbert. DNA'S identity crisis. *Nature*, 464(7287):347–348, 2010.

J. R. Gilder. *Computational methods for the objective review of forensic DNA testing results*. PhD thesis, Wright State University, Ohio, 2007.

P. Gill. Application of low copy number DNA profiling. *Croatian Medical Journal*, 42(3):229–232, 2001.

P. Gill. The presentation and interpretation of DNA evidence. In *Workshop in Forensic Genetics, University of Oslo, Norway*, 2009.

P. Gill and J. Buckleton. Low copy number typing–where next? *Forensic Science International: Genetics Supplement Series*, 2:553 – 555, 2009. Progress in Forensic Genetics 13 - Proceedings of the 23rd International ISFG Congress.

P. Gill and J. Buckleton. A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number. *Forensic Science International: Genetics*, 4(4):221–227, 2010.

P. Gill, A. Jeffreys, and D. Werrett. Forensic application of DNAfingerprints. *Nature*, 318(6046): 577–579, 1985.

P. Gill, P. Ivanov, C. Kimpton, R. Piercy, N. Benson, G. Tully, I. Evett, E. Hagelberg, and K. Sullivan. Identification of the remains of the Romanov family by DNA analysis. *Nature Genetics*, 6(2):130–135, 1994.

P. Gill, B. Sparkes, and B. J.S. Interpretation of simple mixtures of when artefacts such as stutters are present - with special reference to multiplex STRs used by the Forensic Science Service. *Forensic Science International*, 95:213 – 224, 1998a.

P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton. Interpreting simple STR mixtures using allele peak areas. *Forensic Science International*, 91:41–53, 1998b.

P. Gill, J. Whitaker, C. Flaxman, N. Brown, and J. Buckleton. An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International*, 112(1):17–40, 2000.

P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayer, 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. *Forensic Science International*, 160(2-3):90–101, 2006.

P. Gill, A. Kirkham, and J. Curran. LoComatioN: A software tool for the analysis of low copy number DNA profiles. *Forensic Science International*, 166(2-3):128–138, 2007.

P. Gill, R. Puch-Solis, and J. Curran. The low-template-DNA (stochastic) thresholdIts determination relative to risk analysis for national DNA databases. *Forensic Science International: Genectis*, 3(2):104–111, 2009.

S. Greenspoon, J. Ban, L. Pablo, C. Grouse, F. Kist, C. Tomsey, A. Glessner, L. Mihalacki, T. Long, B. Heidebrecht, et al. Validation and implementation of the PowerPlex® 16 BIO System STR multiplex for forensic casework. *Journal of Forensic Sciences*, 49(1):71–80, 2004.

Y. Q. Hu and W. Q. Fung. Evaluation of DNA mixtures involving two pairs of relatives. *International Journal of Legal Medicine*, 119(5):251259, 2005.

A. Jeffreys. 1992 William Allan Award address. *American journal of human genetics*, 53(1):1, 1993.

A. Jeffreys, V. Wilson, and S. Thein. Individual-specific fingerprints of human DNA. *Nature*, 316:76–79, 1985a.

A. Jeffreys, V. Wilson, S. Thein, et al. Hypervariable minisatellite regions in human DNA. *Nature*, 314(6006):67–73, 1985b.

M. A. Jobling and P. Gill. Encoded evidence: DNA: in forensic analysis. *Nature Reviews Genetics*, 5:739–752, 2004.

P. C. Johnson and D. T. Haydon. Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics*, 175: 827–842, 2007.

A. Junge, T. Lederer, G. Braunschweiger, and B. Madea. Validation of the multiplex kit genRESMPX-2 for forensic casework analysis. *International journal of legal medicine*, 117 (6):317–325, 2003.

C. Kimpton, D. Fisher, S. Watson, M. Adams, A. Urquhart, J. Lygo, and P. Gill. Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. *International Journal of Legal Medicine*, 106(6):302–311, 1994.

E. Lander. DNA fingerprinting on trial. *Nature*, 339(6225):501–505, 1989.

E. Lander. Research on DNA typing catching up with courtroom application. *American Journal of Human Genetics*, 48(5):819, 1991.

E. Lander and B. Budowle. DNA fingerprinting dispute laid to rest. *Nature*, 371(6500):735–738, 1994.

R. Lewontin and D. Hartl. Population genetics in forensic DNA typing. *Science*, 254(5039): 1745, 1991.

C. R. Miller, P. Joyce, and L. P. Waits. Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, 160:357–366, Jan 2002.

T. Moretti, A. Baumstark, D. Defenbaugh, K. Keys, J. Smerick, and B. Budowle. Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. *Journal of Forensic Sciences*, 46(3):647–660, 2001.

J. Mortera, A. P. Dawid, and S. L. Lauritzen. Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, 63:191–205, 2003.

N. Morton. Genetic structure of forensic populations. *Proceedings of the National Academy of Sciences of the United States of America*, 89(7):2556, 1992.

D. R. Paoletti, T. E. Doom, C. M. Krane, M. L. Raymer, and D. E. Krane. Empirical analysis of the STR profiles resulting from conceptual mixtures . *Journal of Forensic Sciences*, 50(6): 1361–1366, 2005.

M. Perlin and A. Sinelnikov. An Information Gap in DNA Evidence Interpretation. *PLoS ONE*, 4(12):e8327, 2009.

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

M. W. Perlin. Scientific validation of mixture interpretation methods. *In the Proceedings of Promega's Seventeenth International Symposium on Human Identification*, 2006.

F. Pompanon, A. Bonin, E. Bellemain, and P. Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6:847–859, Nov 2005.

L. Prieto, A. Alonso, C. Alves, M. Crespillo, M. Montesino, A. Picornell, A. Brehm, J. Ramírez, M. Whittle, M. Anjos, et al. 2006 GEP-ISFG collaborative exercise on mtDNA: reflections about interpretation, artefacts, and DNA mixtures. *Forensic Science International: Genetics*, 2(2):126–133, 2008.

R Development Core Team. R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http : //www.Rproject.org/. 2006.

N. Risch and B. Devlin. On the probability of matching DNA fingerprints. *Science*, 255(5045): 717, 1992.

N. Rudin and K. Inman. *An introduction to forensic DNA analysis*. CRC, 2002.

E. Rykiel. Testing ecological models: the meaning of validation. *Ecological modelling*, 90(3): 229–244, 1996.

W. Schmerer, S. Hummel, and B. Herrmann. Optimized DNA extraction to improve reproducibility of short tandem repeat genotyping with highly degraded DNA as target. *Electrophoresis*, 20(8):1712–1716, 1999.

V. Shah, K. Midha, J. Findlay, H. Hill, J. Hulse, I. McGilveray, G. McKay, K. Miller, R. Patnaik, M. Powell, et al. Bioanalytical method validationa revisit with a decade of progress. *Pharmaceutical Research*, 17(12):1551–1557, 2000.

J. Siegel, P. Saukko, and G. Knupfer. *Encyclopedia of forensic sciences*. Academic Press, 2000.

C. Sprecher, C. Puers, A. Lins, and J. Schumm. General approach to analysis of polymorphic short tandem repeat loci. *BioTechniques*, 20(2):266–277, 1996.

C. Strom and S. Rechitsky. Use of nested PCR to identify charred human remains and minute amounts of blood. *Journal of Forensic Sciences*, 43(3):696, 1998.

C. Strom, S. Rechitsky, and Y. Verlinsky. Reliability of gender determination using the polymerase chain reaction (PCR) for single cells. *Journal of Assisted Reproduction and Genetics*, 8(4):225–229, 1991.

P. Taberlet, S. Griffin, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L. Waits, and J. Bouvet. Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, 24(16):3189, 1996.

F. Taroni, J. Lambert, L. Fereday, and D. Werrett. Evaluation and presentation of forensic DNA evidence in European laboratories. *Science & justice*, 42(1):21–28, 2002.

Y. Torres, I. Flore, V. Prieto, M. Lopez-Soto, M. J. Farfan, A. Carracedo, and P. Sanz. Dna mixtures in forensic casework: a 4-year retrospective study. *Forensics Science International*, 134:180–186, 2003.

T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling. Estimating the probability of allelic drop-out of str alleles in forensic genetics. *Forensic Science International: Genetics*, 3(4):222–226, 2009.

U.S. Department of Justice, Federal Bureau of Investigation. *VNTR population data: A world-wide study*. 1993.

R. A. van Oorschot and M. K. Jones. DNA fingerprints from fingerprints. *Nature*, 387:767, 1997.

P. Wandeler, P. Hoeck, and L. Keller. Back to the future: museum specimens in population genetics. *Trends in Ecology & Evolution*, 22(12):634–642, 2007.

T. Wang, N. Xue, and J. Douglas Birdwell. Least-Square Deconvolution: A framework for interpreting short tandem repeat mixtures. *Journal of Forensic Sciences*, 51(6):1284–1297, 2006.

J. Weber and P. May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, 44(3):388, 1989.

B. Weir. Independence of VNTR alleles defined as fixed bins. *Genetics*, 130(4):873, 1992a.

B. Weir. Independence of VNTR alleles defined as floating bins. *American Journal of Human Genetics*, 51(5):992, 1992b.

B. S. Weir, C. M. Triggs, L. Starling, L. I. Stowell, K. A. J. Walsh, and B. J. Interpreting DNA mixtures. *Jounal of Forensic Sciences*, 42(2):213–222, 1997.

J. Whitaker, E. Cotton, and P. Gill. A comparison of the characteristics of profiles produced with the AMPFlSTR® SGM Plus (TM) multiplex system for both standard and low copy number (LCN) STR DNA analysis. *Forensic Science International*, 123(2-3):215–223, 2001.

Numéro d'ordre: 231-2010 Année 2010

THÈSE

Présentée

devant L'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

soutenue le 29 Octobre 2010

par

Hinda HANED

Évaluation de méthodes statistiques pour l'interprétation des mélanges ADN en science forensique

Volume 2: annexes

Directeurs de thèse: Dominique Pontier
Laurent Pène
Frank Sauvage

Jury: Chritisan Biémont (Rapporteur)
Peter Gill (Rapporteur)
Denis Laloë (Examinateur)
Eric Petit (Rapporteur)
Laurent Pène (Co-directeur)
Dominique Pontier (Directrice)
Frank Sauvage (Co-directeur)

Contents

Chapter 1

Appendix A: Forensim Manual

Package 'forensim'

September 12, 2010

Type Package

Title Statistical tools for the interpretation of forensic DNA mixtures

Version 1.1-8

Date 2010-08-28

Author Hinda Haned <haned@biomserv.univ-lyon1.fr>

Maintainer Hinda Haned <haned@biomserv.univ-lyon1.fr>

Suggests gdata,gtools,MASS,mvtnorm,genetics,tcltk,tkrplot

Depends methods

Description Statistical methods and simulation tools for the interpretation of forensic DNA mixtures

License GPL (>= 2)

LazyLoad yes

Collate classes_definitions.R classes_constructors.R accessors.R simufreqD.R simupopD.R zzz.R AuxFunc.R changepop.R PE.R likelihood.R likestim.R mincontri.R Pevid2.R LR.R RMP.R A2.simu.R A3.simu.R A4.simu.R mastermix.R N2Exact.R N2error.R simMixSNP.R wrapdataL.R recordDrop.R DNAproxy.R tabDNAproxy.R recordHeights.R tabSPH.R simPCR2.R simPCR2TK.R PV.R Hbsimu.R dropDB.R

R topics documented:

`forensim-package` *The forensim package*

Description

forensim is dedicated to the interpretation of forensic DNA mixtures through statistical methods. It relies on three S4 classes that facilitate the manipulation and the storage of genetic data produced in forensic casework: tabfreq, simugeno and simumix.
tabfreq objects are used to store allele frequencies, simugeno objects are used to store genotypes and simumix objects are used to store DNA mixtures.
For more information about these classes type 'class ?tabfreq', 'class ?simugeno' and 'class ?simumix'.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

`A2.simu` *A Tcl/Tk graphical user interface for simple DNA mixtures resolution using allele peak heights or areas information when two alleles are observed at a given locus*

Description

The `A2.simu` function launches a Tcl/Tk graphical interface with functionalities devoted to two-person DNA mixtures resolution, when two alleles are observed at a given locus.

Usage

```
A2.simu()
```

Details

When two alleles are observed at a given locus in the DNA stain, seven genotype combinations are possible for the two contributors: (AA,AB), (AB,AB), (AA,BB), (AB,AA), (BB,AA), (AB,BB) and (BB,AB), where A and B are the two observed alleles (in ascending order of molecular weight). Having previously obtained an estimation for the mixture proportion, it is possible to reduce the number of possible genotype combinations by keeping those only supported by the observed data. This is achieved by computing the sum of square differences between the expected allelic ratio and the observed allelic ratio, for all possible mixture combinations. The likelihood of peak heights (or areas), given the combination of genotypes, is high if the residuals are low. Genotype combinations are thus selected according to the peak heights with the highest likelihoods.

The `A2.simu()` function launches a dialog window with three buttons:
`-Plot simulations`: plot of the residuals of each possible genotype combination for varying values of the mixture proportion across the interval [0.1, 0.9]. The observed mixture proportion is also reported on the plot.
`-Simulation details`: a matrix containing the simulation results. Simulation details and

genotype combinations with the lowest residuals can be saved as a text file by clicking the “Save" button. It is also possible to choose specific paths and names for the save files.
-`Genotypes filter`: a matrix giving the mixture proportion conditional on the genotype combination. This conditional mixture proportion helps filter the most plausible genotypes among the seven possible combinations. The matrix can be saved as a text file by clicking the “Save" button. It is also possible to choose a specific path and a name for the save file.

Note

-Linux users may have to download the `libtktable` package to their system before using the `A2.simu` function. This is due to the `Tktable` widget, used in `forensim`, which is not (always) downloaded with the Tcl/Tk package.
-For the computational details, please see forensim tutorial at `http://forensim.r-forge.r-project.org/misc/forensim-tutorial.pdf`.

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

References

Gill P, Sparkes P, Pinchin R, Clayton, Whitaker J, Buckleton J. Interpreting simple STR mixtures using allele peak areas. *Forensic Sci Int* 1998;91:41-53.

See Also

`A3.simu`: the three-allele model, and `A4.simu`: the four-allele model

Examples

```
A2.simu()
```

`A3.simu` *A Tcl/Tk graphical user interface for simple DNA mixtures resolution using allele peak heights or areas when three alleles are observed at a given locus*

Description

The `A3.simu` function launches a Tcl/Tk graphical interface with functionalities devoted to two-person DNA mixtures resolution, when three alleles are observed at a given locus.

Usage

```
A3.simu()
```

Details

When three alleles are observed at a given locus in the DNA stain, twelve genotype combinations are possible for the two contributors: (AA,BC), (BB,AC), (CC,AB), (AB,AC), (BC,AC), (AB,BC), (BC,AA), (AC,BB), (AB,CC), (AC,AB), (AC,BC) and (BC,AB) where A, B and C are the three observed alleles (in ascending order of molecular weights). Having previously obtained an estimation for the mixture proportion, it is possible to reduce the number of possible genotype combinations by keeping those only supported by the observed data. This is achieved by computing the sum of square differences between the expected allelic ratio and the observed allelic ratio, for all possible mixture combinations. The likelihood of peak heights (or areas), given the combination of genotypes, is high if the residuals are low. Genotype combinations are thus selected according to the peak heights with the highest likelihoods.

The `A3.simu()` function launches a dialog window with three buttons:
-`Plot simulations`: plot of the residuals of each possible genotype combination for varying values of the mixture proportion across the interval [0.1, 0.9]. The observed mixture proportion is also reported on the plot.
-`Simulation details`: a matrix containing the simulation results. Simulation details and genotype combinations with the lowest residuals can be saved as a text file by clicking the "Save" button. It is also possible to choose specific paths and names for the save files.
-`Genotypes filter`: a matrix giving the mixture proportion conditional on the genotype combination. This conditional mixture proportion helps filter the most plausible genotypes among the twelve possible combinations. The matrix can be saved as a text file by clicking the "Save" button. It is also possible to choose a specific path and a name for the save file.

Note

-Linux users may have to download the `libtktable` package to their system before using the `A3.simu` function. This is due to the `Tktable` widget, used in `forensim`, which is not (always) downloaded with the Tcl/Tk package.
-For the computational details, please see forensim tutorial at `http://forensim.r-forge.r-project.org/misc/forensim-tutorial.pdf`.

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

References

Gill P, Sparkes P, Pinchin R, Clayton, Whitaker J, Buckleton J. Interpreting simple STR mixtures using allele peak areas. *Forensic Sci Int* 1998;91:41-53.

See Also

`A2.simu`: the two-allele model, and `A4.simu`: the four-allele model

Examples

```
A3.simu()
```

`A4.simu`	*A Tcl/Tk graphical user interface for simple DNA mixtures resolution using allele peak heights or areas when four alleles are observed at a given locus*

Description

The `A4.simu` function launches a Tcl/Tk graphical interface with functionalities devoted to two-person DNA mixtures resolution, when four alleles are observed at a given locus.

Usage

```
A4.simu()
```

Details

When four alleles are observed at a given locus in the DNA stain, six genotype combinations are possible for the two contributors: (AB,CD),(AC,BD),(AD,BC),(BC,AD),(BD,AC) and (CD,AB) where A, B, C and D are the four observed alleles (in ascending order of molecular weights). Having previously obtained an estimation for the mixture proportion, it is possible to reduce the number of possible genotype combinations by keeping those only supported by the observed data. This is achieved by computing the sum of square differences between the expected allelic ratio and the observed allelic ratio, for all possible mixture combinations. The likelihood of peak heights (or areas), given the combination of genotypes, is high if the residuals are low. Genotype combinations are thus selected according to the peak heights with the highest likelihoods.

The `A4.simu()` function launches a dialog window with three buttons:
-`Plot simulations`: plot of the residuals of each possible genotype combination for varying values of the mixture proportion across the interval [0.1, 0.9]. The observed mixture proportion is also reported on the plot.
-`Simulation details`: a matrix containing the simulation results. Simulation details and genotype combinations with the lowest residuals can be saved as a text file by clicking the “Save" button. It is also possible to choose specific paths and names for the save files.
-`Genotypes filter`: a matrix giving the mixture proportion conditional on the genotype combination. This conditional mixture proportion helps filter the most plausible genotypes among the six possible combinations. The matrix can be saved as a text file by clicking the “Save" button. It is also possible to choose a specific path and a name for the save file.

Note

-Linux users may have to download the `libtktable` package to their system before using the `A4.simu` function. This is due to the `Tktable` widget, used in `forensim`, which is not (always) downloaded with the Tcl/Tk package.
-For the computational details, please see forensim tutorial at `http://forensim.r-forge.r-project.org/misc/forensim-tutorial.pdf`.

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

References

Gill P, Sparkes P, Pinchin R, Clayton, Whitaker J, Buckleton J. Interpreting simple STR mixtures using allele peak areas. *Forensic Sci Int* 1998;91:41-53.

See Also

`A2.simu`: the two-allele model, and `A3.simu`: the three-allele model

Examples

```
A4.simu()
```

`Accessors` *Accessors for forensim objects*

Description

Accessors for forensim objects: simugeno, simumix and tabfreq. "\$" and "\$<-" are used to access the slots of an object, they are equivalent to "@" and "@<-".

Value

A simugeno, a simumix or a tabfreq object.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

Examples

```
data(strusa)
class(strusa)

strusa@pop.names
#equivalent
strusa$pop.names
```

`Bates.Database` *Allele counts for first example (Table 1) of Balding and Buckleton, 2010*

Description

The marker allele and counts are given in a data frame.

Usage

```
data(Bates.Database)
```

Format

A data frame with 104 observations on the following 3 variables.

`marker` Marker name

`allele` a numeric vector

`count` a numeric vector

References

Balding DJ, Buckleton J, Interpreting low template DNA profiles, Forensic Science International: Genetics, 4: 1-10, 2009, doi: 10.1016/j.fsigen.2009.03.003. http://www.zebfontaine.eclipse.co.uk/djb.htm

Examples

```
data(Bates.Database)
```

`Bates.DNA` *Alleles of mixture, known contributor, defendant as well as missing (drop-out) allelels. (Case Bates, see http://www.zebfontaine.eclipse.co.uk/djb.htm)*

Description

There are four lines for each marker correspinding to alleles for the (i) the mixture (ii) known contributor (if ther is one) (iii) the defendant and (iv) drop-out alleles. The names and order of the markers should be as for dataBase

Usage

```
data(Bates.DNA)
```

Format

A data frame with 40 observations on the following 6 variables.

`Marker` arker names

`Source` Alleles of `Mixture`, `Known` contributor and `defendant` and missing (drop-out) alleles

`A1` a numeric vector

`A2` a numeric vector

`A3` a numeric vector

`A4` a logical vector

Source

Balding DJ, Buckleton J, Interpreting low template DNA profiles, Forensic Science International: Genetics, 4: 1-10, 2009, doi: 10.1016/j.fsigen.2009.03.003. http://www.zebfontaine.eclipse.co.uk/djb.htm

Examples

```
data(Bates.DNA)
```

`CaseY.Database`	*Allele counts for second example of (Table) 3 of Balding and Buckleton, 2010*

Description

The marker allele and counts are given in a data frame.

Usage

```
data(Bates.Database)
```

Format

A data frame with 104 observations on the following 3 variables.

`marker` Marker name

`allele` a numeric vector

`count` a numeric vector

References

Balding DJ, Buckleton J, Interpreting low template DNA profiles, Forensic Science International: Genetics, 4: 1-10, 2009, doi: 10.1016/j.fsigen.2009.03.003. http://www.zebfontaine.eclipse.co.uk/djb.htm

Examples

```
data(Bates.Database)
```

`CaseY.DNA`	*Alleles of mixture, known contributor, defendant as well as missing (drop-out) allelels for US case (CaseY, see http://www.zebfontaine.eclipse.co.uk/djb.htm)*

Description

There are four lines for each marker correspinding to alleles for the (i) the mixture (ii) known contributor (if ther is one) (iii) the defendant and (iv) drop-out alleles. The names and order of the markers should be as for dataBase.

Usage

```
data(Bates.DNA)
```

Format

A data frame with 40 observations on the following 6 variables.

`Marker` arker names

`Source` Alleles of `Mixture`, `Known` contributor and `defendant` and missing (drop-out) alleles

`A1` a numeric vector

`A2` a numeric vector

`A3` a numeric vector

`A4` a logical vector

Source

Balding DJ, Buckleton J, Interpreting low template DNA profiles, Forensic Science International: Genetics, 4: 1-10, 2009, doi: 10.1016/j.fsigen.2009.03.003. http://www.zebfontaine.eclipse.co.uk/djb.htm

Examples

```
data(Bates.DNA)
```

`changepop` *Function to change population-related information in forensim objects*

Description

The `changepop` function changes population-related information in tabfreq, simugeno and simumix objects

Usage

```
changepop(obj, oldpop, newpop)
```

Arguments

`obj`	a forensim object, either a tabfreq, a simugeno or a simumix object
`oldpop`	a character vector giving the population names to be changed
`newpop`	a character vector giving the new population names

Value

a `forensim` object where the slots containing population-related information have been modified

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

Examples

```
data(strveneto)
tab1 <- simugeno(strveneto,n=100)
tab2 <- changepop(tab1,"Veneto","VENE")
tab1$pop.names
tab2$pop.names
```

`Cmn`	*The number of all possible combinations of m elements among n with repetitions*

Description

The number of all possible combinations of m elements among n with repetitions.

Usage

```
Cmn(m, n)
```

Arguments

`m`	the `m` elements to combine among `n`
`n`	the `n` elements from which to combine `m` elements with repetitions

Details

There are (n+m-1)!/(m!(n-1)!) ways to combine m elements among n with repetitions.

Note

Cmn was implemented as an auxiliary function for the `dataL` function which computes the likelihood of the observed alleles in a mixed DNA stain conditional on the number of contributors.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

`comb` for all possible combinations of m elements among n with repetitions

Examples

```
Cmn(2,3)
comb(2,3)
```

`comb`	*Generate all possible combinations of m elements among n with repetitions*

Description

Generate all possible combinations of m elements among n with repetitions.

Usage

```
comb(m, n)
```

Arguments

`m`	the number of elements to combine
`n`	the number of elements from which to combine the `m` elements

Details

There are (n+m-1)!/(m!(n-1)!) ways to combine m elements among n with repetitions, `combn` generates all these possible combinations.

Value

A matrix of (n+m-1)!/(m!(n-1)!) rows, and n columns, each row is a possible combination of m elements among n .

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

See Also

`Cmn` for the calculation of the number of all possible combinations of m elements among n with repetitions

Examples

```
#combine 2 objects among 3 with repetitions
Cmn(2,3)
comb(2,3)
```

`dataL` *Generic formula of the likelihood of the observed alleles in a mixture conditional on the number of contributors for a specific locus*

Description

The function `dataL` gives the likelihood of a set of alleles observed at a specific locus conditional on the number of contributors that gave these alleles. Calculation is based upon the frequencies of the observed alleles.

Usage

```
dataL(x = 1, p, theta = 0)
```

Arguments

`x`	an integer giving the number of contributors
`p`	a numeric vector giving the frequencies of the observed alleles in the mixture
`theta`	a float in [0,1[. `theta` is equivalent to Wright's Fst. In case of population subdivision, it allows a correction of the allele frequencies in the subpopulation of interest

Note

`dataL` function has several similarities with the `Pevid.gen` function of the *forensic* package which computes the probability of the DNA evidence, `dataL` implements a particular case of this probability. Please see `http://cran.r-project.org/web/packages/forensic/`

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

References

Haned H, Pene L, Lobry JR, Dufour AB, Pontier D. Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count? *J Forensic Sci*, accepted 2010.

Curran JM, Triggs CM, Buckleton J, Weir BS. Interpreting DNA Mixtures in Structured Populations. *J Forensic Sci* 1999;44(5): 987-995

See Also

`lik.loc` and `lik` for calculating the likelihood of a given simumix object

Examples

```
#likelihood of observing two alleles at frequencies 0.1 and 0.01 when the number of
#contributors is 2, in two cases:  theta=0 and theta=0.03
dataL(x=2,p=c(0.1,0.01), theta=0)
dataL(x=2,p=c(0.1,0.01), theta=0.03)
```

DNAproxy	*Approximation of the amount of DNA contributed by a person based on the observed peak heights of the alleles present in the analyzed sample*

Description

DNAproxy gives an estimation of the amount of DNA contributed by a person to a DNA stain based on the observed peak heights of the present alleles. The estimation is performed using data across all available loci, data can either consist of single-contributor or mixed DNA stains. The computation of the DNA proxies from experimental data are described by Tvedebrink et al. (cf. the references sections).

Usage

```
DNAproxy(tab, x)
```

Arguments

tab	a table produced by the recordDrop function, giving the allelic dropouts observations and the corresponding allelic peak heights
x	a character giving the label of the individual for whom the DNA proxy must be specified, this argument is to be specified only when data in tab is made of mixtures. In case data is consist of single-contributor stains, the argument must be left empty. x must match the name given in the tab table.

Note

DNAproxy is an auxiliary function of the tabDNAproxy function that implements the methodology proposed by Tvedebrink et al. to estimate the probability of allelic dropout using experimental DNA mixtures.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics*, 2009, 3(4), 222-226.

See Also

recordDrop, tabDNAproxy

Examples

```
#load the exemple data
data(dropdata)
tabcsv<-dropdata$tabcsv
genot<-dropdata$genot
#individuals' labels are 1 and 2
```

```
#DNA proxy for individual one, when data is composed of a 2-person mixture
DNAproxy(recordDrop(1,2,geno=genot,tabcsv=tabcsv),"c1")
```

`dropdata` *Dropout example data*

Description

`dropdata` gives is an extract of a series of experiments used to determine the probability of dropout

Usage

```
data(dropdata)
```

Format

A list of two components: 'tabcsv' and 'genot'. `tabcsv` is an extract of the validation table of a two-person mixture (Genemapper format) and `genot` is the matrix of genotypes of the individuals contributing to the mixture.

Details

The mixture is characterized using the Applied Biosystems AmpFlSTR Identifiler^TM kit.

Source

Data communicated by Elodie Suzanne and Laurent P\'ene, Laboratoire de Police Scientifique, Ecully, France.

Examples

```
data(dropdata)
names(dropdata)
dropdata$tabcsv
dropdata$genot
```

`dropDB` *Calculates LR allowing for drop-out and drop-in*

Description

This function wraps up David Balding's code.

Usage

```
dropDB(dataBase, DNA, dropHetero = 0.05, alpha = 0.5,
dropIn = 0.05, rel = c(0, 0)/4, maxUnknown = 1,
 adj = 0, fst = 0)
```

Arguments

`dataBase`	An R data frame with header 'marker allele count' with columns giving the names of the marker and the allele and the count (the number of occurences of the allele) in the database.
`DNA`	An R data frame with header 'MarkerSource A1 A2 A3 A4' There are four lines for each marker corresponding to alleles for the (i) the mixture (ii) known contributor (if there is one) (iii) the defendant and (iv) drop-out alleles. The names and order of the markers should be as for dataBase-
`dropHetero`	Probability heterozygous drop-out
`alpha`	Balding and Buckleton's alpha-parameter defining relation between homozygous and heterozygous drop out probabilities.
`dropIn`	Drop-in probability.
`rel`	A vector of length 2 giving the IBD=1 and IBD=2 probabilities and so c(0,0) corrresponds to unrelated contributors.
`maxUnknown`	The maximum number of unknown contributors under the defence hypothesis. Possible values 1 or 2.
`adj`	Parameter to adjust for sampling adjustment. Balding uses 2.
`fst`	Parameter to adjust for coancestry. Balding recommends values between 0.01 and 0.05.

Value

Marker name, LR and numerator and denominator of LR.

Author(s)

Thore Egeland based on David Balding's code.

References

Balding DJ, Buckleton J, Interpreting low template DNA profiles, Forensic Science International: Genetics, 4: 1-10, 2009, doi: 10.1016/j.fsigen.2009.03.003. http://www.zebfontaine.eclipse.co.uk/djb.htm

Examples

```
#First output column of Table 1 in Balding and Buckleton:
data(Bates.Database);data(Bates.DNA)
dropDB(Bates.Database,Bates.DNA,dropHetero=0.05,alpha=0.5,dropIn=0.00,rel=c(0,0)/4,
maxUnknown=1,adj=2,fst=0.02)
#First output column of Table 3 in Balding and Buckleton:
data(CaseY.Database);data(CaseY.DNA)
dropDB(CaseY.Database,CaseY.DNA,dropHetero=0.5,alpha=0.5,dropIn=0.05,rel=c(0,0)/4,
maxUnknown=1,adj=2,fst=0.02)
```

`findfreq` *Finds the allele frequencies of a mixture from a tabfreq object*

Description

The `findfreq` function finds the allele frequencies of a mixture stored in a simumix object, form a given tabfreq object. If the tabfreq object contains multiple populations, a reference population from which to extract the frequencies must be specified.

Usage

```
findfreq(mix, freq, refpop = NULL)
```

Arguments

`mix`	a `simumix` object
`freq`	a `tabfreq` object from which to extract the allele frequencies of the mixture
`refpop`	a factor giving the reference population in `tabfreq` from which to extract the allele frequencies

Value

A list giving the allele frequencies for each locus.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

`simumix`

Examples

```
data(strusa)
s2<-simumix(simugeno(strusa,n=c(0,2000,0)),ncontri=c(0,2,0))
findfreq(s2,strusa,refpop="Cauc")
```

`findmax` *Function to find the maximum of a vector and its position*

Description

The `findmax` function finds the maximum of a vector and its position.

Usage

```
findmax(vec)
```

Arguments

vec	a numeric vector

Details

findmax finds the maximum value of a vector and its position.

Value

A matrix of two columns:
max the position of the maximum in vec
maxval the maximum

Note

findmax is an auxiliary function for the dataL function, used to compute the likelihood of the observed alleles in a mixed DNA stain given the number of contributors.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

Examples

```
findmax(1:10)
```

Hbsimu *A Tcl/Tk simulator of the heterozygous balance*

Description

Hbsimu is a user-friendly graphical interface simulating the heterozygous balance of heterozygous profiles generated according to the simulation model described in Gill et al. (2005)

Usage

```
Hbsimu()
```

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

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

Examples

```
Hbsimu()
```

`lik` *Likelihood of the observed alleles at different loci in a DNA mixture conditional on the number of contributors to the mixture*

Description

The `lik` function computes the likelihood of the observed alleles in a forensic DNA mixture, for a set of loci, conditional on the number of contributors to the mixture. The overall likelihood is computed as the product of loci likelihoods.

Usage

```
lik(x = 1, mix, freq, refpop = NULL, theta = NULL, loc=NULL)
```

Arguments

`x`	the number of contributors to the DNA mixture, default is 1
`mix`	a `simumix` object which contains the mixture to be analyzed
`freq`	a `tabfreq` object from which to extract the allele frequencies
`refpop`	a factor giving the reference population in `tabfreq` from which to extract the allele frequencies. This argument is used only if `freq` contains allele frequencies for multiple populations, otherwise it is by default set to NULL
`theta`	a float from [0,1[giving Wright's Fst coefficient. `theta` accounts for population subdivision while computing the likelihood of the data
`loc`	loci for which the overall likelihood shall be computed. Default (NULL) corresponds to all loci

Details

`lik` computes the likelihood of the alleles observed at all loci conditional on the number of contributors. This function implements the general formula for the interpretation of DNA mixtures in case of population subdivision (Curran et al, 1999), in the particular case where all contributors are unknown and belong to the same subpopulation.
The likelihood for multiple loci is computed as the product of loci likelihoods.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Haned H, Pene L, Lobry JR, Dufour AB, Pontier D. Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count? *J Forensic Sci*, accepted 2010.
Curran JM, Triggs CM, Buckleton J, Weir BS. Interpreting DNA Mixtures in Structured Populations. *J Forensic Sci* 1999;44(5): 987-995

See Also

`lik.loc` for the likelihood per locus, `likestim` and `likestim.loc` for the estimation of the number of contributors to a DNA mixture through likelihood maximization

Examples

```
data(strusa)
#simulation of 1000 genotypes from the African American allele frequencies
gen<-simugeno(strusa,n=c(1000,0,0))
#3-person mixture
mix3<-simumix(gen,ncontri=c(3,0,0))
sapply(1:3, function(i) lik(x=i,mix3, strusa, refpop="Afri"))
```

`lik.loc` *Likelihood per locus of the observed alleles in a DNA mixture conditional on the number of contributors to the mixture*

Description

The `lik.loc` function computes the likelihood of the observed data in a forensic DNA mixture, for each of the loci involved, conditional on the number of contributors to the mixture.

Usage

```
lik.loc(x = 1, mix, freq, refpop = NULL, theta = NULL, loc=NULL)
```

Arguments

`x`	the number of contributors to the DNA mixture
`mix`	a `simumix` object which contains the mixture to be analyzed
`freq`	a `tabfreq` object from which to extract the allele frequencies
`refpop`	a factor giving the reference population in `tabfreq` from which to extract the allele frequencies
`theta`	a float from [0,1[giving Wright's Fst coefficien. `theta` acounts for population subdivision while computing the likelihood of the data.
`loc`	the loci for which the likelihood shall be computed. Default (set to NULL) corresponds to all loci.

Details

`lik.loc` computes the likelihood per locus of the observed alleles. This function implements the general formula for the interpretation of DNA mixtures in case of subdivided populations (Curran et al, 1999), in the particular case where all contributors are unknown and belong to the same subpopulation.
The Fst coefficient given in the `theta` argument allows accounting for population subdivision when all contributors belong to the same subpopulation.

Value

The function `lik.loc` returns a vector, of length the number of loci in `loc`, giving the likelihood of the data for each locus.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Haned H, Pene L, Lobry JR, Dufour AB, Pontier D. Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count? *J Forensic Sci*, accepted 2010.

Curran JM, Triggs CM, Buckleton J, Weir BS. Interpreting DNA Mixtures in Structured Populations. *J Forensic Sci* 1999;44(5): 987-995

See Also

`lik` for the overall loci likelihood, `likestim` and `likestim.loc` for the estimation of the number of contributors to a DNA mixture through likelihood maximization

Examples

```
data(strusa)
#simulation of 1000 genotypes from the Caucasian allele frequencies
gen<-simugeno(strusa,n=c(0,100,0))

#4-person mixture
mix4 <- simumix(gen,ncontri=c(0,4,0))
lik.loc(x=2,mix4, strusa, refpop="Cauc")
lik.loc(x=2,mix4, strusa, refpop="Afri")
#You may also want to try:
#likestim(mix4,strusa,refpop="Cauc")
```

`likestim` *Maximum likelihood estimation of the number of contributors to a forensic DNA mixture for a set of loci*

Description

The `likestim` function gives multiloci estimation of the number of contributors to a forensic DNA mixture using likelihood maximization.

Usage

```
likestim(mix, freq, refpop = NULL, theta = NULL, loc=NULL)
```

Arguments

`mix`	a `simumix` object
`freq`	a `tabfreq` object containing the allele frequencies to use for the calculation
`refpop`	the reference population from which to extract the allele frequencies used in the likelihood calculation. If `tabfreq` contains more than one population, `refpop` must be specified, otherwise, `refpop` is set to default (NULL).

`theta` a float from [0,1[giving Wright's Fst coefficient. `theta` accounts for population subdivision while computing the likelihood of the data.

`loc` loci to be considered in the estimation. Default (set to NULL) corresponds to all loci.

Details

The number of contributors which maximizes the likelihood of the data observed in the mixture is searched in the discrete interval [1,6]. In most cases this interval is a plausible range for the number of contributors.

Value

A matrix of dimension 1 x 2, the first column, `max`, gives the maximum likelihood estimation of the number of contributors, the second column gives the corresponding likelihood value `maxvalue`.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Haned H, Pene L, Lobry JR, Dufour AB, Pontier D. Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count? *J Forensic Sci*, accepted 2010.

Egeland T, Dalen I, Mostad PF. Estimating the number of contributors to a DNA profile. *Int J Legal Med* 2003, 117: 271-275

Curran JM, Triggs CM, Buckleton J, Weir BS. Interpreting DNA Mixtures in Structured Populations. *J Forensic Sci* 1999, 44(5): 987-995

See Also

`likestim.loc` for maximum of likelihood estimations per locus

Examples

```
data(strusa)
#simulation of 1000 genotypes from the Hispanic allele frequencies
gen<-simugeno(strusa,n=c(0,0,100))
#4-person mixture
mix4 <- simumix(gen,ncontri=c(0,0,4))
likestim(mix4,strusa,refpop="Hisp")
```

`likestim.loc` *Maximum likelihood estimation per locus of the number of contributors to forensic DNA mixtures.*

Description

The `likestim.loc` function returns the estimation of the number of contributors, at each locus, obtained by maximizing the likelihood.

Usage

```
likestim.loc(mix, freq, refpop = NULL, theta = NULL, loc = NULL)
```

Arguments

`mix`	a `simumix` object
`freq`	a `tabfreq` object containing the allele frequencies to use for the calculation
`refpop`	the reference population from which to extract the allele frequencies used in the likelihood calculation. Default set to NULL, if `tabfreq` contains more than one population, `refpop` must be specified
`theta`	a float from [0,1[giving Wright's Fst coefficient. `theta` acounts for population subdivision while computing the likelihood of the data.
`loc`	loci to be considered in the estimation. Default (set to NULL) corresponds to all loci.

Details

The number of contributors which maximizes the likelihood of the data observed in the mixture is searched in the discrete interval [1,6]. In most cases this interval is a plausible range for the number of contributors.

Value

A matrix of dimension `loc` x 2. The first colum, `max`, gives the maximum likelihood estimation of the number of contributors for each locus in row. The second column, `maxvalue`, gives the corresponding likelihood value.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Haned H, Pene L, Lobry JR, Dufour AB, Pontier D. Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count? *J Forensic Sci*, accepted 2010.

Egeland T , Dalen I, Mostad PF. Estimating the number of contributors to a DNA profile. *Int J Legal Med* 2003, 117: 271-275

Curran, JM , Triggs CM, Buckleton J , Weir BS. Interpreting DNA Mixtures in Structured Populations. *J Forensic Sci* 1999, 44(5): 987-995

See Also

`likestim` for multiloci estimations

Examples

```
data(strusa)
#simulation of 1000 genotypes from the Hispanic allele frequencies
gen<-simugeno(strusa,n=c(0,0,100))
#4-person mixture
mix4 <- simumix(gen,ncontri=c(0,0,4))
likestim.loc(mix4,strusa,refpop="Hisp")
```

`LR` *Likelihood ratio for DNA evidence interpretation*

Description

The `LR` function calculates the likelihood ratio for a DNA evidence, when two competing hypotheses Hd and Hp, respectively the defence and the prosecution hypotheses, are weighted about the origin of the DNA evidence. The evidence can either be a simple or a mixed stain.

Usage

```
LR(stain, freq, xp=0, xd=0, Tp=NULL, Vp=NULL, Td=NULL, Vd=NULL, theta=0)
```

Arguments

`stain`	a vector giving the set of (distinct) alleles present in the DNA stain
`freq`	vector of the corresponding allele frequencies in the global population
`xp`	the number of unknown contributors to the stain under the prosecution hypothesis Hp. Default is 0.
`xd`	the number of unknown contributors to the stain under the defence hypothesis Hd. Default is 0.
`Tp`	a vector of strings where each string contains two alleles separated by '/', corresponding to one known contributor under the prosecution hypothesis Hp. The length of the vector equals the number of known contributors. Default is NULL.
`Vp`	a vector of strings where each string contains two alleles separated by '/', corresponding to one known non-contributor under the prosecution hypothesis Hp. The length of the vector equals the number of known non-contributors. Default is NULL.
`Td`	a vector of strings where each string contains two alleles separated by '/', corresponding to one known contributor under the defence hypothesis Hd. The length of the vector equals the number of known contributors. Default is NULL.

`Vd` a vector of strings where each string contains two alleles separated by '/', corresponding to one known non-contributor under the defence hypothesis Hd. The length of the vector equals the number of known non-contributors. Default is NULL.

`theta` a float in [0,1[. `theta` is equivalent to Wright's Fst. In case of population subdivision, it allows a correction of the allele frequencies in the subpopulation of interest

Details

`LR` is the implementation of the general formula of Curran et al (1999) for the evaluation of forensic DNA mixtures through likelihood ratios. The likelihood ratio is computed as a ratio of two probabilities of the DNA evidence, E, conditional on the evaluated hypotheses:

$$LR = \frac{P(E|H_p)}{P(E|H_d)},$$

where H_p denotes the prosecution hypothesis and H_d the defence hypothesis.

In case of population subdivision, contributors to the DNA stain are considered to come from the same subpopulation. Allele dependencies within subpopulations are accounted for through Wright's Fst coefficient, denoted here θ.

Note

Please note that the `LR` function is based on functions initially implemented in the forensic package by Miriam Marusiakova http://cran.r-project.org/web/packages/forensic/

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Curran JM, Triggs CM, Buckleton J, Weir BS. Interpreting DNA Mixtures in Structured Populations. *J Forensic Sci* 1999;44(5): 987-995

See Also

the exclusion probability `PE`.

Examples

```
# A rape case in Hong Kong (Hu and Fung, Int J Legal Med 2003)
# The stain shows alleles 14, 15, 17 and 18 at locus D3S1358.
stain =c(14,15,17,18)
# suspect's profile: "14/17"
suspect<-"14/17"
# victim's profile: "15/18"
victim<-"15/18"
# corresponding allele frequencies
freq<-c(0.033,0.331,0.239,0.056)

# Prosecution hypothesis: Contributors were the victim and the suspect
# defence hypothesis: Contributors were the victim and 1 unknown contributor
# Likelihood ratios for DNA evidence for different alternatives:
```

```
LR(stain,freq,xp=0,Tp=c(victim,suspect),Vp=NULL,Td=victim,Vd=suspect,xd=1)
```

`mastermix` *A Tcl/Tk graphical user interface for simple DNA mixtures resolution using allele peak heights/ or areas information*

Description

The `mastermix` function launches a Tcl/Tk graphical user interface dedicated to the resolution of two-person DNA mixtures using allele peak heights/ or areas information. `mastermix` is the implementation of a method developed by Gill et al (see the references section), and previously programmed into an Excel macro by Dr. Peter Gill.

Usage

```
mastermix()
```

Details

`mastermix` is a Tcl/Tk graphical user interface implementing a method developed by Gill et al (1998) for simple mixtures resolution, using allele peak heights or areas information.

This method searches through simulation the most likely combination(s) of the contributors' genotypes. Having previously obtained an estimation for the mixture proportion, it is possible to reduce the number of possible genotype combinations by keeping only those supported by the observed data. This is achieved by computing the sum of square differences between the expected allelic ratio and the observed allelic ratio, for all possible mixture combinations. The likelihood of peak heights (or areas), conditional on the combination of genotypes, is high if the residuals are low. Genotype combinations are thus selected according to the peak heights with the highest (conditioned) likelihoods.

`mastermix` offers a graphical representation of the simulation for three models:
-The two allele model: at a given locus, two alleles are observed in the DNA stain.
-The three allele model: at a given locus, three alleles are observed in the DNA stain.
-The four allele model: at a given locus, four alleles are observed in the DNA stain.

A left-click on each button launches a simulation dialog window for the corresponding model, while a right-click opens the corresponding help page.

Note

-Each implemented model can either be launched using the `mastermix` interface, or the `A2.simu`, `A3.simu` and `A4.simu` functions, depending on the considered model.
-For the computational details, please see forensim tutorial at `http://forensim.r-forge.r-project.org/misc/forensim-tutorial.pdf`.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Gill P, Sparkes P, Pinchin R, Clayton, Whitaker J, Buckleton J. Interpreting simple STR mixtures using allele peak areas. *Forensic Sci Int* 1998;91:41-5.

See Also

`A2.simu`, `A3.simu` and `A4.simu`

Examples

```
mastermix()
```

`mincontri` *Minimum number of contributors required to explain a forensic DNA mixture*

Description

`mincontri` gives the minimum number of contributors required to explain a forensic DNA mixture. This method is also known as the maximum allele count as it relies on the maximum number of alleles showed through all available loci

Usage

```
mincontri(mix, loc = NULL)
```

Arguments

`mix`	a `simumix` object
`loc`	the loci to consider for the calculation of the minimum of contributors, default (NULL) corresponds to all loci

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

`likestim` for the estimation of the number of contributors through likelihood maximization

Examples

```
data(strusa)
#simulation of 1000 genotypes from the African American allele frequencies
gen<-simugeno(strusa,n=c(1000,0,0))
#5-person mixture
mix5<-simumix(gen,ncontri=c(5,0,0))
#compare
likestim(mix5, strusa, refpop="Afri")
mincontri(mix5)
```

`N2error` *Calculates exact error for maximum allele count for two markers*

Description

The maximum allele count principle leads to wrong conclusion for two contributors if only a maximum of one or two alleles is seen. This probability of error is calculated.

Usage

```
N2error(dat)
```

Arguments

`dat`	a data frame, first column gives the alleles size, remaining columns give their frequencies

Value

The probability of error is returned.

Author(s)

Thore Egeland <Thore.Egeland@medisin.uio.no>

Examples

```
#Example based on 15 markers of Tu data
library(forensim)
data(Tu)
N2error(Tu)
```

`N2Exact` *Calculates exact allele distribution for 2 contributors*

Description

The distribution of N, the number of alleles showing is calculated exactly assuming 2 contributors. Theta-correction is not implemented. The function may be used to check accuracy of simulations and indicate required number of simulations for one example.

Usage

```
N2Exact(p)
```

Arguments

`p`	vector of allele frequencies. Must sum to 1. Default: for uniformly distrubted alleles.

Value

Returns(P(N=i) for i=1,2,3,4

Author(s)

Thore Egeland <Thore.Egeland@medisin.uio.no>

Examples

```
#Distribution for a marker with 20 alles of equal frequency
N2Exact(p=rep(0.05,20))
```

`naomitab` *Handling of missing values in a data frame*

Description

`naomitab` handles missing values (NA) in a data frame: it returns a list of the columns where NAs have been removed.

Usage

```
naomitab(tab)
```

Arguments

`tab`	a data frame

Value

Returns a list of length the number of columns in `tab` where each component is a column of `tab`, and the values are the corresponding rows where NAs have been removed.

Note

This function was designed to handle missing values in data frames in the format of the Journal of Forensic Sciences for population genetic data: allele names are given in the first column, and frequencies for a given allele are read in rows for different loci. When a given allele is not observed, the value is coded NA (originally coded "-" in the journal).

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

`tabfreq`

Examples

```
data(Tu)
naomitab(Tu)
```

nball *Number of alleles in a mixture*

Description

nball gives the number of alleles of a simumix object.

Usage

```
nball(mix, byloc = FALSE)
```

Arguments

mix	a simumix object
byloc	a logical indicating whether the number of alleles must be calculated by locus or for all loci (default)

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

simumix

Examples

```
data(strusa)
#simulating 100 genotypes with allele frequencies from the African American population
gaa<-simugeno(strusa,n=c(100,0,0))
#simulating a 4-person mixture
maa4<-simumix(gaa,ncontri=c(4,0,0))
nball(maa4,byloc=TRUE)
```

PE *The random man exclusion probability*

Description

Computes the random man exclusion probability of a mixture stored in a simumix object

Usage

```
PE(mix, freq, refpop = NULL, theta = 0, byloc = FALSE)
```

Arguments

`mix`	a `simumix` object
`freq`	a `tabfreq` object giving the allele frequencies from which to compute the exclusion probability
`refpop`	character giving the reference population, used only if `freq` contains allele frequencies for multiple populations
`theta`	a float from [0,1[giving Wright's Fst coefficient. `theta` accounts for population subdivision while computing the likelihood of the data.
`byloc`	logical, if TRUE, than the exclusion probability is computed per locus, if FALSE (default), the calculations are done for all loci simultaneously

Details

`PE` gives the exclusion probability at a locus, or at several loci when conditions for Hardy Weinberg are met. If this condition is not met in the population, than a value for `theta` must be supplied to take into account dependencies between alleles. The formula of the exclusion probability that allows taking into account departure from Hardy Weinberg proportions due to population subdivision was provided by Bruce Weir, please see the references section.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Clayton T, Buckleton JS. Mixtures. In: Buckleton JS, Triggs CM, Walsh SJ, editors. Forensic DNA Interpretation. CRC Press 2005;217-74

Examples

```
data(strusa)
geno1<-simugeno(strusa,n=c(0,0,100))
mix2 <-simumix(geno1,ncontri=c(0,0,2))
PE(mix2,strusa,"Hisp",byloc=TRUE)
```

`Pevid2` *Conditional profile probabilities*

Description

Calculates the probability of observing a set of DNA profiles conditional on a given hypothesis specifying who were the contributors to the observed profiles. All the individuals involved in the analyzed case are assumed to come from the same subpopulation with a given coancestry coefficient.

Usage

```
Pevid2(stain, freq, x, T = NULL, V = NULL, theta = 0)
```

Arguments

`stain`	vector of distinct alleles (from one specific locus) found in the crime sample.
`freq`	vector of the corresponding allele frequencies in the global population
`x`	the number of unknown contributors to the mixture
`T`	object of class genotype (package **genetics**), or a vector of strings where each string contains two alleles separated by '/', corresponding to one known contributor. The length of the vector equals the number of known contributors. Default is NULL.
`V`	object of class genotype (package **genetics**), or a vector of strings where each string contains two alleles separated by '/', corresponding to one known non-contributor. The length of the vector equals the number of known non-contributors. Default is NULL.
`theta`	a float in [0,1[. `theta` is equivalent to Wright's Fst. In case of population subdivision, it allows a correction of the allele frequencies in the subpopulation of interest

Note

Please note that the `Pevid2` function is an improved version of the `Pevid.gen` function from the forensic package by Miriam Marusiakova (which explains the 2 in the function name). `Pevid2` calls external functions in C code.

Here we define the conditional profile probability as the probability of the profiles under a certain hypothesis stating who gave the observed alleles, hence, Pr(stain="A"|U=0,V=0,T="A/A",H="suspect A/A gave the profile) would equal one rather than 2*p(A)*p(A) in the original formula in Curran et al.

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

References

Curran JM, Triggs CM, Buckleton J, Weir BS. Interpreting DNA Mixtures in Structured Populations. *J Forensic Sci* 1999;44(5): 987-995

See Also

`LR`, `RMP`

Examples

```
# A rape case in Hong Kong (Hu and Fung, Int J Legal Med 2003)
# The stain shows alleles 14, 15, 17 and 18 at locus D3S1358.
stain=c(14,15,17,18)
# suspect's profile: "14/17"
suspect<-"14/17"
# victim's profile: "15/18"
victim<-"15/18"
# corresponding allele frequencies
freq<-c(0.033,0.331,0.239,0.056)
```

```
# Prosecution proposition: Contributors were the victim and the suspect
# defence proposition: Contributors were the victim and 1 unknown contributor
# from the same subpopulation as the victim
# Evaluation of the defence proposition, in case of independence between alleles
Pevid2(stain, freq, x=1, T = victim)

# note that if theta=0, the suspect's profile plays no role in the calculation
#and the same result is obtained
Pevid2(stain, freq, x=1, T = victim, V = suspect)
# In case of allele dependencies, measured by theta=0.03
Pevid2(stain, freq, x=1, T = victim, V = suspect, theta = 0.03)
```

`PV` *Predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture*

Description

The `PV` function implements the predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture

Usage

```
PV(mat, prior)
```

Arguments

`mat`	matrix giving the estimates of the conditional probabilities that the maximum likelihood estimator classifies a given stain as a mixture of i contributors given that there are k contributor(s) to the stain. Estimates i must be given in columns for each possible value of the number of contributors given in rows.
`prior`	numeric vector giving the prior probabilities of encountering a mixture of i contributors. `prior` must be of length the number of rows in `mat`.

Value

Vector of the predictive values

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Haned H., Pene L., Sauvage F., Pontier D., The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, submitted, 2010.

See Also

maximum likelihood estimator `likestim`

Examples

```
# the following examples reproduce some of the calculations appearing
# in the article cited above, for illustrative purpose, the maximum
#number of contributors is set here to 5
#matcondi: Table 2 in Haned et al. (2010)
matcondi<-matrix(c(1,rep(0,4),0,0.998,0.005,0,0,0,0.002,0.937,0.067,0,0,0,0.058,
0.805,0.131,rep(0,3),0.127,0.662,rep(0,3),0.001,0.207),ncol=6)
#prior defined by a forensic expert (Table 3 in Haned et al., 2010)
prior1<-c(0.45,0.04,0.30,0.15,0.06)
#uniform prior, for each mixture type, the probability of occurrence is 1/5,
#5 being the threshold for the number of contributors
prior2<-c(rep(1/5,5))
#predictive values for prior1
PV(matcondi,prior1)
#for prior2
PV(matcondi, prior2)
```

`recordDrop` *Records the allelic dropout events matched with individual DNA proxies*

Description

The `recordDrop` function records the dropout events from experimental data. The function aims to facilitate the manipulation of experimental data used for the estimation of the probability of allelic dropout (cf. the references sections).

Usage

```
recordDrop(x, y, geno, tabcsv,s=40)
```

Arguments

`x`	numeric label of the contributing individual, if the stain is a mixture, `x` should give the label of the first individual contributing to the mixture
`y`	numeric label of the second contributing individual, default is NULL. If the stain is a mixture, `y` should give the label of the second individual contribution to the mixture. This argument is skipped if the stain is not a mixture (default case: `y` set to NULL).
`geno`	a matrix giving the genotypes of the individuals contributing to the analyzed data for each locus. An individual genotype is given in rows for each locus in column. A homozygous carrying allele 9 is coded '9/9', a heterozygous carrying alleles 8 and 9 is coded '8/9'. Individual labels are coded using integers that are simply the order of introduction in the data frame.
`tabcsv`	a matrix giving the validation table of the analysed DNA stain. `tabcsv` must have a "genemapper" validation table structure, namely, information about the present alleles and the corresponding peak heights must be given.
`s`	numeric giving the detection threshold for alleles in Relative fluorescence units (RFU) , default is set to 40 RFUS. An observed allele with a peak height smaller (<) than 40 RFUS is considered as dropped-out.

Value

A list of length the number of analyzed loci, each component of the list is a matrix with the following information: - The names of expected alleles
- The expected allele counts for the first contributor (when date is a mixture)
- The expected allele counts for the second contributor (when date is a mixture)
- The observed alleles
- The observed peak heights
- The dropout variable D, takes 1 if the allele has dropped out, 0 otherwise

Note

`recordDrop` is an auxiliary function of the `tabDNAproxy` function that implements the methodology proposed by Tvedebrink et al. to estimate the probability of allelic dropout using experimental DNA mixtures.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics*, 2009, 3(4), 222-226.

See Also

`DNAproxy`, `tabDNAproxy`

Examples

```
#load the exemple data
data(dropdata)
tabcsv<-dropdata$tabcsv
genot<-dropdata$genot
#individuals' labels are 1 and 2
#record the dropout the surviving peak heights for heterozygotes whith non shared alleles
recordDrop(1,2,geno=genot,tabcsv=tabcsv,s=40)
```

`recordHeights` *Records the peak heights of the alleles present in the analyzed stains*

Description

The `recordHeights` function records the peak heights of the alleles present in the analyzed stains. The function aims to facilitate the manipulation of experimental data used for the estimation of the probability of allelic dropout (cf. the references sections).

Usage

```
recordHeights(x,y=NULL,geno,tabcsv,byloc=FALSE)
```

Arguments

x	numeric label of the contributing individual, if the stain is a mixture, x should give the label of the first individual contributing to the mixture
y	numeric label of the second contributing individual, default is NULL. If the stain is a mixture, y should give the label of the second individual contribution to the mixture. This argument is skipped if the stain is not a mixture (default case: y set to NULL).
geno	a matrix giving the genotypes of the individuals contributing to the analyzed data for each locus. An individual genotype is given in rows for each locus in column. A homozygous carrying allele 9 is coded '9/9', a heterozygous carrying alleles 8 and 9 is coded '8/9'. Individual labels are coded using integers that are simply the order of introduction in the data frame.
tabcsv	a matrix giving the validation table of the analysed DNA stain. tabcsv must have a "genemapper" validation table structure, namely, information about the present alleles and the corresponding peak heights must be given.
byloc	logical indicating whether data should be displayed per locus (TRUE) or overall loci (FALSE, default)

Value

A list of length the number of analyzed loci, each component of the list is a matrix with the following information: - The names of expected alleles - The expected allele counts for the first contributor (when date is a mixture) - The expected allele counts for the second contributor (when date is a mixture) - The observed alleles - The observed peak heights - The dropout variable D, takes 1 if the allele has dropped out, 0 otherwise

Note

recordHeights is an auxiliary function of the tabSPH function that implements the methodology proposed by Gill et al. to estimate the probability of allelic dropout using experimental DNA mixtures.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Gill P, Puch-Solis R, Curran J. The low-template-DNA (stochastic) threshold-Its determination relative to risk analysis for national DNA databases. *Forensic Science International: Genetics*, 2009, 3, 104-111

See Also

recordDrop for an alternative method, tabSPH

Examples

```
#load the exemple data
data(dropdata)
tabcsv<-dropdata$tabcsv
genot<-dropdata$genot
```

```
#individuals' labels are 1 and 2
#peak heights of heterozygote genotypes with non shared alleles
recordHeights(1,2,geno=genot,tabcsv=tabcsv)
```

`RMP` *The Random Match Probability of DNA evidence (RMP)*

Description

`RMP` computes the random match probability of DNA evidence given in a matrix (or data frame) or in a text file. Several situations are handled: the suspect and an unknown offender are unrelated, or are members of the same subpopulation with a given coancestry coefficient theta, or are close relatives. For the latter case, the relationship is described by the kinship coefficients.

Usage

```
RMP(suspect=NULL, filename=NULL, freq, k=c(1,0,0), theta=0,refpop=NULL)
```

Arguments

`suspect`	a matrix or a data frame of dimension L x 2, L being the number of loci involved in the DNA evidence. The first column gives the loci names, and the second column gives the suspect's genotype at each locus. A genotype is coded as a character where each string contains two alleles separated by '/'. The DNA evidence can also be given in a text file, see argument `filename`.
`filename`	the file name from which the input data should be read. Data mut be a matrix of dimension L x 2, L being the number of loci involved in the DNA evidence.The first column gives the loci names, and the second column gives the suspect's genotype at each locus. A genotype is coded as a character where each string contains two alleles separated by '/'.
`freq`	a `tabfreq` object giving the allele frequencies
`k`	vector of kinship coefficients (k0, k1, k2), where ki is the probability that two people (the suspect and an unknown offender) will share i alleles identical by descent, i = 0, 1, 2.
`theta`	a float in [0,1[. `theta` is equivalent to Wright's Fst. In case of population subdivision, it allows a correction of the allele frequencies in the subpopulation of interest
`refpop`	the reference population in `freq` from which to extract the allele frequencies fro the RMP calculation. This argument is obligatory only if `freq` contains allele frequencies from several populations

Details

The match probability is derived from Balding and Nichols (1994) and is computed as:

$$k_2 + k_1 Z_1 + k_0 Z_2$$

where k_0, k_1, k_2 are the kinship coefficients,
Z_1 is the match probability when the suspect an the unknown offender share one allele identical-by-descent.

Z_2 is the match probability in the unrelated case, when the suspect an the unknown offender share 0 allele identical-by-descent.

In the homozygous case, with the allele frequency p_i:

$$Z_1 = \frac{2\theta + (1-\theta)p_i}{1+\theta}$$

$$Z_2 = \frac{[2\theta + (1-\theta)p_i]\,[3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)}$$

In the heterozygous case, with allele frequencies p_i and p_j:

$$Z_1 = \frac{2\theta + (1-\theta)(p_i + p_j)}{2(1+\theta)}$$

$$Z_2 = \frac{2\,[\theta + (1-\theta)p_i]\,[\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)}$$

θ is Wright's Fst coefficient, usually called the coancestry coefficient in forensic studies. Main effects of allele dependencies between loci in the suspect's subpopulation are taken into account though the coancestry coefficient, hence, the match probability at all loci is, to a close approximation, the product of single-locus probabilities.

Value

`RMP` returns a list with the following components:

`RMP.loc`	single-locus match probabilities
`RMP`	multiloci match probability (product of single-locus match probabilities)

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

References

Balding DJ, Nichols RA. DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci I* 1994;64:125-140.

See Also

`LR` for the evaluation of DNA evidence through likelihood ratio

Examples

```
# random match probability
# data input

data <- matrix(c("CSF1PO","FGA","TH01","TPOX","VWA","D3S1358","D5S818",
"D7S820","D8S1179","D13S317","D16S539","D18S51","D21S11","D2S1338","D19S433",
"12/11","22/19","6/7","10/8","17/18","18/17","12/12","8/8","13/13","11/11",
"12/10","14/15","33.2/32.2","23/22","14/14"),nc=2)
colnames(data)<- c('locus','genotype')
#15-locus genotype
data
```

```
#allele frequencies are taken from the strusa data set

data(strusa)

RMP(suspect=data,freq=strusa,refpop="Cauc")

# using a preexisting file from the forensim package
RMP(filename=system.file("files/exprofile.txt", package = "forensim"),
freq=strusa,refpop="Cauc")
```

`simMixSNP` *Simulates SNP mixtures*

Description

Simulates SNP mixtures and outputs optionally file suitable for wrapdataL function for estimation of number of contributors

Usage

```
simMixSNP(nSNP , p , ncont, writeFile, outfile , id )
```

Arguments

`nSNP`	Integer number of SNPs>1
`p`	Minor allele frequency
`ncont`	Number of contributors >= 1
`writeFile`	If TRUE, output written to file
`outfile`	Name of output file
`id`	Column one of output file identifying run

Value

Returns a data frame with columns Id, marker, allele, frequency and height (=1 for now)

Author(s)

Thore Egeland <Thore.Egeland@medisin.uio.no>

Examples

```
simMixSNP()
```

`simPCR2` *Polymorphism chain reaction simulation model*

Description

`simPCR2` implements a simulation model for the polymorphism chain reaction (Gill et al., 2005). Giving several input parameters, `simPCR2` outputs the number of amplified DNA molecules and their corresponding peak heights (in RFUs).

Usage

```
simPCR2(ncells,probEx,probAlq, probPCR, cyc = 28, Tdrop = 2 * 10^7,
probSperm = 0.5, dip = TRUE,KH=55)
```

Arguments

`ncells`	initial number of cells
`probEx`	probability that a DNA molecule is extracted (probability of surviving the extraction process)
`probAlq`	probability that a DNA molecule is selected for PCR amplification
`probPCR`	probability that a DNA molecule is amplified during a given round of PCR
`cyc`	number of PCR cycles, default is 28 cycles
`Tdrop`	threshold of detection: number of molecules (in the total PCR reaction mixture) that is needed to generate a signal, default is set to 2*10^7 molecules
`probSperm`	probability of observing alleles of type A in the initial sample of haploid cells (e.g. sperm cells). Probability of observing allele B is given by 1-`probSperm`
`dip`	logical indicating the cell ploidy, default is diploid cells (TRUE), FALSE is for haploid cells
`KH`	positive constant used to scale the peak heights obtained from the number of amplified molecules (see reference section)

Details

A threshold of `Tdrop` (must be a multiple of 10^7) is needed to generate a signal, then, a log-linear relationship is used to determine the intensity of the signal with respect to the number of successfully amplified DNA molecules. Dropout events occur whenever less than `Tdorp` molecules are generated.

Value

A matrix with the following components:

`HeightA`	Peak height of allele A
`DropA`	Dropout variable for allele A
`HeightB`	Peak height of allele B
`DropB`	Dropout variable for allele B

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Jeffreys AJ, Wilson V, Neumann R and Keyte J. Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. *Nucleic Acids Res* 1988;16: 10953_10971.

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

See Also

simPCR2TK

Examples

```
#simulation of a 28 cycles PCR, with the initial stain containing 5 cells
simPCR2(ncells=5,probEx=0.6,probAlq=0.30,probPCR=0.8,cyc=28, Tdrop=2*10^7,dip=TRUE,KH=55)
```

`simPCR2TK` *A Tcl/Tk graphical interface for the polymorphism chain reaction simulation model*

Description

`simPCR2TK` is a user-friendly graphical interface for the `simPCR2` function that implements a simulation model for the polymorphism chain reaction.

Usage

```
simPCR2TK()
```

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

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

See Also

simPCR2

Examples

```
#launch the graphical interface
simPCR2TK()
```

`simufreqD` *Function to simulate allele frequencies for independent loci from a Dirichlet model*

Description

The `simufreqD` function simulate single population allele frequencies for independent loci. Allele frequencies are generated as random deviates from a Dirichlet distribution, whose parameters control the mean and the variance of the simulated allele frequencies.

Usage

```
simufreqD(nloc = 1, nal = 2, alpha = 1)
```

Arguments

`nloc`	the number of loci to simulate
`nal`	the numbers of alleles per locus. Either an integer, if the loci have the same number of alleles, or an integer vector, if the number of alleles differ between loci
`alpha`	the parameter used to simulate allele frequencies from the Dirichlet distribution. If the `nloc` loci have the same allele number, `alpha` can either be the same for all alleles (default is one: uniform distribution), in this case `alpha` is an integer, or `alpha` can be different between alleles at a given locus, in this case, `alpha` is a matrix of dimension `nal` x `nloc`. When the number of alleles differ between loci, `alpha` can either be the same or differ between alleles at a given locus. In the first case `alpha` is a vector of length `nloc`, in the second case, `alpha` is a matrix of dimensions `nal` x `nloc` where NAs are introduced for alleles not seen at a given locus.

Details

Allele frequencies for independent loci are simulated using a Dirichlet distribution with parameter `alpha`. At a given locus L with n alleles, the allele frequencies are modeled as a vector of random variables p=(p1, ..., pn), following a Dirichlet distribution with parameters:
alpha = (alpha1, ..., alphan) where p1+...+pn=1 and alpha1,..., alphan > 0.

Value

A matrix containing the simulated allele frequencies. The data is presented in the format of the Journal of Forensic Sciences for genetic data: allele names are given in the first column, and frequencies for a given allele are read in rows for the different markers in columns. When an allele is not observed for a given locus, the value is coded NA (instead of "-" in the original format).

Note

The code used here for the generation of random Dirichlet deviates was previously implemented in the gtools library.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Johnson NL, Kotz S, Balakrishnan N. Continuous Univariate Distributions, vol 2. John Wiley & Sons, 1995.

Wright S. The genetical structure of populations. Ann Eugen 1951;15:323-354.

See Also

simupopD

Examples

```
#simulate alleles frequencies for 5 markers with respectively 2, 3, 4, 5, and 6 alleles

simufreqD(nloc=5,na=c(2,3,4,5,6) , alpha=1)
```

---

simugeno *forensim class for simluated genotypes*

Description

The S4 simugeno class is used to store existing or simulated genotypes.

Slots

tab.freq: a list giving allele frequencies for each locus. If there are several populations, tab.freq gives allele frequencies in each population

nind: integer vector giving the number of individuals. If there are several populations, nind gives the numbers of individuals per population

pop.names: factor of populations names

popind: factor giving the population of each individual

which.loc: character vector giving the locus names

tab.geno: matrix giving the genotypes (in rows) for each locus (in columns). The genotype of a homozygous individual carrying the allele "12" is coded "12/12". A heterozygous individual carrying alleles "12" and "13" is coded "12/13" or "13/12".

indID: character vector giving the individuals ID

Methods

names `signature(x = "simugeno")`: gives the names of the attributes of a simugeno object

show `signature(object = "simugeno")`: shows a simugeno object

print `signature(object = "simugeno")`: prints a simugeno object

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

`as.simugeno` for the simugeno class constructor, `is.simugeno`, `simumix` and `tabfreq`

Examples

```
showClass("simugeno")
```

---

simugeno constructor

simugeno constructor

Description

Constructor for simugeno objects.
The function `simugeno` creates a simugeno object from a tabfreq object.

The function `as.simugeno` is an alias for `simugeno` function.

`is.simugeno` tests if an object is a valid simugeno object.

Note: to get the manpage about simugeno, please type 'class ? simugeno'.

Usage

```
simugeno(tab,which.loc=NULL,n=1)
as.simugeno(tab,which.loc=NULL,n=1)
is.simugeno(x)
```

Arguments

`tab`	a tabfreq object created with constructor `tabfreq`
`which.loc`	a character vector giving the chosen loci for the genotypes simulation. The default is set to NULL, which corresponds to all the loci of the `tabfreq` object given in argument
`n`	integer vector giving the number of individuals. If there are several populations, `n` gives the numbers of individuals to simulate per population. For a single population, default is 1.
`x`	an object

Details

At a given locus, an individual's genotype is simulated by randomly drawing two alleles (with replacement) at their respective allele frequencies in the target population.

Value

For `simugeno` and `as.simugeno`, a simugeno object. For `is.simugeno`, a logical.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

"simugeno", and `tabfreq` for creating a tabfreq object from a data file.

Examples

```
data(Tu)
tab<-tabfreq(Tu)
#simulation of 3 individual genotypes for the STR  marker FGA
geno1 <- simugeno(tab,which.loc='FGA', n =100)
geno1@tab.geno
```

`simumix` *forensim class for DNA mixtures*

Description

The S4 `simumix` class is used to store DNA mixtures of individual genotypes along with informations about the individuals poulations and the loci used to simulate the genotypes.

Slots

`ncontri`**:** integer vector giving the number of contributors to the DNA mixture. If there are several populations, `ncontri` gives the number of contributors per population

`mix.prof`**:** matrix giving the contributors genotypes (in rows) for each locus (in columns). The genotype of a homozygous individual carrying the allele "12" is coded "12/12". A heterozygous individual carrying alleles "12" and "13" is coded "12/13" or "13/12".

`mix.all`**:** list giving the alleles present in the mixture for each locus

`which.loc`**:** character vector giving the locus names

`popinfo`**:** factor giving the population of each contributor

Methods

names `signature(x = "simumix")`: gives the names of the attributes of a simumix object

show `signature(object = "simumix")`: shows a simumix object

print `signature(object = "simumix")`: prints a simumix object

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

simugeno, as.simumix, is.simumix, simugeno and tabfreq

Examples

```
showClass("simumix")
data(strusa)
```

`simumix constructor`

simumix constructor

Description

Constructor for simumix objects.
The function `simumix` creates a simumix object from a tabfreq object.

The function `as.simumix` is an alias for `simumix` function.

`is.simumix` tests if an object is a valid simumix object.

Note: to get the manpage about simumix, please type 'class ? simumix'.

Usage

```
simumix(tab,which.loc=NULL,ncontri=1)
as.simumix(tab,which.loc=NULL,ncontri=1)
is.simumix(x)
```

Arguments

`tab`	a simugeno object created with constructor simugeno
`which.loc`	a character vector giving the chosen loci for the genotypes simulation. The default is set to NULL, which corresponds to all the loci of the `simugeno` object given in argument
`ncontri`	integer vector giving the number of individuals. If there are several populations, `ncontri` gives the numbers of individuals to simulate per population. Default is one.
`x`	an object

Details

DNA mixtures are created by randomly drawing individual genotypes with a uniform probability. If there are N individuals in the sample (the `simugeno` object), then each individual has a probability of 1/N to be selected.

Value

For `simumix` and `as.simumix`, a simumix object. For `is.simumix`, a logical.

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

See Also

"`simumix`", `simugeno` for creating a simugeno object.

Examples

```
data(Tu)
tab<-simugeno(tabfreq(Tu),n=1200)
#simulation of a 3-person mixture characterized with markers FGA, TH01 and TPOX
simumix(tab,which.loc=c('FGA','TH01', 'TPOX') , n =3)
```

`simupopD` *Simulate multi-population allele frequencies for independent loci from a reference population, following a Dirichlet model*

Description

Simulate multi-population allele frequencies for independent loci, from a given reference population, following a Dirichlet model. Allele frequencies in the populations are generated as random deviates from a Dirichlet distribution, whose parameters control the deviation of allele frequencies from the values in the reference population.

Usage

```
simupopD(npop = 1, nloc = 1, na = 2, globalfreq = NULL, which.loc = NULL,
alpha1, alpha2 = 1)
```

Arguments

`npop`	the number of populations
`nloc`	the number of loci
`na`	an integer vector giving the numbers of alleles per locus
`globalfreq`	matrix of allele frequencies in the reference population. Data must be given in the format of the Journal of Forensic Sciences for genetic data. Default corresponds to allele frequencies generated form a Dirichlet distribution with parameter `alpha2` for all allele frequencies.
`which.loc`	which loci to simulate from the `globalfreq` matrix, default considers all loci
`alpha1`	a positive float vector of length `npop` giving the variance parameter of the Dirichlet distribution used to generate allele frequencies in the `npop` independent populations
`alpha2`	a positive float giving the parameter to be used to in the Dirichlet distribution to generate allele frequencies for the reference population

Details

In the reference population, allele frequencies for independent loci are simulated using a Dirichlet distribution with parameter `alpha2`.
At a given locus L with n alleles, the allele frequencies are modeled as a vector of random variables p=(p1, ..., pn) following a Dirichlet distribution with a parameter vector of length n, where each component is equal to alpha2, p1+...+pn=1 and alpha2 > 0.
Note that a more sophisticated generation of global allele frequencies is possible using the `simufreqD` function. Similarly, allele frequencies in the independent populations are simulated using a Dirichlet Distribution. For example, for the first population to simulate, at a given locus L with n alleles, the allele frequencies are modeled as a vector of random variables p=(p1, ..., pn) following a Dirichlet distribution with a parameter vector of length n:
(p1(1-a1)/alpha1[1], ..., pn(1-alpha1[1])/alpha1[1]), where p1+...+pn=1 and alpha1[1] > 0.
alpha1[1] is the variance parameter for population 1 and is equivalent to Wright's Fst. The closest this parameter is to one, the more the population allele frequencies are different from the values of the reference population.

Value

The result is stored in a list with two elements :

`globfreq`	a `tabfreq` object giving the allele frequencies of the chosen reference population, with the chosen loci.
`popfreq`	a `tabfreq` object giving the allele frequencies of the simulated populations.

Note

The code used here for the generation of random Dirichlet deviates was previously implemented in the gtools library.

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

References

Nicholson G, Smith AV, Jonsson F, Gustafsson O, Stefansson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J Roy Stat Soc B* 2002;64:695–715

Marchini J, Cardon LR. Discussion on the meeting on "Statistical modelling and analysis of genetic data" *J Roy Stat Soc B*, 2002;64:740-741

Wright S. The genetical structure of populations. *Ann Eugen* 1951;15:323-354

See Also

`simufreqD`

Examples

```
# simulate allele frequencies for two populations
data(Tu)
simupopD(npop=2,globalfreq=Tu, which.loc=c("FGA","TH01","TPOX"),
alpha1=c(0.2,0.3),alpha2=1)
```

`strusa` *Allele frequencies for 15 autosomal short tandem repeats core loci on U.S. Caucasian, African American, and Hispanic populations.*

Description

Allele frequencies for 15 autosomal short tandem repeats loci on three American populations : Caucasians, African Americans and Hispanics. Among the 15 loci, 13 belong to the core Combined DNA Index System (CODIS) loci used by the Federal Bureau of Investigation (USA), in forensic DNA analysis, and two supplementary loci are more commonly used in Europe, see details.

Usage

```
data(strusa)
```

Format

`strusa` is a tabfreq object giving allele frequencies of 15 loci in three American populations.

Details

CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11, belong to the core CODIS loci used in the US, whereas D2S1338 and D19S433 belong to the European core loci.

References

Butler JM, Reeder DJ. http://www.cstl.nist.gov/strbase/index.htm, last visited: May 11th 2009

Butler JM, Schoske R, Vallone MP, Redman JW, Kline MC. Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. *J Forensic Sci* 2003;48(8):908-911.

Examples

```
data(strusa)
strusa
#genotypes simulations from each population
geno<- simugeno(strusa,n=c(100,100,100))
geno
#3-person mixture simulation with the contributors from the 3 populations
mix3<- simumix(geno,ncontri=c(1,1,1))
mix3
```

`strveneto` *Population study of three miniSTR loci in Veneto (Italy)*

Description

Allele frequencies for three short tandem repeats loci D10S1248, D2S441 and D22S1045 in a sample of 198 individuals born in Veneto, Italy. These loci are commonly used in forensic DNA characterization.

Usage

```
data(strveneto)
```

Format

`strveneto` is a tabfreq object

References

Turrina S, Atzei R, De Leo D. Population study of three miniSTR loci in Veneto (Italy). Forensic Sci Int Genetics 2008; 1(1);378-379

Examples

```
data(strveneto)
#allele frequencies
strveneto@tab
```

`tabDNAproxy` *Builds a list of tables that record the dropout events matched with the appropriate DNA proxies*

Description

The `tabDNAproxy` function builds a list of tables that record the dropout events matched with the appropriate "DNAproxies", these are the approximations of the amount of DNA contributed by the individuals in the analyzed DNA stains. Each table is specific to a locus. This function builds the data frames on which the logistic model, proposed by Tvedebrink et al (cf. references section), can be performed.

Usage

```
tabDNAproxy(x, y = NULL, geno, tabcsv)
```

Arguments

`x`	numeric label of the contributing individual, if the stain is a mixture, `x` should give the label of the first individual contributing to the mixture
`y`	numeric label of the second contributing individual, default is NULL. If the stain is a mixture, `y` should give the label of the second individual contribution to the mixture. This argument is skipped if the stain is not a mixture (default case: `y` set to NULL).
`geno`	a matrix giving the genotypes of the individuals contributing to the analyzed data for each locus. An individual genotype is given in rows for each locus in column. A homozygous carrying allele 9 is coded '9/9', a heterozygous carrying alleles 8 and 9 is coded '8/9'. Individual labels are coded using integers that are simply the order of introduction in the data frame.
`tabcsv`	a matrix giving the validation table of the analysed DNA stain. `tabcsv` must have a "genemapper" validation table structure, namely, information about the present alleles and the corresponding peak heights must be given.

Value

A list of length the number of analyzed loci, each component of the list is a matrix with the following information:

`Dloc`	the (per locus) dropout variable D, takes 1 if the allele has dropped out, 0 otherwise
`Hestim`	mean peak heights derived from the DNA proxies, see the references section for further details

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics*, 2009, 3(4), 222-226.

See Also

`recordDrop`, `DNAproxy`

Examples

```
#load the exemple data
data(dropdata)

tabcsv<-dropdata$tabcsv
genot<-dropdata$genot
#individuals' labels are 1 and 2
#lets record the dropout events and the corresponding DNA proxies
tabDNAproxy(1,2,geno=genot,tabcsv=tabcsv)
```

`tabfreq` *forensim class for population allele frequencies*

Description

The S4 `tabfreq` class is used to store allele frequencies, from either one or several populations.

Slots

`tab`**:** a list giving allele frequencies for each locus. If there are several populations, `tab` gives allele frequencies in each population

`which.loc`**:** character vector giving the names of the loci

`pop.names`**:** factor of populations names (optional)

Methods

names `signature(x = "tabfreq")`: gives the names of the attributes of a tabfreq object

show `signature(object = "tabfreq")`: shows a tabfreq object

print `signature(object="tabfreq")`: prints a tabfreq object

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

See Also

`as.tabfreq`, `is.tabfreq` and `simugeno` for genotypes simulation from allele frequencies stored in a `tabfreq` object

Examples

```
showClass("tabfreq")
```

`tabfreq constructor`

tabfreq constructor

Description

Constructor for tabfreq objects.
The function `tabfreq` creates a tabfreq object from a data frame or a matrix giving allele frequencies for a single population in the Journal of Forensic Sciences (JFS) format for population genetic data. Whene multiple populations are considered, data shall be given as a list, where each element is either a matrix or a data frame in the JFS format, and the populations names must be specified.

The function `as.tabfreq` is an alias for the `tabfreq` function.

`is.tabfreq` tests if an object is a valid tabfreq object.

Note: to get the manpage about tabfreq, please type 'class ? tabfreq'.

Usage

```
tabfreq(tab,pop.names=NULL)
as.tabfreq(tab,pop.names=NULL)
is.tabfreq(x)
```

Arguments

`tab`	either a matrix or a data.frame of markers allele frequencies given in the Journal of Forensic Sciences format for population genetic data
`pop.names`	(optional) a factor giving the populations names. For a single population in `tab`, default is set to NULL.
`x`	an object

Value

For `tabfreq` and `as.tabfreq`, a tabfreq object. For `is.tabfreq`, a logical.

Author(s)

Hinda Haned `<haned@biomserv.univ-lyon1.fr>`

See Also

`"tabfreq"`, `simugeno` for creating a simugeno object from a tabfreq object.

Examples

```
data(Tu)
tabfreq(Tu,pop.names=factor("Tu"))
```

`tabSPH` *Builds a matrix of the dropout variable and the corresponding surviving peak heights*

Description

The `tabSPH` function builds a matrix of the dropout variable and the corresponding surviving peak heights, for each available locus or across all loci (default). The constructed matrices have two columns: the dropout variable and the surviving peak heights. The logistic model proposed to model the dropout probability from experimental data (see the references section) can be performed directly on the data yielded by `tabSPH`.

Usage

```
tabSPH(x, y = NULL, geno, tabcsv, byloc = FALSE,s=40)
```

Arguments

x	numeric label of the contributing individual, if the stain is a mixture, x should give the label of the first individual contributing to the mixture
y	numeric label of the second contributing individual, default is NULL. If the stain is a mixture, y should give the label of the second individual contribution to the mixture. This argument is skipped if the stain is not a mixture (default case: y set to NULL).
geno	a matrix giving the genotypes of the individuals contributing to the analyzed data for each locus. An individual genotype is given in rows for each locus in column. A homozygous carrying allele 9 is coded '9/9', a heterozygous carrying alleles 8 and 9 is coded '8/9'. Individual labels are coded using integers that are simply the order of introduction in the data frame.
tabcsv	a matrix giving the validation table of the analysed DNA stain. tabcsv must have a "genemapper" validation table structure, namely, information about the present alleles and the corresponding peak heights must be given.
byloc	logical indicating whether data should be displayed per locus (TRUE) or overall loci (FALSE, default)
s	numeric giving the detection threshold for alleles in Relative fluorescence units (RFU) , default is set to 40 RFUS. An observed allele with a peak height smaller (<) than 40 RFUS is considered as dropped-out.

Details

Both mixed and unmixed samples can be used in tabSPH, setting the y argument to NULL (default) will produce results considering data for x only. In case of mixtures, note that only heterozygote genotypes with no shared alleles are considered.

Value

If argument byloc is TRUE, tabSPH yields a list of length the number of available loci, each elements of the list contain a matrix with two columns:

D	the dropout variable
H	the surviving peak height

If argument byloc is FALSE, tabSPH yields a single matrix with columns D and H described above.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

References

Gill P, Puch-Solis R, Curran J. The low-template-DNA (stochastic) threshold-Its determination relative to risk analysis for national DNA databases. *Forensic Science International: Genetics*, 2009, 3, 104-111

See Also

tabDNAproxy

Examples

```
#load the example data
data(dropdata)
tabcsv<-dropdata$tabcsv
genot<-dropdata$genot
#individuals' labels are 1 and 2
#recording dropout variable matched with the surviving peak heights
#for heterozygotes whith non shared alleles
tabSPH(1,2,geno=genot,tabcsv=tabcsv,s=0)
```

`Tu` *Allele frequencies of 15 autosomal short tandem repeats loci on Chinese Tu ethnic minority group*

Description

Population genetic analysis of 15 STR loci of Chinese Tu ethnic minority group.

Usage

```
data(Tu)
```

Format

a data frame presented in the format of the Journal of Forensic Sciences for genetic data: allele names are given in the first column, and frequencies for a given allele are read in rows for the different markers. When a given allele is not observed, value is coded NA (rather than "-" in the original format).

Details

CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11, belong to the core CODIS loci used in the US, whereas D2S1338 and D19S433 belong to the European core loci.

References

Zhu B, Yan J, Shen C, Li T, Li Y, Yu X, Xiong X, Muf H, Huang Y, Deng Y. (2008). Population genetic analysis of 15 STR loci of Chinese Tu ethnic minority group. *Forensic Sci Int*; 174: 255-258.

Examples

```
data(Tu)
tabfreq(Tu)
```

`virtualClasses` *Virtual classes for forensim*

Description

Virtual classes that are only for internal use in forensim

Objects from the Class

A virtual Class: programming tool, not intended for objects creation.

Author(s)

Hinda Haned <haned@biomserv.univ-lyon1.fr>

`wrapdataL` *ML estimate of number of contributors for SNPs*

Description

Wrap up of dataL in forensim. Given file with columns: "No, Marker, Allele, Frequency and Height" the log likelihood for requested number of contributors is calculated. For now only "Frequency" column is used.

Usage

```
wrapdataL(fil , plotte , nInMixture , tit )
```

Arguments

`fil`	Input file
`plotte`	If T, plot
`nInMixture`	Alternatives for number of contributors, say 1:5
`tit`	Title to be used in plot

Value

Plot (optional) and log likelihoods

Author(s)

Thore Egeland <Thore.Egeland@medisin.uio.no>

Examples

```
aa<-simMixSNP(nSNP=5,writeFile=TRUE,outfile="sim.txt",ncont=3) #Simulates data
res<-wrapdataL(fil="sim.txt") # Calculates and plots
```

Index

Chapter 2

Appendix B: Forensim Tutorial

A tutorial for the R package *forensim*

Hinda Haned

September 12, 2010

Contents

1 Introduction

This tutorial is a presentation of the forensim package for the R software [1, 2]. forensim is dedicated to the interpretation of forensic DNA mixtures through statistical methods. It also provides simulation tools that allow the generation of genetic data commonly encountered in forensic casework.

In this tutorial, I first introduce forensim object classes. Then, I present statistical tools for forensic DNA mixtures interpretation. Finally, various functionalities of forensim are explored. For all addressed topics, practical and reproducible examples are given.

2 Getting started

2.1 forensim installation

The current version of the package is 1.1-8 and is compatible with R 2.11.1 forensim is hosted by R-Forge, the latest version of the package, resulting from the nightly build, can be obtained by typing the following command lines:
Under Windows and Linux

```
> install.packages("forensim",repos="http://r-forge.r-project.org")
```

Under the MacOS system

```
> install.packages("forensim", repos="http://r-forge.r-project.org", type = 'source')
```

Please be aware that this is the development version. To be sure to get the latest stable version, download the forensim package (according to your platform) on forensim web page: `http://forensim.r-forge.r-project.org/`.
Then, the package must be loaded:

```
> library(forensim)
```

```
### forensim 1.1.8 is loaded ###
```

2.2 How to get help

- The mailing list: please ask questions on forensim mailing list, `forensim-help@lists.r-forge.r-project.org`
- The help pages: classes and functions are documented in the help pages, type ?forensim in R to get an overview of the package.
- The forensim package manual: a compilation of all the help pages in a single pdf file, it can be found at: `http://forensim.r-forge.r-project.org/`

3 Generating data in forensim

forensim provides object classes that facilitate the generation and the storage of data that is commonly encountered in forensic casework: population allele frequencies, individual genotypes and DNA mixtures. Thus, three classes of objects are defined in forensim:

- tabfreq objects: used to store allele frequencies
- simugeno objects: used to store genotypes
- simumix objects: used to store DNA mixtures

forensim objects have the particularity that they can either be used to store pre-existing data, such as allele frequencies in a given population, or simulated data. Creating forensim objects is achieved using specific functions, called constructors, that have the same names than the object they are linked to.

3.1 tabfreq objects

In forensim, allele frequencies are stored in tabfreq objects. Importing data into tabfreq objects is achieved using the tabfreq constructor. The input data must be an object of type data frame[1] or matrix. This object must have the format of the *Journal of Forensic Sciences* for Short Tandem Repeat (STR) loci data: allele names (the number of tandem repeats in case of STR loci) are given in the first column, and frequencies for a given allele are read in rows for different loci given in columns. When an allele is not observed for a given locus, value is coded "NA"[2]. Note that even if the requested input format is based on STR data, different kinds of markers can be imported in forensim.
As an example, we will be using a data set included in forensim:

```
> data(Tu)
```

What is the class of object Tu ?

```
> class(Tu)
```

```
[1] "data.frame"
```

Tu is a data frame giving the allele frequencies for 15 STR loci commonly used in forensic studies, in the Tu Chinese population [3] (see ?Tu). Note that the data set is imported using the command data.
Displaying the first rows (command head):

```
> head(Tu)
```

[1] in R a data frame is a collection of variables, possibly of different types
[2] non observed alleles are coded "-" in the Journal of Forensic Sciences

```
  Allele D8S1179 D21S11 D7S820 CSF1PO D3S1358   TH01 D13S317 D16S539 D2S1338
1    6.0      NA     NA     NA     NA      NA 0.1151      NA      NA      NA
2    7.0      NA     NA 0.0033 0.0034      NA 0.2599      NA      NA      NA
3    8.0  0.0098     NA 0.1382 0.0034      NA 0.0559  0.2712  0.0097      NA
4    9.0      NA     NA 0.0493 0.0582      NA 0.4605  0.1503  0.2305      NA
5    9.2      NA     NA 0.0033     NA      NA     NA      NA      NA      NA
6    9.3      NA     NA     NA     NA      NA 0.0691      NA      NA      NA
  DS19S433 vWA   TPOX D18S51 D5S818 FGA
1       NA  NA     NA     NA     NA  NA
2       NA  NA     NA     NA 0.0097  NA
3       NA  NA 0.5359     NA     NA  NA
4       NA  NA 0.1340     NA 0.0487  NA
5       NA  NA     NA     NA     NA  NA
6       NA  NA     NA     NA     NA  NA
```

This data frame is converted into a tabfreq object by the tabfreq constructor:

```
> tupop <- tabfreq(tab = Tu, pop.names = as.factor("Tu"))
```

The population name is specified as a factor in the pop.names argument.

```
> is.tabfreq(tupop)
```

```
[1] TRUE
```

tupop is a tabfreq object:

```
> tupop
```

```
   # Tabfreq object: allele frequencies  #

@tab: list of allele frequencies
@which.loc: vector of  15  locus names
@pop.names:  populations names
```

As a formal class object, tupop is constituted of different 'slots' that contain different types of information. Each slot can be accessed using '@' or the '$' operator that have been implemented for all forensim objects.
Allele frequencies are stored in the @tab slot. For example, frequencies for locus FGA are given by:

```
> tupop$tab$Tu$FGA
```

```
    18     19   19.2     20     21     22   22.2     23   23.2     24     25
0.0392 0.0686 0.0033 0.0458 0.0980 0.1765 0.0033 0.1961 0.0098 0.2222 0.1013
  25.2     26   26.2     27
0.0065 0.0131 0.0065 0.0098
```

Population names are stored in the @pop.names argument:

```
> tupop$pop.names
```

```
[1] Tu
Levels: Tu
```

Finally, locus names appearing in @tab can be accessed elsewhere:

```
> tupop$which.loc
```

```
 [1] "D8S1179"  "D21S11"   "D7S820"   "CSF1PO"   "D3S1358"  "TH01"
 [7] "D13S317"  "D16S539"  "D2S1338"  "DS19S433" "vWA"      "TPOX"
[13] "D18S51"   "D5S818"   "FGA"
```

Note that if several populations are imported in the same tabfreq object, data frames (or matrices) must be given as a list of data frames (or matrices) in the tab argument. In this case, the pop.names argument, which is optional when a single population is handled, becomes obligatory in order to distinguish the populations.

IMPORTANT NOTE: In order to allow reproducibility of the simulations in this tutorial by other users, the random seed is set:

```
> set.seed(123560)
```

3.2 simugeno objects

simugeno objects are used to store simulated genotypes from a tabfreq object. simugeno objects are created from tabfreq objects by specifying the number of individuals to simulate in the n argument. The loci to take into account for the simulation are given in the which.loc argument. For the illustration purpose, 10 individuals are simulated and only three loci are chosen: D8S1179, TH01 and FGA.

```
> tugeno <- simugeno(tab = tupop, n = 10, which.loc = c("D8S1179",
+     "TH01", "FGA"))
```

```
> tugeno
```

```
   # Simugeno object: simulated genotypes #

@which.loc: vector of  3  locus names
@nind: 10
@indID:  vector of the individuals ID
@tab.geno:  10 x 3 data frame of genotypes
@tab.freq:  allele frequencies for the 3  loci

Population-related information:
@pop.names: population names
@popind: factor giving the population of each individual
```

@tab.geno is a matrix of 10 genotypes simulated from the allele frequencies of the Tu population. For instance, the genotypes of the five first simulated individuals are:

```
> tugeno$tab.geno[1:5, ]
```

```
     D8S1179 TH01      FGA
ind1 "15/13" "7/7"     "23/19"
ind2 "14/12" "9/7"     "26/18"
ind3 "15/12" "7/7"     "24/19"
ind4 "11/13" "9.3/9.3" "24/22"
ind5 "16/14" "9/6"     "22/23.2"
```

The genotype of a homozygous individual carrying the allele 9 is coded "9/9". A heterozygous individual carrying alleles 8 and 10 is coded "8/10".

Allele frequencies of the population are stored in the slot @tab.freq:

```
> tugeno$tab.freq
```

```
$Tu
$Tu$D8S1179
     8     10     11     12     13     14     15     16     17
0.0098 0.0784 0.0784 0.1046 0.2876 0.1863 0.1634 0.0719 0.0196

$Tu$TH01
     6      7      8      9    9.3     10
0.1151 0.2599 0.0559 0.4605 0.0691 0.0395

$Tu$FGA
    18     19   19.2     20     21     22   22.2     23   23.2     24     25
0.0392 0.0686 0.0033 0.0458 0.0980 0.1765 0.0033 0.1961 0.0098 0.2222 0.1013
  25.2     26   26.2     27
0.0065 0.0131 0.0065 0.0098
```

simugeno objects also contain information about the simulated individuals, their (default) ID:

```
> tugeno@indID
```

```
 [1] "ind1"  "ind2"  "ind3"  "ind4"  "ind5"  "ind6"  "ind7"  "ind8"  "ind9"
[10] "ind10"
```

and their population names:

```
> tugeno@popind
```

```
 [1] Tu Tu Tu Tu Tu Tu Tu Tu Tu Tu
Levels: Tu
```

3.3 simumix objects

simumix objects store DNA mixtures. Mixtures can be created from simugeno objects using the constructor simumix. The number of contributors is specified in the argument ncontri.

```
> mix2 <- simumix(tugeno, ncontri = 2)
```

Constructor simumix has also a which.loc argument, which is by default set to NULL, corresponding to all loci taken into account.

```
> mix2
```

```
   # Simumix object: simulated mixture #

@which.loc: vector of  3  locus names
@ncontri: 2
@mix.prof:  2 x 3 data frame of the contributors genotypes
@mix.all: list of the alleles found in the mixture
@popinfo: populations of the contributors
```

simumix objects keep two types of information: information usually available when dealing with practical cases of forensic DNA mixtures: the alleles present by locus,

```
> mix2$mix.all


$D8S1179
[1] "12" "13" "14" "16"

$TH01
[1] "6" "7" "9"

$FGA
[1] "22"   "23"   "23.2" "25"
```

and information that is usually not available: the number of simulated contributors

```
> mix2@ncontri


[1] 2
```

and their genetic profiles:

```
> mix2$mix.prof


     D8S1179 TH01  FGA
ind5 "16/14" "9/6" "22/23.2"
ind7 "13/12" "9/7" "23/25"
```

3.4 Allele frequencies simulation

In the following, we denote L a locus with k alleles and the ith allele frequency at this locus, in a given population, is denoted p_i.

3.4.1 The homogeneous population case

In forensim, allele frequencies for a single non subdivided population are simulated using the simufreqD function.

Principle

The vector of allele frequencies at locus L is simulated as a vector of random deviates of the Dirichlet distribution [4] with a vector of parameters $(\alpha_1, ..., \alpha_k)$:

$(p_1, ..., p_k) \rightsquigarrow Dirichlet(\alpha_1, ..., \alpha_k)$

An example

5 loci (argument nloc=5) having 2, 3, 4, 5 and 6 alleles respectively (argument na) are simulated:

```
> simufreqD(nloc = 5, na = c(2, 3, 4, 5, 6), alpha = 1)
```

```
  Allele Marker1 Marker2 Marker3 Marker4 Marker5
1      1    0.21    0.44   0.280   0.650   0.110
2      2    0.79    0.36   0.052   0.096   0.076
3      3      NA    0.20   0.170   0.080   0.032
4      4      NA      NA   0.500   0.100   0.500
5      5      NA      NA      NA   0.068   0.095
6      6      NA      NA      NA      NA   0.190
```

Argument alpha is the parameter of the Dirichlet distribution. Setting a single value for alpha means that all alleles for all loci are simulated with the same value; this can be changed by giving the appropriate values in alpha, for further details please type '?simufreqD'.
Setting alpha to 1, leads to the generation of allele frequencies as random deviates from a uniform Dirichlet distribution, this means that allele frequencies could take any value varying from 0 to 1, with equal probabilities. Note that the simulated data is in the format of the *Journal of Forensic Sciences* for STR loci data.

3.4.2 The subdivided population case

Principle

The simupopD function simulates subpopulations allele frequencies for independent loci, from a given reference population, following a Dirichlet model.
Allele frequencies in the subpopulations are generated as random deviates from a Dirichlet distribution, whose parameters control the deviation of allele frequencies from the values in the reference population.
Each allele frequency is modeled as a random variable; with a parameter $\alpha_i = \frac{p_i(1-\theta)}{\theta}$, where θ is Wright's *Fst* coefficient which allows here accounting for population subdivision [5, 6]. The vector of allele frequencies at a given locus, for a given population, is obtained by:

$$(p_1, ..., p_k) \rightsquigarrow Dirichlet\left(\alpha_1 = \frac{p_1(1-\theta)}{\theta}, ..., \alpha_k = \frac{p_k(1-\theta)}{\theta}\right)$$

An example

In the following example we simulate allele frequencies in two subpopulations: the global population is taken as the Tu Chinese population, and three STR loci are chosen: FGA, TH01 and TPOX. The strength of the deviation from the reference allele frequencies is specified in argument alpha1 for each simulated subpopulation, here we choose 0.01 for the first population and 0.3 for the second one:

```
> simpop1 <- simupopD(npop = 2, globalfreq = Tu, which.loc = c("FGA",
+     "TH01", "TPOX"), alpha1 = c(0.01, 0.3))
```

simpop1 is a list of two tabfreq object; the first one contains allele frequencies used for the simulation (from the Tu population):

```
> simpop1$globfreq
```

```
   # Tabfreq object: allele frequencies  #

@tab: list of allele frequencies
@which.loc: vector of  3  locus names
@pop.names:  - empty -
```

the second tabfreq object contains the subpopulations allele frequencies:

```
> simpop1$popfreq
```

```
   # Tabfreq object: allele frequencies  #

@tab: list of allele frequencies
@which.loc: vector of  3  locus names
@pop.names:  populations names
```

The simulated subpopulations have the following (default) names:

```
> simpop1$popfreq$pop.names
```

```
[1] pop1 pop2
Levels: pop1 pop2
```

4 Statistical methods for forensic DNA mixtures interpretation

Several statistical methods dedicated to the interpretation of forensic DNA mixtures are implemented in forensim:

4.1 The maximum allele count

This method consists in setting the lower bound on the number of contributors to a mixture to the minimum required to explain the observed profiles [7]. For instance, if a mixture shows at three loci, 1, 3 and 4 alleles, then the number of contributors is bounded to 2 $\left(\frac{4}{2}\right)$ contributors.

To exemplify this method, let us simulate a 3-person mixture from the strusa data set, using the allele frequencies from the Caucasian population [8] (see ?strusa):

```
> data(strusa)
> class(strusa)
```

```
[1] "tabfreq"
attr(,"package")
[1] ".GlobalEnv"
```

```
> strusa


   # Tabfreq object: allele frequencies  #


@tab: list of allele frequencies
@which.loc: vector of  15  locus names
@pop.names:  populations names
```

strusa is a tabfreq object that contains multiple populations:

```
> strusa$pop.names


[1] Afri Cauc Hisp
Levels: Afri Cauc Hisp
```

thus, the number of genotypes to simulate must be specified in each population (argument n):

```
> geno <- simugeno(tab = strusa, n = c(0, 100, 0))
```

100 genotypes are simulated from the Caucasian population allele frequencies, no genotypes are simulated from the other two populations.
A 3-person mixture is simulated by randomly drawing three contributors from these 100 simulated individuals. The number of contributors in each population must be specified:

```
> mix3 <- simumix(tab = geno, ncontri = c(0, 3, 0))
```

The minimum number of contributors required is computed by the mincontri function. This number can either be computed from all available loci simultaneously (in this default case, the argument loc is set to NULL),

```
> mincontri(mix3, loc = NULL)


[1] 3
```

or be computed for a specific locus, for example, D8S1179:

```
> mincontri(mix3, loc = "D8S1179")


[1] 2
```

4.2 The maximum likelihood estimator

The main characteristic of this method is that it takes into account allele frequencies in the estimations. The likelihood function is derived from the formula of Curran *et al* [9] for DNA mixtures interpretation, in the particular case where all contributors to the mixture are unknown and there are no typed individuals [10].

4.2.1 Likelihood of the observed alleles at a given locus, conditional on the number of contributors to the mixture

The function lik.loc computes the likelihood of the observed alleles at a given locus, conditional on the number of contributors to the mixture [10]. This function takes in argument the number of contributors x, the mixture as a simumix object, and the allele frequencies given in a tabfreq object. For the previously simulated 3-person mixture mix3,

```
> mix3


   # Simumix object: simulated mixture #


@which.loc: vector of  15  locus names
@ncontri: 3
@mix.prof:  3 x 15 data frame of the contributors genotypes
@mix.all: list of the alleles found in the mixture
@popinfo: populations of the contributors
```

the likelihood per locus of observing alleles given that 1 individual contributed to the mixture is:

```
> lik.loc(x = 1, mix = mix3, freq = strusa, refpop = "Cauc")


   CSF1PO       FGA      TH01      TPOX       VWA   D3S1358    D5S818    D7S820
0.0000000 0.0586168 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  D8S1179   D13S317   D16S539    D18S51    D21S11   D2S1338   D19S433
0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
```

the likelihood that 3 individuals contributed to the mixture is:

```
> lik.loc(x = 3, mix = mix3, freq = strusa, refpop = "Cauc")


     CSF1PO         FGA        TH01        TPOX         VWA     D3S1358
0.015414029 0.001808615 0.163094342 0.095796419 0.071597218 0.099698106
     D5S818      D7S820     D8S1179     D13S317     D16S539      D18S51
0.280534836 0.004101536 0.023984786 0.011244765 0.107510776 0.012642508
     D21S11     D2S1338     D19S433
0.005385985 0.004742859 0.030669330
```

Note here that strusa contains three populations, so the reference population, here Caucasians, must be specified in the refpop argument.
The overall likelihood, for all loci characterized in the mixture can be computed using the function lik:

```
> lik(x = 3, mix = mix3, freq = strusa, refpop = "Cauc")


[1] 1.027420e-24
```

4.2.2 Maximum likelihood estimators

likestim.loc looks for the number of contributors that maximizes the likelihood at each given locus. For the estimations to be biologically plausible, the estimations are restricted to the discrete interval [1,6] [10]. These functions give the number of contributors that maximizes the likelihood (max) and the corresponding likelihood value (maxval). The per locus estimations are:

```
> likestim.loc(mix = mix3, freq = strusa, refpop = "Cauc")
```

```
         max maxval
CSF1PO     5 0.0240
FGA        1 0.0590
TH01       3 0.1600
TPOX       3 0.0960
VWA        4 0.0740
D3S1358    4 0.1200
D5S818     3 0.2800
D7S820     3 0.0041
D8S1179    2 0.0400
D13S317    6 0.0300
D16S539    2 0.1100
D18S51     4 0.0170
D21S11     4 0.0088
D2S1338    3 0.0047
D19S433    2 0.0370
```

and the estimation using all loci simultaneously is:

```
> likestim(mix = mix3, freq = strusa, refpop = "Cauc")
```

```
     max maxval
[1,]   3  1e-24
```

4.3 The exclusion probability

The exclusion probability, also known as the Random Man Not Excluded (RMNE) is implemented in forensim in the function PE.
The PE function takes a simumix object for which to compute the exclusion probability and the allele frequencies given in a tabfreq object. If the latter contains several populations, than the reference population must be specified in the refpop argument. Implementation of the PE function includes the possibility of correcting for deviation from Hardy Weinberg proportions in the population, due to subdivision, using Wright's *Fst* called here theta [11]:

```
> PE(mix3, strusa, refpop = "Cauc", theta = 0, byloc = TRUE)
```

```
           PE_l
CSF1PO  0.2125
FGA     0.8756
TH01    0.3763
TPOX    0.3037
VWA     0.3815
D3S1358 0.3065
D5S818  0.2154
D7S820  0.6526
D8S1179 0.6584
D13S317 0.3037
D16S539 0.4225
D18S51  0.5188
D21S11  0.4474
D2S1338 0.6487
D19S433 0.5482
```

The row PE_l stands for the exclusion probability per locus, read in column. The byloc argument is a logical indicating whether the exclusion probability should be computed per locus (byloc=TRUE) or for all loci (byloc=FALSE):

```
> PE(mix = mix3, freq = strusa, refpop = "Cauc", theta = 0, byloc = FALSE)
```

```
       PE
0.999971
```

4.4 The random match probability

The Random Match Probability (RMP) is computed using the RMP function which implements the formulas gave by Balding and Nichols [12]. The suspect's profile can either be given directly in R as matrix, or be read from a text file.

DNA evidence as a matrix

```
> datas <- matrix(c("CSF1PO", "FGA", "TH01", "TPOX", "VWA", "D3S1358",
+     "D5S818", "D7S820", "D8S1179", "D13S317", "D16S539", "D18S51",
+     "D21S11", "D2S1338", "D19S433", "12/11", "22/19", "6/7",
+     "10/8", "17/18", "18/17", "12/12", "8/8", "13/13", "11/11",
+     "12/10", "14/15", "33.2/32.2", "23/22", "14/14"), nc = 2)
> colnames(datas) <- c("locus", "genotype")
> datas
```

```
      locus     genotype
 [1,] "CSF1PO"  "12/11"
 [2,] "FGA"     "22/19"
 [3,] "TH01"    "6/7"
 [4,] "TPOX"    "10/8"
 [5,] "VWA"     "17/18"
 [6,] "D3S1358" "18/17"
 [7,] "D5S818"  "12/12"
 [8,] "D7S820"  "8/8"
 [9,] "D8S1179" "13/13"
[10,] "D13S317" "11/11"
[11,] "D16S539" "12/10"
[12,] "D18S51"  "14/15"
[13,] "D21S11"  "33.2/32.2"
[14,] "D2S1338" "23/22"
[15,] "D19S433" "14/14"
```

The random match probability in the unrelated case (unknown offender and suspect are not related) and in absence of population subdivision (theta=0,default case) is given by [1]:

```
> RMP(suspect = datas, freq = strusa, refpop = "Cauc")
```

```
$RMP.loc
 CSF1PO     FGA    TH01    TPOX     VWA D3S1358  D5S818  D7S820 D8S1179 D13S317
 0.2200  0.0230  0.0880  0.0600  0.1100  0.0660  0.1500  0.0230  0.0930  0.1200
D16S539  D18S51  D21S11 D2S1338 D19S433
 0.0370  0.0440  0.0045  0.0090  0.1400

$RMP
[1] 6.2e-20
```

[1] RMP calls many functions from the genetics package which is now obsolete. So don't worry if you get a warning message from the genetics package.

In the absence of population subdivision, and in the case where the suspect and an unknown offender are for example siblings, the k argument must be modified from k=(1,0,0) to k=c(1/4,1/2,1/4):

```
> RMP(suspect = datas, freq = strusa, k = c(1/4, 1/2, 1/4), refpop = "Cauc")


$RMP.loc
 CSF1PO     FGA    TH01    TPOX     VWA D3S1358  D5S818  D7S820 D8S1179 D13S317
   0.47    0.32    0.38    0.41    0.40    0.36    0.48    0.33    0.43    0.45
D16S539  D18S51  D21S11 D2S1338 D19S433
   0.35    0.34    0.28    0.29    0.47

$RMP
[1] 4.6e-07
```

DNA evidence read from an existing text file The same data is available in a preexisting file "exprofile.txt" from the forensim package, accessed by the system.file command:

```
> RMP(filename = system.file("files/exprofile.txt", package = "forensim"),
+     freq = strusa, refpop = "Cauc")


$RMP.loc
 CSF1PO     FGA    TH01    TPOX     VWA D3S1358  D5S818  D7S820 D8S1179 D13S317
 0.2200  0.0230  0.0880  0.0600  0.1100  0.0660  0.1500  0.0230  0.0930  0.1200
D16S539  D18S51  D21S11 D2S1338 D19S433
 0.0370  0.0440  0.0045  0.0090  0.1400

$RMP
[1] 6.2e-20
```

4.5 Likelihood ratios

Likelihood ratios are computed using the LR function which implements the general formula of Curran *et al* for forensic DNA mixtures interpretation [13].

An example Consider the following genetic profiles from a rape case in Hong Kong [14]:

Locus	Mixture	Victim	Suspect	Frequency
D3S1358	14		14	0.033
	15	15		0.331
	17		17	0.239
	18	18		0.056

Table 1: Alleles from a DNA stain from a rape case in Hong Kong

Locus D3S1358 shows 4 distinct alleles (14, 15, 17 and 18), thus, the number of contributors to the mixed sample is taken to be 2.

Scenario 1 The following hypotheses are tested:
Prosecution hypotheses Hp: Contributors were the victim and the suspect.
Defense hypotheses Hd: Contributors were 2 unknown people.
First, the genotypes are assigned to the victim and the suspect:

```
> victim <- "15/18"
> suspect <- "14/17"
```

Then, the likelihood ratio is computed using the LR function:

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = NULL, Vd = NULL,
+     xd = 2)
```

```
[1] 285
```

The mixture profile is nearly 285 times more likely if it came from the suspect and the victim than if it came from two unknown unrelated individuals from the population of Hong Kong.

Scenario 2 The following hypotheses are tested:
Prosecution hypotheses Hp: Contributors were the victim and the suspect.
Defense hypotheses Hd: Contributors were the victim and one unknown.

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = victim,
+     Vd = suspect, xd = 1)
```

```
[1] 63.4
```

The mixture profile is 63 times more likely if it came from the suspect than if it came from an unrelated individual from the population of Hong Kong.

5 Two-person DNA mixtures resolution using allele peak heights or areas information: The *mastermix* interface

mastermix is a Tcl/Tk graphical user interface dedicated to the resolution of two-person DNA mixtures using allele peak heights or areas information. mastermix is the implementation of a method developed by Gill *et al* [15] and previously programmed into an Excel macro by Dr. Peter Gill.

This method searches through simulation the most likely combination(s) of the contributors' genotypes. Having previously obtained an estimation for the mixture proportion, it is possible to reduce the number of possible genotype combinations by keeping only those supported by the observed data. This is achieved by computing the sum of square differences between the expected allelic ratio and the observed allelic ratio, for all possible mixture combinations. The likelihood of peak heights

(or areas), given the combination of genotypes, is high if the residuals are low. Genotype combinations are thus selected according to the peak heights with the highest likelihoods. Appendix A gives the formulas for the expected allelic ratios following from [15].

Typing mastermix() in the R console launches a dialog window (Figure 1):

Figure 1: The mastermix interface

mastermix offers a graphical representation of the simulation for three models:

- The two allele model: at a given locus, two alleles are observed in the DNA stain
- The three allele model: at a given locus, three alleles are observed in the DNA stain
- The four allele model: at a given locus, four alleles are observed in the DNA stain

A left-click on each button launches a simulation dialog window for the corresponding model, while a right-click opens the corresponding help page. For instance, a left-click on the "Two-allele model" button yields Figure 2:

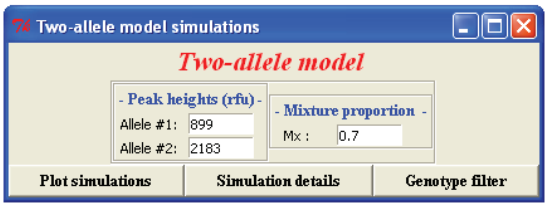


Figure 2: Two-allele model interface.

Note that default values for peak heights and observed mixture proportion are only given for illustration purposes.
As an example, we suppose that a locus showing four distinct alleles gives an estimation for the mixture proportion of 0.70, and that another locus shows two distinct alleles with heights of 899 and 2183 rfus. A left-click on the "Plot simulations" button yields a graphical representation of the residuals of each possible genotype combinations of the peak areas, for varying values of the mixture proportion across the interval [0.1, 0.9].

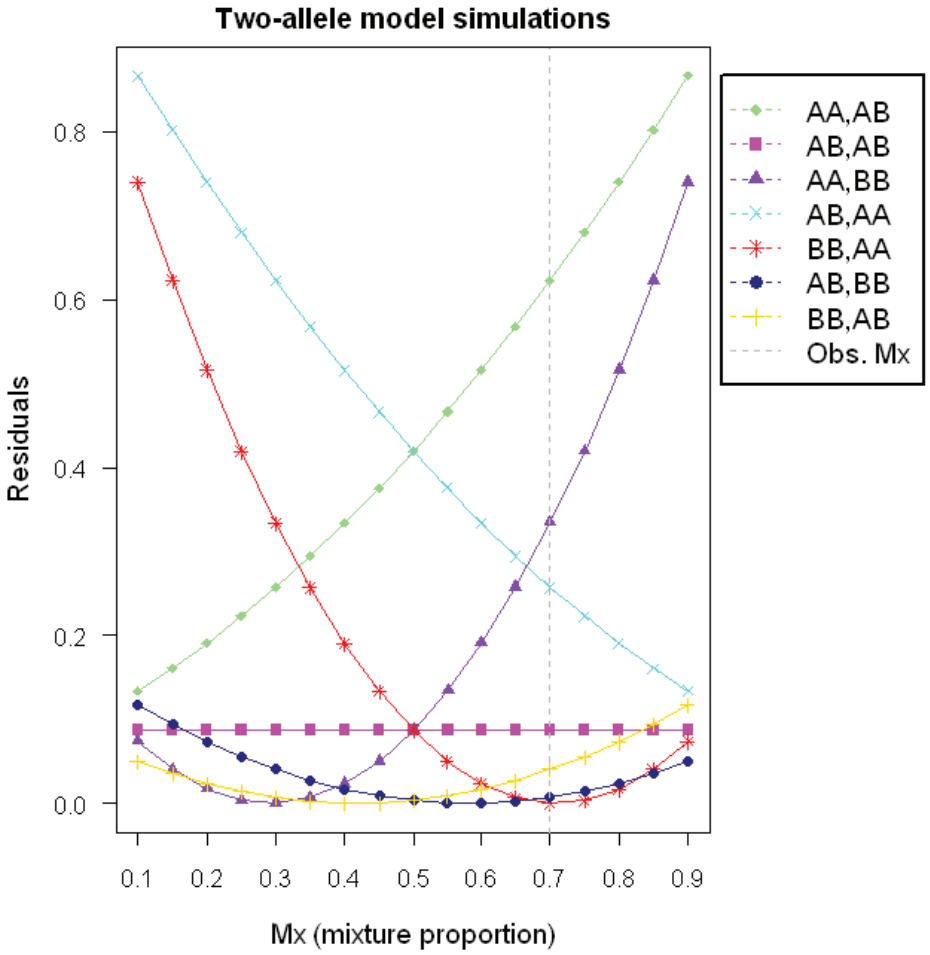
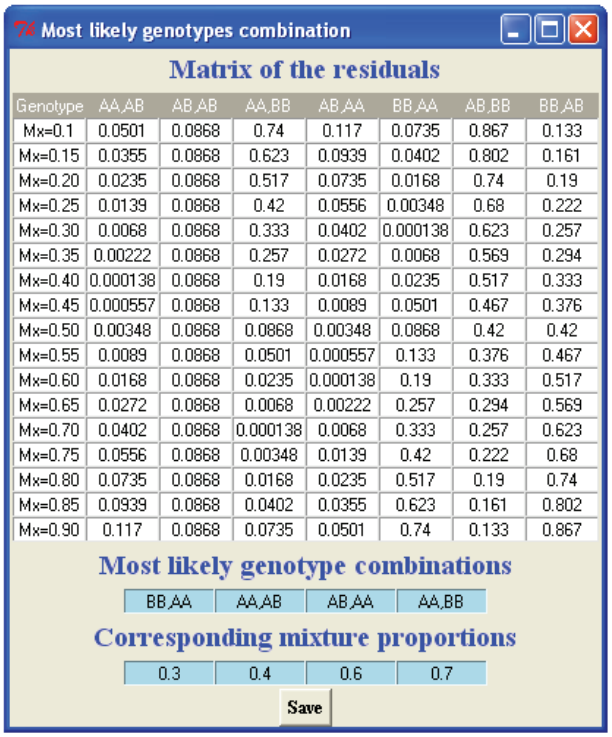


Figure 3: Graphical simulations of the residuals for each possible genotype combination, in a two-allele model, for every possible mixture combination based on variation of the mixture proportion.

The graphical simulation shows that multiple combinations correspond to the lowest residual value. The corresponding numerical results are obtained by clicking the "Simulations details" button:



Genotype	AA,AB	AB,AB	AA,BB	AB,AA	BB,AA	AB,BB	BB,AB
Mx=0.1	0.0501	0.0868	0.74	0.117	0.0735	0.867	0.133
Mx=0.15	0.0355	0.0868	0.623	0.0939	0.0402	0.802	0.161
Mx=0.20	0.0235	0.0868	0.517	0.0735	0.0168	0.74	0.19
Mx=0.25	0.0139	0.0868	0.42	0.0556	0.00348	0.68	0.222
Mx=0.30	0.0068	0.0868	0.333	0.0402	0.000138	0.623	0.257
Mx=0.35	0.00222	0.0868	0.257	0.0272	0.0068	0.569	0.294
Mx=0.40	0.000138	0.0868	0.19	0.0168	0.0235	0.517	0.333
Mx=0.45	0.000557	0.0868	0.133	0.0089	0.0501	0.467	0.376
Mx=0.50	0.00348	0.0868	0.0868	0.00348	0.0868	0.42	0.42
Mx=0.55	0.0089	0.0868	0.0501	0.000557	0.133	0.376	0.467
Mx=0.60	0.0168	0.0868	0.0235	0.000138	0.19	0.333	0.517
Mx=0.65	0.0272	0.0868	0.0068	0.00222	0.257	0.294	0.569
Mx=0.70	0.0402	0.0868	0.000138	0.0068	0.333	0.257	0.623
Mx=0.75	0.0556	0.0868	0.00348	0.0139	0.42	0.222	0.68
Mx=0.80	0.0735	0.0868	0.0168	0.0235	0.517	0.19	0.74
Mx=0.85	0.0939	0.0868	0.0402	0.0355	0.623	0.161	0.802
Mx=0.90	0.117	0.0868	0.0735	0.0501	0.74	0.133	0.867

Figure 4: Numerical results of the graphical simulation.

Genotype combinations having the lowest residuals are highlighted along with the corresponding mixture proportion. The most likely combinations are: (BB,AA), (AA, AB), (AB, AA), (AA, BB) with the corresponding mixtures proportions :0.3, 0.4, 0.5 and 0.7. Note that clinking the "Save" button launches a window where the desired path for the save file can be specified, default creates a text file in the current folder.

The third button, "Genotypes filter" launches a window showing a matrix of the mixture proportion conditional on the genotype combination.

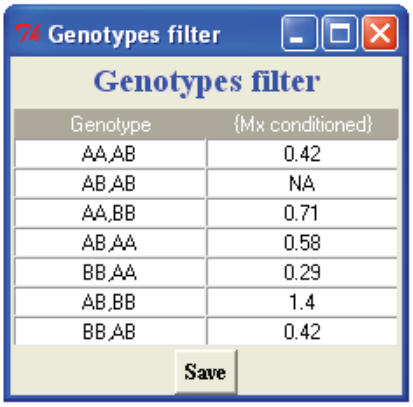


Figure 5: Genotypes filter: Mixture proportion conditional on the genotypes combination.

The mixture proportions conditional on the genotype combination gives a supplementary indication for the reduction of the number of possible combinations: Genotypes with non plausible mixture proportions ranges are not kept. The results confirm that genotypes which have not been already selected during the graphical simulation step, are not supported by the data. Formulas used for the calculations are given in Appendix A.

6 Miscellaneous

6.1 Manipulating forensim objects

forensim objects are mainly formed by lists and data frames. Modification of the slots of an object can easily be done using operators '$' (lists) or '[' (data frame and matrix). For example, we wish to modify the frequencies of a given locus, say FGA, in the tabfreq object tupop:

```
> tupop$tab$Tu$FGA
```

```
    18     19   19.2     20     21     22   22.2     23   23.2     24     25
0.0392 0.0686 0.0033 0.0458 0.0980 0.1765 0.0033 0.1961 0.0098 0.2222 0.1013
  25.2     26   26.2     27
0.0065 0.0131 0.0065 0.0098
```

Frequencies of alleles 18 and 27 are modified from 0.0392 and 0.0098 to 0.01 and 0.03 respectively:

```
> tupop$tab$Tu$FGA[c("18", "27")] <- c(0.01, 0.03)
> tupop$tab$Tu$FGA
```

```
    18     19   19.2     20     21     22   22.2     23   23.2     24     25
0.0100 0.0686 0.0033 0.0458 0.0980 0.1765 0.0033 0.1961 0.0098 0.2222 0.1013
  25.2     26   26.2     27
0.0065 0.0131 0.0065 0.0300
```

6.1.1 How to change population names

Changing population names in any forensim object is achieved using the function changepop. For example, changing the population name in the tabfreq object tupop from "Tu" (argument oldpop) to "Tu2" (argument newpop) is achieved by:

```
> tupop2 <- changepop(tupop, oldpop = "Tu", newpop = "Tu2")
> tupop2@pop.names


[1] Tu2
Levels: Tu2
```

6.1.2 How to find the allele frequencies of a mixture

The allele frequencies of a mixture; stored in a simumix object, can be found using the function findfreq. The tabfreq object from which to extract the allele frequencies must be specified. For instance, allele frequencies in object mix3 are found from the Caucasian population:

```
> temp <- findfreq(mix3, freq = strusa, refpop = "Cauc")
> temp


$Cauc
$Cauc$CSF1PO
     10      11      12      14
0.21689 0.30132 0.36093 0.00828

$Cauc$FGA
     22      23
0.21854 0.13411

$Cauc$TH01
      6       7     9.3
0.23179 0.19040 0.36755

$Cauc$TPOX
      8      10      11
0.53477 0.05629 0.24338

$Cauc$VWA
     16      17      18      19
0.20033 0.28146 0.20033 0.10430

$Cauc$D3S1358
     14      15      16      17
0.10265 0.26159 0.25331 0.21523

$Cauc$D5S818
     11      12      13
0.36093 0.38411 0.14073

$Cauc$D7S820
      7       8       9      10
0.01821 0.15066 0.17715 0.24338

$Cauc$D8S1179
     13      14      15
0.30464 0.16556 0.11424

$Cauc$D13S317
      9      11      12      13      14
0.07450 0.33940 0.24834 0.12417 0.04801

$Cauc$D16S539
```

```
      9      11      12
0.11258 0.32119 0.32616

$Cauc$D18S51
     13      14      15      16      17
0.13245 0.13742 0.15894 0.13907 0.12583

$Cauc$D21S11
     28      29      30    30.2      31
0.15894 0.19536 0.27815 0.02815 0.08278

$Cauc$D2S1338
     19      20      23      24      25
0.11424 0.14570 0.11755 0.12252 0.09272

$Cauc$D19S433
     13      14      16
0.25331 0.36921 0.04967
```

temp is a list of a single element "Cauc", which contains also a list:

```
> class(temp$Cauc)
```

```
[1] "list"
```

Allele frequencies of locus TPOX for example, are given by:

```
> temp$Cauc$TPOX
```

```
      8      10      11
0.53477 0.05629 0.24338
```

6.1.3 The number of alleles in a mixture

The number of alleles in a simumix object can be determined by the function nball. The overall loci number of alleles in the 2-person mixture mix2 is:

```
> nball(mix2, byloc = FALSE)
```

```
[1] 11
```

and the numbers of alleles per locus can be obtained by setting the argument byloc to TRUE:

```
> nball(mix2, byloc = TRUE)
```

```
D8S1179    TH01     FGA
      4       3       4
```

6.2 Forensim vignettes

In addition to the help files accessible through the '?' or the 'help' commands, R packages sometimes provide additional documentation files through *vignettes*. To check whether the forensim package proposes a vignette, type:

```
> vignette(package="forensim")
```

This command launches a window with a list of available vignettes (Figure 6).

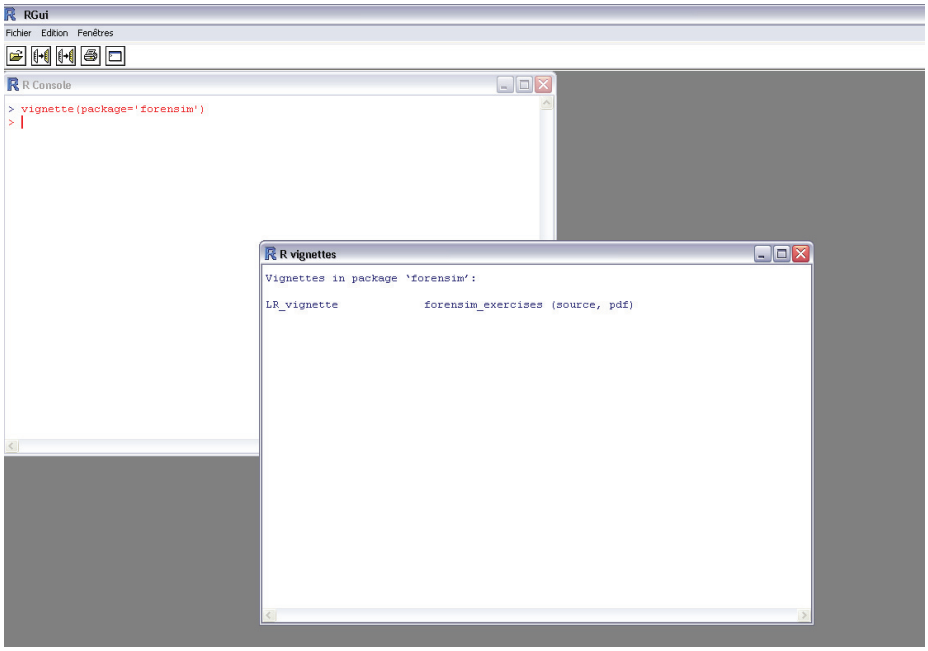


Figure 6: How to check for vignettes in R.

Forensim currently provides one vignette: 'LR_vigentte', in a PDF format, illustrating the use of likelihood ratios through exercises of varying difficulty. To open the vignette from an R shell, simply type:

```
> vignette("LR_vignette", package="forensim")
```

References

[1] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

[2] R Development Core Team. R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http : //www.Rproject.org/. 2006.

[3] B. Zhu, J. Yan, C. Shen, T. Li, Y. Li, X. Yu, X. Xiong, H. Muf, Y. Huang, and Y. Deng. Population genetic analysis of 15 STR loci of Chinese Tu ethnic minority group. *Forensic Science International*, 174:255–258, 2008.

[4] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, vol. 2.* John Wiley & Sons, 1995.

[5] G. Nicholson, A. V. Smith, F. Jónsson, O. Gústafsson, K. Stefánsson, and P. Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society B*, 64:695–715, 2002.

[6] J. Marchini and L. R. Cardon. Discussion on the meeting on "Statistical modelling and analysis of genetic data". *Journal of the Royal Statistical Society B*, 64:740–741, 2002.

[7] D. R. Paoletti, T. E. Doom, C. M. Krane, M. L. Raymer, and D. E. Krane. Empirical analysis of the STR profiles resulting from conceptual mixtures . *Journal of Forensic Sciences*, 50(6):1361–1366, 2005.

[8] J.M. Butler, R. Schoske, M.P. Vallone, J. W. Redman, and M. C. Kline. Allele frequencies for 15 autosomal str loci on u.s. caucasian, african american, and hispanic populations. *Journal of Forensic Sciences*, 48(8):908–911, 2003.

[9] 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.

[10] H. Haned, L. Pene, J. R. Lobry, A. B. Dufour, and D. Pontier. Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count ? *Jounal of Forensic Sciences*, accepted, 2010.

[11] J. Buckleton, C. M. Triggs, and S. J. Walsh. *Forensic DNA evidence interpretation.* CRC PRESS, 2005.

[12] D. J. Balding and R. A. Nichols. DNA profile match probability calculation: how to allow for population stratification, relatedness, databse selection and single bands. *Forensic Science International*, 64:125–140, 1994.

[13] J. Curran, J. Buckleton, and C. M. Triggs. What is the magnitude of the subpopulation effect? *Forensic Science International*, 135:1–8, 2003.

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

[15] P. Gill, P. Sparkes, R. Pinchin, Clayton, J. Whitaker, and J. Buckleton. Interpreting simple STR mixtures using allele peak areas. *Forensic Science International*, 91:41–53, 1998.

[16] T. Clayton and J. Buckleton. *Forensic DNA evidence interpretation*, chapter Mixtures, pages 217–239. CRS PRESS, 2005.

A Appendix: Formulas used in *mastermix*

A.1 Expected allelic ratios

Two-allele model: expected allelic ratios conditional on each possible genotype combination of the contributors to the mixture, when two alleles, A and B (in ascending order of molecular weights) are observed at a given locus, and $\hat{M}_x$ is the proportion of sample from the first contributor [15].

Combination	Alleles	
	A	B
AA,AB	$\frac{\hat{M}_x}{2}$+0.5	$\frac{1-\hat{M}_x}{2}$
AB,AB	0.5	0.5
AA,BB	$\hat{M}_x$	$1-\hat{M}_x$
AB,AA	$1-\frac{\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$
BB,AA	$1-\hat{M}_x$	$\hat{M}_x$
AB,BB	$\frac{\hat{M}_x}{2}$	$1-\frac{\hat{M}_x}{2}$
BB,AB	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}+0.5$

Three-allele model: expected allelic ratios conditional on each possible genotype combination of the contributors to the mixture when three alleles, A, B and C (in ascending order of molecular weights) are observed at a given locus [15].

Combination	Alleles		
	A	B	C
AA,BC	$\hat{M}_x$	$\frac{1-\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$
BB,AC	$\frac{1-\hat{M}_x}{2}$	$\hat{M}_x$	$\frac{1-\hat{M}_x}{2}$
CC,AB	$\frac{1-\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	$\hat{M}_x$
AB,AC	0.5	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$
BC,AC	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	0.5
AB,BC	$\frac{\hat{M}_x}{2}$	0.5	$\frac{1-\hat{M}_x}{2}$
BC,AA	$1-\hat{M}_x$	$\frac{\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$
AC,BB	$\frac{\hat{M}_x}{2}$	$1-\hat{M}_x$	$\frac{\hat{M}_x}{2}$
AB,CC	$\frac{\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	$1-\hat{M}_x$
AC,AB	0.5	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$
AC,BC	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	0.5
BC,AB	$\frac{1-\hat{M}_x}{2}$	0.5	$\frac{\hat{M}_x}{2}$

Four-allele model: expected allelic ratios conditional on each possible genotype combination of the contributors to the mixture when four alleles, A, B, C and D (in ascending order of molecular weights) are observed at a given locus [15].

Combination	Alleles			
	A	B	C	D
AB,CD	$\frac{\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$
AC,BD	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$
AD,BC	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$
BC,AD	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$
BD,AC	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$
CD,AB	$\frac{1-\hat{M}_x}{2}$	$\frac{1-\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$	$\frac{\hat{M}_x}{2}$

A.2 Conditional mixtures proportions

The following tables give the formulas for the mixture proportion conditional on the genotype combinations. The conditional mixture proportions are computed using observed allele peak heights (or equivalently peak areas) [16].

Mixture proportions conditioned on the genotype combination for a locus showing two alleles, A and B (in ascending order of molecular weights), with peak heights ϕ_A and ϕ_B.

Two-allele model

Genotype combination	Conditional mixture proportion
AA,AB	$\frac{\phi_A - \phi_B}{\phi_A + \phi_B}$
AB,AB	No information is present
AA,BB	$\frac{\phi_A}{\phi_A + \phi_B}$
AB,AA	$\frac{2\phi_B}{\phi_A + \phi_B}$
BB,AA	$\frac{\phi_B}{\phi_A + \phi_B}$
AB,BB	$\frac{2\phi_A}{\phi_A + \phi_B}$
BB,AB	$\frac{\phi_B - \phi_A}{\phi_A + \phi_B}$

Mixture proportions conditioned on the genotype combination for a locus showing three alleles,, A , B and C (in ascending order of molecular weights), with peak heights ϕ_A, ϕ_B and ϕ_C.

Three-allele model

Genotype combination	Conditional mixture proportion
AA,BC	$\frac{\phi_A}{\phi_A+\phi_B+\phi_C}$
BB,AC	$\frac{\phi_B}{\phi_A+\phi_B+\phi_C}$
CC,AB	$\frac{\phi_C}{\phi_A+\phi_B+\phi_C}$
AB,AC	$\frac{\phi_B}{\phi_B+\phi_C}$
BC,AC	$\frac{\phi_B}{\phi_A+\phi_B}$
AB,BC	$\frac{\phi_A}{\phi_A+\phi_C}$
BC,AA	$\frac{\phi_B+\phi_C}{\phi_A+\phi_B+\phi_C}$
AC,BB	$\frac{\phi_A+\phi_C}{\phi_A+\phi_B+\phi_C}$
AB,CC	$\frac{\phi_A+\phi_B}{\phi_A+\phi_B+\phi_C}$
AC,AB	$\frac{\phi_C}{\phi_B+\phi_C}$
AC,BC	$\frac{\phi_A}{\phi_A+\phi_B}$
BC,AB	$\frac{\phi_C}{\phi_A+\phi_C}$

Mixture proportions conditioned on the genotype combination for a locus showing four alleles, A , B, C and D (in ascending order of molecular weights), with peak heights ϕ_A, ϕ_B, ϕ_C and ϕ_D.

Four-allele model

Genotype combination	Conditional mixture proportion
AB,CD	$\frac{\phi_A + \phi_B}{\phi_A + \phi_B + \phi_C + \phi_D}$
AC,BD	$\frac{\phi_A + \phi_C}{\phi_A + \phi_B + \phi_C + \phi_D}$
AD,BC	$\frac{\phi_A + \phi_D}{\phi_A + \phi_B + \phi_C + \phi_D}$
BC,AD	$\frac{\phi_B + \phi_C}{\phi_A + \phi_B + \phi_C + \phi_D}$
BD,AC	$\frac{\phi_B + \phi_D}{\phi_A + \phi_B + \phi_C + \phi_D}$
CD,AB	$\frac{\phi_C + \phi_D}{\phi_A + \phi_B + \phi_C + \phi_D}$

Chapter 3

Appendix C: Forensim vignette

Weight of DNA evidence using the forensim package

Thore Egeland Hinda Haned

June 2010

Contents

1 The forensim package: overview

1.1 Available documentation

forensim is an R-package hosted by R-forge dedicated to facilitate the interpretation of forensic DNA mixtures. It also provides simulation tools made to mimick data from case work.

A detailed description of forensim is given in the package tutorial, available from: http://forensim.r-forge.r-project.org/. prepared specifically for potential forensim users who are unfamilliar with R. The present document serves to

- introduce the basic statistical calculations of forensim,
- provided excercises and solutions for a course setting,
- provide examples to verify correct usage and answers. This is mostly done by means of the solution to the mentioned excercises.

1.2 Statistical methods. A worked example

Forensim provides a variety of methods dedicated to evaluating the weight of DNA evidence [1]. Below we focus on the `LR` function for the calculation of likelihood ratios. The `LR` function implements the general formula of Curran et al. for forensic DNA mixtures interpretation [2].

An example Consider the following genetic profiles from a rape case in Hong Kong [3]:

Locus	Mixture	Victim	Suspect	Frequency
D3S1358	14		14	0.033
	15	15		0.331
	17		17	0.239
	18	18		0.056

Table 1: Alleles from a DNA stain from a rape case in Hong Kong

Locus D3S1358 shows 4 distinct alleles (14, 15, 17 and 18). The number of contributors to the mixed sample is taken to be 2.

Scenario 1 The following hypotheses are tested:

- Prosecution hypothesis H_P: Contributors were the victim and the suspect.
- Defence hypothesis H_D: Contributors were 2 unknown people.

Before we start, remember to load the package:

```
> library(forensim)
```

First, the genotypes are assigned to the victim and the suspect:

```
> victim <- "15/18"
> suspect <- "14/17"
```

The likelihood ratio is computed using the `LR` function: Here is a useful extract of this function's help page:

- `stain`: a vector giving the set of (distinct) alleles present in the DNA stain
- `freq`: vector of the corresponding allele frequencies in the global population
- `xp`: the number of unknown contributors to the stain under the prosecution hypothesis Hp. Default is 0.
- `xd`: the number of unknown contributors to the stain under the defence hypothesis Hd. Default is 0.
- `Tp`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known contributor under the prosecution hypothesis Hp. The length of the vector equals the number of known contributors. Default is NULL.
- `Vp`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known non-contributor under the prosecution hypothesis Hp. The length of the vector equals the number of known non-contributors. Default is NULL.
- `Td`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known contributor under the defence hypothesis Hd. The length of the vector equals the number of known contributors. Default is NULL.
- `Vd`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known non-contributor under the defence hypothesis Hd. The length of the vector equals the number of known non-contributors. Default is NULL.
- `theta`: a float in [0,1[. theta is equivalent to Wright's Fst. In case of population subdivision, it allows a correction of the allele frequencies in the subpopulation of interest

The LR is obtained as follows

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = NULL, Vd = c(victim,
+         suspect), xd = 2, theta = 0)
```

```
[1] 285
```

The mixture profile is 285 times more likely if it came from the suspect and the victim than if it came from two unknown unrelated individuals.
Note that as long as theta=0, there is no need to be specify the non-contributing individuals, so the same figure is prodcued with Vd=NULL.

Scenario 2 The following hypotheses are tested:
Prosecution hypo4thesis H_P: Contributors were the victim and the suspect.
Defence hypothesis H_D: Contributors were the victim and one unknown.

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = victim,
+     Vd = suspect, xd = 1, theta = 0)
```

```
[1] 63.4
```

The mixture profile is 63 times more likely if it came from the suspect than if it came from an unrelated individual.

2 Exercises

Some of the problems below are theoretical in the sense that forensim is not used, rather calculation by hand are requested. These excercises may be skipped for those exclusively interested in practising forensim.

2.1 Excercise 1. Likelihood ratios and theta values

Note that in the previous examples, the `theta` argument does not appear in the `LR` function. This means that the argument is set to its default value, which is 0. The problems below extend on scenario 2 of the above example by addressing θ corrections.

1. Change the value of the `theta` argument from 0 to 0.03 and repeat the calculation.
2. Calculate the LR for different values of `theta` taken in the interval [0,0.03].
   - *Tip 1*: use the `seq` function to create a sequence of values for the `theta` argument.
   - *Tip 2*: use the *sapply* function to compute the values of the LR for different values of theta. To get help, type: help('sapply').
3. Represent the obtained results in a plot (use function `plot`).

2.2 Excercise 2. Theoretical continuation of Excercise 1

1. Derive the formulae corresponding to Scenarios 1 and 2 of the worked example (Hong-Kong case). Confirm that the figures obtanined by forensim are correct.
2. Repeat the above problem with θ-correction for scenario 2 ($\theta = 0.03$).

2.3 Excercise 3. LR-calculations for mixtures

The purpose of this exercise is to demonstrate various approaches to LR calculations for a mixture case. The data comes from a proficiency test arranged by GEDNAP `http://gednap.de/`. For simplicity only three markers are considered. There is a mixture (stain), the data is summarised in Table 2 and a reference sample is shown in Table 3.

Locus	Allele
D3S1358	15
D3S1358	16
D3S1358	17
vWA	15
vWA	16
vWA	18
FGA	20
FGA	21
FGA	22
FGA	24
FGA	26

Table 2: The crime scene profile at three STR loci.

Locus	Allele
D3S1358	15/17
vWA	16/18
FGA	20/26

Table 3: The reference sample B

The hypotheses are

- H_P: B and two unknown individuals contributed to the stain
- H_D: Three unknown people contributed to the stain

Calculate the likelihood ratios to weight hypotheses H_P and H_D when $\theta = 0$. For simplicity we assume all allele frequencies to be 0.1.

2.4 Excercise 4. LR: standard and for drop in and out

This example extends on Section 4.4 of [4]. The hypotheses are the usual ones:

- H_P: The DNA came from the suspect.
- H_D: The DNA came from a random man.

Throughout A and B denote alleles with relative frequencies $p_A = 0.2$ and $p_B = 0.1$, and we assume $\theta = 0$.

1. We first consider a standard case with data AB for the suspect and the stain. Derive the formula for the LR and use R to provide the numeric answer. Confirm the above calculation using the LR function of forensim.

2. Repeat the above problem whith data AA for suspect and the stain.

3. Assume markers $1, 2, \cdots, 5$ are as 1 above. Markers 6,7,8,9 are as for 2 above. Calculate the LR for these 9 markers by using the formulae derived above.

4. For the tenth marker the suspect is A and the stain AB. What's the LR for this marker? What's the LR based on all 10 markers?

5. Consider the above problem once assuming that there is a probability D that an allele drops out. According to [4]

$$LR_{10} \approx \frac{D}{(1+D)p_A^2 + 2p_A(1-p_A)D} \tag{1}$$

   Let $D = 0.1$. Use R to find LR_{10} and the LR based on all markers ($LR_{1,10}$). Comment on the answer.

6. Plot LR_{10} as a function of D.

3 Solutions to the excercises

3.1 Excercise 1

1. For a single value, theta=0.03 we find:

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = victim,
+     Vd = suspect, xd = 1, theta = 0.03)
```

```
[1] 37.6
```

2. Define a variable theta, taking different values in the [0,1] interval:

```
> theta <- seq(0, 0.03, by = 0.001)
> theta
```

```
 [1] 0.000 0.001 0.002 0.003 0.004 0.005 0.006 0.007 0.008 0.009 0.010 0.011
[13] 0.012 0.013 0.014 0.015 0.016 0.017 0.018 0.019 0.020 0.021 0.022 0.023
[25] 0.024 0.025 0.026 0.027 0.028 0.029 0.030
```

   To replicate the calculations for different values of theta, we use the `sapply` function.

```
> sapply(theta, function(i) LR(stain = c(14, 15, 17, 18), freq = c(0.033,
+     0.331, 0.239, 0.056), xp = 0, Tp = c(victim, suspect), Vp = NULL,
+     Td = victim, Vd = suspect, xd = 1, theta = i))
```

```
 [1] 63.40 61.83 60.34 58.94 57.61 56.35 55.15 54.01 52.92 51.89 50.90 49.95
[13] 49.05 48.18 47.35 46.56 45.79 45.06 44.35 43.67 43.02 42.39 41.78 41.19
[25] 40.63 40.08 39.55 39.04 38.54 38.06 37.60
```

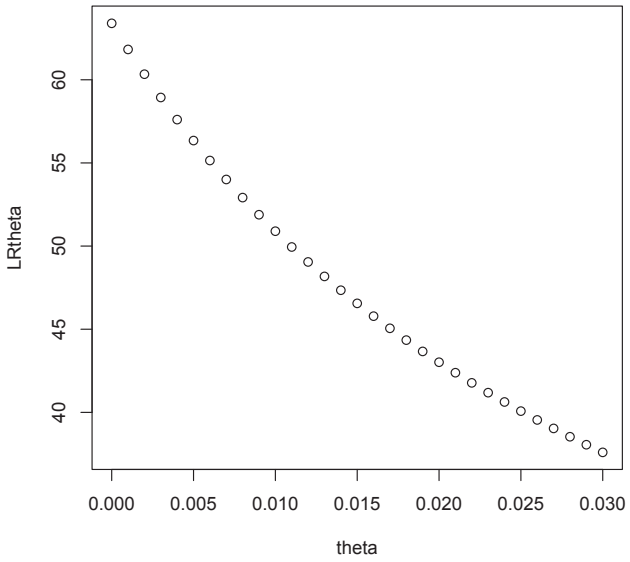
The above command calculates the LR for each value `i` of `theta`.

3. To plot these results, we need first to save them to an object:

```
> LRtheta <- sapply(theta, function(i) LR(stain = c(14, 15, 17,
+     18), freq = c(0.033, 0.331, 0.239, 0.056), xp = 0, Tp = c(victim,
+     suspect), Vp = NULL, Td = victim, Vd = suspect, xd = 1, theta = i))
```

The plot is produced by

```
> plot(theta, LRtheta)
```



3.2 Excercise 2

1. For scenario 1

$$LR = \frac{1}{24p_{14}p_{15}p_{17}p_{18}} = \frac{1}{24 \cdot 0.033 \cdot 0.331 \cdot 0.239 \cdot 0.056}. \tag{2}$$

The numerator is obvious. The denominator can be be obtained by realising that both individuals must be heterozygote and that there are 6 possible combinations, each having probability $4p_{14}p_{15}p_{17}p_{18}$ since (i) Hardy-Weinberrg Equilibrium is assumed to hold and (ii) the individuals are unrelated.

This can be calculated in R as

```
> 1/(24 * 0.033 * 0.331 * 0.239 * 0.056)
```

```
[1] 285.0105
```

For scenario 2

$$LR = \frac{1}{2p_{14}p_{17}} \tag{3}$$

since the suspect must have genotype 14,17.

This can be calculated in R as

```
> 1/(2 * 0.033 * 0.239)
```

```
[1] 63.39546
```

2. Consider first scenario 1. Let $A = 14, B = 15, C = 17, D = 18$. Then the modification of Equation 3 to account for θ-corrections becomes

$$LR = \frac{(1+3\theta)(1+4\theta)}{2(\theta+(1-\theta)p_{14})(\theta+(1-\theta)p_{17})}. \tag{4}$$

```
> (1 + 3 * 0.03) * (1 + 4 * 0.03)/(2 * (0.03 + (1 - 0.03) * 0.033) *
+     (0.03 + (1 - 0.03) * 0.239))
```

```
[1] 37.59529
```

This confirms the forensim value:

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = victim,
+     Vd = suspect, xd = 1, theta = 0.03)
```

```
[1] 37.6
```

3.3 Excercise 3

First, enter the stain profile for each available locus:

```
> stainD3 <- c(15, 16, 17)
> stainv <- c(15, 16, 18)
> stainFGA <- c(20, 21, 22, 24, 26)
```

Second, enter the suspect profile for each available locus:

```
> suspectD3 <- "15/17"
> suspectv <- "16/18"
> suspectFGA <- "20/26"
```

Last, the likelihood ratio:

```
> LRD3 <- LR(stain = stainD3, freq = rep(0.1, 3), xp = 2, Tp = c(suspectD3),
+     Vp = NULL, Td = NULL, Vd = suspectD3, xd = 3, theta = 0)
> LRD3
```

```
[1] 12.04
```

```
> LRDv <- LR(stain = stainv, freq = rep(0.1, 3), xp = 2, Tp = c(suspectv),
+     Vp = NULL, Td = NULL, Vd = suspectv, xd = 3, theta = 0)
> LRDv
```

```
[1] 12.04
```

```
> LRDFGA <- LR(stain = stainFGA, freq = rep(0.1, 5), xp = 2, Tp = c(suspectFGA),
+     Vp = NULL, Td = NULL, Vd = suspectFGA, xd = 3, theta = 0)
> LRDFGA
```

```
[1] 4.667
```

The overall likelihood ratio is obtained by multiplying the above likelihood ratios:

```
> LRD3 * LRDv * LRDFGA
```

```
[1] 676.5358
```

3.4 Excercise 4

1. Note first that $P(data|H_P) = 1$. Next

$$P(data|H_D) = P(\text{culprit is AB}) = 2p_A p_B$$

provided Hardy-Weinbererg Equilibrium holds. First some parameter values are assigned.

```
> D <- 0.01
> pA <- 0.2
> pB <- 0.1
```

A direct calculation in R gives

```
> 1/(2 * pA * pB)
```

```
[1] 25
```

The LR function of forensim gives

```
> LR(stain = c("A", "B"), freq = c(0.2, 0.1), xp = 0, Tp = "A/B",
+     Vp = NULL, Td = NULL, Vd = "A/B", xd = 1, theta = 0)
```

```
[1] 25
```

2. A similar argument gives $LR = 1/p_A^2$ which evaluates to 25. Furthermore, using forensim we find

```
> LR(stain = c("A"), freq = c(0.2), xp = 0, Tp = "A/A", Vp = NULL,
+     Td = NULL, Vd = "A/A", xd = 1, theta = 0)
```

```
[1] 25
```

3.
```
> 25^9
```

```
[1] 3.814697e+12
```

4. The likelihood ratio is 0 for the marker and therefore also the overall LR is 0.

5. With drop-out probability of 0.01 we find

```
> D/((1 + D) * pA^2 + 2 * pA * (1 - pA) * D)
```

```
[1] 0.2293578
```

The LR based on all markers becomes

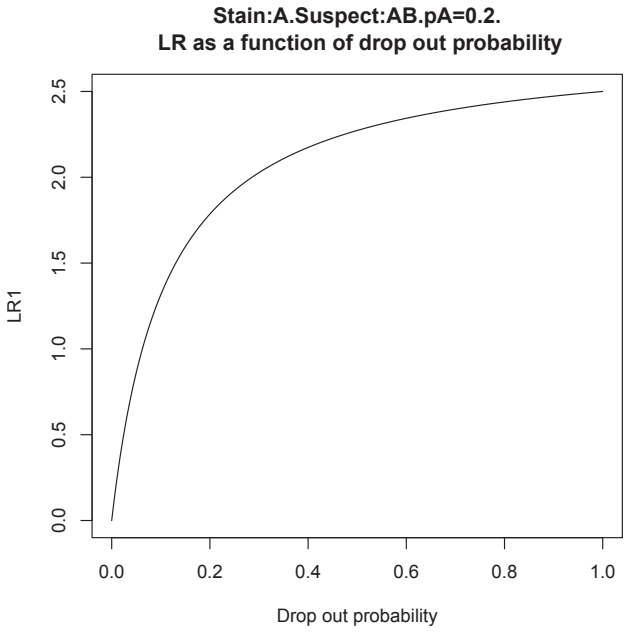
```
> 25^9 * 0.2293578
```

```
[1] 874930572510
```

6. There are several ways to plot. Here's one:

```
> D = seq(0, 1, length = 1000)
> pA = 0.2
> LR1 = D/((1 + D) * pA^2 + D * 2 * pA * (1 - pA))
```

```
> D = seq(0, 1, length = 1000)
> pA = 0.2
> LR1 = D/((1 + D) * pA^2 + D * 2 * pA * (1 - pA))
> plot(D, LR1, type = "l", xlab = "Drop out probability", ylab = "LR1")
> title("Stain:A.Suspect:AB.pA=0.2.\n LR as a function of drop out probability")
```



References

[1] H. Haned. Forensim: an open source initiative for the evaluation of statistical methods in forensic genetics. *Forensic Sci. Int. Genetics*, 2010.

[2] J. Curran, J. Buckleton, and C. M. Triggs. What is the magnitude of the subpopulation effect? *Forensic Science International*, 135:1–8, 2003.

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

[4] P. Gill and J. Buckleton. A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number. *Forensic Science International: Genetics*, 4(4):221–227, 2010.