Statistical Evaluation of Forensic DNA Profile Evidence

Christopher D. Steele and David J. Balding

UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom; email: c.steele.11@ucl.ac.uk, d.balding@ucl.ac.uk

Annu. Rev. Stat. Appl. 2014. 1:361–84

First published online as a Review in Advance on November 6, 2013

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-022513-115602

Keywords

statistical genetics, forensic genetics, forensic identification, weight of evidence

Abstract

The evaluation of weight of evidence for forensic DNA profiles has been a subject of controversy since their introduction over 20 years ago. Substantial progress has been made for standard DNA profiles, but new issues have arisen in recent years with the advent of more sensitive profiling techniques, allowing profiles to be recovered from minuscule amounts of possibly degraded DNA. These low-template DNA profiles suffer from enhanced stochastic effects, including dropin, dropout, and stutter, which pose problems for DNA profile evaluation. These problems are now beginning to be overcome with the emergence of several statistical models and software. We first review the general principles of statistical evaluation of DNA profile evidence, and we then focus on low-template DNA profiles, briefly reviewing the main statistical models and software. We cover methods that use allele presence/absence and those that use electropherogram peak heights, focusing on the likelihood ratio as measure of evidential weight.

1. INTRODUCTION

Electropherogram (epg): a set of graphs conveying DNA profiling results

STR: short tandem repeat

Polymerase chain reaction (PCR): a method used to amplify DNA

Relative fluorescence unit (RFU): unit of epg peak height

Base pair (bp): unit of DNA sequence length

The lack of appropriate methods and software for the statistical analysis of low-template DNA (LTDNA) profiles has been a source of controversy for more than a decade and has hindered the widespread use of potentially very powerful evidence. At long last, substantial progress is now being made, which we review here. We start with a brief review of standard DNA profiles, then discuss the new difficulties for statistical evaluation that accompany the use of LTDNA samples, which often also suffer effects of degradation due to environmental exposure. We then review the statistical evaluation of evidence using likelihood ratios (LRs) and the development of LRs for standard and then LTDNA profiles. Finally, we describe six software programs that are currently available for LTDNA profiling: three that make use of peak heights and three that do not. We hope that this review will help disseminate current best practices in the statistical evaluation of LTDNA evidence, spur further developments, and advertise to the forensic and wider community that robust methods for LTDNA evidence evaluation are now available. We have not verified the software described here, but, where available, we cite validation studies conducted by the authors of each program or package. In our view, courts are now able to avail themselves of the powerful new LTDNA profiling technologies, provided that as much care is taken with the statistical analysis as is necessary for the collection, handling, and analysis of the samples.

1.1. Standard DNA Profiles

The results of forensic DNA profiling are represented in an electropherogram (epg). **Figure 1** shows an epg for a good-quality single-contributor profile. Briefly, profiling focuses on short tandem repeat (STR) regions of the genome, which vary in length because of differing numbers of repeats of a sequence motif such as the four-base motif ACAG. To measure the allele length, the polymerase chain reaction (PCR) is used to amplify a DNA fragment that includes the motif repeats together with some flanking DNA. The amplified fragment is labeled using a fluorescent dye and allowed to traverse a capillary tube under an electric field in a process known as electrophoresis. The time taken for this traverse is measured via a laser that causes the fluorescence to be detected, generating an epg signal peak measured in relative fluorescence units (RFUs). The length of the DNA fragment is then deduced from the time it takes to travel through the capillary. Because the lengths of the flanking DNA fragments are known, typically accurate to the nearest base pair (bp), the number of tandem repeats can be inferred.

Each of the three panels in **Figure 1** corresponds to a different dye color. Choosing flanking regions of different lengths allows the alleles from different loci to be separated on the basis of their size. Because they are also color-separated by the dyes, multiple loci (here, 11) can be tested in a single profiling run. Notice that peak heights tend to decline with fragment length. This can occur for standard profiles, but the rate of decline gives a measure of degradation (see Section 3.5). Pairs of peaks of a similar height that are close together correspond to the two alleles of a heterozygote, whereas the very tall single peaks, such as the two in the middle panel, correspond to homozygotes. The leftmost pair of peaks in the middle panel indicates the presence of both X and Y chromosomes and hence DNA from a male. At other loci, a pair of numbers represents the genotype of an individual. For example, the pair 14,15 represents the genotype for the *D3* locus in **Figure 1** (leftmost peaks in the top panel), and 13,13 represents that for locus *D8*. In practice, some deviations from the simple tandem repeat model occur, such as changes in the repeat motif or a partial repeat. For example, the 9.3 at the *THO1* locus in **Figure 1** indicates a 3-bp partial repeat in addition to nine full 4-bp repeats. Some models of STR mutation and of stutter peak heights use the longest uninterrupted sequence (Brookes et al. 2012).

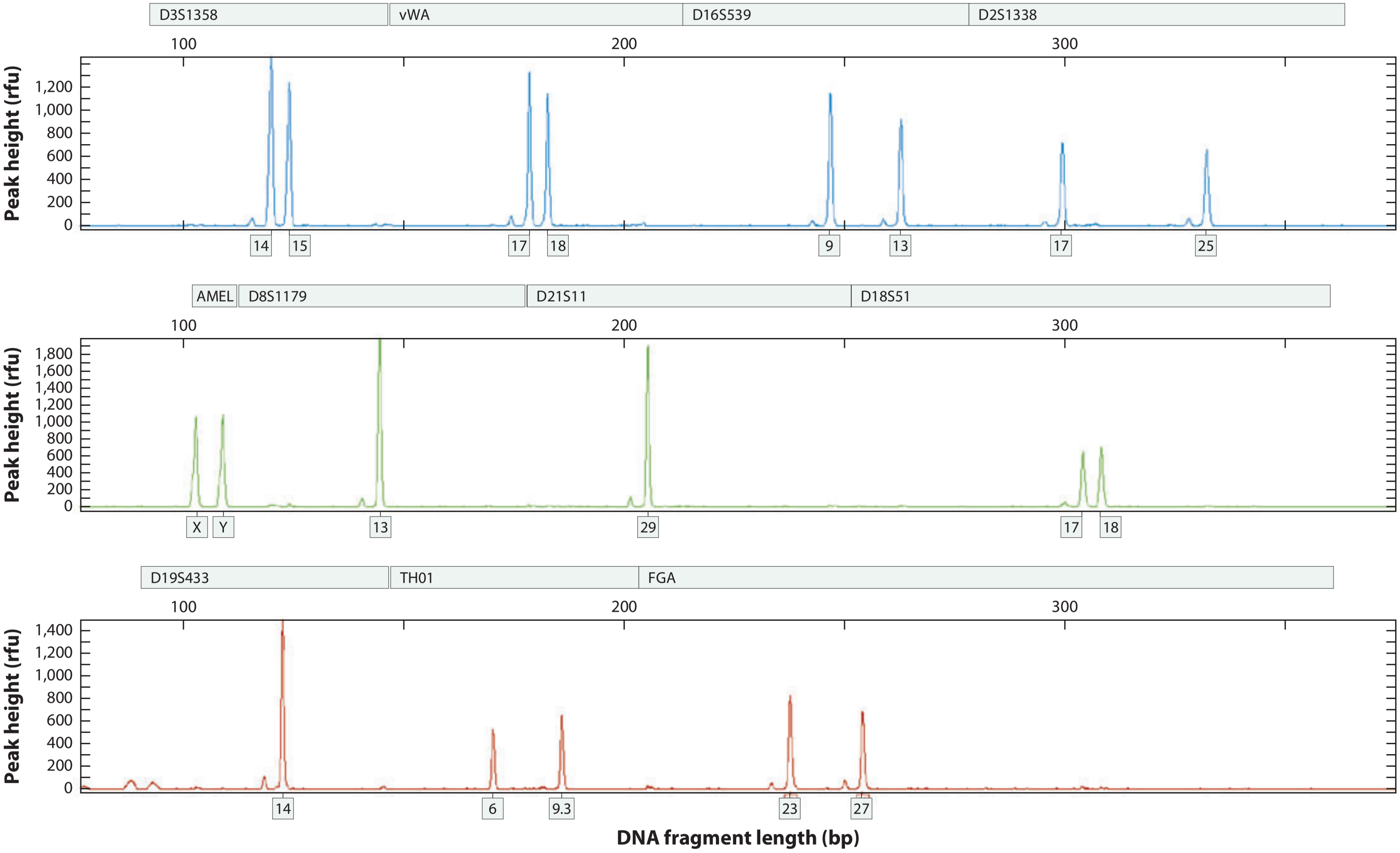

Figure 1

A single-contributor electropherogram using the SGM Plus profiling system, which tests 10 autosomal loci as well as the *Amel* sex-determining locus. Note that the y-axis scales differ. Loci names are shown in the gray banner above each panel. Numerical allele labels are given in the boxes below each peak. Abbreviations: bp, base pairs; rfu, relative fluorescence units.

PEAK IMBALANCE

Peak imbalance refers to variation in the peak heights at the two alleles of a heterozygote, which are expected to be similar in standard DNA profiling, as in **Figure 1**. However, in LTDNA work, large differences in heterozygote peak heights can lead to dropout of one allele. In both panels of **Figure 2**, for example, the *vWA* peak at allele 17 is approximately twice the height of that at allele 18. In the early years of LTDNA interpretation, peak imbalance in LTDNA work was assumed to mean that, except in the case of clearly distinguished major and minor contributors, peak heights convey no useable information about the underlying amount of DNA. Thus, peak heights were only used to decide whether the sample contains a given allele. However, statistical models of peak heights have recently been introduced (see Section 3.8), allowing the extraction of information from the epg about the amount of DNA underlying each peak.

1.2. Low-Template DNA Profiles

Various enhancements to DNA profiling protocols have been introduced in recent years to allow profiling of smaller and more degraded DNA samples than was previously possible. For example, scientists can now profile DNA from a touch that was insufficient to generate a traditional fingerprint. Lohmueller & Rudin (2012) give a description of a complex DNA profiling case and describe some of the laboratory and analysis issues that arise.

The most obvious enhancement is to increase the number of PCR cycles from 28 (the standard number of cycles) to between 29 and 34. The term low copy number (LCN) profiling was initially used, but because of confusion over whether or not this term referred to a specific 34-cycle profiling technique, the more general term low-template DNA (LTDNA) profiling is now preferred. The term template refers to the DNA strands available for copying and is loosely equivalent to the amount of DNA used for profiling, measured in units of mass (picograms, abbreviated pg). The nuclear DNA content of a single human cell is approximately 6.5 pg.

The peak heights in an epg provide a guide to the DNA template used in the profiling run. The epg in **Figure 1** is based on 500 pg of DNA, and heterozygote peaks may be as high as 1,400 RFU. **Figure 2** shows the blue-dye results for two LTDNA profiling runs for the same individual as in **Figure 1**. The starting DNA templates are 31 pg and 15 pg, which generate heterozygote peak heights up to approximately 160 RFU and 80 RFU, respectively. The small number of cells contributing to LTDNA profiles, together with the possibility for highly sensitive techniques to detect extracellular and degraded DNA in a crime stain, often lead to stochastic effects that are negligible at higher DNA template. These effects can include dropin, dropout, peak imbalance, and exaggerated stutter (see sidebars, Dropin, Dropout, Peak Imbalance, and Stutter).

Because there is no clear distinction between LTDNA and standard DNA profiling, any method of analysis for LTDNA profiles should return the same results as would a standard analysis when presented with profiles obtained using optimal DNA template. Standard DNA profiling protocols usually recommend using approximately 1,000 pg of DNA, but they typically perform well when the DNA template is reduced to 300 pg and often lower. **Figure 2** shows that despite some dropout (for a brief description of dropout, see sidebar, Dropout), substantial information can be obtained from LTDNA profiling of only 31 pg of (nondegraded) DNA, and some information can be obtained from profiling as few as 15 pg.

LTDNA: low-template DNA

pg: picogram (10^{-12} gram)

The essential characteristic of LTDNA profiles is that stochastic phenomena are considered to be potentially important. In single-contributor profiles such as that in **Figure 1**, a single peak at a locus indicates a homozygous genotype. However, the single *D3* peak in **Figure 2** (*top*) has a height

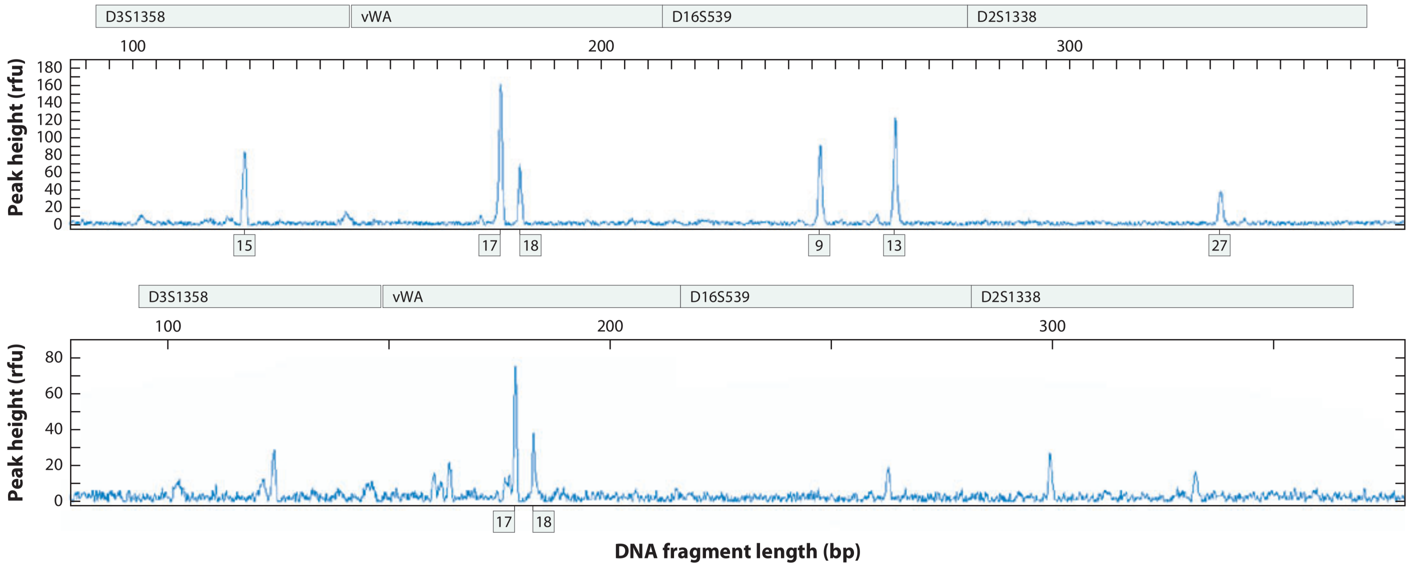

Figure 2

Blue-dye panels for two low-template DNA profiling runs with DNA from the same contributor as in **Figure 1**. These runs have very low DNA template: 31 pg (*top*) and 15 pg (*bottom*). Note the very different y-axis scales; the x-axis scales also differ slightly. Loci names are shown in the gray banner above each panel. Numerical allele labels are given in the boxes below each peak. Abbreviations: bp, base pairs; rfu, relative fluorescence units.

DROPOUT

Dropout arises when the DNA profile of a crime stain does not include an allele from a contributor. Several instances of dropout in **Figure 2** can be noted by comparison with **Figure 1**. For example, in the leftmost locus (*D3*), allele 14 has dropped out in the top panel of **Figure 2**, and alleles 14 and 15 have dropped out in the bottom panel. Dropout also includes peaks that fail to reach the height threshold regarded as sufficient to confidently distinguish an allele from baseline noise. This threshold varies according to laboratory and profiling protocols, but it is typically between 25 and 75 RFU. In **Figure 2** (*bottom*), there is an above-baseline peak (~25 RFU in height) for allele 13 at the position of locus *D16*, which we know from the panel above corresponds to an allele, but as it has not been labeled as such in this run, it is considered a dropout.

of approximately 90 RFU, and an interpreter of the epg should recognize the possibility that an allele has dropped out (which we know is true in this case). A threshold can be adopted, typically 200 to 300 RFU, such that single peaks above this threshold are interpreted as homozygotes (Lohmueller & Rudin 2012). On the basis of such a threshold, the UK National DNA Database encodes a single peak at allele A as AF or AA, where F will match any allele in subsequent searches of the database. Thresholds are somewhat arbitrary and should be selected according to the sample and the profiling system employed. Puch-Solis et al. (2011) propose a method for choosing an appropriate threshold using laboratory- and protocol-specific calibration data.

The effects of DNA degradation due to environmental exposure, and hence the dropout rate, tend to increase with DNA fragment length, so stochastic effects may be important for only some loci. Moreover, most DNA profiles are mixed, and DNA quantification techniques are unable to estimate the amounts of DNA template originating from different contributors. Thus there may be ample DNA from one contributor but little DNA from other contributors. Forensic scientists often reserve the term contamination for the introduction of foreign DNA into a sample after it is recovered. This distinction can be important in a trial, but we cannot determine from the epg the times at which DNA from different sources came to be in the sample. Thus, the term contamination is also used more loosely to refer to any DNA that is not from persons of interest to the investigation, including dropin alleles. It is common to distinguish "gross contamination," in which all of an individual's DNA is introduced into the crime-related sample, from environmental contamination such as dropin (see sidebar, Dropin).

DROPIN

Dropin arises when an allele that is not in the genotype of any assumed contributor to the crime stain is observed in the crime scene profile (CSP). Although a dropin allele must have come from somebody, treating the allele as sporadic may be more appropriate than treating it as one among multiple alleles contributed by an unknown individual, particularly if the DNA from that individual is extremely low template or degraded. The appearance of sporadic peaks in DNA-blank control runs confirms the presence of dropin in LTDNA profiles. These peaks are primarily thought to be a result of airborne DNA fragments, perhaps from previously analyzed samples, hence the term dropin. However, sporadic alleles from degraded DNA may arise from environmental exposure at the crime scene. Some authors only consider dropin due to lab-based contamination, but because the source of a dropin allele cannot be verified, we here regard dropin as referring to any sporadic DNA alleles.

2. PRINCIPLES OF DNA PROFILE EVIDENCE EVALUATION

2.1. Likelihood Ratios

Despite substantial resistance, not least from judges unfamiliar with quantitative evidence evaluation, the use of likelihoods as the primary tool for evidence evaluation has gained ground in recent years (Gill et al. 2012). In simple settings, we can form a likelihood ratio (Evett et al. 1991, Evett & Weir 1998)

$$\mathrm{LR} = \frac{\Pr(E|H_\mathrm{p})}{\Pr(E|H_\mathrm{d})}, \qquad 1.$$

where E denotes the evidence and H_p and H_d represent competing hypotheses corresponding to the prosecution and defense positions, respectively.

Many difficulties may arise in putting Equation 1 into practice, and some proponents of using LRs in the legal process have understated these challenges. However, these difficulties are more easily overcome than are the problems faced by alternative methods of evidence evaluation that lack the sound conceptual basis of likelihood-based analyses (Robertson & Vignaux 1995).

One potential source of confusion for statisticians is that the LR is not considered as a statistic to be tested against a null distribution. Finders of fact (a juror or a judge) can use LRs to update probabilities for H_p and H_d using Bayes theorem. Updating these probabilities requires prior probabilities, which can be based on other evidence, and may therefore lie outside the domain of the DNA expert witness. Such a witness should not apply Bayes Theorem in court, except for illustrative calculations (see Section 5).

Likelihood ratio (LR): in forensic evidence evaluation, the LR is the probability of the evidence assuming H_p to its probability assuming H_d

H_p: hypothesis representing the prosecution allegation; typically unique

H_d: hypothesis representing an alternative to H_p that is consistent with the defense case; typically, many H_d should be considered

Queried contributor (Q): a contributor under H_p but not under H_d

CSP: crime scene profile

2.2. Formulating the Hypotheses

One of the major difficulties in likelihood-based evidence evaluation is the choice of hypotheses, H_p and H_d. Specifying H_p usually poses fewer problems because the prosecution makes a specific allegation. In contrast, the defense is not required to give any account of the events surrounding an alleged crime, making it difficult for a forensic scientist to choose an appropriate H_d. The forensic scientist must usually allow for all reasonable alternatives to H_p. Gill & Haned (2013) propose an exploratory framework for formulating competing hypotheses in the interpretation of complex DNA profiles and discuss many of the associated issues.

Typically, H_p will specify a person of interest Q, such as the defendant or a crime victim, as a source of the DNA that generated the crime scene profile (CSP). Then, it may be reasonable to assume under H_d that Q is replaced by an unknown individual X. Problems can arise in specifying the numbers of known and unknown contributors of DNA, their genetic ancestries (or so-called ethnicities, see Section 2.4), and the levels of relatedness among them. To account for all such possibilities, H_d should be decomposed into exhaustive, mutually exclusive alternatives, each of which is precise enough to allow computation of the likelihood. The defense should seek to identify plausible alternative scenarios that have not been considered.

Thus, the computation of Equation 1 is more complex than at first appears because the denominator is a sum over alternative scenarios that are weighted by their plausibilities (which in turn are based on other evidence). This weighting encroaches on the role of the finder of fact. To avoid such potential encroachment, an LR can be reported for each scenario, which the finder of fact can then combine. The forensic scientist may simplify this task by reporting only one LR for multiple scenarios, the LR that most favors the defense. For example, a forensic scientist may report only the smallest LR for a range of possible ethnicities of X and numbers of contributors of DNA to the CSP. However, strict adherence to the "most favorable to the defendant" principle is not always reasonable. For example, the alternative scenario in which the DNA was left by an identical twin of Q, from whom he was separated at birth, would generate an LR of 1. Such a scenario can rarely be unequivocally disproved, and the forensic scientist must in effect dismiss it as implausible a priori while recognizing that the defense remains free to propose it at trial.

RMNE: random man not excluded

In some cases, the defense may accept that Q could be a source of the crime scene DNA but assert that his DNA became part of the crime stain by some means other than committing the crime, perhaps involving gross contamination of the stain. We do not consider such a defense here: Although the relevance of DNA evidence is often an important issue, it is not easily amenable to statistical analysis. Thus, in the remainder of this review, we assume that the source of the CSP, not the relevance of the DNA evidence, is in dispute between the prosecution and the defense.

2.3. Random Man Not Excluded Probability

The major alternative to the LR approach for evidence evaluation is to report the inclusion probability, often referred to as the Random Man Not Excluded (RMNE) probability. This approach simplifies the DNA profile evidence to the observation that an individual of interest, say Q, cannot be excluded from having contributed DNA to the crime stain. Typically, a large fraction of the population is excluded, so the nonexclusion of Q would be surprising if he were not a true contributor. Because the RMNE probability gives a measure of this surprise, it conveys a sense of the evidential strength that Q is a contributor.

For a single-contributor CSP matching the genotype of Q, the RMNE probability is numerically similar to the LR, but they diverge in more complex scenarios. For example, the RMNE probability does not depend on the genotype of Q, so its value is the same for all nonexcluded individuals. The RMNE probability has two major perceived advantages over LRs: It does not require the number of contributors to the crime stain to be specified, and it is regarded as being easier to explain in court. However, by not addressing the question of direct interest to the court, the RMNE approach wastes information, in addition to having other deficiencies that also apply to standard profiles (Balding 2005, Bille et al. 2013, Gill et al. 2006). It faces particular difficulties in evaluating LTDNA evidence (Buckleton & Curran 2008) because dropout and dropin make the concept of exclusion difficult to define. One attempt to overcome this difficulty (Van Nieuwerburgh et al. 2009) regards an individual Q as excluded if he has more than two alleles that are not observed in the CSP.

Bille et al. (2013) propose an extension to the random match probability (RMP), which is equivalent to the LR in single-source cases, that makes the RMP applicable to mixture CSPs. They propose this approach as a middle ground between the LR and RMNE methods: The RMP approach is similar to the RMNE approach but uses the number of contributors and the peak heights to make more efficient use of the observed CSP data.

2.4. Population Genetics

Most analyses of forensic DNA profiling assume that the genotype of an unprofiled individual not related to any profiled individual is an unordered pair of alleles that correspond to independent random draws according to known allele fractions. In practice, these allele fractions are estimated in one or more populations. The three main databases currently used in the United Kingdom are for Caucasian, Afro-Caribbean, and Indo-Pakistani populations. A sampling adjustment may be made to bias upward the frequency estimates of rare alleles and to avoid allele counts of zero. For example, such an adjustment might involve adding a pseudocount of one or two to the observed allele counts (Balding 1995). There are many deficiencies in current practice in this area: Databases in forensic use are often small and are convenience samples instead of samples resulting from a designed experiment, and ethnic labels are assigned in a manner that is rarely precise and often subjective (e.g., based on a police officer's assessment).

These are long-standing problems in the use of DNA profile evidence. Fortunately, STR allele fractions tend not to vary greatly across populations, however they are defined. Further, incorporating a fixation index (F_{ST}) adjustment into the calculation of the LR can make an allowance for the genetic ancestries of Q, X, and individuals in the database (Balding 2005). Here, we consider F_{ST} to be the population genetics parameter that specifies the coancestry of Q and X relative to the database population. The more similar are the ancestries of Q and X, and the more they differ from those of the database population, the larger the appropriate value of F_{ST}, which tends to reduce the LR. For details of F_{ST} adjustments in a range of non-LTDNA settings, the reader is referred to Fung & Hu (2008). Considering that the value of F_{ST} is never precise, an approximate allowance for coancestry between Q and X can be obtained by replacing each allele fraction p with an adjusted value:

F_{ST}: a population genetics parameter used here to specify the coancestry of Q and X relative to the population in which the allele fractions p are estimated

$$\begin{aligned}&(1-F_{\mathrm{ST}})p/(1+F_{\mathrm{ST}}) \quad \text{for an allele not in the profile of Q,}\\&(F_{\mathrm{ST}}+(1-F_{\mathrm{ST}})p)/(1+F_{\mathrm{ST}}) \quad \text{for a heterozygote allele of Q,}\\&(2F_{\mathrm{ST}}+(1-F_{\mathrm{ST}})p)/(1+F_{\mathrm{ST}}) \quad \text{for a homozygote allele of Q.}\end{aligned}$$

The appropriate value of F_{ST} is typically small, except for relatively few individuals X who have substantial coancestry with Q. However, because coancestry levels among different X are difficult to specify and because of the problems with population databases outlined above, forensic scientists customarily err in favor of defendants by using a relatively large value of F_{ST} (say 2% to 5%) for all X. Such large values can be limited to LRs computed using the database most appropriate for Q: Little F_{ST} adjustment is required for LR calculations using more remote databases. Thus, it is almost always favorable for the defense to report LRs obtained using the database closest to the ancestry of Q, together with a generous value of F_{ST}, even though the alternative X may have a different ancestry from Q.

In US forensic practice, owing to a long-standing misunderstanding introduced by the NRC2 report (Natl. Res. Counc. 1996), F_{ST} (also called θ) is usually used to model only within-individual genetic correlations (i.e., excess homozygosity). However, these correlations are of little relevance to evidential weight. Only between-individual correlations matter in practice, and failing to model them results in LR values that are biased against defendants. Modeling only excess homozygosity, as proposed in the NRC2 report Recommendation 4.1, is essentially irrelevant and gives a false impression of having adjusted for coancestry.

3. COMPUTING LIKELIHOODS

3.1. Standard, Single-Contributor Case

Consider first a standard, single-contributor CSP at a single locus. If the contributors under the competing hypotheses are

$$H_{\mathrm{p}}^{1}: \mathrm{Q} \quad \text{and} \quad H_{\mathrm{d}}^{1}: \mathrm{X},$$

then under the usual assumptions (Balding 2005, Buckleton et al. 2004), the LR is as follows:

$$\mathrm{LR} = \frac{(1+F_{\mathrm{ST}})(1+2F_{\mathrm{ST}})}{2(F_{\mathrm{ST}}+(1-F_{\mathrm{ST}})p_A)(F_{\mathrm{ST}}+(1-F_{\mathrm{ST}})p_B)} \quad \text{if Q} \equiv \text{AB and CSP} = \text{AB}, \qquad 2.$$

$$= \frac{(1+F_{\mathrm{ST}})(1+2F_{\mathrm{ST}})}{(2F_{\mathrm{ST}}+(1-F_{\mathrm{ST}})p_A)(3F_{\mathrm{ST}}+(1-F_{\mathrm{ST}})p_A)} \quad \text{if Q} \equiv \text{AA and CSP} = \text{A}, \qquad 3.$$

Table 1 Likelihoods under H_{p}^{1} and H_{d}^{1} for a single-contributor crime scene profile at a single locus given an alleged contributor Q for whom dropout is considered possible

Crime scene profile	L_{p}^{1} (likelihood under H_{p}^{1}) if Q ≡ AB	L_{p}^{1} (likelihood under H_{p}^{1}) if Q ≡ AA	L_{d}^{1} (likelihood under H_{d}^{1})
A	$D(1-D)$	$(1-D_2)$	$p_A^2(1-D_2)+2p_A(1-p_A)D(1-D)$
Ø	D^2	D_2	$P_{\mathrm{hom}}D_2+(1-P_{\mathrm{hom}})D^2$

Ø denotes no observed alleles.
D and D_2 denote the probabilities of dropout for heterozygote and homozygote alleles, respectively.
P_{hom} is the fraction of the population that has a homozygous genotype.

where ≡ denotes "has genotype" and the p are population allele fractions. Henceforth, we assume the p have already been subjected to the sampling and F_{ST} adjustments described above. Thus, we can ignore F_{ST} and Equations 2 and 3 simplify to $1/(2p_Ap_B)$ and $1/p_A^2$, respectively. Because F_{ST} and, if appropriate, any coefficients measuring direct relatedness of Q and X account for the positive correlations across loci due to shared ancestry between Q and X, full-profile LRs can be computed via multiplication of single-locus LRs, which is standard practice in the assessment of DNA profile evidence (Buckleton et al. 2004). Thus, we focus here on the single-locus case.

3.2. Allowing for Dropout

Suppose that we wish to evaluate an epg displaying low peak heights, which suggest that dropout may have occurred. If at a particular locus Q ≡ AB and CSP = AB then no dropout has occurred and the LR is unchanged from that given by Equation 2. **Table 1** gives likelihoods for other cases. These likelihoods are derived under a standard probability model, similar to that of Gill et al. (2000) and further developed by several subsequent authors (Balding & Buckleton 2009; Gill et al. 2008, 2012), which assumes that allele dropouts are independent Bernoulli events with probability D (D_2 for homozygotes).

Consider the case in which Q ≡ AB and CSP = A (**Table 1**, *top row*). The likelihood L_{p}^{1} is the probability that the B allele of Q has dropped out (D), but the A allele has not ($1-D$). In L_{d}^{1}, either X is AA and no dropout has occurred (first term), or X is heterozygous but the non-A allele has dropped out (second term). Logically, D in L_{p}^{1} differs from D in L_{d}^{1}, but both hypotheses typically support similar values and the values of D in the two likelihoods are often taken to be equal for illustrative LR calculations (Gill et al. 2007).

The case in which Q ≡ AA and CSP = A is consistent with zero dropout ($D = D_2 = 0$). In this case, LR $= 1/p_A^2$, the simplified form of Equation 3 introduced above. Otherwise, if dropout is impossible, $L_{\mathrm{p}} = \mathrm{LR} = 0$.

3.3. Profiled Contributors Not Subject to Dropout

LTDNA profiles are frequently mixed (containing DNA from multiple individuals). In one common scenario, there is ample DNA from a known, profiled contributor K, perhaps a victim, whereas the offender, Q or X, contributes LTDNA. If the genotypes of K and of Q or X have no alleles in common at a locus, then the likelihoods in **Table 1** still apply. Otherwise, K is said to mask one or both alleles of Q or X. **Table 2** gives single-locus likelihoods for some cases of masking under

Table 2 Likelihoods under hypotheses H_{p}^{2} and H_{d}^{2} for a two-contributor crime scene profile at a single locus when K is a profiled contributor not subject to dropout

Crime scene profile	Genotype of contributor K	L_{p}^{2} (likelihood under H_{p}^{2}) if Q ≡ AB	L_{p}^{2} (likelihood under H_{p}^{2}) if Q ≡ AA	L_{d}^{2} (likelihood under H_{d}^{2})
ABC	BC	$1-D$	$1-D_2$	$p_A^2(1-D_2)+2p_A(p_B+p_C)(1-D)+2p_A(1-p_A-p_B-p_C)D(1-D)$
AB	BB	$1-D$	$1-D_2$	$p_A^2(1-D_2)+2p_Ap_B(1-D)+2p_A(1-p_A-p_B)D(1-D)$
BC	BC	D	D_2	$(p_B+p_C)^2+2(p_B+p_C)(1-p_B-p_C)D+P_{\mathrm{het}}D^2+P_{\mathrm{hom}}D_2$
B	BB	D	D_2	$p_B^2+2p_B(1-p_B)D+P_{\mathrm{het}}D^2+P_{\mathrm{hom}}D_2$
AB	AB	1	1	$(p_A+p_B)^2+2(p_A+p_B)(1-p_A-p_B)D+P_{\mathrm{het}}D^2+P_{\mathrm{hom}}D_2$

P_{het} and P_{hom} denote the population fractions of heterozygous and homozygous genotypes that do not include any of the CSP alleles.

the following hypotheses:

$$H_{\mathrm{p}}^{2}: \mathrm{Q}+\mathrm{K} \quad \text{and} \quad H_{\mathrm{d}}^{2}: \mathrm{X}+\mathrm{K}.$$

When Q ≡ AB, K ≡ BC, and CSP = ABC, the LR comparing H_{p}^{2} and H_{d}^{2} can be written in the following form:

$$\mathrm{LR}=\frac{L_{\mathrm{p}}^{2}}{L_{\mathrm{d}}^{2}}=\frac{P(\mathrm{CSP}=\mathrm{ABC}|\mathrm{Q}\equiv\mathrm{AB},\mathrm{K}\equiv\mathrm{BC})}{\sum_{g\in\Gamma}p_g P(\mathrm{CSP}=\mathrm{ABC}|\mathrm{X}\equiv g,\mathrm{K}\equiv\mathrm{BC})} \qquad 4.$$

where Γ denotes the set of possible genotypes and p_g denotes the population fraction of genotype g. The first row of **Table 2** gives explicit expressions for numerator and denominator. The alleles of K must appear in the CSP, and L_{p}^{2} is obtained by noting whether or not each allele of Q has dropped out or whether the occurrence of dropout cannot be determined because of masking by an allele of K. L_{d}^{2} is obtained by summing, over each possible genotype for X, the product of the population fraction of that genotype and dropout, nondropout, or masking terms similar to those in L_{p}^{2}.

K: a possible contributor to the crime stain for whom a reference profile is available

Balding (2013) further extended this model by allowing the allelic status of an epg peak to be uncertain, rather than limited to either present or absent. This status can be applied to alleles of borderline peak height or to peaks that may be attributable to stutter (for a brief description, see sidebar, Stutter) or other artifact. An uncertain allele is treated in the same way as an allele masked

STUTTER

Stutter peaks in electropherograms arise because of imperfect DNA copying during PCR. Most often, one repeat unit of the DNA motif is omitted, generating a stutter peak at a position corresponding to one repeat unit shorter than an allelic peak. Occasionally, two repeat units are omitted (sometimes called double stutter) or a repeat unit is added (overstutter). Although stutter occurs for standard DNA profiles (many small stutter peaks can be observed in **Figure 1**), stutter peak heights are exaggerated in some LTDNA protocols. Stutter is problematic when a sample contains DNA from multiple contributors with different template levels because an allele peak from a minor contributor can be indistinguishable from a stutter peak generated by an allele from a major contributor. Threshold-based decision rules are often employed in practice; these rules suggest treating peaks in stutter positions as stutter peaks if their heights are, say, less than 15% of the peak height at the parent allele. Improved methods of evaluation are now available that allow an "uncertain" designation for peaks near the threshold or that use peak heights continuously, avoiding thresholds altogether (see Section 3.8).

by K: The dropout term has a value of 1 because we do not know whether or not the allele has dropped out.

3.4. Modeling Dropin

Dropins are usually modeled as independent Bernoulli events: At each allele not in the genotype of any hypothesized contributor, a dropin allele occurs with probability C and the allelic type of the dropin is distributed in proportion to the p, after excluding the alleles of the hypothesized contributors. When there are unprofiled contributors, such as X under L_d in Equation 4, a CSP allele not attributable to K is included in the genotype of X for some terms of the sum; for other terms, this allele is treated as a dropin. The probability of dropin at a locus should depend on the relative frequencies of the nondropin alleles because dropins cannot be observed at those alleles. This dependence implies different dropin probabilities for each term of the summation. However, these complexities are often ignored in practice. Both this difficulty and that of modeling the variable effect of degradation with fragment length for dropins provide reasons to avoid explicit modeling of dropin where possible and to assume instead the presence of an additional unprofiled contributor with a low DNA template and hence high dropout. Despite its limitations, the dropin model may provide an acceptable approximation that can reduce computational effort, provided the number of dropins allowed at a locus is limited. When unlimited dropout and dropin are allowed, every CSP is consistent with every hypothesis for the contributors.

3.5. Multidose Dropout and Degradation

One consequence of the bottom row of **Table 1** is that a null CSP (no observed alleles) is informative (LR $\neq$ 1). This occurs because D_2, the probability of dropout for a homozygous allele, is in general not equal to the probability of dropout for two heterozygous alleles (D^2). Balding & Buckleton (2009) noted that we expect $D_2 < D^2$, and, on the basis of limited data, they used $D_2 = D^2/2$ in their numerical calculations. Tvedebrink et al. (2009) criticized an obvious weakness of this approximation: D_2 can never exceed 0.5. They proposed an alternative model in which dropout probabilities were determined as a function of an average peak height, which was used as a proxy for DNA dose. A general formula is needed when multiple contributors are subject to dropout because the model must then allow for individuals to contribute DNA template for the same allele.

In Tvedebrink et al. (2009), $D(k)$, the dropout probability for dose k of DNA can be written

$$\frac{D(k)}{1 - D(k)} = (\alpha_s k)^\beta, \qquad 5.$$

where s indicates the locus. For k large,

$$\frac{D(2k)}{D(k)^2} \approx \left(\frac{2}{\alpha_s k}\right)^\beta > 1,$$

implying that a homozygous dropout can be more likely than the independent dropout of both alleles, which is implausible. However, because this inequality holds only for low dropout probabilities (Tvedebrink et al. 2012a), the defect of the model is unimportant in practice.

Cowell et al. (2013) and Puch-Solis et al. (2013) define the dropout probability as corresponding to the lower tail of a gamma distribution. Cowell et al. (2013) show that their definition implies, for typical parameter values, a slower increase in dropout rate than that given by Equation 5 as the DNA dose decreases. They note that their threshold-based definition of dropout must satisfy $D_2 < D^2$, but no empirical comparison of goodness-of-fit has been made for these models.

DNA degrades over time at a rate that depends on temperature, humidity, and environmental exposure. This degradation is manifested in an approximately exponential decline of epg peak heights as fragment lengths increase (Bright et al. 2013a). Tvedebrink et al. (2012b) proposed a geometric model to estimate the effective amount of DNA at a given allele fragment length: The longer the fragment, the smaller the effective dose. An STR allele consists of flanking regions and tandem repeats, so the number of repeats that characterizes the allele is not a good proxy for fragment length. Fragment length for many DNA profiling systems can be obtained from the Short Tandem Repeat DNA Internet DataBase (STRbase) website (**http://www.cstl.nist.gov/strbase/**).

U: an unprofiled possible contributor of DNA to the crime stain

3.6. Additional Contributors Subject to Dropout

Given a multidose dropout model, likelihoods can be specified for any number of contributors, each of whom may or may not be profiled and/or subject to dropout. For example, in the scenario of Equation 4, if instead of K, we posit an unprofiled contributor U who is subject to dropout, then we also need to sum the numerator and denominator over all possibilities for the genotype of U, multiplying each term by the genotype probability, which yields the following equation:

$$\mathrm{LR} = \frac{\sum_{g\in\Gamma} p_g P(\mathrm{CSP} = \mathrm{ABC}|\mathrm{Q} \equiv \mathrm{AB}, \mathrm{U} \equiv g)}{\sum_{g1,g2\in\Gamma} p_{g1} p_{g2} P(\mathrm{CSP} = \mathrm{ABC}|\mathrm{X} \equiv g1, \mathrm{U} \equiv g2)}. \qquad 6.$$

Because two contributors are now subject to dropout, we need information about the relative amounts of DNA from each contributor to compute the likelihood. This information can come from the heights of peaks believed to have originated from only one individual (Tvedebrink et al. 2009), or DNA template levels can be unknown parameters to be eliminated via integration or likelihood maximization (Balding 2013, Cowell et al. 2013).

3.7. Replicates

Because of the stochasticity of LTDNA profiles, attempting to replicate the profiling process seems natural in order to distinguish robust signals that appear in each replicate from unreplicated artifacts. Currently in the United Kingdom, most LTDNA work uses between two and four profiling runs. Mitchell et al. (2012) report that they use three profiling runs if the total amount of extracted DNA is <300 pg and one or two otherwise.

An early recommendation for the evaluation of replicate epgs was to first derive a single consensus profile (Gill et al. 2000). Benschop et al. (2011) developed optimal strategies for constructing a consensus from multiple replicates. They recognized that a full statistical analysis could obviate the need for a consensus profile, but they felt that the statistical models and software available when they were writing (in 2010) were not yet sufficiently advanced and available to be a practical option. The software review below suggests that in 2013, statistical analysis of replicates is a practical option.

There are arguments for and against replication: Some view seeking verification of results by replication wherever possible as fundamental to the scientific method. Pfeifer et al. (2012) advocate replication to overcome the interpretation problems inherent in LTDNA profiles and to benefit from combining the strengths of different profiling technologies. Grisedale & van Daal (2012) oppose this view, noting that replication can divide an already minuscule sample, and for very small samples it is preferable to use all available DNA to get the best profile possible from a single run. By sharing information across alleles and loci under an appropriate statistical model, its parameters can be estimated, and likelihoods can be obtained from a single profiling run.

Continuous model (algorithm): here, a model (algorithm) that uses epg peak height information, resulting in an LR that varies continuously with peak height

Curran et al. (2005) extended LR calculations to multiple replicates in the presence of dropout and dropin. They assumed the replicates were independent, conditional on the genotypes of any unknown contributor. For example, under the hypothesis Q + U, the likelihood for multiple replicates can be expressed in the form

$$L = \sum_{g \in \Gamma} p_g \prod_r P(\mathrm{CSP}_r | Q, U \ \equiv g), \qquad 7.$$

where CSP_r denotes the set of alleles observed in the rth replicate. Note that whenever the hypothesis specifies an unknown contributor, replicates are not unconditionally independent: The assumption of independence only applies conditional on the unknown contributor genotypes (i.e., for each term in the summation, not the overall sum).

3.8. Using Peak Heights

So far, we have assumed that the information in the epg is summarized as a list of alleles called as present (possibly also a list of uncertain alleles). For mixed DNA profiles, epg peak heights contain information about allele dose, as **Figure 3** illustrates for a CSP with negligible dropout. There are four peaks with roughly equal heights at loci *D3* and *D2*, suggesting two contributors and equal probabilities for the six possible pairings of the four alleles. At locus *vWA*, however, there are only three peaks, and the peak at allele 17 is approximately double the height of those at alleles 14 and 19, indicating that either the two contributors are 17,17 and 14,19, or they are 14,17 and 17,19. At *D16*, it appears that either both contributors are 11,13, or one is 11,11 and the other is 13,13.

In LTDNA work, the variability of peak heights makes categorical inferences of allele counts infeasible. However, an appropriate statistical model for peak height as a function of DNA template can lead to useful inferences about the latter and hence about contributor genotypes (Pascali & Merigioli 2012, Perlin & Sinelnikov 2009, Perlin & Szabady 2001). Although DNA profiling systems also report peak areas (Evett et al. 1998, Gill et al. 1998), they are highly correlated with peak heights (Tvedebrink et al. 2010), and the latter are usually preferred.

Continuous models (which use peak heights) lead to likelihoods of the same form as, for example, Equation 7, but the CSP now consists of a peak height for each allele at a locus, rather than an indicator of presence or absence. The lognormal or gamma distributions can provide models for peak heights in which the means are specified by the DNA template and the variances estimated by deviations from the mean over the whole profile (Cowell et al. 2007a). Puch-Solis et al. (2013) also introduced a gamma model for stutter peak height, in which the mean height was computed as a fraction of the mean at the parent allele, fixed over loci but estimated from the observed profile (not fixed over runs).

Because baseline noise generates peaks across the entire epg, some so-called continuous models are not fully continuous. Rather, they use a threshold of detection below which any peak is ignored. Thus, the distribution of peak heights is continuous above the threshold but has an atom of probability mass corresponding to dropout (below-threshold peak height).

Continuous methods are expected to outperform discrete methods because they exploit additional information in the peak heights (Perlin & Sinelnikov 2009). However, peak heights in LTDNA profiles are highly variable, and the pattern of variability may be sensitive to details of the DNA profiling protocol, hindering their usefulness (Gill & Haned 2013). Parameter estimation for some continuous models depends on only the CSP data (Cowell et al. 2013, Perlin et al. 2011),

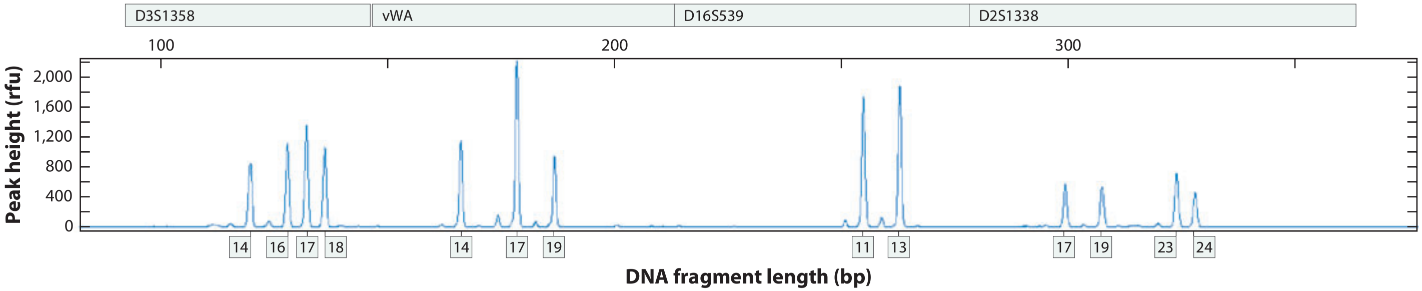

Figure 3

The blue-dye panel of an electropherogram from standard DNA profiling of a mixture with equal amounts of DNA template from two contributors. Loci names are shown in the gray banner above each panel. Numerical allele labels are given in the boxes below each peak. Abbreviations: bp, base pairs; rfu, relative fluorescence units.

whereas other models depend on calibration data generated under the same conditions that were employed for the CSP being evaluated (Puch-Solis et al. 2013, Taylor et al. 2013). In the latter case, a lack of calibration data may make evaluation impossible, whereas in the former case, models may be sensitive to the details of the DNA profiling protocol. Although discrete-model LRs may have similar drawbacks, the simpler data and modeling assumptions on which they are based diminish such concerns, possibly allowing these models to enjoy an advantage of robustness to laboratory-specific details in return for a loss of statistical efficiency. To our knowledge, no work has assessed the robustness of any LTDNA model to varying the DNA profiling platform.

The gain in statistical efficiency from using peak height data is most important for single-run CSPs. Puch-Solis et al. (2013) show results from two single-run CSPs: Both are two-person mixtures, and in each analysis the contributor other than Q or X is regarded as unknown. These authors find that their continuous model yielded a weight of evidence (WoE) in favor of a true hypothesis that was approximately 2–3 decibans per locus higher than the WoE for a discrete model. However, as the number of replicates increases, the LR computed from an LTDNA sample under a discrete model often converges to the LR available from a CSP not subject to dropout, in which case a continuous model can convey little advantage. When the dropout rate is low, only two or three replicates can suffice to come close to this optimal LR.

WoE: weight of evidence

Discrete model (algorithm): here, a model (algorithm) that uses epg peak heights only indirectly after peaks are classified as allelic or nonallelic; peak status may also be uncertain in some models

Ban: unit of WoE used when comparing two specified hypotheses; $\log_{10}(\text{LR}) = x$ bans

4. CURRENT SOFTWARE

4.1. Overview of Software Programs

The currently available programs for computing LTDNA profile LRs under discrete models are Forensim, Forensic Statistical Tool (FST), and likeLTD. TrueAllele, DNAmixtures, and STRmix compute LTDNA profile LRs under continuous models. **Table 3** summarizes the features of these programs, and the programs are described briefly in subsequent subsections. Another continuous model has been published by Puch-Solis et al. (2013), but no software has been released.

All six programs discussed here sum over unknown contributor genotypes. The population genetics models that specify the prior distribution are similar across the six programs but disagree over the use of F_{ST}. DNAmixtures represents the genotypes of the unknown contributors and the peak height data in Bayesian networks and uses efficient network algorithms to perform the integrations. (These algorithms are implemented in HUGIN and accessed through the R

Table 3 Summary of key features of software programs available for low-template DNA profile analysis

Program	Open source	Model	Allows relatedness[a]	Parameter elimination method[b]	Maximum number of unprofiled contributors under H_d
Forensim	✓	Discrete	✗	User	3
FST	✗	Discrete	✗	Plug-in	4
likeLTD	✓	Discrete	✓	Maximization	3
TrueAllele	✗	Continuous	✓	Integration	6
DNAmixtures	✓[c]	Continuous	✗	Maximization	6
STRmix	✗	Continuous	✓	Integration	4

[a] Symbols indicate whether or not the program takes into account close relatedness between Q and X.

[b] The method used to eliminate the model parameters from the likelihood (all programs use integration for the unprofiled contributor genotypes).

[c] The DNAmixtures code is open source, but running it requires HUGIN, which is not.

package Rhugin.) Another R package, Rsolnp, is then used to maximize over the model parameters. DNAmixtures as described here requires a licensed copy of the commercial software HUGIN (**http://www.hugin.com**).

MCMC: Markov chain Monte Carlo

Forensim requires user-supplied estimates for the model parameters, and FST uses estimates generated from in-house training data. Similar to DNAmixtures, likeLTD maximizes likelihoods over the model parameters, avoiding the need to specify prior distributions, but for some parameters, it uses penalty functions, which are analogous to prior distributions.

TrueAllele and STRmix implement fully Bayesian inferences. This implies that a prior distribution is assumed for all unknowns, leading to a posterior distribution that reflects both the prior and the DNA profile evidence. The specification of prior distributions can be problematic in an adversarial courtroom environment. In addition, performing Bayesian inference may appear to conflict with the requirement that expert witnesses do not present to courts a probability that the prosecution case is true, as that would require an assessment of all the evidence. However, it is possible to report an LR relating only to the DNA evidence by focusing on the posterior distribution for the genotype of the contributor of interest (X) under H_d. The LR is then evaluated in terms of the posterior probability assigned to the genotype of Q.

The prior distribution for the genotype of X is given by the standard population genetics model described above in Section 2.4. In the single-contributor setting discussed in Section 3.1, the posterior probability that X has the genotype of Q is equal to 1. The LR is then the ratio of the posterior to prior probabilities for that genotype and is equal to the inverse of the prior probability. In more complex LTDNA settings, the posterior probability that X has the genotype of Q is less than 1, and the LR is reduced accordingly.

TrueAllele and STRmix approximate the posterior distribution for the genotype of X using Markov chain Monte Carlo (MCMC) algorithms to perform the required integrations over all other unknowns, including the genotypes of other unprofiled contributors and model constants such as prior distribution parameters. MCMC generates a sequence of output vectors, each of which is treated as a sample from the joint posterior distribution of all unknowns. These outputs facilitate the approximation of marginal posterior distributions for unknowns of interest other than the genotype of X, such as the genotypes of other unprofiled contributors. Inferring the genotypes of all contributors to a DNA mixture is sometimes called deconvolution.

Because computations are performed assuming H_d, the above approach is only feasible when H_p is a special case of H_d, which is usually but not always the case. An exception arises when the prosecution and defense propose different numbers of contributors to the CSP. Perlin et al. (2011) noted that because the computations are performed under H_d, any possibility of prosecution bias is avoided. This is also true whenever the algorithm for computing the LR is chosen without case-specific knowledge.

Half of the maximum number of distinct alleles observed over a set of loci provides a lower bound on the number of contributors to the CSP, but no upper bound exists. Both L_p and L_d must be nondecreasing in the assumed number of contributors, and these likelihoods are difficult to compute for large numbers of contributors (**Table 3**). Intuitively, one would expect the LR to change little if at all as the number of contributors increases beyond the minimum number required to explain the observed alleles. Cowell et al. (2013) use DNAmixtures to verify this intuition for a CSP requiring at least three contributors. The calculated WoE of 14.09 bans assuming three contributors decreases slightly as the number of assumed contributors increases, reaching 14.04 bans for eight contributors. Although this increase means that limiting the number of contributors to the minimum required to explain the observed alleles is unfavorable to a defendant, the effect size is negligible, and adjustments intended to favor the defendant, such as generous F_{ST} and sampling adjustments, easily compensate for it.

4.2. Features of Current Software Packages

The following sections discuss salient features and computation methods of current software packages for computing LTDNA LRs.

4.2.1. Forensim. Forensim (Haned et al. 2012) is an R package for forensic DNA profile simulation and analysis that includes functions to compute LTDNA LRs in a manner similar to LoComatioN (Gill et al. 2007). The LR calculation in this program is built around the recommendations of Gill et al. (2006, 2012), and Gill & Haned (2013) propose its use as a basic model by which other models may be assessed. Forensim includes a function, LRmix, which allows users to enter parameter values through a graphical user interface (GUI) and to select files containing profile information, as well as an allele frequency database. The GUI makes LRmix more accessible to forensic scientists who are not familiar with scientific computing, but users can also choose to use likEvid, which has no GUI but offers additional model flexibility (Haned et al. 2012). For example, LRmix, but not likEvid, assumes that D is the same under H_p and H_d, and for all replicates and all contributors, possibly leading to misleading inferences for complex mixtures. Forensim does not estimate parameter values; it requires the user to enter values for dropout rates, dropin rates, and F_ST. However, it can generate a plot of LR values as a function of the dropout rate D in addition to a 95% confidence interval for D.

4.2.2. FST. FST (Mitchell et al. 2012) uses empirically estimated dropout rates from laboratory experiments that varied the number of PCR cycles, STR locus, number of contributors, and mixture ratio (approximately equal or not). The calibration data to estimate dropout rates used 2,000 amplifications of 700 DNA samples with between one and three contributors and between 6.25 pg and 500 pg of DNA template. The amplifications were run using 31 PCR cycles and in triplicate for samples up to 300 pg of DNA and with two 28-cycle PCR iterations otherwise.

For the analysis of a crime scene DNA sample, an accurate DNA quantitation system is used, and dropout rates are estimated by interpolation from the lab-generated dropout rates. Separate rates are estimated for homozygous dropout, partial heterozygous dropout, and complete heterozygous dropout, but no dropout rate is specified for cases in which multiple low-template contributors have the same allele. Locus-specific dropout rates are adjusted according to the estimated degree of degradation of a sample, from moderate to severe; sample degradation is estimated on the basis of the ratio of the peak heights for the longest and shortest loci in each dye color. The dropin rate is estimated as a function of the number of PCR cycles. Nonallelic stutter peaks are treated as dropin. FST is not currently transportable outside the laboratory of the Office of the Chief Medical Examiner, New York City, but Mitchell et al. (2012) state that they intend to make it more widely available.

4.2.3. likeLTD. likeLTD (Balding 2013) is an R package that maximizes penalized likelihoods over the model parameters, including the dropout probabilities for a reference individual in each replicate, relative DNA templates for other contributors subject to dropout, degradation and (optional) dropin parameters, and locus adjustment and slope parameters for the dropout model in Equation 5. This tool allows for relatedness between Q and X, specified by two coefficients, and it allows allelic positions to be classified as uncertain in addition to present or absent. likeLTD takes the following as its inputs: the genotypes of Q and any profiled possible contributors; a list of allelic and uncertain CSP calls for each locus and each replicate; and parameter values such as sampling adjustment, F_ST and relatedness coefficients, the numbers of unprofiled contributors under H_d and H_p, a parameter indicating whether or not dropin is being modeled, and the dropin and degradation penalties. The current version of likeLTD (5.0) has evolved from the model described in Balding & Buckleton (2009). An earlier version has been integrated into the Lab Retriever software (**http://scieg.org/lab_retriever.html**).

4.2.4. TrueAllele.

TrueAllele (Perlin et al. 2011, 2013; Perlin & Sinelnikov 2009) models background noise (peak heights in the absence of any allele), allowing it to use epg peak heights at all allelic positions without considering dropout. The underlying model is similar to that of Cowell et al. (2007b). At each locus, the DNA template has a diffuse normal prior truncated at zero. The relative DNA templates from all but one contributor are assigned a multivariate normal prior truncated to a hypercube; the prior has constant variance and means equal to a priori uniform locus-independent weights. The total and relative DNA template parameters, together with the genotypes of the hypothesized contributors (temporarily assigned for unprofiled contributors), specify the expected peak heights at all allelic positions. The peak height variances have one contribution that is linear with respect to expected peak height and another contribution that specifies the variance of the baseline peak heights. All three variance parameters are assigned inverse-gamma priors. Given the means and variances outlined above, the peak heights are independent and have normal distributions truncated at zero. TrueAllele can also account for stutter and some other artifacts, DNA degradation and coancestry, using F_{ST}. In addition to assessing the identity of a contributor, TrueAllele can be used to assess paternity and other types of relatedness, for example, in familial database searches or disaster victim identification.

4.2.5. DNAmixtures.

The R package DNAmixtures (Graversen 2013a,b) is based on the model proposed by Cowell et al. (2013), which is a recent extension of earlier models (Cowell 2009; Cowell et al. 2007a, 2011). Graversen & Lauritzen (2013) further describe computational aspects. Allele peak heights are gamma distributed: The means are a function of the DNA template, the fractions from different contributors, the genotypes of the hypothesized contributors, and the fraction of PCR product that generates stutter peaks (assumed to be beta distributed). Dropin is not explicitly modeled. Silent alleles are handled through the addition of an extra allele that does not result in a peak. Each locus is represented in a Bayesian network, which facilitates likelihood integration over the genotypes of the unprofiled contributors. This representation is possible because loci are independent (contributors are assumed to be unrelated). The program deals with the dependence of expected peak height across neighboring alleles due to stutter by expressing the genotype in a Markov structure using partial allele sums. DNAmixtures does not currently adjust for F_{ST}, relatedness, variation over loci in the peak height model, or degradation. However, it does allow for multiple profiling runs on the same DNA sample in addition to allowing for the analysis of multiple DNA samples assumed to have the same set of contributors who can contribute different proportions of DNA template to each sample. DNAmixtures also has facilities for deconvolution, reporting uncertainty in parameter estimates, and model diagnostics.

4.2.6. STRmix.

STRmix is the result of a collaboration between Environmental Science & Research (ESR) in New Zealand, Forensic Science South Australia, and the Australian National Institute of Forensic Science. Its underlying models are outlined in Bright et al. (2013a,b) and Taylor et al. (2013). Model parameters include DNA template and degradation for each contributor, as well as amplification efficiency for each locus. Input data include the genotypes of profiled possible contributors, and, for the CSP, allelic designations, peak heights, molecular weights (fragment lengths), and expected stutter ratios. STRmix can incorporate an F_{ST} adjustment and can handle multiple replicates of differing intensities (assuming each replicate has the same contributors in the same proportions). The rate of dropout (which corresponds to a subthreshold peak height) is controlled by the peak height variances, which are obtained from calibration data. Two dropin models are available: One uses a fixed penalty, and the other uses a function of observed peak height. STRmix allows different profiling kits and populations. Users can generate an LR for each population, or a single LR can be obtained by weighting over populations. Sampling variation is taken into account using the highest posterior density.

5. DISCUSSION

5.1. Quality of Results

Laboratory procedures to measure a physical quantity such as a concentration can be validated by showing that the measured concentration consistently lies within an acceptable range of error relative to the true concentration. Such validation is infeasible for software aimed at computing an LR because it has no underlying true value (no equivalent to a true concentration exists). The LR expresses our uncertainty about an unknown event and depends on modeling assumptions that cannot be precisely verified in the context of noisy CSP data.

Some progress can be made in evaluating the validity and performance of software. Courts need these kinds of evaluations to have confidence in the results of software-based forensic analyses. Open source software is highly desirable in the court environment because openness to scrutiny by any interested party is an invaluable source of bug reports and suggestions for improvement.

Cowell et al. (2013) noted that the LR implicating Q as a contributor to a crime stain can never exceed the inverse of the match probability for a good-quality single-contributor profile. Thus, they proposed the ratio of the actual LR to this theoretical upper bound as a measure of evidential efficiency. The LR for a sequence of noisy replicate profiling runs of a single-contributor stain should converge to the maximum efficiency of one. A sequence of runs from a mixed-source stain can also reach this bound, provided that the DNA contributions from different individuals are very different, allowing contributors to be distinguished either by different dropout rates for the nonshared alleles or by different contributions to expected peak height. Ballantyne et al. (2013) propose subsampling to generate different mixture ratios in the replicates as a strategy to assist mixture deconvolution. The upper bound on efficiency provides a means to check a statistical model for LTDNA profiles that include many replicates.

Gill & Haned (2013) promote the use of performance tests using a false H_p, which can be a useful way to check whether a model or program implementing it is performing as expected and to compare the behaviors of different models. Performance tests may improve understanding for those unfamiliar with LRs, but they have no direct bearing on the strength of evidence in a specific case. Forensim and FST both facilitate the comparison of an LR with values obtained for the same CSP and hypotheses, but the genotype of Q is replaced with a genotype chosen randomly according to the assumed population genetics model. This permits forensic scientists to make statements similar to the following: "The reported LR is greater than 99.9% of LRs calculated in the same way but replacing Q with a random noncontributor." Unless the number of observed alleles in the CSP is large, a randomly chosen false Q almost always generates a small LR and is therefore effectively excluded under almost any reasonable model. Thus, many simulated examples may generate few situations in which a false Q is a plausible contributor given the CSP. However, both Forensim and FST are computationally fast, so the computational cost is not an issue as it may be for other programs. Balding (2013) reports LRs computed for randomly generated Q genotypes, and noted that a false Q usually explains few CSP alleles, so H_p often requires one more contributor than does H_d.

Owing to the recent emergence of new programs and updates to existing programs, few systematic comparisons of the performance of the programs described above have been published. Such comparisons are a high priority now that the field is beginning to mature. From results that are available, the various programs often do generate different results when comparing the same hypothesis pair. The most important differences are typically between the continuous and discrete algorithms, as the former exploit peak height information. The resulting WoEs can differ substantially, by as much as several bans for single-replicate profiles (less for multiple replicates). Smaller differences arise between programs within these two classes owing to different modeling assumptions and approaches to eliminating nuisance parameters. These differences appear to be small relative to the 10 or more orders of magnitude over which LRs range in practical applications. It follows that it is fruitless to insist on very precise likelihood calculations under any specific model. An error of 1 deciban (∼26% on the natural scale) should be regarded as negligible; different, reasonable, modeling assumptions often have larger implications than this, so bans should be reported to at most one decimal place.

5.2. Presentation of Results: A New Scale for Evidence?

One way to assist a court in interpreting an LR is through illustrative calculations. In court, we have used the following language to describe a WoE of 6 bans comparing Q with an unrelated X:

> Consider the hypothetical scenario in which, on the basis of the non-DNA evidence, a juror considered that there were 1,000 men who could be the questioned DNA source: Mr. Q and 999 men unrelated to him. If each is initially considered equally likely to be the source, the effect of the DNA evidence would be to change the probability that Mr. Q is indeed the correct source from 1 in 1,000 up to 99.9%.

Another possible template sentence that may be helpful is to say that the evidence is as strong as it would be if an eyewitness had chosen Mr. Q from a lineup of 1 million men, and we have to decide whether the eyewitness has picked out the culprit without error or has chosen completely at random. However, neither of these formulations allows for possible relatedness between Q and X.

Because the range of LRs that are reported to courts spans more than 10 orders of magnitude, reporting LRs on a logarithmic scale is convenient. Here we use the ban, a WoE unit introduced by Alan Turing (Good 1979) where $\log_{10}(\mathrm{LR}) = x$ bans. The ban is not currently used in courts, but it may prove useful. There is a convenient analogy between these units of WoE and of the Richter scale for earthquake magnitude. Both are logarithmic scales for which typical values range up to about 10 and are reported to at most one decimal place. There is no maximum value for the WoE, but in practice in the United Kingdom, any value above 9 bans (LR $>$ 1 billion) is reported as "over a billion." The crucial difference between earthquakes and evidence is that WoE depends on the hypotheses compared. Courts therefore need to be reminded that WoE is specific to the stated pair of hypotheses.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Matthew Greenhalgh of Cellmark Forensic Services UK for providing **Figures 1–3** and John Buckleton, Hinda Haned, Steffen Lauritzen, Adele Mitchell, Mark Perlin, Roberto Puch-Solis, and Duncan Taylor for helpful comments on a manuscript draft. C.D.S. is funded by a PhD studentship from the UK Biotechnology and Biological Sciences Research Council and Cellmark Forensic Services.

LITERATURE CITED

Balding DJ. 1995. Estimating products in forensic identification using DNA profiles. *J. Am. Stat. Assoc.* 90:839–44

Balding DJ. 2005. *Weight-of-Evidence for Forensic DNA Profiles*. New York: Wiley

Balding DJ. 2013. Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proc. Natl. Acad. Sci. USA* 110:12241–46

Balding DJ, Buckleton J. 2009. Interpreting low template DNA profiles. *Forensic Sci. Int. Genet.* 4:1–10

Ballantyne J, Hanson EK, Perlin MW. 2013. DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information. *Sci. Justice* 53:103–14

Benschop C, van der Beek C, Meiland H, van Gorp A, Westen A, Sijen T. 2011. Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results. *Forensic Sci. Int. Genet.* 5:316–28

Bille T, Bright JA, Buckleton J. 2013. Application of random match probability calculations to mixed STR profiles. *J. Forensic Sci.* 58:474–85

Bright JA, Taylor D, Curran J, Buckleton J. 2013a. Degradation of forensic DNA profiles. *Aust. J. Forensic Sci.* In press. doi: 10.1080/00450618.2013.772235

Bright JA, Taylor D, Curran J, Buckleton J. 2013b. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Sci. Int. Genet.* 7:296–304

Brookes C, Bright JA, Harbison S, Buckleton J. 2012. Characterising stutter in forensic STR multiplexes. *Forensic Sci. Int. Genet.* 6:58–63

Buckleton J, Curran J. 2008. A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Sci. Int. Genet.* 2:343–48

Buckleton J, Triggs CM, Walsh SJ. 2004. *Forensic DNA Evidence Interpretation*. Boca Raton, FL: CRC Press

Cowell RG. 2009. Validation of an STR peak model. *Forensic Sci. Int. Genet.* 3:193–99

Cowell RG, Graversen T, Lauritzen SL, Mortera J. 2013. Analysis of DNA mixtures with artefacts. arXiv:1302.4404v1 [stat.ME]

Cowell RG, Lauritzen SL, Mortera J. 2007a. A gamma model for DNA mixture analyses. *Bayesian Anal.* 2:333–48

Cowell RG, Lauritzen SL, Mortera J. 2007b. Identification and separation of DNA mixtures using peak area information. *Forensic Sci. Int.* 166:28–34

Cowell RG, Lauritzen SL, Mortera J. 2011. Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Sci. Int. Genet.* 5:202–9

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

Evett I, Buffery C, Wilcott G, Stoney D. 1991. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. *J. Forensic Sci. Soc.* 31:41–47

Evett I, Gill P, Lambert J. 1998. Taking account of peak areas when interpreting mixed DNA profiles. *J. Forensic Sci.* 43:62–69

Evett I, Weir B. 1998. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer

Fung WK, Hu YQ. 2008. *Statistical DNA Forensics: Theory, Methods and Computation*. Sussex, UK: Wiley

Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, et al. 2006. DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. *Forensic Sci. Int.* 160:90–101

Gill P, Curran J, Neumann C, Kirkham A, Clayton T, et al. 2008. Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Sci. Int. Genet.* 2:91–103

Gill P, Gusmão L, Haned H, Mayr W, Morling N, et al. 2012. DNA commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Sci. Int. Genet.* 6:679–88

Gill P, Haned H. 2013. A new methodological framework to interpret complex DNA profiles using likelihood ratios. *Forensic Sci. Int. Genet.* 7:251–63

Gill P, Kirkham A, Curran J. 2007. LoComatioN: A software tool for the analysis of low copy number DNA profiles. *Forensic. Sci. Int.* 166:128–38

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

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

Good IJ. 1979. Studies in the history of probability and statistics. XXXVII AM. Turing's statistical work in World War II. *Biometrika* 66:393–96

Graversen T, Lauritzen S. 2013a. Computational aspects of DNA mixture analysis. arXiv:1307.4956v1 [stat.ME]

Graversen T, Lauritzen S. 2013b. Estimation of parameters in DNA mixture analysis. *J. Appl. Stat.* 40:2423–36

Grisedale K, van Daal A. 2012. Comparison of STR profiling from low template DNA extracts with and without the consensus profiling method. *Investig. Genet.* 3:1–9

Haned H, Slooten K, Gill P. 2012. Exploratory data analysis for the interpretation of low template DNA mixtures. *Forensic Sci. Int. Genet.* 6:762–74

Lohmueller K, Rudin N. 2012. Calculating the weight of evidence in low-template forensic DNA casework. *J. Forensic Sci.* 12017:1–7

Mitchell AA, Tamariz J, O'Connell K, Ducasse N, Budimlija Z, et al. 2012. Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. *Forensic. Sci. Int. Genet.* 6:749–61

Natl. Res. Counc. 1996. *The Evaluation of Forensic DNA Evidence.* Washington, DC: Natl. Acad. Press

Pascali VL, Merigioli S. 2012. Joint Bayesian analysis of forensic mixtures. *Forensic Sci. Int. Genet.* 6:735–748

Pascali VL, Merigioli S. 2014. 'Stochastic' effects at balanced mixtures: a calibration study. *Forensic Sci. Int. Genet.* 8:113–125

Perlin MW, Belrose JL, Duceman BW. 2013. New York State TrueAllele Casework Validation Study. *J. Forensic Sci.* 58:1458–66

Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, et al. 2011. Validating TrueAllele® DNA mixture interpretation. *J. Forensic Sci.* 56:1430–47

Perlin MW, Sinelnikov A. 2009. An information gap in DNA evidence interpretation. *PLoS ONE* 4:e8327

Perlin MW, Szabady B. 2001. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *J. Forensic Sci.* 46:1372–78

Pfeifer CM, Klein-Unseld R, Klintschar M, Wiegand P. 2012. Comparison of different interpretation strategies for low template DNA mixtures. *Forensic Sci. Int. Genet.* 6:716–22

Puch-Solis R, Kirkham AJ, Gill P, Read J, Watson S, Drew D. 2011. Practical determination of the low template DNA threshold. *Forensic Sci. Int. Genet.* 5:422–27

Puch-Solis R, Rodgers L, Mazumder A, Pope S, Evett I, et al. 2013. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Sci. Int. Genet.* 7:555–63

Robertson B, Vignaux T. 1995. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester, UK: Wiley

Taylor D, Bright JA, Buckleton J. 2013. The interpretation of single source and mixed DNA profiles. *Forensic Sci. Int. Genet.* 7:516–28

Tvedebrink T, Eriksen PS, Asplund M, Mogensen HS, Morling N. 2012a. Allelic drop-out probabilities estimated by logistic regression—further considerations and practical implementation. *Forensic Sci. Int. Genet.* 6:263–67

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. 2012b. Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Sci. Int. Genet.* 6:97–101

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. 2009. Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Sci. Int. Genet.* 3:222–26

Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. 2010. Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. *Appl. Stat.* 89:855–74

Van Nieuwerburgh F, Goetghebeur E, Vandewoestyne M, Deforce D. 2009. Impact of allelic dropout on evidential value of forensic DNA profiles using RMNE. *Bioinformatics* 25:225–29

RELATED RESOURCES

International Society for Forensic Genetics (ISFG) software page: **http://www.isfg.org/software**
European Forensic Genetics (EUROFORGEN) Network of Excellence: **http://euroforgen.com**