Evaluation of mixed-source, low-template DNA profiles in forensic science

David J. Balding[1]

University College London Genetics Institute, University College London, London WC1E 6BT, United Kingdom

Edited by Terence P. Speed, University of California, Berkeley, CA, and accepted by the Editorial Board May 31, 2013 (received for review November 13, 2012)

Enhancements in sensitivity now allow DNA profiles to be obtained from only tens of picograms of DNA, corresponding to a few cells, even for samples subject to degradation from environmental exposure. However, low-template DNA (LTDNA) profiles are subject to stochastic effects, such as "dropout" and "dropin" of alleles, and highly variable stutter peak heights. Although the sensitivity of the newly developed methods is highly appealing to crime investigators, courts are concerned about the reliability of the underlying science. High-profile cases relying on LTDNA evidence have collapsed amid controversy, including the case of Hoey in the United Kingdom and the case of Knox and Sollecito in Italy. I argue that rather than the reliability of the science, courts and commentators should focus on the validity of the statistical methods of evaluation of the evidence. Even noisy DNA evidence can be more powerful than many traditional types of evidence, and it can be helpful to a court as long as its strength is not overstated. There have been serious shortcomings in statistical methods for the evaluation of LTDNA profile evidence, however. Here, I propose a method that allows for multiple replicates with different rates of dropout, sporadic dropins, different amounts of DNA from different contributors, relatedness of suspected and alternate contributors, "uncertain" allele designations, and degradation. R code implementing the method is open source, facilitating wide scrutiny. I illustrate its good performance using real cases and simulated crime scene profiles.

forensic genetics | forensic identification | statistical genetics | weight of evidence

Reliability of Low-Template DNA Profiling

Problems with the courtroom use of low-template DNA (LTDNA) profiles were brought into sharp focus in the United Kingdom in 2007, with the collapse of a trial arising from the Omagh bombing in Northern Ireland in 1998. This crime killed 29 people and injured many more; consequently, early termination of the trial and acquittal of the defendant attracted widespread adverse publicity. The judge gave several reasons, but it was his critical appraisal of the LTDNA evidence that captured headlines. In response to the controversy, a report reviewing LTDNA evidence (1) was commissioned by the UK Forensic Science Regulator. The report found the underlying science to be "sound" and LTDNA profiling to be "fit for purpose," although admitting that there was lack of agreement "on how LTDNA profiles are to be interpreted."

I suggest that these comments are somewhat contradictory: Without valid methods of assessing evidential strength, a technique cannot be fit for purpose in the criminal justice system. Fig. 1 shows part of the electropherogram (epg) giving the results from replicate LTDNA profiling runs in a crime investigation. The two epgs show substantial similarity yet also important differences: For example, the 17 allele at locus D19 is detected in Fig. 1 (*Left*) but not in Fig. 1 (*Right*), yet the reverse is true for the 11 allele. Is a technology that produces such variable results reliable? This is often asked by legal commentators, but the term "reliable" is too vague for the question to be useful. What is evident is that there is substantial, but imperfect, information in these results about the genotypes of individuals contributing DNA to the sample. The important question is whether or not we can extract that information with enough statistical efficiency for it to be useful while avoiding overstatement of evidential strength. Fortunately, progress has been made on this front since publication of the report by Caddy et al. (1), and I propose here a methodology for robust and efficient analyses of LTDNA evidence that is incorporated in a freely available suite of R functions.

Case of Knox and Sollecito

Table 1 shows three interpretations of the DNA evidence at five loci from exhibit 165B of the trial in Perugia, Italy, in 2009. The exhibit includes the clasp of a bra, attached to some apparently blood-stained fabric, that was found near the murdered woman, Meredith Kercher. The report (2), written by two academic experts from the Sapienza Università di Roma, was highly critical of the prosecution's DNA evidence at trial and led to the convictions of Amanda Knox and Raffaele Sollecito being overturned on appeal. Here, I will use "interpretation" for the process of deciding which epg peaks are allelic and "evaluation" for the calculation of numerical measures of evidential weight for an interpretation.

The interpretation by the Italian Scientific Police presented at trial identified exactly the alleles of the victim and one of the coaccused, Sollecito, in the DNA profiling results. Using methods described below, I computed a weight of evidence (WoE) in favor of the contributors of DNA being Kercher and Sollecito, rather than Kercher and an unknown man, of >15 bans. The ban is the unit of WoE introduced by Alan Turing (3): x bans means $\log_{10}(\mathrm{LR}) = x$, where LR is the likelihood ratio, such that 6 bans means an LR of 1 million. In reviewing the evidence, Vecchiotti and Conti (2) agreed with the alleles originally identified but also reported many additional epg peaks. They cited recommendation 6 of Gill et al. (4) in concluding that all peaks in stutter positions should be regarded as allelic. Of the 24 additional peaks identified by Vecchiotti and Conti (2), of which 6 had heights below the threshold of 50 relative fluorescence units, 9 are included in the profile of the other codefendant, Knox, providing apparent support for the presence of DNA from her. However, four of her alleles were not observed, including two homozygotes, which are less prone to dropout.

These interpretations pose problems for standard methods of evidence evaluation because of the alleles not attributable to any of the profiled individuals, uncertainty over whether or not Knox is a contributor, and the need to allow for the possibility that subthreshold peaks may be allelic. The number of above-threshold alleles recorded at any locus is six or less, which implies three

Author contributions: D.J.B. designed research, performed research, analyzed data, and wrote the paper.

The author declares no conflict of interest.

Freely available online through the PNAS open access option.

This article is a PNAS Direct Submission. T.P.S. is a guest editor invited by the Editorial Board.

[1]E-mail: d.balding@ucl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1219739110/-/DCSupplemental.

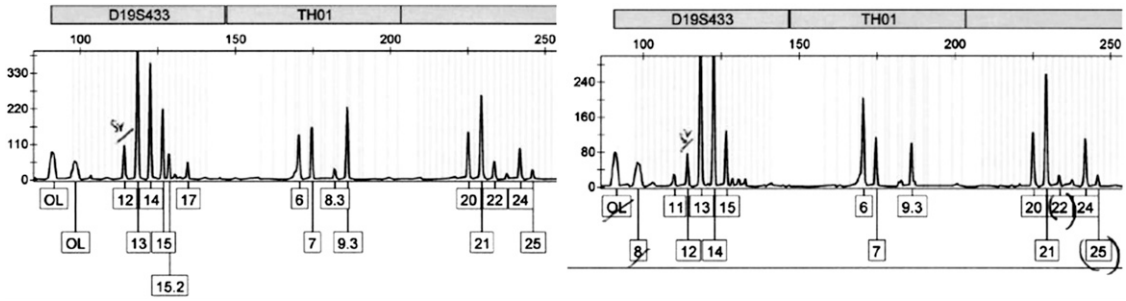

Fig. 1. Illustrative epgs from a swab of a handgun magazine. Two replicates are shown at three loci: D19, THO1, and FGA. Note the different *y*-axis scales, chosen automatically, in units of relative fluorescence units; the *x* axis shows fragment length in base pairs. Allele labels in boxes are assigned automatically but can be overridden by a forensic expert taking into account factors like peak morphology and potential stutter. Some manual annotations are shown, indicating subthreshold peaks in ( ) as well as possible artifacts, such as stutter.

or more contributors of DNA. However, if Knox is assumed to be a contributor, the alleles not attributable to her still imply three or more other contributors. I first compare these prosecution (H_p) and defence (H_d) hypotheses for the contributors of DNA:

H_p: Kercher, Knox, Sollecito, and one unknown individual

H_d: Kercher, Knox, and two unknown individuals

I introduce an innovation to likelihood-based analyses to allow for an "uncertain" allele designation. In previous formulations (5–10), the likelihood at a locus in a profiling run is the product over all allelic positions in the epg of one of four possible terms, according to whether or not the corresponding allele is represented in the crime scene profile (CSP) and whether or not it is included in the profiles of any of the hypothesized contributors (*Materials and Methods*). I introduce here a fifth possibility corresponding to an absence of information about whether the allele is present, irrespective of whether or not it is included in the profile of a hypothesized contributor. An assumption of no information is appropriate if there is substantial uncertainty, for example, due to borderline peak height or the possibility that a peak is due to stutter or other artifact.

Using this uncertain designation for the six subthreshold alleles, the estimated dropout rate for Knox is close to 100%. A separate analysis with her as the queried contributor returned an LR < 1, also favoring a conclusion of no DNA from her. I reran the analysis excluding Knox from both H_p and H_d, and obtained an LR in favor of H_p of 42 million (WoE = 7.6 bans). Thus, although the additional alleles have, by providing evidence for an additional contributor, weakened the evidence implicating Sollecito by a massive 8 bans, this evidence nevertheless remains strong. Moreover, Gill et al. (4) did not consider uncertain designations for peaks that are potentially due to stutter. After reclassifying as uncertain all peaks below 15% of the height at one extra repeat unit, a common stutter guideline (4), there remain four alleles not attributable to either Sollecito or Kercher and the WoE is increased to 10.7 bans.

Note that I cannot address here issues of how the DNA came to be in the exhibit: Possible contamination was an issue in the trial and appeal. I only consider whether there is DNA from Sollecito for which the evidence remains very strong after allowing for the additional alleles identified by Vecchiotti and Conti (2) and the possibility that apparent stutters are allelic.

Table 1. Allele calls at 5 of 15 loci in the DNA profile obtained from exhibit 165B (case of Knox and Sollecito)

Locus	Trial*	Appeal†	New‡
D8	13, 15, 16	11, 12, 13, 14, 15, 16	11, 12, 13, 14, 15, 16
D21	30, 32.2, 33.2	29, 30, 32.2, 33.2	29, 30, 32.2, 33.2
D7	8, 11	8, 10, 11	8, 10, 11
CSF	10, 12	10, 11, 12	10, 11, 12
D3	14, 16, 17, 18	14, 15, 16, 17, 18	14, 15, 16, 17, 18
LR§	7×10^{15} (15.8 bans)	4×10^{7} (7.6 bans)	5×10^{10} (10.7 bans)

*Alleles reported at the original trial.

†Alleles identified by Vecchiotti and Conti (2); underlined alleles have peak heights <50 relative fluorescent units.

‡Apparent stutters are also underlined (peaks with a height <15% of the peak height at one extra repeat unit).

§LR for Sollecito to be a contributor of DNA, given that Kercher is a contributor, based on all 15 loci [*x* bans means $\log_{10}(\mathrm{LR}) = x$].

LikeLTD Software

The probability model used to calculate these LRs is implemented in the likeLTD (likelihoods for LTDNA profiles) software, which computes likelihoods for hypotheses, such as H_p and H_d, that specify the contributors to a sample of DNA, some or all of whom may have contributed low levels of possibly degraded DNA. For mixed-source profiles, epg peak heights are potentially informative beyond simply indicating whether or not an allele is present because they can reflect the amount of DNA, which may differ among contributors. However, this information can be difficult to exploit because peak heights for LTDNA are highly variable and this variability can be sensitive to the details of the profiling protocol used. The data input into likeLTD are the reference profiles, together with the CSP, coded as present/uncertain/absent at each allelic position in each replicate. Peak height information is used by the forensic scientist when deciding these classifications, for example, when assessing whether a peak in a stutter position should be regarded as allelic or uncertain. The full set of present/uncertain/absent indicators, combined over alleles, loci, and replicates, is highly informative about the amount of DNA from different contributors, and hence about dropout probabilities, permitting powerful and robust evaluations of evidential weight without the need to use sensitive peak height information.

In this article, I describe the probability model underlying likeLTD and assess its performance on real and artificial CSPs. I show that likeLTD provides a good solution to the problem of evaluating LTDNA profiles with up to two unprofiled contributors in addition to the queried contributor. The DNA evidence for Knox and Sollecito was criticized by Vecchiotti and Conti (2) because only a single DNA profiling run was performed. For any "noisy" scientific process, replicate analyses are desirable. This is broadly true for LTDNA evidence, but replication does have the potential disadvantage of dividing an already minuscule sample, which may adversely affect the results (11). As long as the noise is adequately modeled, which is possible by combining information over alleles and loci, replication is not a prerequisite for valid evaluation of the evidence. In Table 1, the extra uncertainty due to lack of replication has led to a much lower WoE than might have been realized had replicate PCR assays been successfully analyzed. In other words, a penalty for lack of replication

arises automatically in likelihoods that model stochastic phenomena, such as dropin or dropout.

Results

Hammer Case. The profile data in Table 2, consisting of two CSP replicates and reference profiles from a queried contributor, Q, and two victims, K1 and K2, are taken from Table 2 of a study by Gill et al. (7), which did not consider the possibility of uncertain allele calls. There is some variability across the two replicates, a symptom of low-template and/or degraded DNA, such that 12 alleles are observed in only one of the two replicates. There are a total of 6 alleles, ≤2 per locus and all of them unreplicated, that are not from Q, K1, or K2. This suggests a comparison of the following two hypotheses for the contributors of DNA to the sample:

$$H_p'\text{: Q + K1 + K2 + U1,}$$

$$H_d'\text{ : X + K1 + K2 + U1,}$$

where X and U1 are both unprofiled individuals. The distinction between them is that X is the alternative to Q; thus, the ethnic backgrounds of X and Q, and the degree of relatedness between them, can have important impacts on the WoE, whereas U1 plays the same role under both H_p' and H_d'.

Every CSP allele attributable to K2 could also come from K1 or Q (Table 2); thus, under H_p', there is no evidence for DNA from K2. However, under H_d', the DNA of Q is not present, leaving three CSP alleles that can be attributed to K2 but not to K1. Nevertheless, likeLTD estimates 100% dropout of the alleles of K2 in both replicates under both hypotheses. This is because the three alleles attributable to K2 under H_d' are all replicated, whereas seven other alleles of K2 do not appear at all, indicating very high dropout; thus, likeLTD finds that attribution of the three alleles to K2 is unlikely. Although K2 cannot be excluded from contributing any DNA to the sample, these results indicate that including K2 in the analysis brings no explanatory power and so has a negligible impact on the WoE implicating Q as a contributor.

After removing K2 from H_p' and H_d', the WoE is 10.6 (SD = 0.10). H_p' is favored over H_d' at every locus except D18 (−0.6 bans); the most incriminating locus (2.5 bans) was D19, where Q has two rare alleles that appeared in both CSP replicates.

The WoE of 10.6 bans computed here is stronger than that obtained by Gill et al. (the maximum of the blue solid curve in figure 1 of ref. 7 is just over 9 bans). The extra discrimination power of likeLTD results from its extra flexibility, for example, allowing different dropout rates per replicate and per contributor.

I recoded eight CSP alleles as uncertain and observed differing effects at individual loci, depending, for example, on whether the uncertain allele is in the reference profile of Q. The resulting changes in the computed WoE match intuition; for example:

Locus D16, CSPa, allele 11: This is an allele of Q not shared with K1; thus, changing its status from present to uncertain reduces the WoE, but only slightly, because the allele is called in CSPb. The single-locus WoE decreases from 1.17 to 1.14 bans (Table 2, column 4).

Locus D21, CSPb, alleles 29 and 30: These are both alleles of K1 not shared with Q, and they are not called as alleles in CSPa. Thus, changing the allele call to uncertain has a bigger impact, although still modest. The WoE is increased by just over a deciban because of the reduction in possible genotypes for X.

There are also indirect effects on all loci, because the changes in allele calls have an impact on the support for dropout parameter values. Overall, the evidence is weakened but remains very strong at 9.3 bans (Table 2, bottom row).

Table 2. Hammer case DNA profiles and results from two analyses

Locus	D3	D16	D2	D8	D21
CSPa*	14	11^u, 13	20, 23	11^u, 12	28
	16	14	24, 25	13, 15	31
CSPb	14	11, 13	20, 24	11, 12	28, 29^u, 30^u
	16	14	25	13, 15^u	31, 31.2
Q[†]	**14, 16**	**11, 14**	**24, 25**	**12, 13**	**28, 31**
K1	**16, 16**	**13, 13**	**20, 20**	**11, 15**	*29, 30*
K2	15, 17	12, **13**	18, **25**	**11, 13**	*29, 30*
Other[‡]			*23*		*31.2*
WoE[§], bans					
Mean (SD)	1.23 (0.057)	1.17 (0.033)	0.91 (0.084)	0.88 (0.029)	1.48 (0.14)
unc[¶]	1.10	1.14	0.95	0.94	1.59

*The crime scene DNA sample was profiled in duplicate (CSPa and CSPb). Results from 5 of 10 loci are shown.

[†]The profiles of the two uncontested possible contributors of DNA, K1 and K2, and the queried contributor, Q, are shown using the notations: **replicated alleles**, *unreplicated alleles*, and unobserved alleles.

[‡]CSP alleles not attributable to any of Q, K1, or K2.

[§]WoE for Q to be a contributor, given that K1 and one unprofiled individual are also contributors. The mean 10-locus WoE from 25 likeLTD runs is 10.6 bans (SD = 0.10).

[¶]Mean WoE based on 10 likeLTD runs when eight alleles were reclassified as "uncertain," of which five were at the displayed loci and are indicated with ^{u}. The mean 10-locus WoE is 9.3 bans (SD = 0.15).

Simulated Profiles. Detailed results for a range of tests of likeLTD on DNA profiles subject to a number of modifications, such as artificial dropin and dropout, as well as modifications to the modeling assumptions underlying likeLTD, are provided in *SI Text*. I summarize here the main conclusions.

For a one-contributor, two-replicate CSP with no dropin or dropout (*SI Text*, section S2), likeLTD returns almost exactly the same WoE as the standard match probability formula whether or not dropin is explicitly modeled (because the dropin rate is estimated at zero) and whether X is unrelated to Q or is a brother of Q. If I wrongly hypothesize two contributors rather than one, the dropout rate for the additional contributor is estimated at 100% and the WoE is unaffected. If the CSP is modified, the WoE at individual loci changes in line with expectations and the overall WoE is reduced. For a CSP affected by four dropouts and two dropins over the two replicates, repeat likeLTD runs with a search length (n) of 1,000 simulated annealing iterations (*Materials and Methods, Parameter Estimation*) give WoE values with a SD less than half of a deciban (about 11% on the LR scale). For n = 5,000 and n = 10,000, the WoE is precise to <1% on the LR scale. When the alleged contributor Q was chosen at random, such that the prosecution hypothesis was false, the dropout rate for the noncontributor Q was estimated to be very high and, consequently, the WoE was usually negative and always low.

Proceeding to a two-contributor CSP (*SI Text*, section S3), with neither contributor known, a series of experiments introducing 50% dropout to one or both contributors, as well as uncertain allele calls due to stutter, gave satisfactory results in that parameter estimates and the WoE varied in accord with intuition. With three unknown contributors (*SI Text*, section S4), the larger number of parameters implies less precision in evaluating the WoE. With n = 1,000 simulated annealing iterations, the overall WoE has an SD of 0.4 bans (a factor of 2.5 on the LR scale), which is reduced to 0.24 bans (a factor of 1.7 on the LR scale) and 0.03 bans (about 6%) for n = 5,000 and n = 10,000, respectively. Taking a different three-contributor CSP, this time with one contributor a known and profiled individual, I investigated (*SI Text*, section S5) high and low extremes for the degradation parameters, the variance of the locus-specific parameters, and the dropout model power parameter. The WoE was relatively stable under these extreme perturbations, varying

in the range of 10.3–10.7 bans, compared with a standard analysis WoE of 10.6 bans.

Discussion

There is no "gold standard" test of an LR calculation for LTDNA profiles. Likelihoods reflect uncertainty, and even when the profiles of the true contributors are known in an artificial simulation, this does not tell us what is the appropriate level of uncertainty justified by a given observation affected by stochastic phenomena. Likelihoods depend on modeling assumptions, and there can be no "true" statistical model for a phenomenon as complex as an LTDNA profile.

I have shown here good performance of likeLTD in analyzing a wide range of crime scene DNA profiles involving complex mixtures, uncertain allele designations, dropin and dropout, degradation, stutter, and relatedness of alternative contributors. It behaves consistently over replicate analyses and agrees with well-established formulas in simple settings. The parameter estimates and WoE change in a coherent and interpretable manner under artificial modifications of the CSPs, and they are robust to major modifications of the modeling assumptions. For $n = 5{,}000$ iterations of the simulated annealing algorithm, the reported WoE values are reasonably precise when the hypotheses involve both U1 and U2 (SD of about 0.25 bans) and very precise when U2 is not required.

The analysis of LTDNA profiles embodied in likeLTD has elements in common with existing methods (6–10, 12, 13). It goes beyond these methods by eliminating nuisance parameters automatically via maximization of penalized likelihoods, avoiding the use of external calibration data specific to the profiling protocol used, as required by other methods (9). Even with extensive calibration data, estimation of dropout and dropin rates for the specific conditions of a crime sample cannot be precise, but precise estimates are not required: "Best fit" (in the sense of maximum penalized likelihood) values under each of the competing hypotheses provide a fair evaluation of the WoE. To achieve this, likeLTD adopts a multidose dropout model (14–16) that uses information across all replicates, loci, and contributors. The model underlying likeLTD is highly flexible, allowing both amounts of DNA and level of degradation to vary over contributors, as well as locus- and replicate-specific dropout rates. In particular, the contribution of DNA from different individuals is estimated and can be zero, such that additional profiled or unprofiled contributors can be proposed with little error arising if, in fact, there is no DNA from those individuals.

As well as providing strong WoE in favor of true contributors in simulation experiments, I showed in examples that likeLTD identified no support for the presence of DNA even when there superficially appeared to be some support and that the WoE declined appropriately as dropins and dropouts were introduced or allele calls were altered to uncertain.

The problem of how to make a numerical expression of the WoE meaningful to judges or jurors is common to all evaluations of complex DNA evidence. The problem is not insurmountable, and illustrative examples can be helpful (17).

An early "consensus" method approach to the analysis of LTDNA profiles took account only of alleles that appear in both of two DNA profiling replicates (12). This method is often claimed to be conservative, but this is not necessarily the case because it allows alleles that are inconvenient for the prosecution case to be "swept under the carpet." The analysis proposed here makes use of all the results in every DNA profiling run. The consensus method served a useful purpose when few alternative approaches for the analysis of LTDNA profiles were available, but it is no longer best practice.

Methods of analysis that directly use epg peak height information have been developed (18, 19), but software is not currently freely available. These have potential advantages over the method proposed here, in which peak heights are used to classify every allele as present/uncertain/absent in each replicate. However, peak heights can be highly variable, and their statistical properties can depend sensitively on details of the experimental protocol. Thus, our freely available R code likeLTD may remain useful as a robust and efficient approach to the analysis of LTDNA profiles, even if peak height-based methods can be more statistically efficient, given relevant calibration data. Previous versions of likeLTD have already been used in many criminal investigations, with results presented as evidence in UK and US courts (20).

Materials and Methods

Consider a single crime stain that may have been profiled multiple times from replicate PCR assays of the sample. Forensic DNA profiling predominantly assays autosomal short tandem repeat (STR) loci, using technology in which an allele in the profiled sample is represented by a peak in an epg (5), such as those shown in Fig. 1.

I assume that a reference profile is available for a queried contributor (Q) and the goal is to evaluate the LR for two competing hypotheses, one including Q as a contributor (the "prosecution hypothesis," H_p), whereas the "defence hypothesis," H_d, has an unprofiled individual X in place of Q. Both hypotheses may include additional unprofiled contributors [in practice, I can handle up to two (U1 and U2)], as well as profiled possible contributors, for example, the victim or a bystander (K1, K2,...). The contribution of DNA from each proposed contributor is estimated, and this estimate can be zero, such that including an individual in H_p or H_d does not imply that the individual contributed DNA to the sample.

The LR can depend on, for example, the assumed ethnicity of X and his/her relatedness to Q (the more genetically similar X is to Q, the smaller is the LR). The likeLTD software program allows close relatedness of X to Q, specified with two relatedness coefficients, whereas all other hypothesized contributors must be mutually unrelated and unrelated to X and Q. In addition, remote shared ancestry ("coancestry") of X with Q is modeled using the population genetics parameter F_{ST} (17). Typically, in US forensic practice, F_{ST} (also called θ) is only used to model intraindividual genetic correlations (i.e., excess homozygosity) (9). However, intraindividual correlations are of little relevance to evidential weight. Only between-individual correlations matter in practice, and failing to model them gives WoE values that are biased against defendants. This deficiency affects some alternative methods for analyzing DNA profiles. Because the relatedness coefficients and F_{ST} account for the positive correlations across loci due to shared ancestry of X and Q, it is reasonable to compute full-profile LRs by multiplication of single-locus LRs, which is standard practice in the assessment of DNA profile evidence (5). I thus focus below on the single-locus case.

Unless otherwise stated, all analyses reported here use $n = 5{,}000$ iterations of the simulated annealing algorithm within likeLTD. The allele frequencies from a standard database of ~200 UK Caucasians have undergone sampling and F_{ST} adjustments as described in *SI Text*, section S2.

Single-Locus LR with Dropout. Consider first a single profiling run, with a single contributor who is Q under H_p and X under H_d. If Q ≡ AB, where "≡" denotes "has genotype," but the CSP shows only A and low epg peak heights suggest that dropout is possible, then the possibility that B has dropped out must be considered. Under a standard model (8, 10), the LR can be written as

$$\frac{D(1-D)}{p_A^2(1-D_2)+2p_A(1-p_A)D(1-D)}, \quad [1]$$

where D and D_2 denote the probabilities of dropout for heterozygote and homozygote alleles, respectively. The numerator is the probability that the B allele of Q has dropped out (D), whereas the A allele has not ($1 - D$). In the denominator, either X is AA and there has been no dropout (first term) or (second term) X is heterozygous but the non-A allele has dropped out. Logically, D in the numerator of the LR is different from D in the denominator; however, typically similar values are supported under both hypotheses, and they are often taken to be equal for illustrative calculations (7).

Effect of an Uncertain Allele Designation. If I now assume CSP = A[B], where [] denotes an uncertain allele designation, and, again, Q ≡ AB, the LR becomes

$$\frac{1-D}{p_A^2(1-D_2)+2p_Ap_B(1-D)+2p_A(1-p_A-p_B)D(1-D)}. \quad [2]$$

In the numerator, I know that Q's A allele has not dropped out ($1 - D$) but not whether the B allele has dropped out. In the denominator, the three

terms correspond to X ≡ AA, AB, and AZ, respectively, where Z is any allele other than A or B.

Fig. 2 (solid curves) shows LRs as functions of D for a locus with $p_A = p_B = 0.1$ (after adjustment). As expected, the LR for CSP = A[B] (Fig. 2, red curve) is always intermediate between those for CSP = AB (Fig. 2, black) and CSP = A (Fig. 2, green). When D is high, the red and green curves in Fig. 2 are similar because in the presence of high dropout, both an uncertain designation and an absent designation for B convey little information about whether or not X has a B allele. However, when D is small, the two LRs differ substantially because CSP = A is inconsistent with X ≡ AB, whereas CSP = A [B] is consistent with both X ≡ AA and X ≡ AB.

Next, consider the LRs when there is a second replicate that gives A[B] in each case (Fig. 2, dashed curves). I assume the same D for both replicates. When CSP = AB + A[B], I must have X ≡ AB (I ignore dropin here, as discussed below). When CSP = A + A[B], the LR is

$$\frac{D(1-D)^2}{p_A^2(1-D_2)^2 + 2p_A p_B D(1-D)^2 + 2p_A(1-p_A-p_B)D^2(1-D)^2},$$

whereas for CSP = A[B] + A[B], it is

$$\frac{(1-D)^2}{p_A^2(1-D_2)^2 + 2p_A p_B (1-D)^2 + 2p_A(1-p_A-p_B)D^2(1-D)^2}.$$

Note that I assume the different replicates are independent, conditional on the genotypes of all contributors (6).

I see from Fig. 2 that observing A[B] in the second replicate increases both LRs when D is small but decreases them when D is large. In fact, when D is very high, observing either A or A[B] in just one replicate yields LR > 1, favoring H_p, whereas observing two such observations in independent replicates gives LR < 1, against H_p. This is because X ≡ AA under H_d then provides a better explanation of the replicate observations than H_p, because homozygotes are much less likely to drop out than a heterozygote allele.

Additional Contributors. LRs, such as those in Eqs. **1** and **2**, can be rewritten more generally as

$$\mathrm{LR} = \frac{P(\mathrm{CSP}|\mathrm{Q} \equiv \mathrm{AB})}{\sum_{g\in\Gamma} p_g P(\mathrm{CSP}|\mathrm{X} \equiv g)}, \qquad [3]$$

where Γ denotes the set of possible genotypes and p_g denotes the population fraction of genotype g. Eq. **3** makes explicit the requirement to sum over all possible genotypes for the unprofiled contributor X. When there is an additional unprofiled contributor U1, it is necessary to sum over all possibilities for the unknown genotypes, multiplying each term by the genotype probability:

$$\mathrm{LR} = \frac{\sum_{g\in\Gamma} p_g P(\mathrm{CSP}|\mathrm{Q} \equiv \mathrm{AB}, \mathrm{U1} \equiv g)}{\sum_{g1,g2\in\Gamma} p_{g1} p_{g2} P(\mathrm{CSP}|\mathrm{X} \equiv g1, \mathrm{U1} \equiv g2)}. \qquad [4]$$

Each term in these sums follows the same well-established rules used for Q and X above, now applied additionally to the current genotype for U1.

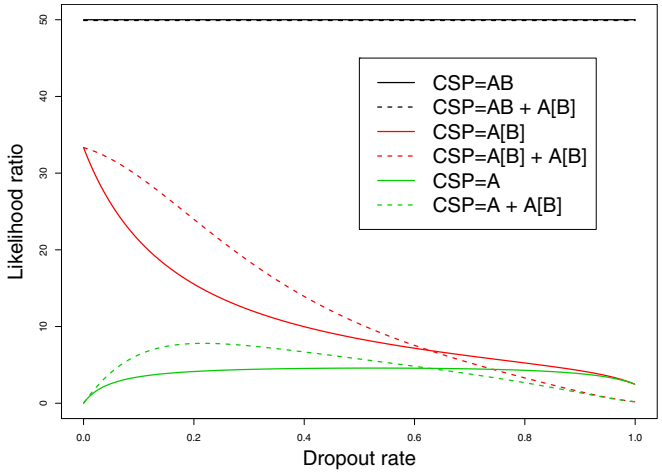

Fig. 2. Single-locus, single-contributor LRs for three CSPs with one replicate (solid curves) and three CSPs with two replicates (dashed curves). The LRs are expressed as functions of the dropout rate D, assumed to be the same for all alleles in both the numerator and denominator. In the legend box, "+" separates the two replicates and [ ] denotes an uncertain allele call. Allele A is observed in every replicate; the designation of allele B is uncertain in the second replicate, whereas it varies over present, uncertain, and absent in the first replicate.

Multidose Dropout Model. Individuals contribute different amounts of DNA to a mixed-source sample, and multiple individuals can have one or two copies of a given allele. Thus, given $D(1)$, the dropout probability for a unit "dose" of DNA, it is necessary to evaluate $D(k)$, the dropout probability for dose k of DNA. I adopt the model of Tvedebrink et al. (14), which can be written as

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

where s indicates the locus. I choose the scale by fixing the mean over loci of α_s at 1. I take $k = 1$ to correspond to a single heterozygote allele of a reference individual, usually X or Q.

The estimates obtained by Tvedebrink et al. (14) from experimental nondegraded LTDNA profiled at the 10 loci of the SGM+ system imply an SD for α_s of 0.141. Because they may depend sensitively on the experimental protocol used, I do not use the estimates of Tvedebrink et al. (14) directly; instead, I estimate the α_s under each hypothesis from the observed CSP. To keep the estimates realistic, I impose a γ-distribution prior on the α_s, with mean = 1 and SD = 0.141 (a different SD may be appropriate, for example, in highly degraded samples). For the global parameter β, I adopt here the estimate $\hat{\beta} = -4.35$ (14). Fig. 3 illustrates $D(k)$ as a function of $D(1)$ for several values of k, evaluated by substituting $\alpha_s^{\beta} = D(1)/(1-D(1))$ in Eq. **5**. Note, for example, that if $D(1) = 0.5$, $D(1.2) \approx 0.3$ and $D(0.8) \approx 0.7$; thus, a 20% change in DNA dose can have a large impact on dropout probabilities.

The problem of calculating likelihoods for LTDNA profiles was not addressed by Tvedebrink et al. (14); they validated their model by comparing theoretical and empirical dropout rates. To achieve this, they estimated the amount of DNA from each contributor using the heights of peaks due only to that contributor over the whole profile. This is problematic for calculating LTDNA likelihoods because it ignores information from allele peaks with multiple contributors and requires alleles of individual contributors to be distinguished, which is frequently not possible. Here, I directly specify the likelihood for each replicate in terms of $D(k)$ at every allelic position, with k calculated according to the contributions of DNA and the genotypes of all the hypothesized contributors. I thus use present/uncertain/absent information at every allelic position to provide information about amounts of DNA, with the contributions from different contributors being estimated by maximum likelihood.

Degradation Model. DNA degrades over time at a rate that depends on temperature, humidity, and environmental exposure. In forensic DNA profiling, degradation is manifested as higher dropout rates for alleles with large fragment lengths. Our model for the effect of degradation is based on that of Tvedebrink et al. (15), who posited a geometrical distribution for the effective amount of DNA as a function of allele fragment length. Thus, the average allele dose k from the ith contributor subject to dropout is modified at an allele with fragment length l base pairs (centered to have mean zero) according to $k' = k(1+\gamma_i)^{-l}$, where $\gamma_i > 0$. Shorter fragments ($l < 0$) correspond, in effect, to an enhanced allele dose, whereas longer fragments generate a smaller effective allele dose. An STR allele consists of flanking regions in addition to the tandem repeats; thus, the repeat number that characterizes the allele is not a good proxy for fragment length, which can be obtained for many DNA profiling systems at www.cstl.nist.gov/div831/strbase/.

In the spirit of shrinkage regression methods, likeLTD incorporates a weak penalty (exponential, mean = 0.02) on each γ_i. The effect of this penalty is a slight tendency to shrink the parameter estimates toward zero, which is usually negligible but avoids inflated values when there is very little information.

Dropin. Dropin refers to an allele in the CSP that is not included in the genotype of any hypothesized contributor, profiled or unprofiled. Dropin alleles can arise from individuals contributing a very low level of DNA to the sample, for example, via environmental contamination either in the laboratory

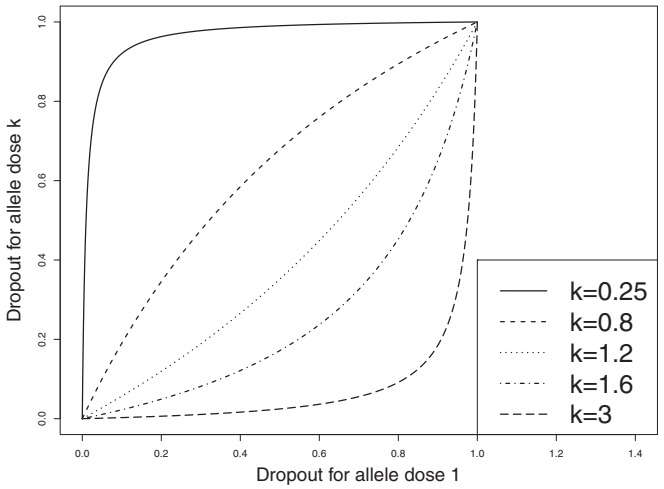

Fig. 3. Dropout probabilities for dose k of DNA (y axis) against those for a unit dose (x axis). The values of k are shown in the legend box.

or at the scene of recovery of the item. They can be generated from tiny fragments of the DNA molecule that persist for some time after the death and decay of a cell. Forensic scientists often restrict "dropin" to laboratory-based alleles, the rate of which can be measured by control runs and is usually found to be low. However, I cannot usually distinguish laboratory-based dropin from alleles arising at the crime scene (12).

Each dropin allele does come from an individual, but it may arise from very few and possibly degraded cells. It is computationally inefficient to sum over all possible genotypes, as in Eq. **4**, for such low-level contributors; thus, I allow the possibility of modeling dropin more simply as independent Bernoulli trials (6). Dropin is nondropout of an allele of a low-level contributor; thus, I model the dropin probability as a constant (c) times the nondropout rate for each replicate. As for the γ_i, I impose a weak penalty on c (exponential, mean = 0.5) to discourage solutions with a large c, reflecting background information that dropin is usually rare.

Parameter Estimation. To compute the LR, it is necessary to deal with the "nuisance" parameters under each hypothesis. These are the $D(1)$ (one per replicate), the α_s (one per locus), possibly a dropin parameter c (see above), the contributions of DNA relative to the reference individual, and the γ_i (one for each contributor subject to dropout). The likeLTD program seeks to maximize a penalized likelihood over these parameters, with penalties on α_s, γ_i, and c as described above. The penalties can be thought of as prior distributions, but I do not use a Bayesian approach because I maximize over unknown parameters rather than integrate. The primary purpose of the penalty is to discourage the maximization algorithm from exploring unrealistic regions of the parameter space.

I use a simulated annealing algorithm (21) to maximize in an approximate manner the penalized likelihood L. Starting with L computed at any set of parameter values, the algorithm repeatedly takes a random step in parameter space; compute the penalized likelihood L'; and accept the new state with probability $\exp((L'-L)/Lt)$, where t is the temperature, computed here as $t=(1-i/n)^3$, with i and n being the current and total numbers of iterations. Our goal is to obtain the maximized L under each hypothesis: $\widehat{L_p}$ and $\widehat{L_d}$. Estimates of the nuisance parameters are available as a byproduct. They may not be precise, because there are some regions of the space of nuisance parameters over which L varies little, particularly when both U1 and U2 are included in the hypotheses being compared, because their genotypes cannot easily be distinguished. However, the assessment of evidential weight uses only $\widehat{L_p}/\widehat{L_d}$ and does not require precise estimates of the nuisance parameters.

Simulated annealing is a well-established algorithm with good properties, but there is no guarantee that it will find the exact maximum value of L. A larger n generally produces better approximations to the maximum; however, beyond a certain value, the improvement may be negligible. In *SI Text*, section S2 and S3, I show that $n = 5{,}000$ provides good precision for hypotheses involving both U1 and U2, as well as excellent precision when U2 is not required.

Computing Times. Using $n = 5{,}000$, likeLTD requires a few minutes if neither U1 nor U2 is included in the hypotheses being compared, a few hours for U1 only, and several days for both U1 and U2. Other parameters affecting computing times include the number of replicates and whether dropin is modeled. The runs of likeLTD for the Hammer case and the Knox and Sollecito case reported here required just over 10 min per locus on a standard desktop machine. A much faster implementation of the algorithm is under development.

ACKNOWLEDGMENTS. I thank Ben Lanham of Cellmark Forensic Services for providing Fig. 1 and Professor Carla Vecchiotti of the Sapienza Università di Roma for providing Fig. S1, which is a higher quality version of a figure in ref. 2. I gratefully acknowledge help with computations from Chris Steele and Adrian Timpson, both of University College London, and I also thank Dr. Torben Tvedebrink of Aalborg University and Dr. Norah Rudin, a forensic DNA consultant from California, for helpful comments on a draft of the manuscript.

1. Caddy B, Taylor G, Linacre A (2008) *A Review of the Science of Low Template DNA Analysis* (UK Home Office, London).
2. Vecchiotti C, Conti S (2011) [DNA evidence in the case against Amanda Knox and Raffaele Sollecito. Corte di Assise di Appello di Perugia], trans komponisto (English translation available at knoxdnareport.wordpress.com, accessed November 12, 2012).
3. Good I (1979) Studies in the history of probability and statistics. XXXVII AM Turing's statistical work in World War II. *Biometrika* 66(2):393–396.
4. Gill P, et al.; DNA commission of the International Society of Forensic Genetics (2006) DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Sci Int* 160(2-3):90–101.
5. Buckleton J, Triggs C, Walsh S (2004) *DNA Evidence* (CRC, Boca Raton, FL).
6. Curran JM, Gill P, Bill MR (2005) Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure. *Forensic Sci Int* 148(1):47–53.
7. Gill P, Kirkham A, Curran J (2007) LoComatioN: A software tool for the analysis of low copy number DNA profiles. *Forensic Sci Int* 166(2-3):128–138.
8. Balding DJ, Buckleton J (2009) Interpreting low template DNA profiles. *Forensic Sci Int Genet* 4(1):1–10.
9. Mitchell AA, et al. (2012) Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. *Forensic Sci Int Genet* 6(6):749–761.
10. Gill P, 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(6): 679–688.
11. Grisedale KS, van Daal A (2012) Comparison of STR profiling from low template DNA extracts with and without the consensus profiling method. *Investig Genet* 3(1):14.
12. 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(1):17–40.
13. Haned H (2011) Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics. *Forensic Sci Int Genet* 5(4):265–268.
14. 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(4): 222–226.
15. Tvedebrink T, Eriksen PS, Mogensen HS, Morling N (2012) Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Sci Int Genet* 6(1):97–101.
16. Tvedebrink T, Eriksen PS, Asplund M, Mogensen HS, Morling N (2012) Allelic drop-out probabilities estimated by logistic regression—Further considerations and practical implementation. *Forensic Sci Int Genet* 6(2):263–267.
17. Balding DJ (2005) *Weight of Evidence for Forensic DNA Profiles* (Wiley, New York).
18. Perlin MW, et al. (2011) Validating TrueAllele® DNA mixture interpretation. *J Forensic Sci* 56(6):1430–1447.
19. Cowell RG, Lauritzen SL, Mortera J (2011) Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Sci Int Genet* 5(3):202–209.
20. Lohmueller KE, Rudin N (2013) Calculating the weight of evidence in low-template forensic DNA casework. *J Forensic Sci* 58:(Suppl 1):S243–S249.
21. Kirkpatrick S, Gelatt CD, Jr., Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680.