DE GRUYTER Stat. Appl. Genet. Mol. Biol. 2016; 15(5): 431–445

Christopher D. Steele*, Matthew Greenhalgh and David J. Balding

Evaluation of low-template DNA profiles using peak heights

DOI 10.1515/sagmb-2016-0038

Abstract: In recent years statistical models for the analysis of complex (low-template and/or mixed) DNA profiles have moved from using only presence/absence information about allelic peaks in an electropherogram, to quantitative use of peak heights. This is challenging because peak heights are very variable and affected by a number of factors. We present a new peak-height model with important novel features, including over- and double-stutter, and a new approach to dropin. Our model is incorporated in open-source `R` code `likeLTD`. We apply it to 108 laboratory-generated crime-scene profiles and demonstrate techniques of model validation that are novel in the field. We use the results to explore the benefits of modeling peak heights, finding that it is not always advantageous, and to assess the merits of pre-extraction replication. We also introduce an approximation that can reduce computational complexity when there are multiple low-level contributors who are not of interest to the investigation, and we present a simple approximate adjustment for linkage between loci, making it possible to accommodate linkage when evaluating complex DNA profiles.

Keywords: DNA mixtures; forensic; likelihood ratio; `likeLTD`; low-template DNA; peak heights.

1 Introduction

The computation of likelihood ratios (LRs) for complex forensic DNA evidence has progressed in recent years from using only presence/absence of alleles inferred from an electropherogram (epg), (Gill et al., 2000, 2008, 2012; Balding and Buckleton, 2009; Balding, 2013) to the use of quantitative peak heights (Perlin et al., 2011; Bright et al., 2013b; Puch-Solis et al., 2013; Graversen and Lauritzen, 2014; Cowell et al., 2015; Bleka et al., 2016). The LR approach to evaluating weight of evidence has long been preferred for standard DNA profiles (Gill et al., 2006, 2012), and for complex profiles there appears to be no realistic alternative. It takes the form:

$$\mathrm{LR}=\frac{\Pr(E|H_p)}{\Pr(E|H_d)}, \tag{1}$$

where E is the DNA evidence, consisting of an epg representing the crime scene profile (CSP) and the reference profiles of at least one possible contributor, while H_p is a hypothesis corresponding to the prosecution case that is contrasted with a defense hypothesis H_d. H_p includes a profiled individual, Q, as a contributor of DNA to the CSP. H_d is often the same as H_p except that Q is replaced by an unprofiled individual. If there are multiple queried contributors then a series of LRs can be computed each contrasting a queried contributor with an unprofiled alternative.

If Q is a contributor of DNA to the CSP then peaks are expected in the epg corresponding to the alleles in the reference profile of Q. However, if Q is a low-template contributor peaks can be sub-threshold or absent

***Corresponding author: Christopher D. Steele,** University College London – UGI, Darwin Building Gower Street, London WC1E 6BT, United Kingdom of Great Britain and Northern Ireland, e-mail: c.steele.11@ucl.ac.uk.
http://orcid.org/0000-0002-6110-0086
Matthew Greenhalgh: Orchid Cellmark Ltd., Abingdon Business Park, Blacklands Way, Abingdon OX14 1YX, UK
David J. Balding: University of Melbourne – Centre for Systems Genomics, School of BioSciences and School of Mathematics and Statistics, Melbourne, Victoria, Australia

for some alleles, which is known as dropout. For mixed CSPs, contributors may share alleles making it difficult to evaluate evidence for the presence of DNA from Q. Interpretation is further complicated by experimental artifacts such as stutter and dropin (see below). Peak height information can help reduce the impact of these issues. For example, dropout is only plausible if the heights of the observed peaks indicate low DNA mass from that contributor. Further, consider a CSP with peak heights 80, 790, 640 and 90 at alleles 13, 14, 15 and 16, respectively. The peak heights support a major contributor with genotype 14, 15. They also indicate that the 13 allele may be partly or entirely due to stutter from the 14 peak, and statistical modeling can generate probabilities for a minor contributor genotype to be either 13, 16 or 14, 16, with some other possibilities also having non-zero probabilities, such as 16, 16 or 16, F, where F denotes a dropped-out allele.

While there are multiple models and software now available for computing LRs using peak heights, our new model has important features not currently available, as well as modeling choices that differ from other programs (see Table 1 for a summary). Moreover our `likeLTD` software is open-source and easily accessible from the comprehensive `R` archive network (CRAN). Because of the importance of DNA profile analysis to society and the lack of a definitive test of validity, it is important to have alternative models available for study and comparison by researchers and practitioners.

A full comparison of the available models is beyond the scope of this article, but we highlight here some important distinctions. Stutter models range in complexity from a constant stutter fraction across the whole epg, through models that have a locus-specific linear relationship between stutter rate and the longest uninterrupted sequence (Brookes et al., 2012; Bright et al., 2013b; Kelly et al., 2014), to models that account for multiple uninterrupted sequences (MUS; Taylor et al., 2016). `likeLTD` uses the middle approach, but fixes the intercept to zero, which we found to improve performance by reducing the number of parameters requiring estimation. Moreover `likeLTD` appears to be unique in modeling double-stutter. In addition, `likeLTD` has a more realistic dropin model: dropin is modeled as a contribution to expected peak height at every allele, in proportion to the population allele fraction. An important difference between models is the choice of probability distribution for peak heights: most models employ a gamma distribution, whereas `STRmix` adopts the lognormal and `TrueAllele` a truncated normal distribution. Some models do not incorporate the effects of DNA degradation on peak heights. All models that do include degradation, `likeLTD` among them, assume an exponential decline of expected peak height with allele fragment length. Lastly, `likeLTD` and `EuroForMix` are the only fully open-source software.

We validate the `likeLTD` peak-height model using 108 laboratory-generated mixtures. We show that it behaves as predicted by theory in relation to probability intervals for peak heights, inference of contributor genotypes and with additional replicates (Steele et al., 2014a).

Table 1: Summary of current software for evaluation of complex DNA profiles using peak heights.

Program	Peak height dist.	Param. elim.	Stutter model	Dropin model	*D*	*O*	Deg model	Open source
`DNAmixtures`[1]	Γ	Max.	Constant	Extra *U*	×	×	×	Partial
`EuroForMix`[2]	Γ	Both	Constant	exp(dropin PH)	×	×	✓	✓
`LiRa`[3]	Γ	Max.	Linear (bp)	Γ (dropin PH)	×	×	×	×
`likeLTD`	Γ	Max.	Linear (LUS)	Dropin dose	✓	✓	✓	✓
`STRmix`[4]	log *N*	Int.	MUS	exp(dropin PH)	×	✓	✓	×
`TrueAllele`[5]	$\mathcal{N}$	Int.	✓	✓	?	?	✓	×

Distributions: Γ, gamma; log *N*, lognormal; $\mathcal{N}$, truncated normal. Parameter elimination methods: maximization (Max.) or integration (Int.). Stutter models: the expected fraction of parent peak height lost to stutter is either constant, linear or varies with all uninterrupted sequences in the amplicon (MUS), in the middle case the linearity is either with the length of an allele in base pairs (bp), or with longest uninterrupted sequence (LUS). Dropin can be modeled as an extra unknown contributor (*U*), or the dropin peak heights (PH) have an exponential (exp) or gamma (Γ) distribution, or a dropin dose is added to every allelic position. *D*, Double-stutter; *O*, over-stutter; Deg, degradation. `DNAmixtures` is partly open source but requires the commercial software `HUGIN`. For `TrueAllele` ticks indicate that the phenomenon is modeled but details are unknown, while question marks indicate that we are not aware if the phenomenon is modeled. For all other models ticks and crosses indicate that the phenomenon is or is not modeled. [1]Graversen and Lauritzen (2014), [2]Bleka et al. (2016), [3]Puch-Solis et al. (2013), [4]Bright et al. (2013b), [5]Perlin et al. (2011).

Replication is often viewed as a cornerstone of the scientific method, and if it can be performed without cost it is clearly desirable, for example to guard against failure of a profiling run. DNA extraction protocols typically produce a fixed volume which exceeds that required for PCR, so that post-extraction replication is available "for free." Some protocols may not give this free replication, such as purifying a low-concentration extract through dialysis (Williams et al., 1994), filtering through a spin column (McCord et al., 1993; Ruiz-Martinez et al., 1998), or alcohol/salt precipitation (Nathakarnkitkool et al., 1992). If replication is achieved at the cost of splitting an already low quantity of DNA, for example prior to DNA extraction, then its merits are less clear. Although each replicate profile will be of lower quality than a single profile that uses all the available DNA, statistical analysis that combines information across replicates can recover information lost in individual replicates, and possibly exploit additional information because the replicate samples will have (slightly) different ratios of DNA mass from different contributors, leading to better overall discrimination of their alleles (Steele et al., 2014a). Here we simulate pre-extraction replication by splitting DNA samples with x pg DNA into n samples with x/n pg DNA each, in order to assess its merits when analysis is performed using a statistically-efficient peak-height model.

We investigate reducing the computational complexity of likelihood calculations by modeling an unknown minor contributor as dropin, thus reducing the number of genotypes that must be inferred.

We present a simple adjustment to the LR that accounts for linkage between loci when X is assumed closely related to Q, excluding parent-offspring relationships. This has become a concern with the adoption of STR typing kits with multiple loci on a single chromosome.

The LR will be reported here in terms of the Information Gain Ratio (IGR=$\log_{10}$ (LR)/$\log_{10}$ (IMP)). IGR allows for easy comparison of LRs across different Q, as max(IGR)=1.0 for every Q.

2 Materials and methods

2.1 The `likeLTD` peak-height model

Computations are performed separately under H_p and H_d. Let C denote the set of contributors under a given hypothesis. Suppose that the CSP replicates are indexed by the elements of a set R, and include loci in the set L, while I_l denotes the set of possible alleles at locus $l \in L$. Each element of G_l is an allocation of genotypes at locus l to each $c \in C$. The genotype of Q is constant over G_l, and similarly for other c with reference profile available, but the elements of G_l vary according to the genotypes allocated to unprofiled c. Population genotype probabilities are assumed given. In practice, allele probabilities are obtained from a database, possibly using a sampling adjustment, and genotype probabilities are derived as products of allele probabilities assuming Hardy Weinberg equilibrium, possibly with an F_{ST} adjustment (Balding and Steele, 2015).

Let χ_c denote the effective DNA mass at a heterozygote allele of $c \in C$ in the first replicate, expressed in RFU, a unit of peak height. To compute the expected contribution from c to the height of an epg peak at allele $i \in I_l$ for a given $g \in G_l$, we first adjust for the genotype of c specified by g, the replicate $r \in R$, and DNA degradation:

$$P_{l,r,g,c,i} = \frac{n_{g,c,i}\rho_r\chi_c}{(1+\delta_c)^{f_i}}, \tag{2}$$

where $n_{g,c,i} \in \{0, 1, 2\}$ indicates the number of i alleles in the genotype of c and ρ_r denotes a replicate adjustment (ρ_1=1), while δ_c is a parameter measuring the degradation of DNA from c and f_i is the mean adjusted length of allele i in base pairs. Each $P_{l,r,g,c,i}$ must next be adjusted for the fractions that stutter to allelic position i–1 (S), double-stutter to i–2 (D) or over-stutter to i+1 (O). Whereas D and O are global constants, because these are rare events and it would be difficult to parametrise the relationship, we propose a zero-intercept linear model for S:

$$S_{l,i} = \alpha_l u_i.$$

Here, α_l is the locus-specific coefficient of u_i, the longest uninterrupted sequence (LUS) of allele i (Brookes et al., 2012; Bright et al., 2013b; Kelly et al., 2014). To compute the expected peak height at allele i in replicate r for a given g, each $P_{l,r,g,c,i}$ is incremented with any stutter contribution from allele i+1, double stutter from i+2 and over-stutter from i–1, and summed over contributors c. Finally, a contribution from dropin is added. This gives the expected peak height as:

$$E_{l,r,g,i}=\frac{\lambda p_i}{(1+\delta)^{f_i}}+\sum_{c\in C}(OP_{l,r,g,c,i-1}+(1-S_{l,i}-D-O)P_{l,r,g,c,i}+S_{l,i+1}P_{l,r,g,c,i+1}+DP_{l,r,g,c,i+2}), \tag{3}$$

where p_i is the population allele fraction and λ is a dropin parameter, in RFU. Note that dropin of an allele is assumed to occur in proportion to its population frequency, and is adjusted for degradation with a dropin-specific rate δ.

The peak height at allelic position i is then assumed to have a gamma distribution with expectation $E_{l,r,g,i}$ and variance $\sigma E_{l,r,g,i}$. The scale parameter σ is a global constant, so that values of l, r, g and i affect peak-height variance only through the mean. In `likeLTD` we treat peak heights as discrete: observed values are recorded to the nearest integer RFU value, say j, and we compute the corresponding probability as the gamma probability mass between j–0.5 and j+0.5. The dropout probability is the gamma probability mass assigned to the interval (0, t_l–0.5), where t_l is the detection threshold (the smallest recordable peak height).

In `likeLTD`, alleles that are not observed in any CSP replicate or any reference profile of an assumed contributor are combined into a single allelic class. When the unprofiled contributors are assigned >1 allele in this class, these are assumed to be distinct: unprofiled contributors are assumed not to share any unobserved allele.

In order to encourage the optimisation algorithm to search in realistic regions of the parameter space, the penalty terms shown in Table 2 are imposed. Large values of δ and σ are penalized, while for both D and O a zero value is excluded but a broad range of positive values is supported. Two separate penalties on the α_l are intended to allow flexibility for its mean while limiting its variance over loci. Incorporation of these penalty terms into the likelihood function is analogous to imposing a prior distribution, but our approach is not Bayesian: elimination of nuisance parameters is achieved via maximization and not integration, which is for example the approach adopted by `STRmix`, implemented using Markov chain Monte Carlo.

The probability assigned to allelic position i, whether or not there is an observed above-threshold peak, is computed as a gamma probability mass as described above. Denoting this probability $a(l, r, g, i, \sigma)$, the penalized likelihood is computed by multiplying over alleles and replicates, summing over genotype allocations each multiplied by the product of genotype probabilities for the unprofiled contributors, and then multiplying over loci including the penalty term:

$$\prod_{l\in L}\pi_l\sum_{g\in G_l}\left[\prod_{c\in C}\Pr(\mathcal{G}_{g,c})\right]\prod_{r\in R}\prod_{i\in I_l}a(l, r, g, i, \sigma) \tag{4}$$

where $\mathcal{G}_{g,c}$ denotes the genotype allocated to c in g, while π_l is the combined penalty on the likelihood at locus l given the values for α_l, D, O, σ and the δ. (4) is then maximized over these parameters. `likeLTD` uses

Table 2: Penalties applied to the parameters of the peak-height model.

Parameter	Distribution	Mean	SD
$E[\alpha_l]$	*N*	0.013	0.010
$\log_{10}(\alpha_l/E[\alpha_l])$	*N*	0	0.300
D	Γ	0.02	0.019
O	Γ	0.02	0.019
δ	*e*	0.02	0.020
σ	*e*	100	0.010

Distributions: *N*, normal; Γ, gamma; *e*, exponential. The degradation parameters *d* have the same penalty for each contributor and for dropin.

a genetic algorithm `DEoptim` that simulates mutation, recombination and selection on parameter vectors to search for the vector that maximizes the penalized likelihood (Mullen et al., 2011). Maximization is performed separately under H_p and H_d and the LR is the ratio of the maximized values.

2.2 Validation studies

Many validation checks for forensic DNA software have been proposed. We have previously proposed using simulated or laboratory-generated replicate profiling runs (Steele et al., 2014a). It uses the fact that the inverse match probability (IMP) gives an upper bound on the LR, and the bound should be closely approached in certain settings. Bright et al. (2015) suggest generating artificial mixtures based on the assumptions of the model, to check that parameter estimates are consistent with those used to generate the CSP. Taylor et al. (2015) propose checking that the mean LR for a given CSP over many randomly-generated Q is close to the expected value of 1, noted by Alan Turing (Good, 1950). This is a refinement of the false Q validation method of (Gill and Haned, 2013).

It remains the case that no one test can fully validate a model or its implementation in software. We have therefore devised an extensive range of checks on `likeLTD`, which we now describe.

2.2.1 Simulated two-person mixtures

First, we compared the performance of a simplified version of the peak-height model with a discrete model, also implemented in `likeLTD`, that classifies peaks as allelic/uncertain/non-allelic (Balding, 2013). Comparisons were conducted when inferring the single-locus genotypes of two contributors to a CSP, with varying mixture ratios. The contributor genotypes were both heterozygous, sharing one allele. The expected peak height for the unshared allele of the first contributor was 600 RFU (no degradation), and mixture ratios were considered ranging from 0.1 to 10. The following model simplifications were introduced to aid interpretability of changes in genotype probabilities resulting from changes in mixture ratio, without fundamentally altering the model. The stutter fraction was always 0.1, irrespective of LUS. All observed peak heights were taken to be equal to the expected values, and those above t_l=50 RFU were recorded in the CSP. For the peak-height model, the expected heights $E_{l,r,g,i}$ were calculated assuming D, O, δ and λ all equal to zero, and S constant across alleles. Contributor doses, χ_c, were assumed equal to the values used to generate the CSP and we fixed σ=10. For the discrete model, all allelic and non-allelic peaks were correctly designated as such in the data input. Dropout probabilities were calculated using the model of Tvedebrink et al. (2009):

$$\Pr(D|H)=\frac{\exp(\beta_0+\beta_1\log H_T)}{1+\exp(\beta_0+\beta_1\log H_T)}, \tag{5}$$

where β_1=−4.35, as estimated by Tvedebrink et al., and β_0=18.556 which is the mean of the locus estimates reported in Tvedebrink et al. (2009). The combined doses for a peak H_T are H_1 and $2H_1$, for an unshared heterozygous and homozygous allele of the first contributor, respectively, and H_1+H_2 for a heterozygous allele shared by the two contributors. H_1 and H_2 are estimated from unshared alleles of each contributor.

2.2.2 Laboratory-generated validation data

Cheek swab samples were collected from 36 volunteer donors. DNA was extracted using a PrepFiler Express BTA Forensic DNA Extraction Kit and the Life Technologies Automate Express Instrument as per the manufacturer's recommendations.

Single-contributor and multi-contributor mock crime samples were created from 36 DNA samples as shown in Table 3. These crime samples were amplified using the AmpFℓSTR® NGMSelect® PCR kit as per

Table 3: Laboratory protocol for generation of single-contributor and multi-contributor CSPs from 36 donated DNA samples.

# Cont	# Samples	Condition	DNA mass (pg)
1	9	250 pg	250
	9	62 pg	62
	9	16 pg	16
	9	4 pg	4
2	12	Maj/min	266 (250:16)
	12	Equal	62 (31:31)
3	6	Unequal	328 (250:62:16)
	6	Equal	93 (31:31:31)

DNA masses are given as a total, with individual contributions in brackets. These are target values, realized values can vary.

the manufacturer's recommendations on a Veriti® 96-Well Fast Thermal Cycler. The amplified PCR products were size separated by capillary electrophoresis using an ABI 3130 Sequencer, with 1 µl of the PCR product, 10 second injections and 3 kV voltage. The results were analysed using GeneMapper® ID v3.2 with a detection threshold t_l=20 RFU for all $l \in L$; all peaks above the detection threshold were recorded.

For one of the three-contributor mixtures, we compared the observed peak heights with the probability distributions generated under the model, in order to verify that the probability distributions are well calibrated.

2.2.3 Comparison with discrete model

Next, we used the laboratory-generated data to compare the performance of the `likeLTD` peak-height model with that of the discrete model. For multi-contributor CSPs (see Table 3), each contributor was queried in turn, leading to 36, 48 and 36 evaluations for the single-, two- and three-contributor CSPs, respectively. To convert the laboratory-generated epgs into appropriate input data for the discrete model, interpretation rules set out in Table 4 were used. If there were multiple possible designations, "non-allelic" was adopted if it is one of the possibilities, otherwise "uncertain" is the default. For example, if the CSP shows alleles 13, 14 and 15 with peak heights 800, 35 and 600, respectively, the 14 allele would be called as non-allelic when considered as an *O* of the 13 allele (x=0.044), but uncertain when considered as an *S* of the 15 allele (x=0.058), and so the final call would be non-allelic.

2.3 Replication

To mimic pre-extraction replication, the mixtures described in Table 3 were created multiple times, but with DNA contributions of approximately x/n pg in each of n samples, successively for n=2, 3 and 4 (Table 5). PCR amplification, capillary electrophoresis and genotype analysis were performed for each replicate as described above.

Table 4: Interpretation rules for epg peaks in positions that could correspond to stutter (*S*), double-stutter (*D*) or over-stutter (*O*).

Designation	*S*	*D* and *O*
Non-allelic	$x<0.05$	$x<0.05$
Uncertain	$0.05 \le x < 0.15$	$0.05 \le x < 0.1$
Allelic	$x \ge 0.15$	$x \ge 0.1$

x is the ratio of heights of the possible stutter peak to the parent peak. These rules are used to generate input data for discrete-model LRs computed to compare with the LRs generated by the `likeLTD` peak-height model.

Table 5: Experimental design for investigating the relative merits of pre-extraction replication.

# Cont	Condition	Unsplit DNA mass (pg)	# Samples	# Reps	Split DNA mass (pg)
2	Equal	62 (31:31)	4	2	31 (16:16)
			4	3	21 (10:10)
			4	4	16 (8:8)
	Maj/min	266 (250:16)	4	2	133 (125:8)
			4	3	89 (83:5)
			4	4	67 (63:4)
3	Equal	93 (31:31:31)	2	2	47 (16:16:16)
			2	3	31 (10:10:10)
			2	4	23 (8:8:8)
	Unequal	328 (250:62:16)	2	2	164 (125:31:8)
			2	3	109 (83:21:5)
			2	4	82 (63:16:4)

Target DNA masses are rounded to the nearest picogram (pg), and are given as a total, with individual contributions in brackets.

Both the replicated and unreplicated two- and three-contributor CSPs (see Table 5) were evaluated assuming each contributor as Q in turn, to investigate whether pre-extraction replication holds any benefit over profiling a single sample. Next, we implemented the validity checks for a low-template DNA LR algorithm that we previously proposed (Steele et al., 2014a): the two-contributor replicated CSPs were evaluated with sequential addition of replicates, to check that the LR with the peak-height model approaches, but does not exceed, the IMP.

We also used the replicate CSPs to assess the approach of the WoE towards the IMP as the number of replicates increases (here, up to 4) as proposed in (Steele et al., 2014a).

2.4 Model extensions

2.4.1 Minor contributors modeled as dropin

The single-replicate, unequal two- and three-contributor CSPs were re-evaluated assuming one less contributor to the CSP. For these analyses Q was never the minor contributor. Under the peak-height model, any low peak not attributable to one of the hypothesized contributors will be explained as dropin. Because of peak-height variability, the algorithm will often assign positive probability to several different sets of peaks designated as dropin; note that `likeLTD` has no definitive classification of peaks as dropin or non-dropin as all allelic peaks are hypothesized to have some contribution from dropin.

2.4.2 Linkage adjustment

Linkage can lead to non-independence of loci when the alternative to Q under H_d, say X, is a close relative (other than parent or offspring). The number of loci used in DNA profiling kits has increased in recent years, so that two loci on the same chromosome arises in many of these kits; specifically the 17-locus system recently adopted in the UK has two pairs of linked loci: vWA and D12S391 on chromosome 12, and D2S1338 and D2S441 on chromosome 2. While it is possible to account fully for linkage and population structure for each genotype allocation when calculating the LR (Bright et al., 2013a), the full computation is complex and current practice is either to omit one of each pair of linked loci, which tends to understate evidential strength if Q is indeed a contributor, or to ignore the linkage which tends to overstate the evidence. We propose instead a simple adjustment to the LR:

$$\mathrm{LR}' = \mathrm{LR}\frac{\Omega_l}{\Omega_u} \tag{6}$$

where Ω_l is the IMP assuming linkage (Bright et al., 2013a), and Ω_u is the IMP ignoring linkage. The result of our adjustment normally lies between the values resulting from the two current practices, and should not be systematically biased towards either prosecution or defense.

To verify these expectations, a three-contributor CSP was evaluated, with the 16 pg contributor as Q, and the 250 pg contributor as K (reference profile available). The LR was computed six times, with H_d specifying a sibling of Q, with:

1. No linkage adjustment
2. Removal of vWA and D2S441
3. Removal of vWA and D2S1338
4. Removal of D12S391 and D2S441
5. Removal of D12S391 and D2S1338
6. Linkage adjustment (6)

All likelihood evaluations were performed with `likeLTD` v6.1. Table 6 gives the hypothesis pairs evaluated for each condition. All evaluations assumed F_{ST}=0.03, t_l=20 for every locus l, a sampling adjustment of one, and a Caucasian population database for all unknown contributors (Steele and Balding, 2014; Steele et al., 2014b).

3 Results

3.1 Model validation

3.1.1 Simulated two-person mixtures

Ideally an epg interpretation model would assign probability one to the correct genotype allocation for the unknown contributors. The red dotted line in the left panel of Figure 1A shows that this is the case for a wide range of mixture ratios for simplified, simulated CSPs with two unknowns. Correct genotype inference is not possible for mixture ratios close to one, because there is no information to distinguish the alleles of the two contributors, nor for mixture ratios close to zero because of allele dropout affecting the minor contributor. Correct genotype inference is never possible for mixtures under a discrete model, because by definition it

Table 6: Hypothesis pairs evaluated for the CSPs generated from the mixtures in Tables 3 and 5.

# Contributors	Condition	Hypotheses	Figure
2	Single rep	*U*1+dropin	1B, 3A, 4A,B
	Multiple reps	*U*1+dropin	3A
	Minor dropin	dropin	4B
3	Data fit	*K*1 (250 pg)+*U*1	2
	Single rep	*U*1+*U*2	3B, 4B
	Multiple reps	*U*1+*U*2	3B
	Minor dropin	*U*1+dropin	4B

K and *U* denote contributors with and without a reference profile available. To the contributors stated here, *Q* was added under H_p and an unrelated individual *X* was added under H_d. For the "Minor dropin" conditions the LR was evaluated for all true contributors other than the minor. For the "Data fit" condition *Q* was always the 16 pg contributor. For other conditions, each contributor was queried in turn.

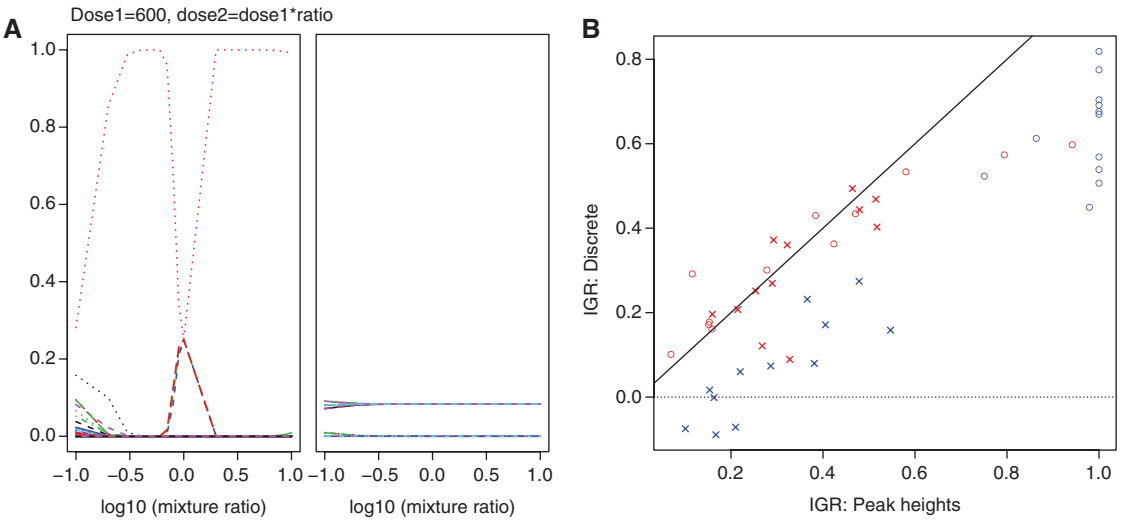

Figure 1: Simulated two-person mixtures. (A) Theoretical: probabilities assigned to possible genotype allocations for two unknown contributors, one with DNA dose corresponding to 600 RFU, while the other has DNA dose=600×ratio, where the mixture ratio varies from 0.1 to 10 (*x*-axis, $\log_{10}$ scale). The left panel corresponds to a simplified peak-height model while the right panel gives results for a discrete model. Each line corresponds to an allocation of the pair of genotypes, the red dotted line denoting the correct allocation which has probability close to one for most mixture ratios under the peak-height model. The true genotypes have one allele in common and 12 possible ordered genotype pairs are consistent with three distinct alleles. The discrete model assigns probability close to 1/12 to each of these for most of the range of ratios. (B) Laboratory: Gives the information gain ratio (WoE/$\log_{10}$ (IMP)) for 12 two-contributor equal-contribution CSPs (red, 31 pg for each contributor) and 12 two-contributor major/minor CSPs (blue, 16 pg minor, 250 pg major) using both the peak height (*x*-axis) and discrete (*y*-axis) models. Both contributors to each CSP were queried in separate calculations, with circles and crosses distinguishing the two contributors.

uses no information that could distinguish the alleles of the two contributors. The right panel of Figure 1A shows that the discrete model performs as well as can be expected: for all but very small mixture ratios it assigns probability close to 1/12 for each of the 12 genotype pairs consistent with three observed alleles, with deviations for low ratios arising because of dropout. However, even in the equal-contributions case (mixture ratio=1), the peak-height model does better than the discrete model because it can recognize which allele is represented twice among the two genotypes, and so assigns equal probability to each of four genotype allocations, rather than 12 under the discrete model.

3.1.2 Laboratory data: model fit

For one of the three-contributor mixtures, evaluated assuming the major was a known contributor and with the minor as the queried contributor (see Table 6, Data fit), we found that the proportion of observed peak heights within the 95% probability interval computed under the peak-height model was 0.94, while the proportion within the inter-quartile range was 0.51 (Figure 2), indicating that the model is well calibrated for this example.

3.1.3 Comparison with discrete model

Despite the superiority of the peak-height model in a simplified setting with no peak-height variability (Figure 1A), when querying laboratory-generated two-equal-contributor low-template CSPs the WoE supporting a true H_p appears on average no higher when computed under the peak-height model than under a discrete model (Figure 1B, red). In effect, the additional information potentially available from peak heights is lost due

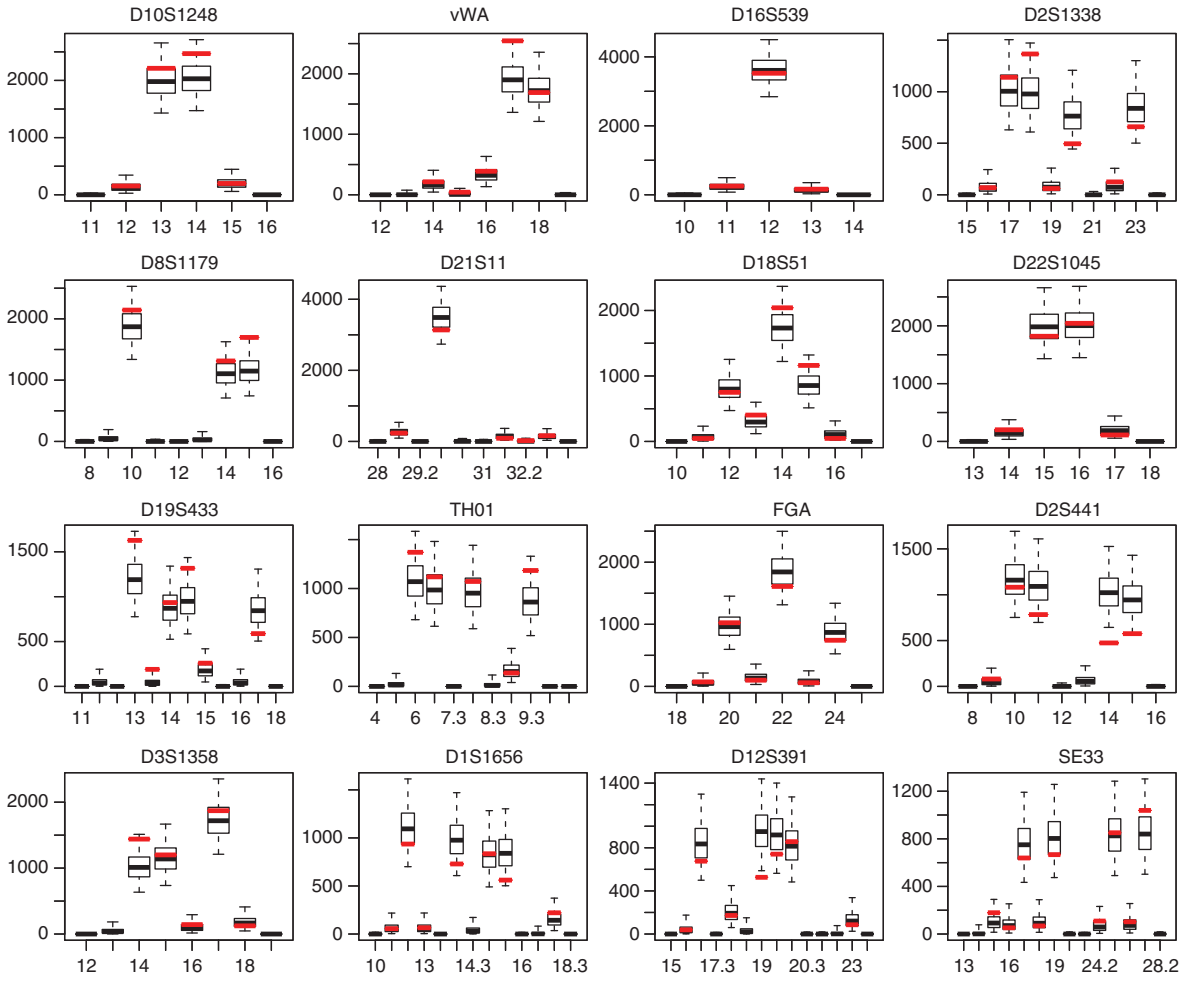

Figure 2: Observed and fitted peak heights under H_d for a CSP assuming a 250 pg K and two unknown contributors. Boxes show the central 50% (inter-quartile range) of the gamma distribution for each hypothesized peak, whiskers represent the 95% equal-tailed probability interval and red bars show observed peak heights. The y-axes gives peak height in RFU, while each boxplot corresponds to an allele.

to peak-height variability at the low template used here (31 pg per contributor). Even the two red x in Figure 1B that seem to indicate better performance of the peak-height model for low-template profiles in fact have been verified by manual inspection to reflect unequal contributions, apparently due to pipetting error.

When instead the Maj/min CSPs are queried, the peak-height model does perform better than the discrete model (Figure 1B, blue). In four cases the peak height IGR for the minor (crosses) supports H_p, while the discrete IGR supports H_d even though H_p is true. However, the peak-height IGR for the major (blue circles) is almost always ≈1.0 (the two exceptions have been verified by manual inspection to have a lower than expected contribution from the major, once again due to pipetting variability). This means that the discrepancy in DNA mass between the two contributors is so large that the genotype of the major can be confidently inferred by the peak-height model, which in practice implies that it can also be inferred manually. Therefore the superior performance of peak-height over discrete model for these Maj/min CSPs is of limited benefit, since in practice the discrete model may be applied after manually inferring the genotype of the major. However, even when treating the major contributor as known, there remains an advantage of the peak-height model (results not shown) largely because it has some ability to distinguish dropin peaks from minor contributor alleles. Further, manual deconvolution of a major is often problematic in practice because it is hard to delineate exactly the circumstances under which this can be done with high confidence.

3.2 Replication

When a sample containing x pg of DNA is split into n replicates, each with x/n pg DNA, the IGR for multiple replicates is on average about the same as for a single replicate for both two- (Figure 3A) and three-contributor CSPs (Figure 3B). These results show that with efficient statistical analysis splitting a sample to achieve replication does not lose information. We discuss potential advantages below.

If replication is "free" in the sense of not exhausting the supply of DNA then it is potentially always advantageous. However, there are costs involved and a declining return from additional replicates. Figure 4A shows the increase in IGR with sequential addition of replicates from major/minor mixtures. When querying the major contributor (solid blue lines), the IGR reaches 1.0 for nine out of 12 CSPs, and never exceeds 1.0.

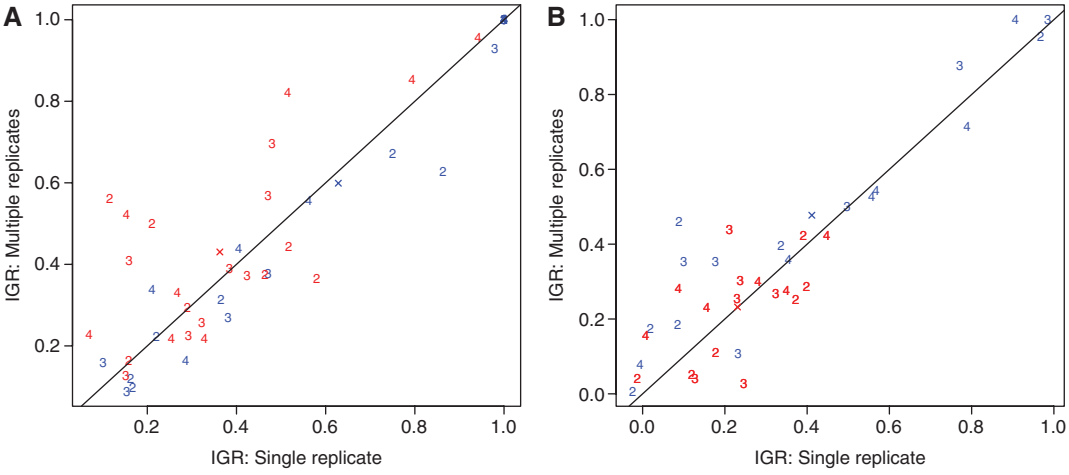

Figure 3: Information gain ratio (IGR) for (A) 24 two-contributor CSPs and (B) 12 three-contributor CSPs using a single replicate (x-axis) or splitting the sample into n replicates (y-axis). The CSPs had either equal contributions (red, 31 pg for each contributor) or unequal contributions (blue, 16 pg minor, 64 pg middle for three contributor only, 250 pg major). The plotted values indicate the number of replicates, with crosses indicating mean values for each color. Each of the contributors was queried in turn, leading to 48 and 36 data points.

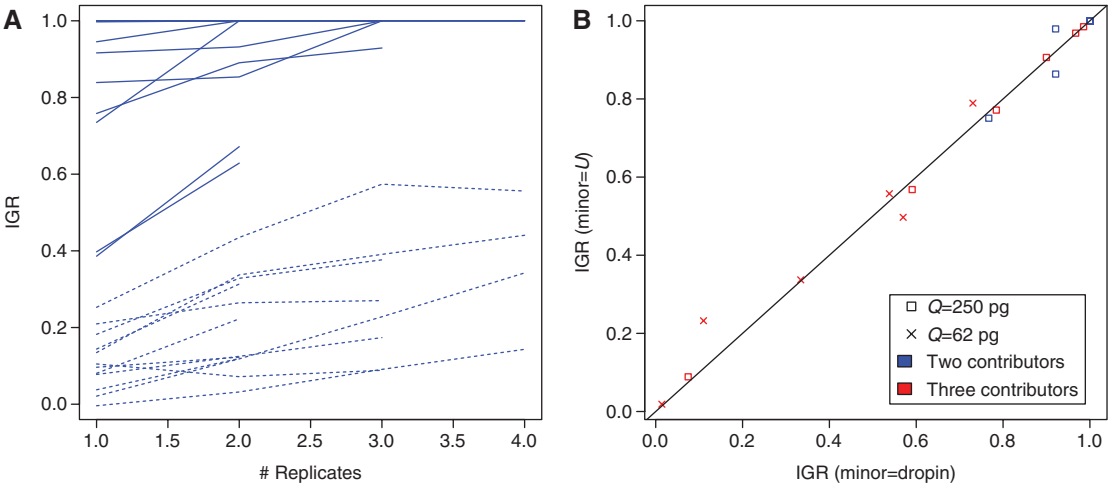

Figure 4: Information gain ratio (IGR) for (A) 12 major/minor two-contributor CSPs with sequential addition of replicates, dashed and solid lines correspond to minor and major contributor, respectively; (B) 12 two- and/or 6 three-contributor CSPs (blue and red, respectively) treating the minor contributor as dropin (x-axis) and as an additional contributor (y-axis).

3.3 Model extensions

3.3.1 Minor contributors modeled as dropin

The IGR when treating all contributors to an unequal-contributions mixture as unknowns under H_d is approximately equal to that with one fewer unknown contributor under H_d so that the minor contribution is modeled as dropin (Figure 4B). Because it can be difficult to decide whether additional low-level peaks in an epg should be modeled as dropin or as an additional contributor, it is important to establish that the result of the analysis is little affected by this choice. Moreover there can be computational advantages to treating as dropin any low-level contributors that are not the contributor of interest.

3.3.2 Adjustment for linkage

When the same three-contributor CSP as in Figure 2 is evaluated, but now proposing as X a sibling of Q, the LR with our proposed linkage adjustment lies, as predicted, between the no-adjustment LRs with and without removal of one locus from each linked pair (Table 7). The IMP is also affected by linkage adjustment and locus removal, and its values satisfy the same ordering as the LR. Note that ignoring linkage tends to be unfavorable to defendants, while with locus removal the LR varies substantially with the choice of loci to be removed. So both standard practices have serious defects which are avoided by our simple adjustment.

4 Discussion

We have presented a novel statistical model for evaluation of complex (low-template and/or mixed) DNA profiles using peak-height information, implemented in open-source software `likeLTD`. We have investigated its performance using a series of validation tests, including comparison with an established discrete model, and we have used it to investigate the advantages of pre-extraction replication. We further proposed two useful extensions of the model, to deal with low-level contributors and linked loci.

Our peak-height model incorporates a number of important features lacking from comparable software (Table 1). These include modeling both double- and over-stutter. Over-stutter is commonly seen at the trinucleotide locus D22, now a part of the DNA17 set of loci routinely used in the UK, while double-stutter is sporadically observed across all loci. If these phenomena are not modeled, it may be necessary to increase the detection threshold, which risks losing minor peaks of interest, or else explain any observations as dropin, yet this feature is not incorporated in dropin models. `likeLTD` is the only software that models a contribution

Table 7: WoE and $\log_{10}$ (IMP) for a three-contributor CSP with and without our proposed linkage adjustment (6), in the latter case using all loci, and with all possible combinations of removing one of each pair of linked loci.

Linkage adjustment (6)	Loci removed	WoE	$\log_{10}$ (IMP)
No	None	0	7.3
Yes	None	−0.2	7.1
No	vWA and D2S441	−1.2	6.4
No	vWA and D2S1338	−0.5	6.4
No	D12S391 and D2S441	−1.8	6.4
No	D12S391 and D2S1338	−0.8	6.3

WoE, Weight of Evidence=$\log_{10}$ (LR); IMP, inverse match probability. Here, H_d specifies a brother of Q as the alternative source of the DNA, which is false in this example but because Q is a low-level minor contributor (16 pg), the results show that there is no information to distinguish Q from a sibling (WoE is zero or weakly negative).

from dropin at every allelic position, whether or not a peak is observed, which reflects reasonable intuition that if dropin is feasible it can potentially contribute to any observed peak. The `likeLTD` runtime for the 48 two-contributor single-replicate evaluations ranged from 7 to 18 min, while the 36 three-contributor single-replicate evaluations ranged from 18 to 200 min.

Regarding the validation tests, first we showed that the peak-height model performs well in inferring the genotypes of the two contributors to each of 24 simulated two-person mixtures (Figure 1). Next, we verified that probability intervals for peak heights under the model fitted to a three-contributor CSP are well calibrated (Figure 2). We further verified that the WoE increases towards the IMP with additional replicates but does not exceed the IMP (Figure 4A), thus implementing for the peak-height model a validity check that we previously applied to a discrete model (Steele et al., 2014a).

In our equal-contributor CSPs we found little benefit of a peak-height model over a discrete model, for either two or three contributors. This seems counter-intuitive because peak heights are potentially informative about shared alleles (either homozygosity or shared across contributors) and can also deal better with possible stutter than a discrete model, but against this is the high variability of peak heights for low DNA template. There was a noticeable gain in information for the unequal-contributor CSPs (Figure 1), supporting the results of Bright et al. (2015) who also found a gain in information from peak heights for unequal contributors but not for equal contributors.

We found that when analysed with our peak-height model, replication on average entails no loss of information even when it requires splitting a low-template sample (Figure 3), and there may be a small overall gain in information. Replication implies additional profiling costs, but it may provide additional reassurance to a court and it can guard against failure of a profiling run. Using the LRmix discrete model Benschop et al. (2015) found that pre-extraction splitting a sample into four subsamples for PCR and subsequent profiling provided additional information to identify the major contributor but led to a substantial loss of information when the minor contributor was queried, due to high levels of drop-out and also masking. This contrasts with our finding of no systematic gain or loss of information due to replication for either contributor which may be due to our use of a peak-height model and also our low detection threshold.

Thanks to the novel dropin model of `likeLTD`, which is conceptually simple yet more realistic than other dropin models, we showed that it can be a valid strategy to reduce computational complexity by modeling as dropin any low-level contributors not of interest to the investigation (Figure 4B). Conceptually, dropin is modeled like a shower of alleles that fall in proportion to population frequencies. This could be a valid model for any contributor but it does not permit inference of the genotypes of individual contributors, which is why it is only appropriate for low-level contributors not including the contributor of interest. The fact that hypotheses contrasted in an LR specify the number of contributors, whereas this is often unknown and can be difficult to infer (Haned et al., 2011; Manabe et al., 2013), is sometimes used as a criticism of the use of LRs as a measure of evidential weight (Buckleton and Curran, 2008). However, if multiple low-level contributors can be modeled as dropin it is unnecessary to specify the number of contributors exactly.

Not adjusting for linked loci tends to favor prosecutions, while the degree that removing one locus from linked pairs favors defenses can depends on the loci chosen for removal. Our proposed adjustment avoids both of these problems, is conceptually appealing and easy to compute, avoiding exact full computation of linked LRs (Bright et al., 2013a; Dørum et al., 2016). We showed that our adjustment behaves as expected in an example (6), returning an LR between that with no adjustment and those with removal of linked loci (Table 7).

Inference for complex DNA profiles has advanced impressively in recent years, from a situation prior to about 2010 when such profiles were regularly being presented in court without valid evaluation techniques being available, to the current availability of multiple models and software offering a range of modeling options. This has increasingly allowed minuscule, mixed and degraded samples to be presented in court accompanied by robust and meaningful measures of evidential weight. We hope that this will render obsolete the retrograde Dlugosz judgment that permitted in the courts of England and Wales subjective, qualitative assessments of complex evidence based only on an expert's experience (Champod, 2013). However, there remains room for further progress in understanding and reducing differences among the different models, although preliminary indications suggest that such differences are rarely if ever important in practice.

Funding: Cellmark Forensic Services, (Grant/Award Number: "CMD-PHD1") Biotechnology and Biological Sciences Research Council, (Grant/Award Number: "507493").

References

Balding, D. J. (2013): "Evaluation of mixed-source, low-template DNA profiles in forensic science," Proc. Natl. Acad. Sci. USA 110, 12241–12246.

Balding, D. J. and J. Buckleton (2009): "Interpreting low template DNA profiles," Forensic Sci. Int.-Gen., 4, 1–10.

Balding, D. J. and C. D. Steele (2015): Weight-of-evidence for Forensic DNA Profiles, 2nd Ed., London: John Wiley & Sons.

Benschop, C. C. G., S. Y. Yoo and T. Sijen (2015): "Split DNA over replicates or perform one amplification?," Forensic Sci. Int.-Gen. Supplement Series, 5, e532–e533.

Bleka, Ø., G. Storvik and P. Gill (2016): "EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts," Forensic Sci. Int.-Gen., 21, 35–44.

Bright, J.-A., J. M. Curran and J. S. Buckleton (2013a): "Relatedness calculations for linked loci incorporating subpopulation effects," Forensic Sci. Int.-Gen., 7, 380–383.

Bright, J.-A., D. Taylor, J. M. Curran and J. S. Buckleton (2013b): "Developing allelic and stutter peak height models for a continuous method of DNA interpretation," Forensic Sci. Int.-Gen., 7, 96–304.

Bright, J.-A., I. W. Evett, D. Taylor, J. M. Curran and J. Buckleton (2015): "A series of recommended tests when validating probabilistic DNA profile interpretation software," Forensic Sci. Int.-Gen., 14, 125–131.

Brookes, C., J.-A. Bright, S. Harbison and J. Buckleton (2012): "Characterising stutter in forensic STR multiplexes," Forensic Sci. Int.-Gen., 6, 58–63.

Buckleton, J. and J. Curran (2008): "A discussion of the merits of random man not excluded and likelihood ratios," Forensic Sci. Int.-Gen., 2, 343–348.

Champod, C. (2013): "DNA transfer: informed judgment or mere guesswork?," Front. Genet., 4, 300.

Cowell, R. G., T. Graversen, S. L. Lauritzen and J. Mortera (2015): "Analysis of forensic DNA mixtures with artefacts," J. Roy. Stat. Soc. C-App., 64, 1–48.

Dørum, G., D. Kling, A. Tillmar, M. D. Vigeland and T. Egeland (2016): "Mixtures with relatives and linked markers," Int. J. Legal Med., 130, 621–634.

Gill, P. and H. Haned (2013): "A new methodological framework to interpret complex DNA profiles using likelihood ratios," Forensic Sci. Int.-Gen., 7, 251–263.

Gill, P., J. Whitaker, C. Flaxman, N. Brown and J. Buckleton (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.

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

Gill, P., J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker and J. Lambert (2008): "Interpretation of complex DNA profiles using empirical models and a method to measure their robustness," Forensic Sci. Int.-Gen., 2, 91–103.

Gill, P., L. Gusmão, H. Haned, W. R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. M. Schneider and B. S. Weir (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.-Gen., 6, 679–688.

Good, I. J. (1950): Probability and the weighing of evidence, Ann Arbor, MI, USA: JSTOR.

Graversen, T. and S. Lauritzen (2014): "Computational aspects of DNA mixture analysis," Stat. Comput., 25, 527–541.

Haned, H., L. Pene, J. R. Lobry, A. B. Dufour and D. Pontier (2011): "Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count?," J. Forensic Sci., 56, 23–28.

Kelly, H., J.-A. Bright, J. S. Buckleton and J. M. Curran (2014): "Identifying and modelling the drivers of stutter in forensic DNA profiles," Aust. J. Forensic Sci., 46, 194–203.

Manabe, S., C. Kawai and K. Tamaki (2013): "Simulated approach to estimate the number and combination of known/unknown contributors in mixed DNA samples using 15 short tandem repeat loci," Forensic Sci. Int.-Gen. Supplement Series, 4, e154–e155.

McCord, B. R., J. M. Jung and E. A. Holleran (1993): "High resolution capillary electrophoresis of forensic DNA using a non-gel sieving buffer," J Liq. Chromatogr. R. T., 16, 1963–1981.

Mullen, K. M., D. Ardia, D. L. Gil, D. Windover, and J. Cline (2011): "DEoptim: An R package for global optimization by differential evolution," J. Stat. Softw., 40, 1–26.

Nathakarnkitkool, S., P. J. Oefner, G. Bartsch, M. A. Chin and G. K. Bonn (1992): "High-resolution capillary electrophoretic analysis of DNA in free solution," Electrophoresis, 13, 18–31.

Perlin, M. W., M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose and B. W. Duceman (2011): "Validating TrueAllele DNA mixture interpretation," J. Forensic Sci., 56, 1430–1447.

Puch-Solis, R., L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran and D. Balding (2013): "Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters," Forensic Sci. Int.-Gen., 7, 555–563.

Ruiz-Martinez, M. C., O. Salas-Solano, E. Carrilho, L. Kotler and B. L. Karger (1998): "A sample purification method for rugged and high-performance DNA sequencing by capillary electrophoresis using replaceable polymer solutions. A. Development of the cleanup protocol," Anal. Chem., 70, 1516–1527.

Steele, C. D. and D. J. Balding (2014): "Choice of population database for forensic DNA profile analysis," Sci. Justice, 54, 487–493.

Steele, C. D., M. Greenhalgh and D. J. Balding (2014a): "Verifying likelihoods for low template DNA profiles using multiple replicates," Forensic Sci. Int.-Gen., 13, 82–89.

Steele, C. D., D. S. Court and D. J. Balding (2014b): "Worldwide F_{ST} estimates relative to five continental-scale populations," Ann. Hum. Genet., 78, 468–477.

Taylor, D., J. Buckleton and I. Evett (2015): "Testing likelihood ratios produced from complex DNA profiles," Forensic Sci. Int.-Gen., 16, 165–171.

Taylor, D., J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut and J. Buckleton (2016): "Validating multiplexes for use in conjunction with modern interpretation strategies," Forensic Sci. Int.-Gen., 20, 6–19.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen and N. Morling (2009): "Estimating the probability of allelic drop-out of STR alleles in forensic genetics," Forensic Sci. Int.-Gen., 3, 222–226.

Williams, P. E., M. A. Marino, S. A. Del Rio, L. A. Turni and J. M. Devaney (1994): "Analysis of DNA restriction fragments and polymerase chain reaction products by capillary electrophoresis," J. Chromatogr. A, 680, 525–540.