Statistical issues surrounding the analysis of forensic low-template DNA samples

Author:

Christopher D. Steele

Supervisors:

Prof. David Balding

Dr. Denise Syndercombe Court

Matthew Greenhalgh

UCL Genetics Institute
Department of Genetics, Evolution and Environment

A thesis submitted to University College London for the degree of Doctor of Philosophy
August 24, 2016

Declaration of originality

I, Christopher David Steele, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Increased sensitivity of forensic DNA profiling over the last decade has led to increased stochasticity in the resulting profiles, causing difficulties for interpretation that were acknowledged by the Caddy Report [Caddy et al., 2008]. These difficulties were largely overcome with the adoption of statistical models allowing for dropout and dropin, but interpretation issues remain, several of which are tackled in this thesis. One such issue concerns the choice of allele frequency databases when the ethnic background of the true source of the crime scene DNA is unknown. I propose a heuristic for choosing a single database and adjusting the likelihood ratio calculations to allow for the possibility that a different database may be more appropriate. Another issue in general, and specifically for the database choice heuristic, is the choice of an appropriate value for the population genetics parameter F_{ST} to account for distant relatedness between the alleged contributor and an alternative source of the DNA. I present empirical estimates of F_{ST} in worldwide populations, relative to the continental-scale reference databases that are used for UK forensic DNA profiles. In the last few years many software packages for the evaluation of low-template DNA samples have emerged, including likeLTD originally developed by my supervisor Prof Balding but greatly improved and reprogrammed by myself as part of my PhD work. There remains little consensus on how to validate these software packages. I present a method of validation based on the use of multiple-replicate crime stain profiles. It relies on the intuition that sufficient replicates of even very noisy DNA profiling runs eventually generate the same information as a single high-quality replicate. I show that likeLTD performs well when assessed by this approach. Finally, I present a new statistical model that extends likeLTD to incorporate the peak height information in a crime scene profile. I show results based on simulation and laboratory trials verifying the good performance of the new model in improved discrimination between true and false hypotheses.

List of Figures

List of Tables

Contents

List of Abbreviations

F_{ST}	Fixation index
H_d	Defence hypothesis
H_p	Prosecution hypothesis
K	Profiled/known individual
µl	microlitre
Q	Queried individual
U	Unprofiled/unknown individual
X	Random individual that replaces Q under H_d
AMEL	Amelogenin
bp	base pairs
CCD	charge-coupled device
CRAN	Comprehensive R Archive Network
CSP	Crime scene profile
DS	Double-stutter
E	Evidence (CSP)
epg	Electropherogram
FSS	Forensic Science Service
HGDP	Human Genome Diversity Project
IGR	Information gain ratio = WoE/$\log_{10}$IMP
IMP	Inverse match probability
LCN	Low copy number
LR	Likelihood ratio
LTDNA	Low-template DNA
ltIGR	Low-template information gain ratio
ltLR	Low-template likelihood ratio

LUS	Longest uninterrupted sequence
MCMC	Markov chain Monte Carlo
mixIGR	Mixture information gain ratio
mixLR	Mixture likelihood ratio
MUS	Multiple uninterrupted sequences
ng	nanogram
OS	Over-stutter
PC	Principal component
PCA	Principal components analysis
PCR	Polymerase chain reaction
pg	picogram
PH	Peak height
RFU	Relative fluorescence unit
S	Stutter
SNP	Single nucleotide polymorphism
STR	Short-tandem repeat
unc	Uncertain
WoE	Weight-of-Evidence = $\log_{10}$LR

Chapter 1

Introduction to forensic low-template DNA analysis

Much of the information in this chapter, specifically regarding low-template DNA, has been published in Steele and Balding [2014b], see Appendix B, and Balding and Steele [2015].

1.1 History of low-template DNA analysis

DNA profiling was first developed by Prof. Sir Alec Jeffreys in 1984 based on variable number tandem repeats, and was quickly used in the courts to convict the killer of Lynda Mann and Dawn Ashworth in 1988 through matches to semen samples found on the victims' bodies. DNA analysis became routine in court, and moved to short tandem repeats, but some high profile cases highlighted some deficiencies in thinking about DNA weight-of-evidence (WoE). An early example was the case of Raymond Easton, who was found to match a burglary crime scene profile at six loci that was obtained 200 miles from his home through a database search. Easton had severe tremors as a result of Parkinson's disease, and was unable to perform simple tasks. Despite this exculpatory evidence, Easton was charged with burglary based solely on the DNA match. He was later exonerated on the basis of a 10 locus profile which did not match the crime scene profile.

Over time the sensitivity of DNA profiling increased through various means, so that usable profiles could be obtained from very small amounts of DNA, termed low-template DNA (LTDNA) evidence. In a similar vein to DNA evidence, a number of high profile cases cast doubt on some LTDNA analysis practices. Sean Hoey was arrested and accused of murder in 2003, five years after the 1998 Omagh bombing. During his acquittal in 2007, it was claimed that LTDNA methods were unreliable and that contamination was a serious issue for interpretation [NICC, 2007]. These claims lead to the temporary suspension of LTDNA sample analysis. As a direct result of the Hoey acquittal, a report was commissioned by the UK Forensic

Science Regulator [Caddy et al., 2008] which concluded that LTDNA methods are "fit for purpose", paving the way for analysis of LTDNA samples to resume. However, the report highlighted that questions remained over how best to analyse and interpret LTDNA results, and recommended that a consensus should be agreed upon for LTDNA interpretation. A further ruling in a separate case clarified that cases with more than 100-200 picograms (pg) of DNA, approximately equivalent to between 16 and 34 cells, should no longer be challenged on the grounds of being low-template [EWCA, 2009], however, this left open questions regarding the analysis of DNA samples that have less than 100-200 pg DNA. This was addressed in EWCA [2010], in which the court ruled that despite the increased incidence of stochastic effects and the potentially debatable probative value of DNA evidence at less than 100-200 pg of DNA, LTDNA evidence at such low levels should still be admissible. Interpretation challenges still remain, as demonstrated in the trial of Raffaele Sollecito and Amanda Knox, in which allelic calls from a DNA profile were altered between appeals [Balding, 2013], and further questions of contamination were raised. See Naughton and Tan [2011] for a history of some important cases regarding DNA evidence, and Bentley and Lownds [2011] for some of the case history surrounding LTDNA evidence.

Forensic evidence can generally be partitioned into a four level hierarchy of propositions:

- Offence level i.e. has an offence been committed?
- Activity level i.e. what activity has taken place?
- Source level i.e. who or what is the source of an item of evidence?
- Sub-source level e.g. who contributes to a DNA sample?

It is possible to have an item of evidence for which being a sub-source contributor does not imply that you are the source of the item of evidence e.g. a blood stain from a murder victim may contain DNA from the perpetrator; the victim is the source of the blood stain, but the perpetrator is a sub-source contributor to the DNA within the blood stain. Note that LTDNA evidence introduces difficulties to interpretation at an activity level, such as secondary/indirect transfer, that may be important to the WoE against a suspect, but that are not considered in this thesis, as I consider DNA evidence only at a source or sub-source level.

As the sensitivity of DNA typing technology increased, the WoE to be presented in court moved away from the certainty of full match, to the WoE given some partial matching to a suspect. With such potential partial matches, the method of calculating the WoE of a particular profile moved from the rudimentary probabilities of inclusion [Buckleton and Curran, 2008], through likelihood ratios (LRs) based on presence/absence of alleles [Evett and Weir, 1998, Evett et al., 1991, Gill et al., 2000, 2008, 2012, Balding and

Buckleton, 2009, Balding, 2013] in an electropherogram (epg), to LRs based on continuous data [Graversen and Lauritzen, 2014, Cowell et al., 2015, Bleka et al., 2016, Puch-Solis et al., 2013, Bright et al., 2013c, Perlin et al., 2011], usually incorporating the peak heights (PHs) of the epg. As the models proposed to calculate the WoE have become more sophisticated, so too have the software packages that perform the calculations.

1.2 Goals

The case history of DNA and particularly LTDNA evidence clearly demonstrates challenges in the interpretation of such evidence in court. This thesis will address some of these issues, and provide evidence to inform the consensus for LTDNA analysis recommended by the Caddy report.

A new PH model will be presented in Chapter 5. While this PH model utilises more of the information available in an epg than previous models, perhaps the most important benefit of a PH model in relation to the challenges previously described is the minimal input by the forensic scientist. Presence/absence models require manual designation of peaks as either allelic or non-allelic, which can be challenging when, for example, a potential minor allelic peak is in the same position as a major allelic peak. A PH model removes this need, as the program determines the most likely genotypes automatically, through maximisation or integration of model parameters. This would have removed some of the difficulties seen in the Knox/Sollecito case, and allowed the court to focus on more important questions than the designation of peaks.

As the sophistication of WoE models has increased, methods to verify the validity of a model, or an implementation of a model, have not kept pace. Chapter 2 will present a method for validating forensic WoE software that depends on multi-replicate epgs, which will be used to validate a presence/absence model. Further, in Chapter 7, this method will be used to model a PH model. Both of these models are available in the Comprehensive R Archive Network (CRAN) package, likeLTD. Further validation of the PH model will be presented in Chapter 6, which will largely present results of the PH model on laboratory-generated epgs.

A common theme throughout the case challenges described earlier was the issue of contamination. While many of the challenges concerned contamination of evidence samples with DNA from a queried contributor, contamination by individuals not of interest can also be problematic leading to complex samples and challenging interpretation. Chapter 7 will present strategies that are available when employing a PH model to reduce this complexity, either through assuming a major contaminant as a known contributor, or assuming a minor contaminant as environmental contamination.

Population genetics phenomena, such as population substructure, that have an important effect on the WoE will be investigated in Chapters 3 and 4. This will provide estimates of the population genetics

parameter, F_{ST}, that are more widespread and more exact than has previously been available. This work will tie in to the use of population databases when calculating the WoE, the effects of mis-assigning a database, and possibilities for reducing computational complexity through guidance on the what database to use for a given calculation.

1.3 Forensic DNA typing

The most commonly used marker type in forensic DNA profiling is the short tandem repeat (STR), or microsatellite. STRs are sections of the genome in which a short motif is repeated multiple times; forensic markers commonly have repeat units of four base pairs (a notable exception is the D22 locus included in the NGMSelect® profiling kit, which instead has repeat units of three base pairs) that vary from around three to 50 or more repetitions. Partial repeats are possible, but are usually less common than full repeats. An STR with five repeat units (e.g. $[ACGA]_5$) is defined as allele 5, while an allele with five repeat units, and a partial repeat of three base pairs (e.g. $[ACGA]_2ACG[ACGA]_3$) is defined as allele 5.3. The allele definition is determined by the length of the allele in base pairs, so both $ACGA[ACG]_5ACGA$ and $[ACGA]_2ACG[ACGA]_3$ are allele 5.3, despite the difference in internal structure of the two sequences. Smaller differences between repeat structures of the same allele, such as point mutations, are also possible without changing the allele designation. These variants within alleles cannot be detected by current STR typing using capillary electrophoresis, however, full sequencing of STRs is currently in development, which will allow for discrimination between these within-allele variants. Repeat sections of STRs are flanked by non-repeat sections, the flanking regions, that are targeted by primers for amplification by polymerase chain reaction (PCR). Thus, the total length of an amplified STR in base pairs is the length of both flanking regions in addition to the length of the repeat section. STR kits that are less susceptible to some PCR and/or case circumstance artefacts (see Section 1.7) have been developed by reducing the size of these flanking regions.

In current practice, after DNA has been extracted from a sample, PCR is used to amplify the STRs of interest. The primers used in PCR are fluorescently tagged, and are incorporated into the amplified product during each elongation step. After PCR, the amplified product is drawn through a capillary by applying an electric field across it (capillary electrophoresis); DNA is negatively charged at neutral pH due to the phosphate groups in the sugar-phosphate backbone, so travels towards the positive end of the dipole. At a certain point, a laser is shone through a window in the capillary, which causes the fluorescent tag incorporated into each primer to fluoresce as PCR product passes the laser, which is detected by a charge-coupled device (CCD) camera. CCD cameras are capable of detecting multiple wavelengths of light at the

same time. Longer DNA molecules travel more slowly through the capillary due to increased mechanical obstruction by the media in the capillary, leading to size separation of the PCR product over time. Multiple loci can be interrogated in a single reaction (multiplexing) by:

1. size separating loci e.g. locus A spans 100-150 base pairs (bp) while locus B spans 200-250 bp.
2. wavelength separating loci e.g. two loci both spanning 100-150 bp, but locus A is tagged with a dye that fluoresces at 494 nm while locus B is tagged with a dye that fluoresces at 705 nm.

See [Butler et al., 2004] for a full review of forensic DNA typing using capillary electrophoresis.

1.4 Forensic DNA evidence

A crime stain and a sample from one or more reference individuals are STR typed, generating a crime stain profile (CSP) and one or more reference profiles. The capillary electrophoresis generates an epg (Figure 1.1), in which each panel shows the loci tagged with the different dyes, blue, green, yellow (displayed as black here) and red from top to bottom, the y-axis of each panel shows relative fluorescence units (RFU), a proxy for the amount of DNA in each peak, and the x-axis shows time in seconds, a proxy for the length of alleles in base pairs. Grey boxes above each panel indicate the size range of each locus in that panel, while vertical grey bars indicate the size of each common allele within that locus. Red vertical bars indicate the peaks from a fourth unshown panel, the allelic ladder, that is used to correctly size each allele. Coloured boxes below prominent peaks show the allelic designation of each peak that has been automatically called as allelic.

In Figure 1.1 the sex determining locus, amelogenin (AMEL), shows a large X peak and a small Y peak, suggesting a mixture containing a large amount of female-origin DNA and a small amount of male-origin DNA. The epg suggests a minimum of two contributors due to a maximum of four alleles being observed at any locus, with many loci displaying peak heights that suggest a major/minor mixture e.g. vWA and D16. A plausible scenario for generating such an epg would be a vaginal swab from a rape victim, although the case circumstances of the epg shown in Figure 1.1 are unknown.

1.5 Likelihood ratio (LR)

The court is interested in whether or not a given individual is guilty of a crime. However, a forensic DNA scientist cannot directly answer this question for a number of reasons:

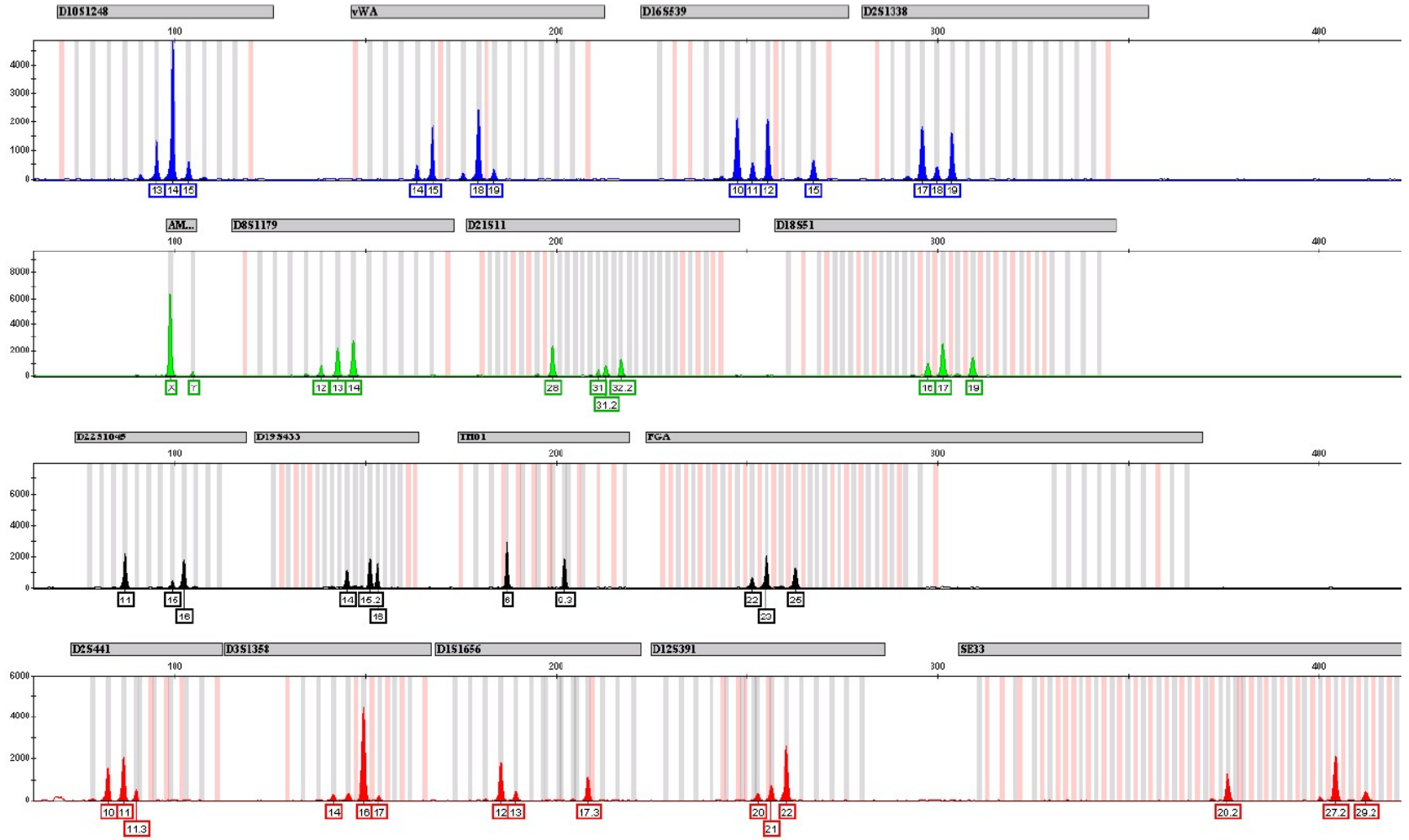


Figure 1.1: An epg showing a mixed STR profile obtained from a DNA sample with two contributors. The major profile is from a female and the minor component from a male. Each panel corresponds to a different dye colour (from top to bottom): Blue, Green, Yellow (displayed as black) and Red. The DNA sample was amplified with the NGM SElect® STR kit. The amplified fragments were separated on the ABI 3130xL Genetic Analyzer and analysed using the GeneMapper® 3.2 software. Image supplied courtesy of Cellmark Forensics. ©2014 Cellmark Forensics.

The "ultimate issue" rule: In the UK, the duty of the jury is to determine the guilt or innocence of a suspect. Expert witnesses are barred from commenting on this question to avoid them unduly influencing the jury. Additionally, non-DNA evidence may exist that has a direct bearing on the possible guilt of the suspect, such as an alibi, eye-witness account, CCTV footage, or general case circumstances. The expert witness generally has no experience in evaluating this qualitative non-scientific evidence, so cannot incorporate this non-DNA evidence with the DNA evidence to ascertain an overall probability of guilt.

Source level evidence: DNA can often only determine if someone has contributed to a DNA sample; this may not imply whether or not they committed a crime. DNA does not always suffice to place a suspect at the scene of a crime, if secondary DNA transfer is believed to have occurred [Meakin and Jamieson, 2013].

The DNA expert is charged with evaluating the evidence for whether a given individual, termed the queried individual (Q), is a contributor to a CSP. Q is often the alleged culprit of a crime, however, in some cases Q may be some other individual, such as the victim of a crime. Q is always the individual who has the greatest bearing on whether or not the suspect is guilty of the crime e.g. Q may be the victim of a rape if a penile swab has been taken from the alleged culprit as whether or not the victim's DNA is represented in the DNA from the penile swab is most relevant to whether or not the alleged culprit is guilty or not.

Once an appropriate Q has been determined, the evidence evaluation can be formulated as a likelihood ratio (LR) [Evett and Weir, 1998, Evett et al., 1991]:

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

where E is the CSP evidence, H_p is the prosecution hypothesis which asserts that Q is represented in the CSP and H_d is the defence hypothesis which asserts that anyone other than Q, who will be termed X, is represented in the CSP. Typically the specification of H_p is relatively simple, as the prosecution must propose a particular set of events by which the suspect is guilty of the crime in question. The defence has no such obligation, so specification of H_d is more problematic. H_p does not need to be a special case of H_d, where Q has replaced X i.e. H_p can posit two contributors while H_d posits three contributors. However, all hypotheses pairs presented throughout this thesis will replace X with Q under H_p. Gill and Haned [2013] suggest a framework for determining appropriate hypothesis pairs in complex scenarios.

For artificial CSPs, as presented throughout this thesis, H_1 and H_2 would be more appropriate designations as there is no prosecution or defence hypothesis in an artificial setting, however, H_p and H_d will be used throughout for convenience. LRs will be presented throughout as $\log_{10}(\text{LR})$, which gives the weight-of-evidence (WoE) in units of bans, which was first proposed by Alan Turing in 1940 during his code breaking work in WWII [Good, 1979].

Each separate likelihood, L_p and L_d, at a single locus is given by the following equation, adapted from Curran et al. [2005]:

$$\sum_{j=1}^{n} \left[\prod_{r \in R} Pr(r|K, U_j, \phi) \right] Pr(U_j), \tag{1.2}$$

where R is the set of replicates that make up a CSP, K are the alleles of any profiled individuals, which will always include Q under H_p, U_j are the jth set of hypothesised unknown contributor alleles that can explain the CSP, which will always include X under H_d, n is the number of unknown genotype allocations, and ϕ are any model parameters. This notation is often simplified to $\sum_{j=1}^{n} Pr(E|G_j)Pr(G_j|K)$, where G_j is $\{K, U_j\}$, and the product over replicates is implicit in E.

1.6 Good-template DNA

Here, (1.2) will be built up from first principles using the LR framework. In the simplest possible setting of a good-template CSP with a single observed peak, A ($\mathcal{C}$=A), assuming a single contributor with no population genetics effects the LR for a Q with genotype A,A ($\mathcal{G}_Q$=AA) is:

$$LR = \frac{1}{p_A^2}, \tag{1.3}$$

where p_A is the population allele probability for allele A. Similarly, if $\mathcal{C}$=AB and $\mathcal{G}_Q$=AB, then LR $= 1/2p_Ap_B$. The numerator is 1 if Q matches the CSP, and 0 otherwise, because $P(E|G = \mathcal{G}_Q)$ is 1 if $G = \mathcal{G}_Q$ and 0 if $G \neq \mathcal{G}_Q$, so for $\mathcal{G}_Q$=AC, LR $= 0/2p_Ap_C = 0$. The denominator is the probability of observing a genotype matching the CSP alleles in the given population assuming Hardy-Weinberg equilibrium.

Now consider a two-contributor CSP where both individuals are good-template, $\mathcal{C}$=ABC, and $\mathcal{G}_Q$=AB. If the second contributor is unprofiled (U), then the LR is:

$$LR = \frac{2p_Ap_C + 2p_Bp_C + p_C^2}{12(p_A^2p_Bp_C + p_Ap_B^2p_C + p_Ap_Bp_C^2)} = \frac{2p_A + 2p_B + p_C}{12p_Ap_B(p_A + p_B + p_C)}, \tag{1.4}$$

where each term under the unsimplified H_d comprises both a homozygote/heterozygote mix and a heterozygote/heterozygote mix with both orderings of the individuals e.g. the first H_d term has genotype combinations {AA,BC}, {BC,AA}, {AB,AC} and {AC,AB}. Under H_p only unknown genotypes are given a population probability, as any known profile is not being drawn from the population. Any genotype that does not explain the CSP is not included e.g. $\mathcal{G}_U$=BB under H_p.

Because peak heights are a proxy for the dose of DNA, they can often be used as a guide to the allele count of a peak. This information can reduce the number of reasonable combinations by suggesting which peaks come from which contributors, and which are homozygous/heterozygous e.g. for D10 from Figure 1.1 the peak heights suggest the most parsimonious genotype is 13,15 for the minor contributor and 14,14 for the major contributor, so if $\mathcal{G}_Q$=13,15, the LR would be:

$$LR = \frac{p_{14}^2}{4p_{13}p_{14}^2p_{15}} = \frac{1}{4p_{13}p_{15}}, \tag{1.5}$$

where H_d includes both orderings of genotypes 13,15 and 14,14. Compared to (1.4) the number of genotypes that are considered has been reduced from three to one under H_p, and from 12 to two under H_d, by using peak heights to inform plausible genotype combinations. Note that this simplification of the LR by utilising peak height information may be possible for simple good-template CSPs, but may not possible for low-template CSPs as high peak height variability at low DNA levels may lead to incorrect inference of allele counts.

Returning to the scenario that generated (1.4), $\mathcal{C}$=ABC, $\mathcal{G}_Q$=AB, but assuming that instead of an unprofiled second contributor, we have a profiled individual, $\mathcal{G}_K$=AC, the LR becomes:

$$LR = \frac{1}{2p_Ap_B + p_B^2 + 2p_Bp_C}, \tag{1.6}$$

where we can see once again that only genotypes that are able to explain the CSP fully are considered.

1.7 Low template DNA

Over the last decade or so the sensitivity of STR typing has improved, leading to the typing of samples comprised of smaller and smaller amounts of DNA; it is now possible to imperfectly type samples that include less DNA than is found in a single cell, approximately 6 pg DNA.

Multiple routes to increased sensitivity for a low-template sample are available:

PCR cycles: Increase the number of PCR cycles.

Input concentration: Increase the concentration of PCR product introduced into capillary electrophoresis.

Voltage: Increase the voltage across the capillary, so more of the PCR product is drawn into the capillary.

Injection time: Increase the injection time, increasing the amount of DNA drawn into the capillary.

Purification: Purify the PCR products, which removes unused primers and non-target DNA, so that a greater proportion of the DNA in the sample is relevant.

The low levels of DNA that can now be analysed, and the methods used to increase sensitivity, lead to increased stochastic effects in the profiles that are generated, so the interpretation of LTDNA profiles can be more difficult than described in Section 1.6. These difficulties were highlighted by the Caddy Report [Caddy et al., 2008], after the acquittal of Sean Hoey [NICC, 2007]. Stochastic effects introduced into LTDNA profiles that are generally not problematic at good-template are:

Heterozygote imbalance: For good-template samples the peak height for the two peaks of a heterozygous individual are expected to be approximately equal. This expectation is no longer reasonable for low-template samples due to the stochastic nature of PCR, and sampling variance during pipetting. For example, if five cells worth of DNA enter PCR, and two copies of one allele but all five copies of the other happen to be amplified in the first cycle, then there are four and ten copies of each available for amplification in the second cycle, compounding the initial stochastic imbalance in the first cycle amplification. Extreme heterozygote imbalance can lead to one of the alleles having a peak height below the detection threshold of the genotyping technology, termed dropout. The possibility of dropout means that during computation of the LR it may no longer be sufficient to only propose genotypes that consist of alleles that were observed in the CSP.

Dropin: Dropin is low-level environmental contamination of an allele from DNA at the crime-scene or in the laboratory at very low levels. Other authors distinguish between contamination before and after sample collection, however, these phenomena are not treated separately here, as the epg contains no information on the nature of the contamination. The designation of both dropin and dropout depend on the hypothesised genotypes being assessed. There is no upper limit to the number of dropin alleles, however, more than two per replicate may be better modelled as an extra unknown contributor to the CSP, unless the dropin originates from a police officer or laboratory scientist, at which point they can be included as a known contributor. Any replicated allele should not be treated as dropin, but rather

as having originated from an extra unknown contributor, as the probability of the same allele dropping in at multiple replicates is small. Dropin and gross contamination (contamination of a full or near-full profile) are two extremes of the same process, contamination, but dropins are treated as independent events to reduce computational complexity, whereas gross contamination is often treated as an extra unknown contributor.

LTDNA samples also lead to difficulties for analysis that arise from artefacts that are present in good-template samples, but which cannot confidently be determined to be artefactual under low-template:

Stutter: If there is a large peak in the epg at position x bp, then a small peak can often be seen at position $x-n$ bp, where n is the length of the repeat unit at that locus, which is termed stutter (S). More rarely peaks can be seen at $x-2n$ bp, termed double-stutter (DS), and $x+n$ bp, termed over-stutter (OS). These stutter peaks are believed to originate from an error in the replication of DNA during PCR, DNA polymerase slippage, where the DNA polymerase enzyme skips a repeat (DS and S) or replicates a repeat more than once (OS). These stutter peaks are not a problem for good-template DNA samples e.g. the small peak at D10 allele 12 in Figure 1.1, however, when there is a low-level contributor it can be difficult to distinguish between a stutter peak of a major contributor and an allelic peak of a minor contributor.

Degradation: DNA from crime scenes is often degraded due to exposure to the environment for some time before sample collection; humidity [McCord et al., 2011], bacterial metabolism [Cotton et al., 2000] and ultraviolet exposure [Diegoli et al., 2012] have all been shown to degrade DNA. Long stretches of DNA are more susceptible to degradation; Bright et al. [2013c] showed an exponential decay of epg peak heights with increasing size of an allele in base pairs. Therefore peaks on the right side of an epg are often lower than those on the left for degraded samples e.g. the peaks at FGA of Figure 1.1 are slightly smaller than those at D22. High levels of degradation can lead to an overabundance of dropout for long alleles, especially when the sample is already low-template.

Along with these extra artefacts that need to be accounted for, the increased sensitivity of LTDNA methods means observation of multiple contributors to a CSP is more likely, often due to gross contamination, so mixtures are more often encountered for LTDNA samples than good-template samples (a reference profile is usually a good-template high-quality sample).

1.8 Dropout model

The dropout model for forensic low-template DNA treats dropouts as Bernoulli events, and has been iterated on by multiple authors [Gill et al., 2000, 2008, 2012, Balding and Buckleton, 2009, Balding, 2013].

1.8.1 Single contributor

Returning to the scenario that generated (1.3), $\mathcal{C}$=A, $\mathcal{G}_Q$=AA, but instead assuming that the observed peak was low-level then it is no longer obvious that the single peak is from a homozygous individual. (1.3) now becomes:

$$LR = \frac{1 - D_2}{p_A^2(1 - D_2) + 2p_A p_Z D(1 - D)}, \tag{1.7}$$

where D and D_2 are the dropout probabilities for a heterozygote and homozygote allele respectively, and Z is all alleles other than A. D has an inverse relationship with the RFU of a peak. The numerator is now the P($E|\mathcal{G}_Q$) for which there has been a non-dropout of a homozygote A allele to explain the CSP, $1-D_2$, and the denominator is now P($E|\mathcal{G}_X$=AA)+P($E|\mathcal{G}_X$=AZ) allowing genotypes that correspond to a homozygote non-dropout, $\mathcal{G}_X$=AA, and a heterozygote dropout, $\mathcal{G}_X$=AZ. Note that when dropout is allowed, $\mathcal{G}_Q$=AB would be able to explain the CSP with probability $D(1-D)$, where previously the probability of H_p would have been 0 because it was previously assumed that D=0. Logically, D is different between H_p and H_d, however, a single D is shown under both hypotheses here for illustration purposes, as is common practice [Gill et al., 2007]. Note that if D=0 and D_2=0, (1.7) simplifies to (1.3), the good-template scenario.

Degradation reduces the peak height for long alleles, which leads to a higher probability of dropout. If an individual's dose of DNA is k then the degradation adjusted dose, k', for each of their alleles is modelled as:

$$k' = k(1 + \delta)^{-f}, \tag{1.8}$$

where δ is the degradation parameter, and f is the mean adjusted length of the allele in base pairs. This up-weights the dose for shorter than average alleles, and down-weights the dose for longer than average alleles.

1.8.2 Multiple contributors

With multiple contributors to a CSP, each individual is expected to contribute a different amount of DNA, and to have a different dropout rate. Tvedebrink et al. [2009] published a model that estimates dropout rates from average peak heights of a single contributor using logistic regression. From this model, average peak heights can be converted into doses, k, where an average single heterozygote peak is given dose 1, giving $D(k)$ as the dropout rate for dose k, leading to:

$$\frac{D(k)}{1-D(k)} = (\alpha_l k)^{\beta}, \tag{1.9}$$

where l is the locus, α is proportional to Tvedebrink's $\exp(\beta_{0,s}/\beta_1)$ with some proportionality constant that converts doses, k, to peak heights, H in Tvedebrink et al., and β is Tvedebrink's β_1 parameter. This allows for the $D(k)$ to be calculated when $D(1)$ is known, where $D(1)$ can be an integrated or maximised parameter alongside contributions of all hypothesised contributors relative to a specified contributor with fixed dose 1. The dropout rate for a single-contribution homozygous allele, D_2, can now be thought of $D(2k)$ where $D(k)$ is the dropout rate for a single dose of the allele. This gives the inequality:

$$\frac{D(2k)}{D(k)^2} \approx \left(\frac{2}{\alpha_l k}\right)^{\beta} > 1, \tag{1.10}$$

which implies that dropout of a homozygous allele can be more likely than locus dropout of two heterozygote alleles, which is incorrect. This inequality is only relevant for low dropout probabilities [Tvedebrink et al., 2012], so the incorrect inference is irrelevant for practical purposes.

Returning to the scenario that gave (1.4), $\mathcal{C}$=ABC, $\mathcal{G}_Q$=AB, but now assuming both contributors are low-template, then the multi-dose dropout model assuming no degradation gives the LR in Table 1.1, where L_p is $\sum \mathrm{P}(E|\mathcal{G}_1,\mathcal{G}_2)\mathrm{P}(\mathcal{G}_p)$ over all red rows, and L_d is $\sum \mathrm{P}(E|\mathcal{G}_1,\mathcal{G}_2)\mathrm{P}(\mathcal{G}_d)$ over all rows. With such a multi-dose dropout model $\mathrm{P}(E|\mathcal{G}_1,\mathcal{G}_2)$ is no longer the same for unordered genotypes when each individual is not a heterozygote non-dropout e.g. $\mathrm{P}(E|\mathcal{G}_1 = AA, \mathcal{G}_2 = BC)$ differs from $\mathrm{P}(E|\mathcal{G}_1 = BC, \mathcal{G}_2 = AA)$ and $\mathrm{P}(E|\mathcal{G}_1 = AZ, \mathcal{G}_2 = BC)$ differs from $\mathrm{P}(E|\mathcal{G}_1 = BC, \mathcal{G}_2 = AZ)$ if the two contributors have different dropout rates. Note that if all D=0, the LR from Table 1.1 simplifies to (1.4), the good-template case.

1.8.3 Dropin

Dropin, as discussed in Section 1.7, is the observation of epg peaks that cannot be explained by one of the hypothesised contributors. Dropin events are treated as independent Bernoulli events, with probability I.

$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathrm{P}(E\|\mathcal{G}_1,\mathcal{G}_2)$	$\mathrm{P}(\mathcal{G}_p)$	$\mathrm{P}(\mathcal{G}_d)$
AA	BC	$(1-D_{2,1})(1-D_{1,2})^2$	$2p_Bp_C$	$2p_A^2p_Bp_C$
BC	AA	$(1-D_{1,1})^2(1-D_{2,2})$	p_A^2	$2p_A^2p_Bp_C$
AB	AC	$(1-D_{1,1})^2(1-D_{1,2})^2$	$2p_Ap_C$	$4p_A^2p_Bp_C$
AC	AB	$(1-D_{1,1})^2(1-D_{1,2})^2$	$2p_Ap_B$	$4p_A^2p_Bp_C$
BB	AC	$(1-D_{2,1})(1-D_{1,2})^2$	$2p_Ap_C$	$2p_Ap_B^2p_C$
AC	BB	$(1-D_{1,1})^2(1-D_{2,2})$	p_B^2	$2p_Ap_B^2p_C$
AB	BC	$(1-D_{1,1})^2(1-D_{1,2})^2$	$2p_Bp_C$	$4p_Ap_B^2p_C$
BC	AB	$(1-D_{1,1})^2(1-D_{1,2})^2$	$2p_Ap_B$	$4p_Ap_B^2p_C$
CC	AB	$(1-D_{2,1})(1-D_{1,2})^2$	$2p_Ap_B$	$2p_Ap_Bp_C^2$
AB	CC	$(1-D_{1,1})^2(1-D_{2,2})$	p_C^2	$2p_Ap_Bp_C^2$
AC	BC	$(1-D_{1,1})^2(1-D_{1,2})^2$	$2p_Bp_C$	$4p_Ap_Bp_C^2$
BC	AC	$(1-D_{1,1})^2(1-D_{1,2})^2$	$2p_Ap_C$	$4p_Ap_Bp_C^2$
AZ	BC	$D_{1,1}(1-D_{1,1})(1-D_{1,2})^2$	$2p_Bp_C$	$4p_Ap_Bp_Cp_Z$
BC	AZ	$D_{1,2}(1-D_{1,1})^2(1-D_{1,2})$	$2p_Ap_Z$	$4p_Ap_Bp_Cp_Z$
BZ	AC	$D_{1,1}(1-D_{1,1})(1-D_{1,2})^2$	$2p_Ap_C$	$4p_Ap_Bp_Cp_Z$
AC	BZ	$D_{1,2}(1-D_{1,1})^2(1-D_{1,2})$	$2p_Bp_Z$	$4p_Ap_Bp_Cp_Z$
CZ	AB	$D_{1,1}(1-D_{1,1})(1-D_{1,2})^2$	$2p_Ap_B$	$4p_Ap_Bp_Cp_Z$
AB	CZ	$D_{1,2}(1-D_{1,1})^2(1-D_{1,2})$	$2p_Cp_Z$	$4p_Ap_Bp_Cp_Z$

Table 1.1: Likelihood ratio for a CSP with peaks at alleles A, B and C using a multi-dose dropout model assuming a single unknown contributor under H_p. $D_{x,y}$ is the dropout probability for x copies of an allele for the yth contributor. Each likelihood is the sum of the column products of $\mathrm{P}(E|\mathcal{G}_1,\mathcal{G}_2)$ and the corresponding $\mathrm{P}(\mathcal{G})$, where L_p is summed over only the rows highlighted in red, while L_d is summed over all rows.

Returning again to the example that gave (1.4) and Table 1.1, $\mathcal{C}$=ABC, and $\mathcal{G}_Q$=AB, but instead assuming the CSP has a single contributor, the CSP cannot be explained in full without invoking dropin. Allowing dropin, restricted to one dropin event, the LR is:

$$LR = \frac{p_C(1-D)^2 I}{6p_Ap_Bp_C(1-D)^2 I}, \tag{1.11}$$

where the denominator is a summation over three genotypes, AB, AC and BC, each with the missing allele being a dropin event with probability p_xI, where p_x is the population probability of the dropin allele.

If both dropout and dropin are being modelled, and dropin is not restricted to a single event, then any possible hypothesised genotype allocation is able to explain any possible CSP, although many hypothesised genotype allocations may be very improbable.

1.8.4 Replicates

A CSP may consist of more than one profiling run performed on the same sample; for LTDNA samples these replicates will often have different allelic peaks observed in each profiling run, as peak heights, and thus dropout events, are highly stochastic at low template. The information from any and all replicates can be

incorporated into a single LR for the overall profile, as shown through the product over replicates in (1.2). Suppose a CSP consisting of two replicates, $\mathcal{C}_1$=A and $\mathcal{C}_2$=B, with $\mathcal{G}_Q$=AB, while modelling dropout but no dropin and assuming a single contributor to the CSP. Then:

$$LR = \frac{D^2(1-D)^2}{2p_A p_B D^2(1-D)^2}, \tag{1.12}$$

where under H_d the only genotype that can explain the CSP is AB, because both of those alleles were observed over the multi-replicate CSP, and that P($E|G$) is composed of the first replicate with probability $D(1-D)$ and the second replicate, also with probability $D(1-D)$. P(G) remains the same as for a single replicate under both H_p and H_d. Replicate profiling runs will be further discussed in Chapter 2 in the context of validating software packages through the expected behaviour of the WoE when multiple replicate runs are performed, and in Chapter 7 in the context of comparing the efficacy of splitting a sample into multiple replicate runs, or running a single replicate with the maximum amount of DNA available.

1.9 Population genetics

1.9.1 F_{ST}

F_{ST} was first proposed by Wright [1949] as a measure of genetic variation between subpopulations relative to that in the total population. An F_{ST} value of 1 indicates that all subpopulations are at total fixation, so in each subpopulation a single allele has probability 1 while all other alleles have probability 0, where the fixed allele can be different between different subpopulations. This implies that all genetic variation can be explained by population structure. An F_{ST} value of 0 indicates that there is no effect of population structure on genetic variation, so the total population can be thought of as a single interbreeding population, rather than n distinct subpopulations i.e. the subpopulation allele probabilities for every allele are identical between subpopulations, and identical to the total population allele probabilities.

A second interpretation of the F_{ST} parameter, which is more directly relevant to forensic genetics, is that F_{ST} measures the extent of relatedness of individuals within the various subpopulations relative to that in the total population. More relatedness within subpopulations compared to the total population leads to a high variability in allele probabilities across the subpopulations, and therefore a high value of F_{ST}.

In forensic genetics, Q and X may be believed to have shared distant ancestry, in which case the WoE against Q will depend on any population structure that exists within their population; it is possible that Q and X share an allele due to it being common in their population rather than because they are the same

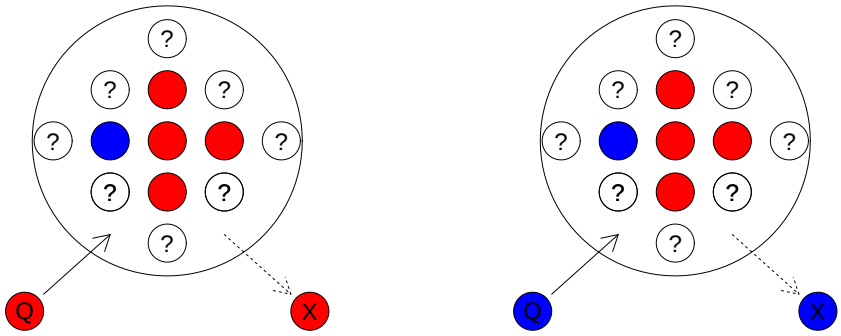


Figure 1.2: Representation of how the probability of drawing an allele from a population depends on the previous observations of that allele in the population, and F_{ST}, when Q has a common allele (left) or a rare allele (right). Central blue and red circles represent a database of alleles in the population, white circles represent the rest of the unsurveyed population. Q is assumed to originate from the population, while X is drawn from the population.

individual, which would imply that H_p is true (Figure 1.2). So an F_{ST} correction [Balding and Nichols, 1994] must be applied to the LR acting through the allele probabilities of Q [Weir, 2007], which should decrease the WoE against Q with increasing F_{ST}. The F_{ST} adjustment commonly used [Balding and Nichols, 1994] is derived from a population genetics sampling formula that assumes that population allele probabilities are Dirichlet distributed, and that drawing a specific allele from a population then makes it more likely to draw the same allele in subsequent attempts [Balding and Nichols, 1994]. The sampling formula is given as:

$$P(A|m,n,F_{ST}) = \frac{mF_{ST} + (1-F_{ST})p_A}{1+(n-1)F_{ST}}, \tag{1.13}$$

which gives the probability of drawing an A allele from the population when m A alleles have already been drawn out of n total alleles drawn. If n=10, F_{ST}=0.03 and p_A=0.2, and no A allele has been observed (m=0) then the probability of drawing an A allele on the 11th draw needs to be down-weighted as the previous draws provide evidence that A is more rare than p_A implies, $P(A|m,n,F_{ST})$=0.15, down from p_A=0.2. If instead all observed were A (m=10) then the probability that the 11th draw will be an A allele needs to be up-weighted, $P(A|m,n,F_{ST}) = 0.39$.

In a forensic context, the alleles of Q have already been observed, so when the alleles of X are drawn from the population of Q, the fact that some alleles in that population have already been observed needs to be accounted for. This is shown in Figure 1.2, where on the left, with F_{ST}=0.03, P(red) is 0.78, down

slightly from the database P(red)=0.80, while on the right P(blue)=0.22, up from the database P(blue)=0.2.

A sample size of four has a special significance in forensic genetics, as this accounts for having observed both alleles of Q in the population, and then drawing the two alleles of X. In the example that gave (1.3), $\mathcal{C}$=A, $\mathcal{G}_Q$=AA, two alleles have been observed from the population, the two alleles of Q, and they were both A as Q is homozygous. For a match to the CSP, X must be AA (assuming good-template DNA), so the alleles of X are sequentially drawn with probability:

$$\frac{2F_{ST}+(1-F_{ST})p_A}{1+F_{ST}}\frac{3F_{ST}+(1-F_{ST})p_A}{1+2F_{ST}}. \tag{1.14}$$

This is the match probability for a single contributor homozygous locus, a probability with special relevance in forensic genetics, which has a corresponding probability for a heterozygous genotype, $\mathcal{G}$=AB, $\mathcal{C}$=AB:

$$2\frac{F_{ST}+(1-F_{ST})p_A}{1+F_{ST}}\frac{F_{ST}+(1-F_{ST})p_B}{1+2F_{ST}}. \tag{1.15}$$

These both give the defence likelihood in each case, so the LR for Q is the inverse of this match probability, because L_p remains as 1. If $\mathcal{G}_X$=AA, $\mathcal{C}$=A and p_A=0.2, without an F_{ST} adjustment LR=25, but with F_{ST}=0.03 the LR is reduced to 15.1, mirroring our reduced confidence that this match originates because Q contributes to the DNA sample rather than through allele sharing from the population. Note that if F_{ST}=0, both inverse match probabilities (IMPs) simplify to the good-template single contributor LRs, $1/p_A^2$ and $1/2p_Ap_B$. The IMP acts as a theoretical maximum for the LR, and will be utilised to validate the behaviour of the dropout model in Chapter 2. Throughout the thesis, F_{ST} adjustments will be applied to calculations, except where stated, but F_{ST} notation will largely be suppressed.

Choice of an appropriate F_{ST} value will be discussed in Chapter 3, while using F_{ST} to allow for misattributed population databases will be discussed in Chapter 4. A full treatment of F_{ST} in forensic genetics can be found in Fung and Hu [2008].

1.9.2 Sampling adjustment

The population allele probabilities in databases must be adjusted for sampling variance during the survey process. Databases may be biased because they are typically convenience samples rather than random samples, and because they typically have a low sample size. To demonstrate the need for accounting for these biases, assume Q possesses an extremely rare allele, W, that has not been observed when collating the

allele database. Without a sampling adjustment p_W=0, so any LR with no Ks would be undefined if W has not dropped out, as all H_d genotypes must explain W (as either dropin or allelic), so all L_d terms would be 0. This situation is clearly erroneous as W has been observed in Q themselves, so $p_W > 0$. For this reason, the population allele probability is adjusted [Balding, 1995] as:

$$p_i' = \frac{a + sn}{b + 2s}, \tag{1.16}$$

where a is the number of observations of allele i in the database of sample size b alleles, s is the sampling adjustment, and n indicates the number of alleles of type i that Q possesses. Various authors use values of 1 or 2 for s, but the choice of s has little effect compared to the choice of other parameters, such as F_{ST}. With s=1 this can be thought of as adding the alleles of the Q to the database, in effect assuming H_p is true, while with s=2 the alleles of both Q and the CSP are being added to the database, in effect assuming H_d is true. If b=100, this adjustment increases the probability of an unobserved allele to 0.98% and 1.96% for the homozygous and heterozygous single contributor case respectively with s=1, and to 1.92% and 3.85% respectively with s=2.

1.10 Population allele probabilities

Considering each likelihood as $\sum_{j \in n} P(E|G_j)P(G_j)$, it is clear that the likelihood depends on the probability of a given individual's genotype, especially under H_d. To determine the value of $P(G_j)$ individuals from a population are sampled, from which an estimate of p_Z for all alleles represented in that population is obtained. As discussed previously, these estimates are subject to sampling error, meaning that more rare alleles will be observed as the sample size increases, and the accuracy of estimates will increase as sample size increases. Sample sizes in the past were typically in the order of hundreds, however, with the adoption of new loci, and new databases, sample sizes can range from several hundred to several thousand observations for a population.

While estimating $P(G_j)$ seems trivial once estimates of p_Z in a population have been obtained, there are subtleties that complicate the process.

Firstly, an individual may not fit well into any of the available databases. This may be because an individual comes from a relatively isolated population, that we know to differ from an available database e.g. individuals from Sardinia or Iceland would most appropriately fit a Caucasian database, but are known to differ genetically from the majority of Europeans. Alternatively if an individual is admixed, and for example

has one African parent and one Caucasian parent, they will not fit either the African or Caucasian database well. Additionally, there may be no suitable database available; if it is believed that a Fijian individual committed a crime, and there is no Austronesian population database available, it becomes difficult to assign an appropriate database.

Secondly, the act of assigning an individual to a database can be difficult. In the UK, the arresting officer is responsible for assigning what global population an individual comes from, through their physical appearance. Individuals from some subpopulations may not be closest genetically to the population that the police office would assign them to, some individuals may be misassigned, and classifying admixed individuals is difficult.

Thirdly, when using population allele probabilities, we are necessarily considering unprofiled individuals, either X or U. As these individuals are unknown to the forensic scientist, how should it be determined what population to assign to them for evaluation of $P(G_j)$? The prevailing method is to compute the LR for all relevant databases, and to report the most conservative LR to the court. However, this method becomes impractical with either a large number of unknown individuals, or a large number of relevant databases. Moreover, this may lead to an analysis that is incongruent with the known case circumstances e.g. a crime was committed in rural Cornwall, but an East Asian database gives the most conservative LR.

These issues are discussed further in Chapter 4, where a heuristic for the choice of database is presented, that attempts to alleviate some or all of these difficulties while remaining conservative to avoid miscarriages of justice.

1.11 Multiple loci

The LR at any locus will be affected by shared relatedness, which was described for distant relatedness in Section 1.9.1. This means that the LRs at multiple loci are not independent, as they all depend on some shared ancestry.

Similarly, with the introduction of 16-locus STR kits, each locus is no longer located on a separate chromosome, meaning some loci are no longer independent for close relatives due to the effects of linkage. Once the probability of recombination between two loci is low enough, they can be inherited as a single block, and so the second locus has no bearing on the WoE against Q once the WoE for the first locus has been evaluated. This is not problematic for LRs where Q is assumed unrelated to X, or when Q and X are assumed as a parent-child relationship, but can have a significant impact on the LR when Q and X are closely related e.g. siblings. A full approach for calculating linked-locus LRs is given by Bright et al. [2013a],

however, I instead propose to adjust the LR by a factor of:

$$\frac{\Omega_l}{\Omega_u}, \tag{1.17}$$

where Ω_l is linked match probability, and Ω_u is the unlinked match probability.

Once relatedness and linkage have been taken into account the LRs at multiple loci become approximately conditionally independent, so can be combined by taking the product over loci [Buckleton et al., 2005], often termed the "product rule".

1.12 likeLTD

likeLTD is an open source software package published on the Comprehensive R Archive Network (CRAN) that implements a dropout model to evaluate the WoE against Q for LTDNA samples [Balding, 2013].

1.12.1 Uncertain designation

In addition to the allelic and non-allelic calls that other dropout models allow, likeLTD provides an uncertain designation that ameliorates the difficulty of making definite allele calls for LTDNA samples. Uncertain designations are particularly useful for positions where it is thought that the stutter of a major contributor peak may be masking an allelic peak of a minor contributor. Uncertain calls can also be used in any situation where the true nature of the peak is in dispute, which can include potential pull-up peaks as well as peaks just above or below the detection threshold. Returning to the example that generated (1.7), $\mathcal{G}_Q$=A, but instead using $\mathcal{C}$=A[B], where [B] indicates that a B allele has been called as uncertain, the LR becomes:

$$LR = \frac{1 - D_2}{p_A^2(1 - D_2) + 2p_A p_B(1 - D) + 2p_A p_Z D(1 - D)}, \tag{1.18}$$

where Z is now all alleles other than A and B. When a B allele has been hypothesised (middle term under H_d), the dropout term for that allele is not included in the LR, as the uncertain designation of [B] implies that it is unknown whether B has dropped out or not. With D=0 and D_2=0, the LR no longer simplifies to the good-template LR (1.3) as $\mathcal{G}_X$=AB is being considered, and neither would it simplify to the good-template scenario if $\mathcal{G}_Q = AB$, as $\mathcal{G}_X = AA$ would still be being considered. Here, the dropout model including the uncertain extension will be termed the discrete model.

$\mathcal{G}_Q$	$\mathcal{G}_X$	Allelic adjustment	Genotypic adjustment
AA	AA	$\frac{(2F_{ST}+(1-F_{ST})p_A)(2F_{ST}+(1-F_{ST})p_A)}{(1+F_{ST})(1+F_{ST})}$	$\frac{(2F_{ST}+(1-F_{ST})p_A)(3F_{ST}+(1-F_{ST})p_A)}{(1+F_{ST})(1+2F_{ST})}$
AA	AB	$2\frac{(2F_{ST}+(1-F_{ST})p_A)(1-F_{ST})p_B}{(1+F_{ST})(1+F_{ST})}$	$2\frac{(2F_{ST}+(1-F_{ST})p_A)(1-F_{ST})p_B}{(1+F_{ST})(1+2F_{ST})}$
AB	AB	$2\frac{(F_{ST}+(1-F_{ST})p_A)(F_{ST}+(1-F_{ST})p_B)}{(1+F_{ST})(1+F_{ST})}$	$2\frac{(F_{ST}+(1-F_{ST})p_A)(F_{ST}+(1-F_{ST})p_B)}{(1+F_{ST})(1+2F_{ST})}$
AB	AA	$\frac{(F_{ST}+(1-F_{ST})p_A)(F_{ST}+(1-F_{ST})p_A)}{(1+F_{ST})(1+F_{ST})}$	$\frac{(F_{ST}+(1-F_{ST})p_A)(2F_{ST}+(1-F_{ST})p_A)}{(1+F_{ST})(1+2F_{ST})}$
AB	AC	$2\frac{(F_{ST}+(1-F_{ST})p_A)(1-F_{ST})p_C}{(1+F_{ST})(1+F_{ST})}$	$2\frac{(F_{ST}+(1-F_{ST})p_A)(1-F_{ST})p_C}{(1+F_{ST})(1+2F_{ST})}$
AB	CD	$2\frac{((1-F_{ST})p_C)((1-F_{ST})p_D)}{(1+F_{ST})(1+F_{ST})}$	$2\frac{((1-F_{ST})p_C)((1-F_{ST})p_D)}{(1+F_{ST})(1+2F_{ST})}$

Table 1.2: F_{ST} adjusted genotype probabilities for all combinations of matching/non-matching alleles between Q and X, with both the allelic F_{ST} adjustment used in likeLTD and the full genotypic F_{ST} adjustment. This table is included in the likeLTD guide which is included with the package.

1.12.2 F_{ST} adjustment

likeLTD does not implement a full genotypic F_{ST} adjustment (see Section 1.9.1), but rather an allelic adjustment. In (1.13) the genotypic F_{ST} adjustment depends on both the genotype of Q and the genotype of X, so separate allele probabilities, p_x, would have to be computed for every G_j under H_d, as every G_j considers a different X. likeLTD instead implements an allelic adjustment, where the database allele probabilities are adjusted as:

$$\begin{aligned}
(1-F_{ST})p/(1+F_{ST}) &\qquad \text{non-}Q\text{ allele}\\
(F_{ST}+(1-F_{ST})p)/(1+F_{ST}) &\qquad \text{heterozygote } Q \text{ allele}\\
(2F_{ST}+(1-F_{ST})p)/(1+F_{ST}) &\qquad \text{homozygote } Q \text{ allele}
\end{aligned}$$

Allelic adjustments need only be applied once before computation, so simplify the overall computation of the LR. Other choices of allelic adjustment are possible, for example some authors use $(1+2F_{ST})$ as the denominator, however, this difference will have negligible impact on the final LR, especially compared to the value of F_{ST} chosen.

Table 1.2 gives the full genotypic and allelic F_{ST} adjustments for all possible combinations of shared/non-shared alleles between Q and X. Genotypic adjustments are derived from (1.13). In all genotype pairings a denominator $(1+F_{ST})$ in the allelic adjustment is replaced with $(1+2F_{ST})$ in the genotypic adjustment. Otherwise the only difference between the two occurs when X is homozygous, where an F_{ST} in the right hand of the numerator in the allelic adjustment is replaced with $2F_{ST}$ in the genotypic adjustment, so the greatest difference between the two will be seen when X is homozygous. Note that not only will using an allelic adjustment alter the LR, but it will also alter the match probability (rows 1 and 3 in Table 1.2), and therefore the maximum possible LR. Throughout this thesis both LRs and IMPs will be calculated with an allelic adjustment rather than a genotypic adjustment.

Validation of likeLTD is presented in Chapter 2, while in Chapter 5 a peak heights model is added to likeLTD, which is itself validated in Chapter 6. See Gill et al. [2015] for a general review of forensic practices in STR DNA analysis, and Steele and Balding [2014b] for a review specific to low-template DNA analysis.

Chapter 2

Verifying likelihoods for low template DNA profiles using multiple replicates

Work in this chapter has been published in Steele et al. [2014a], see Appendix B. I generated all of the laboratory CSPs, simulated all of the simulation CSPs and performed all analyses. The likeLTD discrete model used to evaluate the CSPs was developed by my supervisor, Prof. David Balding.

2.1 DNA profiling replicates

As described in Chapter 1, multiple profiling runs of LTDNA samples are often performed to assess the stochastic effects that differ between epgs, such as dropin, dropout and stutter [Steele and Balding, 2014b]. Joint likelihoods for multiple replicates are obtained by assuming that the replicates are independent conditional on the genotypes of all contributors and the model parameters, ϕ, such as the amounts and degradation levels of DNA from each contributor [Curran et al., 2005], given in (1.2).

There is currently no consensus on an approach to verify the validity of an implementation of a model to calculate LTDNA LRs (ltLRs). One possible approach is to evaluate the ltLR for the CSP, but to repeatedly replace Q with a randomly generated profile [Gill and Haned, 2013]. With a random profile as Q, H_p is almost always false, so the majority of computed ltLRs are expected to be small. A method proposed by [Taylor et al., 2015] extends that of Gill and Haned to verify that the mean ltLR over all random profiles for Q is 1.0 when H_d is true, as stated by Alan Turing [Good, 1950]. This method gives an indication of the validity of a model, but is primarily concerned with the reliability of a specific LR. In this chapter, a method is explored which instead proposes a performance indicator for ltLR algorithms when H_p is true, rather than when H_d is true. Under H_d, it may occur that $\mathcal{G}_X{=}\mathcal{G}_Q$; this occurs with probability $\pi_Q{=}\Pr(\mathcal{G}_X{=}\mathcal{G}_Q)$, the

match probability for Q. Since $\Pr(E|H_d, \mathcal{G}_X{=}\mathcal{G}_Q) = \Pr(E|H_p)$, it follows that [Cowell et al., 2015]

$$\text{ltLR} = \frac{\Pr(E|H_p)}{\Pr(E|H_d, \mathcal{G}_X{=}\mathcal{G}_Q)\pi_Q \ + \Pr(E|H_d, \mathcal{G}_X{\neq}\mathcal{G}_Q)(1{-}\pi_Q)} \leq \frac{1}{\pi_Q}. \tag{2.1}$$

Therefore $1/\pi_Q$, the IMP, acts as an upper bound on the LR. It is then possible to describe an LR in terms of a fraction of the IMP, which in this chapter and subsequent chapters will be termed the information gain ratio (IGR) and defined as $\log_{10}(\text{LR})/\log_{10}(\text{IMP})$. This measure normalises the LR between different CSPs, and between different queried contributors, so that the maximum IGR is 1.0, and support for H_d is negative. Other measures considered were the log ratio, $\log_{10}(\text{LR/IMP})$, or the ratio, LR/IMP. The log ratio measure has maximum 0 and supports H_d at $< \log_{10}(\text{IMP})$. The ratio measure has maximum 1.0, and supports H_d at $< 1/\text{IMP}$. With both alternative measures, support for H_d depends on IMP, and so both are less suitable for comparing across CSPs and queried contributors.

If Q is the major contributor to a good-template profile, then E implies that $\mathcal{G}_X{=}\mathcal{G}_Q$ and equality should be achieved in (2.1). If instead Q is the major contributor to a LTDNA profile, if H_p is true then increasing numbers of LTDNA replicates should provide increasing evidence that $\mathcal{G}_X{=}\mathcal{G}_Q$, so the ltLR should converge to the IMP, and the IGR should converge to 1.0. Even for mixtures the ltLR should approach the IMP, since differential dropout rates should allow for deconvolution of the alleles of Q from multiple replicates. However, the ltLR may be prevented from approaching the IMP due to possible inadequacies in the mathematical model or approximations being amplified with additional replicates. This expected convergence of the ltLR towards the IMP as the number of replicates increases can be used as an indicator of the validity of an algorithm to compute the ltLR when Q is the major contributor.

For a minor contributor Q, it may not be possible to determine the genotype of Q, even with many replicates, so the ltLR may not reach the IMP. However, the bound, (2.1), is still valid for minor contributors, so the ltLR should continue to approach the IMP with additional replicates. To give some indication of the extra information available through multiple replicates, the ltLR can be compared to the LR for a high-quality mixture where all contributors are fully observed (mixLR, or mixIGR) computed using only the presence or absence of alleles, so no uncertain designations [Weir et al., 1997].

Suppose the hypotheses for $\mathcal{C}$=ABC, $\mathcal{G}_Q$=AB, are of the form:

$$H_p\text{: } Q + U,$$
$$H_d\text{: } X + U.$$

Then the mixLR is:

$$\begin{aligned}\text{mixLR} &= \frac{\sum_{j=1}^{n_p} \Pr(\mathcal{C}{=}\text{ABC},\mathcal{G}_Q{=}\text{AB}|Q,U_j)\Pr(U_j|\mathcal{G}_Q)}{\sum_{j=1}^{n_d} \Pr(\mathcal{C}{=}\text{ABC},\mathcal{G}_Q{=}\text{AB}|X_j,U_j)\Pr(X_j,U_j)} \\ &= \frac{\Pr(\mathcal{G}_U \text{ is one of AC,BC,CC})}{\Pr((\mathcal{G}_X,\mathcal{G}_U) \text{ is one of (AA,BC),(AC,BB),(AB,CC),(AB,AC),(AB,BC),(AC,BC))}}, \quad (2.2)\end{aligned}$$

where within-pair ordering is ignored in the denominator for simplicity. Under the standard population genetics model [Gill et al., 2006, 2012] and setting $F_{ST} = 0$, the mixLR is then:

$$\frac{\sum_{j=1}^{n_p} \Pr(\mathcal{C}{=}\text{ABC},\mathcal{G}_Q{=}\text{AB}|Q,U_j)\Pr(U_j|\mathcal{G}_Q)}{\sum_{j=1}^{n_d} \Pr(\mathcal{C}{=}\text{ABC},\mathcal{G}_Q{=}\text{AB}|X_j,U_j)\Pr(X_j,U_j)} = \frac{2p_A{+}2p_B{+}p_C}{12p_Ap_B(p_A{+}p_B{+}p_C)} < \frac{1}{2p_Ap_B}, \quad (2.3)$$

which is the same as (1.4), as the contributor ordering is not ignored here, and is less than the IMP = $1/2p_Ap_B$. See [Balding, 2005] and Chapter 1 for further details and examples. Because no peak height information is utilised to generate the mixLR, it can be thought of as the LR for an equal-contributions mixture where all contributors are good-template.

The ltLR with multiple replicates should not only reach the mixLR, due to identification of all alleles present in any contributor, but should also exceed the mixLR, because differential dropout rates should allow the alleles of different contributors to be deconvoluted, partially if not fully. In fact, subsampling has been proposed by Ballantyne et al. [2013] to enhance mixture deconvolution by generating divergent mixture ratios in distinct low-template replicates. This proposal is explored in Section 2.5.3 by evaluating a real-world multi-replicate CSP where each replicate was profiled with one of two different sensitivities. The more general behaviour of the ltLR in relation to the mixLR and IMP will be investigated throughout this chapter, utilising both laboratory-generated CSPs and simulated CSPs.

2.2 Experimental protocol

2.2.1 Laboratory replicates

Cheek swab samples were obtained from five volunteers, and 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. The samples were then quantified using the life Technologies Quantifiler® Human DNA Quantification kit as per the manufacturer's recommendations.

Each sample was serially diluted after extraction and then amplified using the AmpFℓSTR® SGM

Cond.	Contributor	Init. conc. ng μl^{-1}	Dilution (%)	Volume (μl)	Mass (pg)	Approx. cell equiv.	Cycles	Product (μl)	Formamide : ROX	F/ROX Mixture (μl)
(i)	B	31.0	1	1.6	500	83	28	1	17:1	9
(ii)	B	31.0	0.1	2.0	60	10				
(iii)	B	31.0	0.01	5.0	15	3				
(iv)	A	23.0	1	17.6	500	83	28	1	17:1	9
	C	18.1	0.1	16	30	5				
(v)	A	23.0	0.1	22.4	60	10	28	1	17:1	9
	C	18.1	1	22.0	500	83				
(vi)	A	23.0	0.1	2.7	60	10	28	1	17:1	9
	B	31.0	0.1	2.0	60	10				
	C	18.1	0.1	3.5	60	10				
(vii)	A	23.0	0.1	2.7	60	10	28	1	600 : 1	9
	B	31.0	0.1	2.0	60	10				
	C	18.1	0.1	3.5	60	10				
(viii)	A	23.0	0.1	2.7	60	10	28	9	366 : 1	11
	B	31.0	0.1	2.0	60	10				
	C	18.1	0.1	3.5	60	10				
(ix)	A	23.0	0.1	2.7	60	10	30	1	17:1	9
	B	31.0	0.1	2.0	60	10				
	C	18.1	0.1	3.5	60	10				

Table 2.1: Sample preparation (left) and genotyping protocol (right) for all conditions examined in the lab-based experiments (described in Table 2.2). Cond. gives the condition. Each condition was replicated eight times. The initial DNA concentration (column 3), dilution (column 4) and volume (column 5) generate approximately the DNA mass indicated in column 6, with approximate cell equivalent in column 7; concentration$\times$(dilution/100)$\times$ volume. Columns 8 and 9 show the number of PCR cycles and the volume of PCR product added to each well for the genotyping. Columns 10 and 11 show the ratio of Hi-Di^{TM} formamide to GeneScan^{TM} 400HD ROX^{TM} and the volume of the mixture added to each well.

Study type	# Conts.	Cond.	Contributions A : B : C (pg)	PCR cycles	Enhance. strat.	Hypotheses tested	
						H_p	H_d
Lab-based	1	(i)	0 : 500 : 0	28	-	Q (B)	X
		(ii)	0 : 60 : 0	28	-	Q (B)	X
		(iii)	0 : 15 : 0	28	-	Q (B) + dropin	X + dropin
	2	(iv)	500 : 0 : 30	28	-	Q (A) + dropin	X + dropin
						Q (A) + U1	X + U1
						Q (C) + U1	X + U1
		(v)	60 : 0 : 500	28	-	Q (C) + dropin	X + dropin
						Q (C) + U1	X + U1
						Q (A) + U1	X + U1
	3	(vi)	60 : 60 : 60	28	-	Q (A) + U1 + U2	X + U1 + U2
		(vii)	60 : 60 : 60	28	Phase 1	Q (A) + U1 + U2	X + U1 + U2
		(viii)	60 : 60 : 60	28	Phase 2	Q (A) + U1 + U2	X + U1 + U2
		(ix)	60 : 60 : 60	30	-	Q (A) + U1 + U2	X + U1 + U2

Table 2.2: Experimental conditions and hypotheses compared to investigate replication in the laboratory. Cond. gives the condition index from Table 2.1, Contributions gives the approximate DNA contributions for donors A, B and C in pg. Enhance. strat. gives the enhancement strategy for the condition. Under Hypotheses tested Q denotes the queried contributor, who is one of A, B or C as indicated in parentheses, X is an unknown alternative to Q under H_d, while U1 and U2 are unknown contributors under both H_p and H_d.

Study type	# Conts.	Pr(D) A : B : C	Pr(C)	Pr(unc) $\upsilon \sim$Pois($\lambda = 1$)	Hypotheses tested	
					H_p	H_d
Simulation	1	1.0 : 0.0 : 1.0	0.00	-	Q (B)	X
		1.0 : 0.4 : 1.0	0.05	-	Q (B) + dropin	X + dropin
		1.0 : 0.8 : 1.0	0.05	-	Q (B) + dropin	X + dropin
		-	-	0.8	Q (B)	X
		-	-	0.4	Q (B)	X
	2	0.2 : 1.0 : 0.8	0.00	-	Q (A) + dropin	X + dropin
					Q (A) + U1	X + U1
					Q (C) + U1	X + U1
		0.2 : 1.0 : 0.6	0.00	-	Q (A) + dropin	X + dropin
					Q (A) + U1	X + U1
					Q (C) + U1	X + U1
	3	0.8 : 0.5 : 0.2	0.00	-	Q (A) + U1 + U2	X + U1 + U2
		0.5 : 0.5 : 0.5	0.00	-	Q (A) + U1 + U2	X + U1 + U2
		0.2 : 0.5 : 0.8	0.00	-	Q (A) + U1 + U2	X + U1 + U2

Table 2.3: Simulation parameters and hypotheses compared to investigate replication *in silico*. Pr(D) denotes the probability of dropout for a heterozygote allele for donors A, B and C; where Pr(D)=1.0, that contributor was not included in the simulation. Pr(C) denotes the probability of dropin. Pr(unc) indicates the probability of designating a CSP allele as uncertain. υ indicates the number of uncertain dropins per locus per replicate; see text for further details. Q denotes the queried contributor, who is one of A, B or C as indicated in parentheses. X is an unknown alternative to Q under H_d, while U1 and U2 are unknown contributors under both H_p and H_d. Profiles were simulated from the profiles of the same donors that were investigated in the laboratory (see Tables 2.1 and 2.2).

Study type	# Conts.	Sampled replicates	Hypotheses tested	
			H_p	H_d
Real-world	≥ 3	Standard and sensitive	Q + U1 + U2	X + U1 + U2
		Standard only	Q + U1 + U2	X + U1 + U2
		Sensitive only	Q + U1 + U2	X + U1 + U2

Table 2.4: Sampling strategy and hypotheses compared to investigate replication for a real-world crime sample. Q denotes the queried contributor, X is an unknown alternative to Q under H_d, while U1 and U2 are unknown contributors under both H_p and H_d. The sampled replicates were generated in the course of casework investigation of a real crime.

Plus® PCR kit as per the manufacturer's recommendations on a Veriti® 96-Well Fast Thermal Cycler.

An ABI 3130 Sequencer was used to analyse 1 µl of the PCR products, with 10 second injections at 3 kV; these settings were used for all subsequent analyses. The results returned from the 3130 sequencer were analysed using GeneMapper® ID v3.2 to determine which samples were suitable for further use.

For the one-contributor investigation eight replicates of each of three conditions were created (Table 2.1). The conditions were created to investigate increasing dropout rate. For the 500 pg and 60 pg conditions, one-contributor hypotheses were compared, B under H_p and X under H_d, while for the 15 pg condition dropin was also modelled under both hypotheses (Table 2.2). Note that throughout this chapter, the DNA quantity refers to approximate DNA quantity per replicate.

For the two-contributor investigation eight replicates of each of two conditions were created (Table 2.1). The major and minor contributors were reversed between conditions, with an increased DNA contribution from the minor. These samples were amplified and analysed as described previously. Two-contributor hypotheses were compared, with each of A and C in turn playing the role of Q, while the other contributor was treated as unknown. Additionally one-contributor-plus-dropin hypotheses were compared, with only the major contributor playing the role of Q (Table 2.2).

For the three-contributor investigation eight replicates of each of four conditions were created (Table 2.1). The conditions were created to investigate different profiling protocols. The Phase 1 and Phase 2 conditions are post-PCR purification protocols designed to enhance the sensitivity of detection of the standard protocol [Roeder et al., 2009], and both involve concentrating the post-PCR product using an Amicon® PCR microcon unit according to the manufacturer's recommendations. Phase 1 enhancement increases the amount of formamide in the mixture compared to the manufacturer's recommendations, while Phase 2 enhancement increases the amount of DNA, formamide and ROX compared to Phase 1. For all four conditions (30 cycles, 28 cycles, Phase 1, and Phase 2), three-contributor hypotheses were compared, with A playing the role of Q and the other contributors treated as unknown (Table 2.2). Dropin was not modelled

under either hypothesis, although dropin was included in the simulations. This reflects a realistic challenge for few replicates with multiple contributors, whereby any dropin alleles may be wrongly attributed to one of the contributors. However the incorrect model will lead to deterioration of inferences for larger numbers of replicates.

2.2.2 Simulated replicates

All of the conditions described in this section were simulated in eight replicates, with the whole simulation being performed five times. Initially a number of single-contributor CSPs were simulated using the profile of individual B. The first condition investigated was a "perfect match", in which all eight replicates generated exactly the profile of B. Next mild dropout (Pr(D)=0.4) and severe dropout (Pr(D)=0.8) of the alleles of B were introduced, in each case with dropins included at rate Pr(C)=0.05 (at most one dropin per locus per replicate). The homozygous dropout probability was set equal to $\Pr(D)^2/2$, as suggested by [Balding and Buckleton, 2009]. The effect of uncertain allele designations was then investigated by randomly designating some alleles of B as uncertain, first with Pr(unc)=0.4 and then Pr(unc)=0.8. In both conditions, at each locus and in each replicate a Poisson mean one number of alleles not in the profile of B was also designated as uncertain, with types randomly selected according to frequencies in the UK Caucasian database. For all these simulated profiles, one-contributor hypotheses were compared, B under H_p and X under H_d.

Next two-contributor CSPs were simulated, based on the profiles of A and C. Two conditions were simulated, both used $\Pr_A(D) = 0.2$, while $\Pr_C(D)$ was initially 0.8 and then 0.6. Dropin was not simulated. For shared alleles the dropout probability was the product of the dropout probabilities for each contributor having that allele. Two-contributor hypotheses were compared, with each of A and C in turn taking the role of Q, while the other was treated as unknown in the analysis. Additionally one-contributor-plus-dropin hypotheses were compared, only for A playing the role of Q (Table 2.3).

Three-contributor CSPs were then simulated under three conditions, with dropout probabilities for Donors A, B and C as shown in Table 2.3. Dropin was included as for the one-contributor simulations. Three-contributor hypotheses were compared, with A playing the role of Q and the other two contributors being treated as unknown.

2.2.3 Crime case replicates

A CSP from an actual crime investigation was explored, consisting of five replicates: two using standard SGM+ profiling and three generated using a low copy number (LCN) protocol with 34 PCR cycles. Only

Locus	Sensitive profiling			Standard profiling	
	Run 1	Run 2	Run 3	Run 4	Run 5
D3	16, [15]	16, [15]	16, 18, [15]	16	16
vWA	15, 16, [17]	15, [14]	15, 18, [14]	15	15
D16	9	9	9, 11, [10]	9	9
D2	17, 19, 24	16, 17, 24,[23]	17, [16]	24	24
D8	8, 13, 15, 16	8, 12, 13, 16, [15]	8, 13, 14, 16, [15]	[8]	
D21	30, 32, 33.2	32, 32.2, 33.2	32, 32.2, 33.2, 34, [31]	[32], [32.2]	[33.2]
D18	12, 17	12, 17, 19	12, 17, [11], [16]	[17]	17
D19	14, 21, [13]	11, 14, [13]	14, [13]	14	14
TH01	6, 9.3	6, 9.3	6, 8, 9.3	[6], [9.3]	[6]
FGA	21	21, [20]	21, 20	21	

Table 2.5: Five replicates of a crime scene profile, three from a sensitive LTDNA profiling technique and two from standard DNA profiling. Alleles shown in [ ] were called as uncertain.

limited information about the profiling protocol was provided by the profiling lab. Extraneous details are not required by likeLTD because it estimates the unknown parameters from the CSP allele designations. The five actual replicates were re-sampled to generate simulated profiles with up to eight replicates, consisting of standard replicates only, sensitive replicates only, or both. Six distinct alleles were observed at locus D8, but no more than three replicated alleles were observed at any locus, so the minimum number of contributors is three. Therefore, three-contributor hypotheses were compared, with all contributors unknown under H_d, and no dropin (Table 2.4).

2.3 Single-contributor results

2.3.1 Lab-based

For the good-template experiments (500 pg) full information (IGR=1.0) is obtained with just a single replicate, and the IGR does not exceed one with the addition of replicates two through eight (Figure 2.1, left, red). This is expected for a good-template single-contributor sample, and demonstrates that there is no deterioration in the modelling assumptions with a large number of replicates in this simple scenario.

Low DNA template (60 pg) reduces the IGR to approximately 0.9 at one replicate, however, the IGR is close to 1.0 at two replicates, and IGR≈1.0 with subsequent additional replicates (Figure 2.1, left, purple), but never exceeds IGR=1.0. The total DNA contribution at eight replicates is roughly equal to that of a single replicate of the good-template condition (480 pg for eight replicates of the 60 pg, 500pg for a single replicate of the 500 pg), and both obtain full information at this level of total DNA. The low DNA template results suggest that full information should be available from a single replicate of 120pg DNA, equalling the

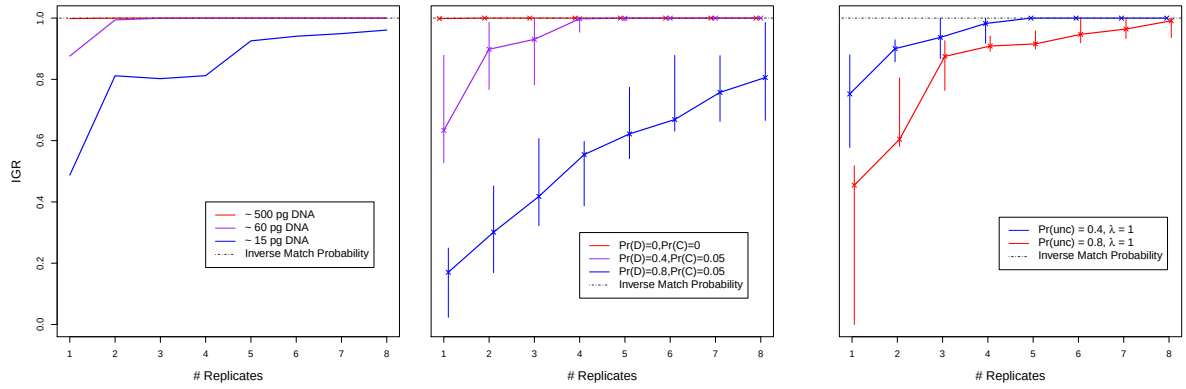


Figure 2.1: The low-template information gain ratio (ltIGR) from one-contributor CSPs evaluated using from one up to eight replicates. Left: lab-based replicates, with DNA template (in pg) as shown in the legend box. Middle: simulated replicates with dropout (probability Pr(D)) and dropin (probability Pr(C)); the plotted points represent the median from five repetitions of the simulation, and the vertical bars show the range. Right: simulated replicates with uncertain allele calls probability=Pr(unc) for a true allele to be uncertain, and a Poisson (rate λ=1) number of non-alleles labelled as uncertain at each locus.

DNA contribution from 20 cells (one cell $\approx$ 6 pg DNA).

For very low DNA template (15 pg) IGR $\approx$ 0.5 for a single replicate, which constitutes between two and three cells worth of DNA (Figure 2.1, left, blue). Replicate profiling brings the IGR substantially closer to 1.0, but not entirely, with IGR $\approx$ 0.95 at eight replicates. Note that the 60 pg condition (purple) was able to reach IGR$\approx$1.0 with 120 pg total contribution (two replicates), while 15 pg condition (blue) was unable to reach IGR=1.0 at 120 pg total contribution (eight replicates). At four replicates the approximate total DNA contribution is equal to that of a single replicate of the low DNA template condition (60 pg); the IGRs for each are similar, but with a lower IGR for the replicated very low template condition ($\approx$0.9 for 1$\times$60 pg, $\approx$0.8 for 4$\times$15 pg). Similarly, at eight replicates the total DNA contribution is equal to the two replicates 60pg condition, with a similar IGR but slightly lower in the condition with more replicates ($\approx$1.0 for 2$\times$60 pg, $\approx$0.95 for 8$\times$15 pg). Comparisons of experiments with approximately equal DNA contributions, such as these, are analogous to investigating pre-extraction replicates, where a sample that contains x pg DNA is split into n replicates of x/n pg DNA. All such scenarios here return a lower IGR in the condition with more replicates, suggesting that the reduced per-replicate DNA contribution introduces increased stochasticity that decreases the IGR from the same total DNA contribution. Alternatively, the difference in IGR may result from pipetting variability, as the sample size here is small, with just three paired comparisons. The scenario of splitting a sample into replicates will be further investigated in Chapter 7 using a continuous model for analysis.

2.3.2 Simulation

Similar behaviour is seen in the simulation studies as was observed in the laboratory studies. The median IGR rises to 1.0 with a small number of replicates, but does not exceed it (Figure 2.1, middle) for both the perfect match (Pr(D)=0, red) and mild dropout (Pr(D)=0.4, purple) conditions. For the severe dropout (Pr(D) = 0.8, blue) the median IGR rises towards 1.0 but does not reach it by eight replicates. By using IGR to display the WoE, it is possible to estimate a Pr(D) that roughly corresponds to a DNA contribution; Pr(D)=0.4 is somewhere between 15 and 60 pg DNA, while Pr(D)=0.8 is equivalent to less than 15 pg DNA. However, the Pr(D) implemented for these simulations was uniform across alleles, whereas in reality Pr(D) is expected to increase with allele length in base pairs due to the effects of degradation. Degradation would be limited when considering laboratory-generated samples such as those in Section 2.3.1.

As described in Chapter 1, the availability of uncertain allele designations is a novel feature of likeLTD, which mitigates the problem of choosing a detection threshold, as highlighted by [Budowle et al., 2009], as an all-or-nothing call is no longer necessary. IGR≈1.0 is reached at five and eight replicates for the low and high rates of uncertain calls respectively (Figure 2.1, right), despite up to 80% of true alleles being designated as uncertain and inclusion of multiple uncertain non-alleles. Neither condition exceeds the bound of IGR=1.0, even with many replicates. However, Pr(unc) was uniform across alleles, but should vary with allele length in a real-world scenario as calling a peak as either uncertain or dropout depends on the peak height, which is affected by degradation and/or amplification efficiency.

2.4 Two-contributor results

2.4.1 Lab-based

When the minor contributor of a laboratory generated two person mixture provides only 30 pg of DNA (Figure 2.2, top left panel), then if Q is the major contributor the IGR is very close to 1.0 for all numbers of replicates (solid and dashed blue lines), whereas if Q is the minor contributor then the IGR remains substantially lower than 1.0, even at eight replicates (solid red line). However, even with this very low template for the minor, the ltIGR exceeds the mixIGR with six or more replicates. When the major and minor contributors are reversed, and the amount of DNA from the minor is doubled (Figure 2.2, bottom left), then if Q is the minor contributor the ltIGR exceeds mixIGR from six replicates again, and rises to within 0.2 of IGR=1.0 at eight replicates. Under both conditions, the two-contributor analysis gives a very similar result to the one-contributor-with-dropin analysis.

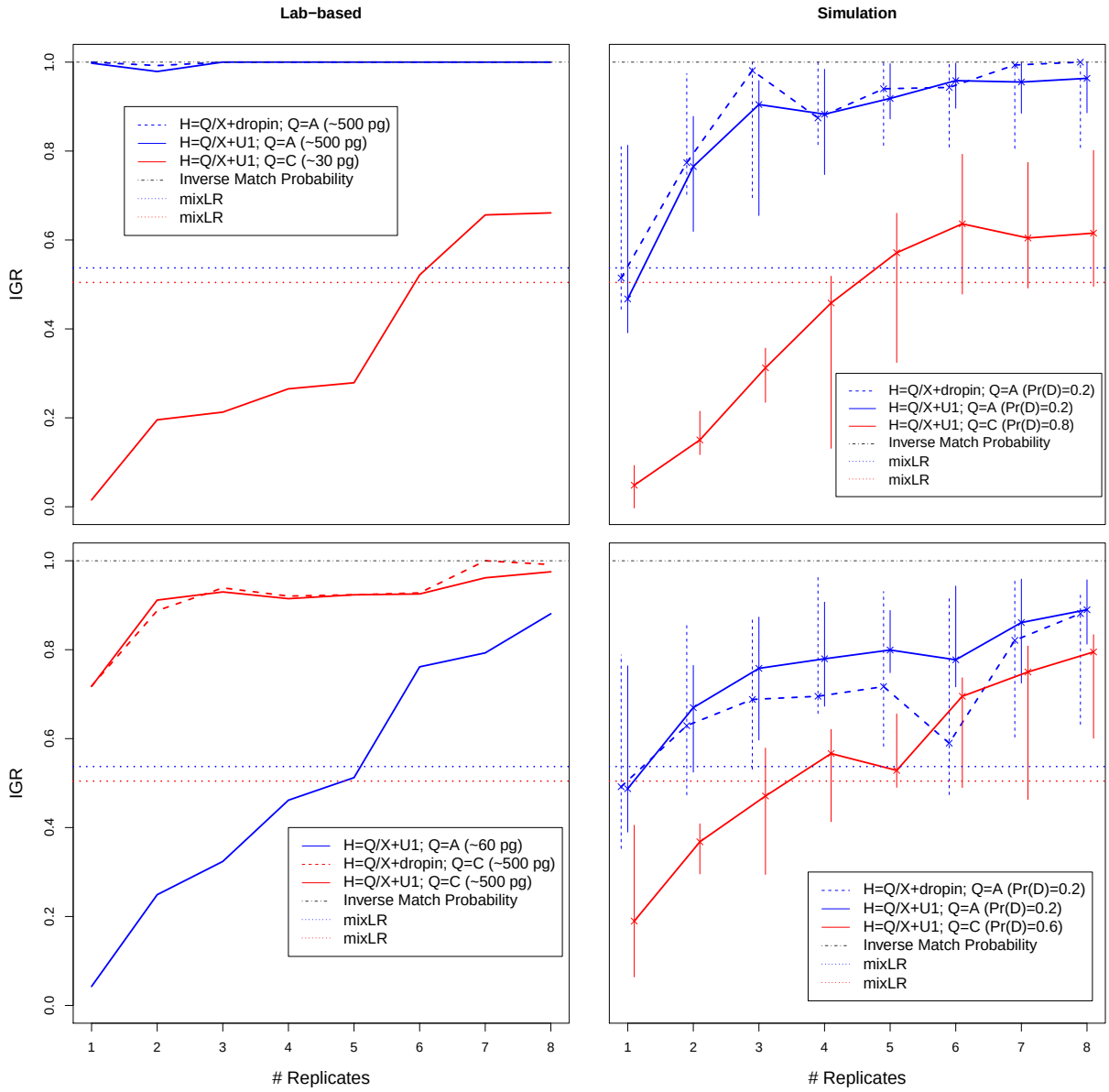


Figure 2.2: The low-template information gain ratio (ltIGR) from two-contributor CSPs profiled at up to eight replicates. Left: lab-based replicates, with the DNA template from the minor contributor greater in the lower panel (see legend boxes). Right: simulation-based replicates, with the minor contributor having reduced dropout in the lower panel. The simulated CSPs were generated from the profiles of Donors A and C, and the line colours on the graph indicate whether the queried individual (Q) is A (blue) or C (red). Solid lines indicate a two-contributor analysis, with the non-Q individual regarded as unknown (U1). Dashed lines indicate a one-contributor analysis that also allows for dropin (only for Q the major contributor). The IMP is shown with a grey dot-dash line at IGR=1.0. The IGR equivalent to mixLR is shown with dotted lines, coloured according to Q. In the legend boxes, H indicates the hypotheses with X an unknown alternative to Q, and Pr(D) indicates the probability of dropout.

At eight replicates of 30 pg the total DNA contribution from donor C (red) is ≈ 240 pg, and yet the ltIGR is similar to that of a single replicate when the donor C contributes 500 pg of DNA from a single replicate (both IGR≈0.7). Conversely at eight replicates of 60 pg the total DNA contribution from donor A (blue) is ≈ 480 pg, similar to the 500 pg at a single replicate when donor A is the major contributor, and the IGR for multiple replicates is lower than that for a single replicate (IGR for 1×500 pg≈1.0, IGR for 8×60 pg≈0.9). This suggests that deconvoluting the major and minor contributors is more difficult when they contribute a more similar amount of DNA (minor contributes ∼6% of DNA at 30 pg, and ∼11% of DNA at 60 pg).

2.4.2 Simulation

When the minor contributor of a simulated two person mixture is subject to high dropout (Figure 2.2, top right), then if Q is the major contributor the ltIGR exceeds the mixIGR with two or more replicates, with the median ltIGR rising rapidly to approximately 0.9 IGR with three replicates, but rising towards 1.0 only slowly with additional replicates. The one-contributor-plus-dropin analysis gives ltIGRs that are broadly similar to the two-contributor analysis, but with a wider range, especially at many replicates, indicating greater variability; with many replicates many of the minor contributors alleles will be replicated, but are being explained as dropin with the one-contributor-plus-dropin analysis. If Q is the minor contributor, the median ltIGR increases rapidly from a low base, and stabilises after about five replicates, at approximately 0.6 IGR, which exceeds the mixIGR by approximately 0.1. The range increases after three replicates, and remains high up to eight replicates. The expected number of observed alleles for a fully heterozygous genotype is 16 for the major and 4 for the minor, so the minor alleles are unlikely to mask the major alleles.

With reduced dropout for the minor contributor (Figure 2.2, bottom right), deconvoluting the genotype of the major contributor Q is harder because of additional masking by alleles of the minor contributor. The median ltIGR in both the two-contributor and one-contributor-plus-dropin analyses reaches ≈ 0.9 IGR at eight replicates, with the latter showing a greater range again. Conversely, the lower dropout rate leads to improved inference for a minor contributor Q, with the median ltIGR rising to ≈ 0.8 IGR at eight replicates, which exceeds the mixIGR from four replicates onwards. Interestingly, from six replicates onwards the range of the minor contributor ltIGR overlaps the range for the major contributor. The expected number of observed alleles for a fully heterozygous genotype are 16 for the major, and 8 for the minor, approximately doubling the probability that the minor may share an allele with the major.

Due to the interplay between the two contributors' genotypes, through masking, it is more difficult

to infer what Pr(D) equates to what DNA contribution, however, Pr(D)=0.2 is equivalent to less than 500 pg DNA. Pr(D)=0.8 is approximately equivalent to 30 pg DNA, while it is difficult to estimate what DNA contribution Pr(D)=0.6 is equivalent to.

The one-contributor-plus-dropin analyses show a good approximation to the two-contributor analyses, but with a greater divergence than was seen in the laboratory-generated CSPs.

The genotypes were simulated with a single Pr(D) for all alleles, when in reality Pr(D) is altered by allele length acting through degradation, which likeLTD assumes when calculating the WoE. Therefore the model used to generate the data simulated here does not match the model used to evaluate the WoE, which may have caused some of the high variances, the lack of reaching the IGR, or the divergence between the one-contributor-plus-dropin and two-contributor analyses.

2.5 Three-contributor results

2.5.1 Lab-based

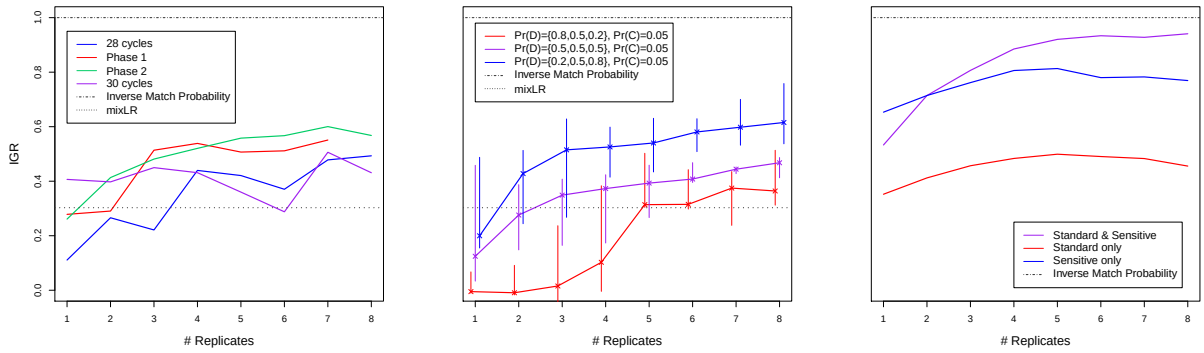


Figure 2.3: The low-template information gain ratio (ltIGR) for three-contributor crime stains profiled with one to eight replicates. Left: laboratory replicates using four lab techniques indicated in the legend box and described further in Materials and Methods. Middle: simulated replicates with dropout rates for the three contributors as shown in the legend box against Pr(D), the first value being for the queried contributor. Pr(C) is the dropin probability. Right: re-sampled actual crime-stain replicates; the original data are two standard profiling replicates, and three replicates using enhanced sensitivity. The ltIGR returned from a perfect replicate of the contributors (consisting of every allele from each contributor) is shown with dotted lines; this is not possible for the real-world case, as the true contributors are unknown.

The laboratory-based three-contributor CSPs were generated specifically to compare different LTDNA profiling techniques, which differs from the laboratory investigations presented previously. The three contributors were included at equal DNA contributions, approximately 60pg each, to make this test the most

challenging possible, as the genotypes of different individuals are difficult to deconvolve when the contributors are at equal contributions.

The 30 PCR cycles condition gives an ltIGR of ∼0.4 at one replicate but little improvement with additional replicates, with an ltIGR at eight replicates of ∼ 0.4 (Figure 2.3, left, purple); this exceeds the mixIGR for all evaluations other than the six replicates evaluation. The ltIGR does increase with increasing numbers of replicates for the other amplification methods, but in no case did the ltIGR exceed 0.6. As expected, the ltIGR for both Phase 1 (red) and Phase 2 (green) enhancement exceeds that for standard 28 PCR cycles across the range of replicates, with Phase 2 enhancement typically giving a slightly larger ltIGR than Phase 1 enhancement, and both exceeding the ltIGR with three or more replicates. The 28 PCR cycles (blue) ltIGR exceeds the mixIGR with four or more replicates.

These results suggest that 28 cycle PCR (regardless of enhancement) is preferable to 30 cycle PCR beyond one replicate. It is known that increasing the number of PCR cycles introduces more stochasticity in the results, as stated in the AmpFℓSTR® SGM Plus® PCR Amplification Kit user guide. Post-PCR enhancement provides extra information over an unenhanced sample, with Phase 2 enhancement providing a small further improvement over Phase 1. These results support those of Forster et al. [2008], who demonstrated that increasing PCR cycles increases the size of stutter peaks and the incidence of dropin; the observation here of no improvement in the WoE for 30 PCR cycles is possibly due to these increased stochastic effects.

2.5.2 Simulation

All three curves in Figure 2.3 (middle) show an increasing ltIGR with an increasing number of replicates, with the median ltIGR being in the expected order throughout (median ltIGR for low>medium>high dropout rates). The median ltLR exceeds the mixIGR after one, two and four replicates for the low, medium and high dropout conditions respectively. The ltIGR is strongly dependent on the details of the specific simulation, leading to a high range for the ltIGR; in particular the degree of allele sharing across simulated contributors.

2.5.3 Real-world case

The standard-only and sensitive-only ltIGRs show a similar trend, increasing up to five replicates before falling slightly, with the standard-only ltIGR being approximately 0.2 below the sensitive-only ltIGR throughout (Figure 2.3, right). The ltIGR using both standard and sensitive replicates exceeds that for both the other conditions after two replicates, reaching ltIGR>0.9 after eight replicates. This may be due to the lim-

ited pool of replicates available from the actual crime case, but suggests that employing different sensitivities in the profiling replicates may allow different contributors to be better distinguished, and so may result in a higher ltIGR than using the same number of replicates but with a single sensitivity.

2.6 Overview

In all conditions tested, the ltIGR has been bounded at 1.0, so the likeLTD ltLR has been bounded by the IMP, as predicted by (2.1). The mathematical model underlying the dropout model (see Chapter 1), and its implementation in likeLTD are validated by the tight bounding of the ltLR when a high-level contributor is queried (Figures 2.1 and 2.2). Moreover, these results demonstrate that the full genotype of a contributor can be deconvoluted from a mixture using multiple noisy profiling replicates [Schneps and Colmez, 2013], rather than detrimentally compounding the noise from the replicates.

Furthermore, the mixLR was exceeded in all 19 of the tested conditions for which a mixLR could be computed, and was often exceeded after only a small number of replicates. The inference is that a single replicate good-quality profile correctly representing the alleles of all contributors (mixLR) provides weaker evidence than multiple LTDNA replicates (ltLR) in all conditions. This is because differential dropout rates between replicates allow different contributors to be partially distinguished, which is not possible with a fully represented high-quality sample, supporting the proposition of Ballantyne et al. [2013] to perform multiple replicates at divergent mixture ratios.

2.6.1 Use of replicates

Pfeifer et al. [2012] advocate the use of multi-replicate CSPs to overcome the inherent variability of LTDNA analysis, while Grisedale and van Daal [2012] instead support performing a single profiling run with as much DNA as is available. Note that the conclusion of Grisedale and van Daal is based on a comparison with analysing a consensus sequence obtained from multiple replicates, which is a less efficient approach than analysing all replicates individually in a single CSP. While the results presented throughout this chapter demonstrate increased IGR with increasing replicates, and therefore support replication, this does not inform on the relevant question; both Pfeifer et al. and Grisedale and van Daal were commenting on the use of a single run of x pg DNA versus the use of n replicates each with x/n pg DNA. The small number of such comparisons that are available from the data in this chapter, six, suggest that the process of splitting a sample into replicates reduces the information content of the CSP (replicated IGR$\leq$unreplicated in five out of six

comparisons), however, the low number of comparisons and small differences mean that that this may be noise. Further results are presented in Chapter 7 that directly investigate the effect of splitting a sample into multiple replicates. Post-extraction replication often does not require splitting a sample because standard extraction produces a volume of extract that is a few times larger than is optimal for PCR allowing for a few replicates (between four and six); the results presented here support the use of post-extraction replicates to maximise the information available in a CSP. However there are strategies that may be employed that remove this availability of post-extraction replicates. It is possible to purify the extract either 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], after which it may be possible to run the whole extract through PCR. However, these methods may be unsuitable for a mixture with a good-template major contributor and one or more low-template contributors, as enhancing the signal for the minor contributors may lead to oversaturation of the major contributor.

Eight replicates were used here to rigorously test the behaviour of the ltIGR returned by likeLTD in relation to the IMP and mixIGR. Taberlet et al. [1996] have suggested that seven replicates are required for low-template samples to generate a high-quality profile, however, seven replicates are rarely available in real world low-template crime samples [Budowle et al., 2009].

2.7 Improvements

As mentioned previously, Pr(D) and Pr(unc) in reality will vary with the fragment length of the dropout or uncertain allele if degradation has occurred. However, the Pr(D) and Pr(unc) used to simulate CSPs throughout were uniform across fragment lengths. Similarly, if degradation has occurred, Pr(unc) should behave differently whether an observed true allele is being called as uncertain or whether an unobserved false allele is being inserted as uncertain; Pr(unc) for a true allele should increase with fragment length, while Pr(unc) for a false allele should decrease with fragment length. A future study that wished to implement fragment length adjusted probabilities of dropout and uncertains would need to include conditions with multiple levels of simulated degradation to investigate how degradation affects the results observed here.

Chapter 3

Worldwide F_{ST} estimates relative to five continental-scale populations

Work in this chapter has been published in Steele et al. [2014b], see Appendix B. I performed data munging, all analyses, and coded the direct method from preliminary code provided by Prof. David Balding. The data were collected and provided by Dr. Denise Syndercombe Court. The indirect method was implemented in BayesFST by Prof. David Balding.

3.1 F_{ST} in forensics

F_{ST} adjustments are widely used in forensic genetics during analyses of mixed and low-template DNA profiles, and can have a substantial impact on the WoE. As described in Chapter 1, an F_{ST} adjustment can be formulated using the sampling formula, (1.13), to account for the fact that Q may share alleles with X due to shared ancestry, rather than because Q contributes to the CSP. This is a direct interpretation of the F_{ST} parameter, measuring genetic variability between subpopulations compared to the genetic variability in the total population. It is also possible to view an F_{ST} adjustment as allowing for the fact that any available database will not fit the case circumstances exactly. This introduces extra uncertainty, which reduces confidence in any result from that database, so the F_{ST} adjustment reduces the LR with increasing F_{ST}. See Chapter 1 for further background to the formulation of the F_{ST} adjustment used in forensic work.

It is common in population genetics to estimate an F_{ST} value relative to some hypothetical ancestral population (see Figure 3.1). In forensic casework a database of allele probabilities is available from a population survey; in this case the most relevant F_{ST} value is relative to the surveyed database rather than to a hypothesised ancestral population. If a database is used that is directly appropriate for Q and X, then a small value of F_{ST} may be sufficient even when Q and X share a very similar ethnic background. The

most appropriate F_{ST} value increases with an increasing dissimilarity between the database used and the ethnic background of Q and X [Steele and Balding, 2014b]. Regardless of the fit of the database chosen to Q and X, a large value of F_{ST} is usually applied to any X that shares a population with Q, adhering to the maxim "innocent until proven guilty" by generating a conservative LR. A small value is applied when X and Q do not share a population, as they are expected to share little coancestry relative to the database chosen for X e.g. when Q is Caucasian and an Afro-Caribbean X is considered. F_{ST} accounts for the fit of the chosen database to the ethnicity of X, not Q. Issues surrounding the choice of an appropriate database for a necessarily unknown individual, X, are discussed in Chapter 4. Extra uncertainty is introduced into the forensic estimates of F_{ST} as each alternative contributor has a different ethnic background, and because F_{ST} is usually estimated at a scale that is not suitable for forensic analysis.

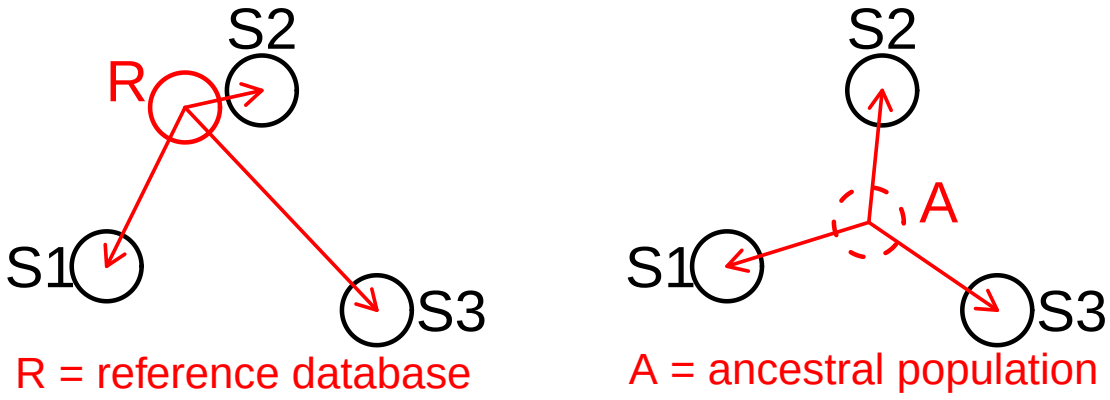


Figure 3.1: Visual representation of the difference between a forensic focussed (direct; left) and a population genetics focussed (indirect; right) F_{ST} formulation. S1-3 signify three subpopulations (black circles), A and R signify an ancestral population (dashed red circle) and a reference database (solid red circle) respectively, while red arrows signify genetic distance.

The origins of the study subjects in this Chapter are recorded at a national level, without reference to sub-national ethnic identities e.g. India is treated as a subpopulation of a broader South Asian population. This ignores the genetic variation among different ethnic groups within India, which can be substantial. This problem is most prevalent for countries that span large geographic areas e.g. native individuals from eastern and western Russia (>9000 km separation) will be more genetically distinct from each other than native individuals from northern Scotland and Southern England (965 km separation). As mentioned previously,

it is appropriate to consider a separate F_{ST} for each possible X. If instead a single F_{ST} value is employed, it should be taken from the upper tail of the distribution of F_{ST} across alternative contributors, to ensure that the prosecution is not unduly favoured. As a result, posterior 97.5 percentile estimates of F_{ST} will be utilised when considering forensic applications, while posterior median estimates will be utilised otherwise.

Two extensive studies estimating F_{ST} from human STR loci have previously been published, focussing on well-defined ethnic groups [Pemberton et al., 2013] and worldwide forensic databases [Silva et al., 2012]. The data presented in this chapter, and that in Silva et al. [2012], mainly constitute large ethnically mixed populations in contrast to Pemberton et al. [2013]. Here, F_{ST} is estimated at both within-continent and between-continent scales, and is estimated using both inferred (indirect) and observed (direct) reference populations. The estimates here provide posterior quantiles, and account for variable sample size through the use of likelihood based estimation. They are directly relevant to forensic casework, and aid understanding of human genetic variation at national, regional and continental scales in general populations.

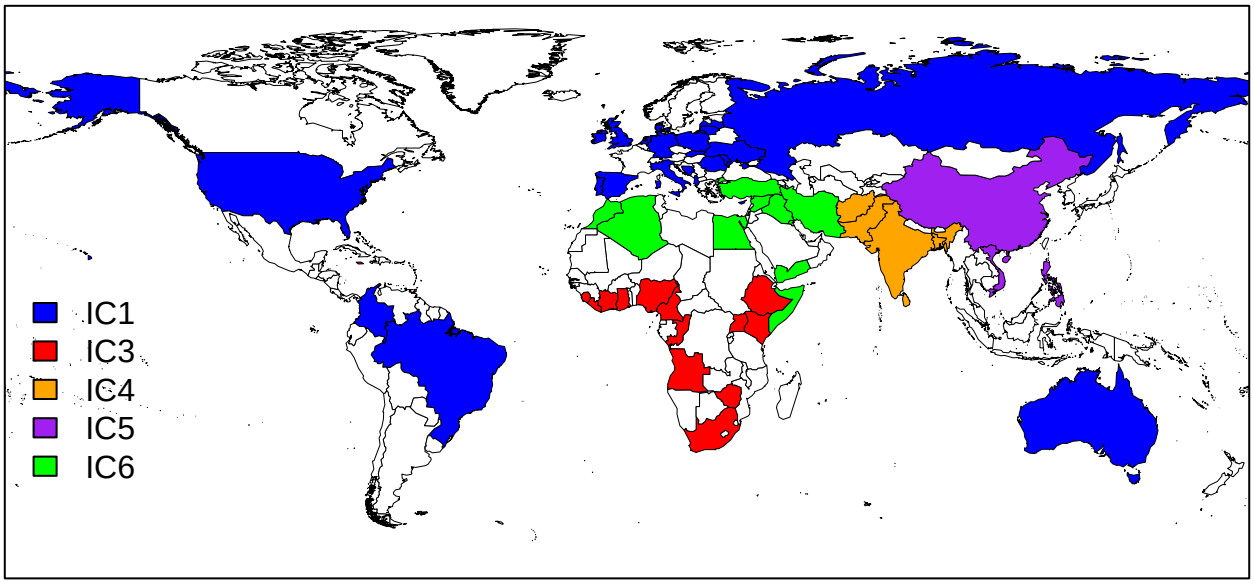


Figure 3.2: Countries of origin of the individuals included in the study, coloured according to the population that provides the best fit according to the indirect method (see text). White indicates countries represented by fewer than five individuals.

3.2 Dataset and munging

3.2.1 Database

The dataset used here includes the STR profiles of 7 121 individuals living in the UK or Eire, or applying to migrate to the UK on the basis of a either relatedness to or a relationship with a UK resident, which will be termed the DNA17 dataset. They are all genotyped by the same laboratory at up to 16 STR loci. The individuals are self-identified into one of six populations: White (IC1 and IC2, with IC2 including darker-skinned individuals of European origin), Black African/Caribbean (IC3), South Asian (IC4), East/South-East Asian (IC5) or Middle Eastern/North African (IC6). They are further classified into subpopulations, in most cases defined at the national level. The worldwide coverage of individuals is extensive (Figure 3.2), but some large populations are not included, such as Japan and Indonesia, while Latin America only has small sample sizes. The analyses here use only allele counts and not individual genotypes. In a few instances a single allele was observed at a locus because the peak intensity was insufficient to confirm homozygote status, so total allele counts are not always even integers (Table 3.1).

Observations	IC1	IC2	IC3	IC4	IC5	IC6	Total
D3S1358	7013	162	5200	704	625	226	13930
TH01	6953	158	5177	694	624	226	13832
D21S11	7006	162	5198	704	624	225	13919
D18S51	6944	157	5180	704	626	226	13837
D16S539	6951	162	5183	694	626	226	13842
VWA	7013	162	5194	704	626	226	13925
D8S1179	7007	162	5200	704	626	226	13925
FGA	6988	162	5196	700	626	226	13898
D19S433	6836	158	5122	687	621	226	13650
D2S1338	6575	152	4995	667	620	220	13229
D22S1045	1822	56	3478	523	506	162	6547
D1S1656	1835	56	3509	528	511	162	6601
D10S1248	1823	56	3497	516	506	118	6516
D2S441	1808	56	3458	521	501	160	6504
D12S391	1869	56	3531	551	507	162	6676
SE33	376	4	1039	308	396	140	2263

Table 3.1: Number of alleles typed per locus and population.

3.2.2 Data munging

All subpopulations with > 40 individuals sampled were included in the analyses. Some subpopulations of particular interest were also included despite having sample size < 40, while other subpopulations with small

sample sizes were removed or merged. Study participants self-identified both population and subpopulation labels, however, in some cases a different population classification was more appropriate for the identified subpopulation, and was changed accordingly (see below). These decisions introduce some subjectivity into the classification as no canonical classification scheme for human populations exists. Where possible, subpopulations with small sample sizes were combined on the basis of the United Nations geo-scheme for the relevant continent [United Nations Statistics Division, 2014].

IC1 and IC2

IC2 individuals from Europe were moved to IC1. Two national subpopulations were kept distinct, Eire and Great Britain, while the remaining European subpopulations were merged according to the United Nations geo-scheme for Europe [United Nations Statistics Division, 2014]:

Eastern Europe: Hungary, Moldova, Poland, Romania, Russia, Slovakia, Ukraine.

Northern Europe: Denmark, Latvia, Lithuania, Sweden.

Southern Europe: Albania, Bosnia, Croatia, Cyprus, Greece, Italy, Kosovo, Malta, Macedonia, Portugal, Spain, Yugoslavia.

Western Europe: Belgium, France, Germany, Netherlands.

IC2 individuals from Argentina, Bolivia, Brazil, Columbia, Mexico and Venezuela were combined ("Latin America"), as were IC1 individuals from Australia, New Zealand, and USA ("Anglo New World"). Those with no subpopulation identified, and those from Jersey, Northern Ireland or South Africa, were removed.

IC3

Six national subpopulations were kept distinct: Ghana, Jamaica, Kenya, Nigeria, Sierra Leone and Uganda. The following subpopulations were created from mergers according to the United Nations geo-scheme for Africa [United Nations Statistics Division, 2014], with Middle and Southern Africa combined as Central/Southern Africa:

Other W Africa: Benin, Gambia, Guinea, Guinea-Bissau, Ivory Coast, Liberia, Mali, Togo.

Other C/S Africa: Angola, Chad, Congo, Cameroon, South Africa.

Other E Africa: Burundi, Ethiopia, Eritrea, Malawi, Rwanda, Sudan, Tanzania, Zambia, Zimbabwe.

Other Caribbean: Barbados, Bermuda, Dominica, Guyana, Grenada, Monserrat, St Lucia, Virgin Islands, Trinidad.

Individuals with missing subpopulation were included as 'Unknown IC3'. Those with origin not in Africa or the Caribbean were removed (Eire, GB, USA). Algeria, Egypt, Morocco and Somalia were all included with IC6 (see Section 3.3.4).

IC4

Four national subpopulations were kept distinct: Afghanistan, Bangladesh, India, Pakistan. Individuals with missing subpopulation, or if the subpopulation was Nepal or Sri Lanka, were included as 'Unknown IC4'. Mauritius was removed.

IC5

SE Asian subpopulations were merged (Cambodia, Indonesia, Philippines, Thailand, Vietnam). Mongolia and South Korea were merged with the much larger China sample to form NE Asia. Fiji was removed.

IC6

Iran, Iraq, Somalia and Turkey were kept as separate national subpopulations. Other subpopulations were merged into N Africa (Algeria, Egypt, Morocco) or Middle East (Jordan, Kuwait, Lebanon, Palestine, Qatar, Syria, Yemen, UAE). Those from Georgia or with no subpopulation identified were removed. Afghanistan was moved to IC4.

Databases of STR frequencies at 10 loci were previously collated by the UK Forensic Science Service (FSS) [Foreman and Evett, 2001] in six populations with similar definitions to the DNA17 dataset presented here: EA1 (Caucasian), EA2 (Mediterranean), EA3 (Afro-Caribbean), EA4 (South Asian), EA5 (East Asian) and EA6 (Middle East/North Africa). These databases are small (<2 000 individuals combined) and do not include subpopulation labels. EA5 and EA6 both have sample sizes varying over loci, with the average sample size reported below. Based on the analyses presented in this chapter, the UK NDNAD chose to include the following populations in the UK DNA17 allele database:

NDU1: UK Caucasian (analogous to IC1 here).

NDU2: African + Afro Caribbean (analogous to IC3 here).

NDU3: South Asian (analogous to IC4 here).

NDU4: East Asian (analogous to IC5 here).

NDU6: African (subset of IC3 here).

NDU7: Afro Caribbean (subset of IC3 here).

EA1-6 were the reference databases used in most DNA forensics in the UK while the 10-locus SGM+ kit was standard in the UK, but since the adoption of the 17-locus kit NDU1-7 have become the reference databases that are most commonly used. Note that the IC population codes refer to the 16-locus DNA17 dataset presented here, while the EA codes refer to the historic FSS 10-locus dataset.

Filtering out possible relatives

Pairwise allele sharing was measured within all subpopulations, counting only loci for which both individuals were genotyped, only including all pairs of individuals that had at least four genotyped loci in common. If $> 75\%$ of alleles were shared, the individual with the fewest loci typed was removed; this is analogous to removing one of each pair of individuals suspected to be highly inbred first degree relatives. For subpopulations with < 100 individuals, the threshold for removal was reduced to 50% allele sharing; this is analogous to removing one of most pairs suspected to be first degree relatives. Both filtering thresholds will additionally remove some anomalous individuals that arose through database errors i.e. some pairs were found with 100% allele sharing which arose through duplicate entries of the same individual.

3.3 Estimation of F_{ST}

3.3.1 F_{ST} definition

There are various ways to define, estimate and interpret F_{ST} [Bhatia et al., 2013]. The original definition [Wright, 1949] compared the variance of an allele fraction over subpopulations (S) to its variance in the total population (T):

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\overline{p}(1-\overline{p})}, \tag{3.1}$$

where $\overline{p}$ denotes the population allele fraction. The total population used in this formulation is usually a hypothetical ancestral population, from which observed subpopulations are assumed to have descended [Weir, 2001], as illustrated in Figure 3.1. However, in forensic work it is necessary to compare the subpopulation of a suspect with the population from which the available allele frequency database has been surveyed. Thus

the reference population allele fractions are observed rather than inferred [Balding and Nichols, 1997]. These two approaches to estimation of F_{ST} will be referred to here as the indirect and direct methods, respectively.

Likelihood-based estimation of F_{ST} is used here instead of moment-based estimation [Bhatia et al., 2013], as it provides high precision, correct accounting for sample size and interpretable intervals and quantiles [Balding, 2003, 2005]. The maximum likelihood estimation of F_{ST} used here is based on the multinomial-Dirichlet distribution [Mosimann, 1962], as opposed to a normal approximation to the multinomial proposed by Weir and Hill [2002], as it does not assume a large sample size. Given a locus with k distinct alleles, the multinomial-Dirichlet has $k-1$ parameters specifying the population allele fractions, which are replaced with observed values in the direct method and are unknown parameters in the indirect method. The remaining parameter ψ specifies the variance, with $F_{ST} = 1/(1+\psi)$.

3.3.2 Direct method estimation

The multinomial-Dirichlet likelihood is used for allele counts in a subpopulation, with reference allele fractions obtained from reference database counts, adjusted by adding a pseudocount of one for each allele in order to avoid zero values. The direct analyses in this chapter only use the 10 loci in common between the DNA17 dataset and the historic FSS database, which are the loci with total allele counts $> 10^4$; D3S1358-D2S1338 (Table 3.1).

When using a uniform prior on F_{ST}, the likelihood curve for F_{ST} can then be interpreted as a posterior density for F_{ST}. Another possibility for an uninformative prior on F_{ST} that was not investigated here is Jeffreys prior. Previous work with small sample sizes [Balding and Nichols, 1997] suggested F_{ST} typically takes values below 4%. To formulate an informative prior this information was incorporated into a beta prior distribution for F_{ST}, with median 2.3% and 95% credible interval (CI) from 0.26% to 8.0%, which was given a larger spread than suggested by Balding and Nichols [1997] due to consideration of more diverse subpopulations in the DNA17 dataset than were included in the Balding and Nichols study.

To illustrate the effects of sample size, direct estimation under both the uniform and beta priors was performed using different sample sizes. Multinomial allele counts were simulated based on allele fractions that were Dirichlet-distributed, with means given by the EA4 allele fractions and $\psi = 99$ so that $F_{ST} = 1\%$. The 95% CI includes 1% at all sample sizes, and becomes tighter as the sample size is increased (Figure 3.3). For small sample sizes, the beta prior leads to slightly smaller posterior interval widths than the uniform, and the posterior median moves towards the prior value.

Figure 3.4 shows that the choice of prior has a noticeable effect on the posterior for Iran (n=12),

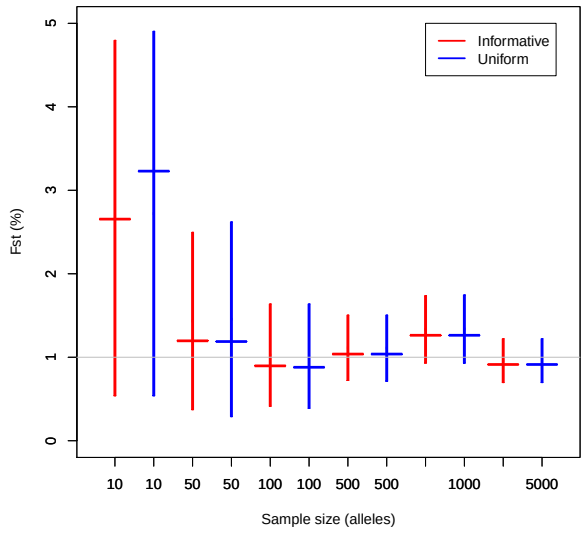


Figure 3.3: F_{st} posterior 95% interval using: (red) a beta prior with median 2.3% and 95% CI (0.26%,8.0%); (blue) the uniform prior. Sample sizes are shown on x-axis. Data were simulated to have $F_{ST} = 1\%$ (horizontal line). The vertical lines indicate the 95% equal-tailed CI, and medians are indicated with horizontal segments.

and less so for Afghanistan (n=42), in both cases the informative prior shifts the F_{ST} posterior distribution to slightly higher values compared with the uniform prior. Of the 33 subpopulations studied here, 10 have sample sizes $\leq$ Afghanistan, the majority of which are in IC6, while only Iran has a sample size ≤ 12. This suggests that the results of F_{ST} estimation presented here will be relatively invariant to the choice of prior distribution, with the most sensitive population being IC6.

3.3.3 Indirect method estimation and locus dependence

While the direct method is the most appropriate for forensic applications due to the role of the reference database in F_{ST} estimation matching its role in computing DNA profile likelihoods, the indirect method requires no such reference database. This allows 15 of the 16 available loci to be analysed using the indirect method, as this method is not constrained by the historic 10-locus FSS databases. SE33 was not included in the analyses presented here due to small sample sizes in the available DNA17 dataset (Table 3.1), which would lead to poor estimation of F_{ST} at this locus.

In the indirect method, the reference population is not observed, but is assumed to be a hypothetical ancestral population from which two or more observed subpopulations have descended independently (see Figure 3.1). The BayesFST software [Beaumont and Balding, 2004] implementation of the indirect method

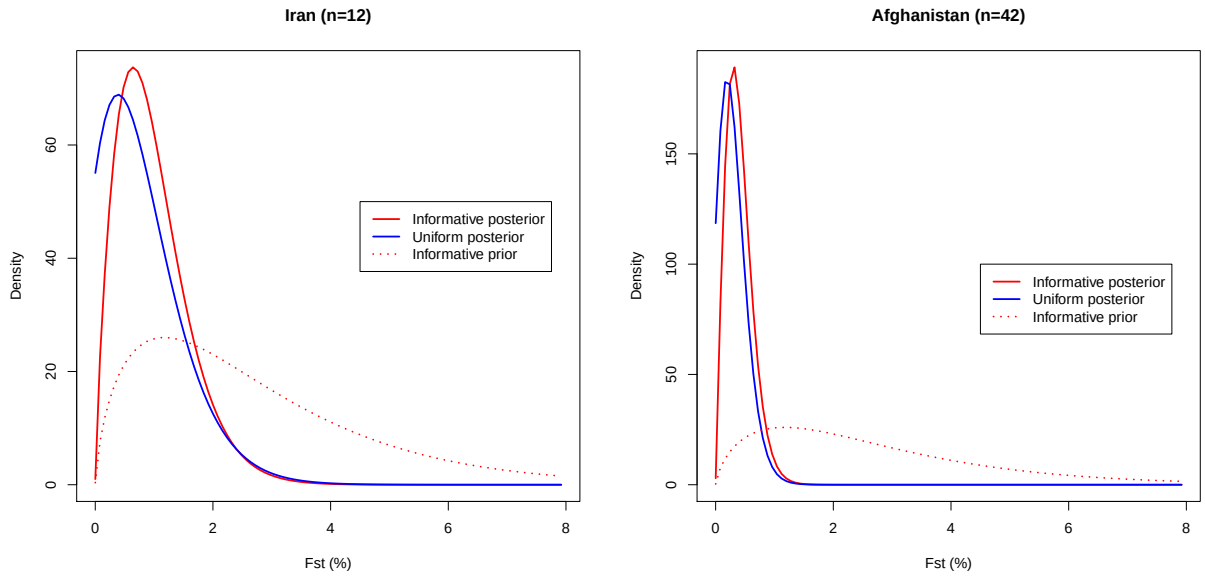


Figure 3.4: F_{ST} posterior densities (solid lines) using the direct method, given a uniform prior (blue) and an informative beta prior (red). Dotted red lines show the beta prior density. The subpopulations analysed are Iran (left) and Afghanistan (right), with EA6 (Middle East/North Africa) and EA4 (South Asia) respectively as reference populations.

was used, which samples from the posterior distribution of F_{ST} in each subpopulation given the allele counts using a Markov Chain Monte Carlo method. BayesFST assigns a jointly uniform prior distribution to the ancestral allele fractions at each locus, using the model:

$$F_{ST}^{i,j} = \frac{e^{a_i+b_j}}{1+e^{a_i+b_j}} \tag{3.2}$$

where a_i and b_j denote locus and population effects, respectively. All inferences reported here are based on

Locus	**Percentile**		**Locus**	**Percentile**	
	2.5	97.5		2.5	97.5
D3	-1.72	-0.2	D19	-0.62	0.62
TH01	0.11	1.58	D2	-0.59	0.62
D21	-0.85	0.45	D22	-0.06	1.32
D18	-0.79	0.38	D1	-0.7	0.52
D16	-1.3	0.15	D10	-0.87	0.6
vWA	-0.93	0.42	D2	-0.21	1.15
D8	-0.73	0.6	D12	-0.71	0.56
FGA	-1.04	0.23			

Table 3.2: Posterior 95% intervals for locus effect parameters using the indirect method. The analysis used all 7 121 individuals with IC1 through IC6 treated as six subpopulations.

150 000 posterior samples, with 7 500 "burn-in" samples performed beforehand which are then discarded.

A pre-analysis was performed to investigate the variation of F_{ST} estimates across loci, treating IC1 through IC6 as six subpopulations of the hypothetical global ancestral population. Each subpopulation parameter b_j was assigned an N$(-3, 1.8)$ prior, while the locus parameters a_i were assigned an N(0,1) prior. The resulting prior distribution for F_{ST} using (3.2) has a prior median 4.7%, with 95% CI from 0.02% to 92%, which is a very loose prior spanning almost the whole range of possible values for F_{ST}. Table 3.2 shows that the posterior 95% CI for the a_i include zero for 13 of the 15 loci (D21-D12), while D3 has a posterior 95% CI < 0 and TH01 has a posterior 95% CI > 0. This implies that D3 has lower F_{ST} values than the whole profile average, TH01 has higher F_{ST} values than the whole profile average, while all other loci do not have F_{ST} values significantly different to the whole profile average. In view of this limited evidence for locus heterogeneity, the locus effect parameter was set close to zero for all subsequent analyses in order to estimate an average F_{ST} over loci, allowing for greater comparability across analyses. The implied prior median with $a_{1...15} \approx 0$ is then 4.7%, with 95% CI from 0.1% to 63%, a tighter prior than that including locus parameters in full.

All 15-locus analyses were repeated with only the 10 loci used in the direct analyses; the resulting inferences were similar with each, but on average more precise with 15 loci (10-locus results not shown here). Thus, the differences reported below between direct and indirect F_{ST} values for a subpopulation are almost entirely due to the different reference population, rather than the different number of loci used.

3.3.4 Best population fit

Each subpopulation defined above was assigned to the FSS database giving the "best fit" (lowest median F_{ST} under the direct method), for both direct and indirect method analyses below. The majority of allocations were as expected: most European subpopulations fit best with EA1, most African and Caribbean subpopulations with EA3, all South Asian subpopulations fit best with EA4, both East Asian subpopulations fit best with EA5 and most Arab subpopulations fit best with EA6. Three subpopulations close to the Middle East fit EA6 equally or slightly better than their nominal population: Southern Europe (EA1), Afghanistan (EA4) and Kenya (EA3). The nominal classification was retained in each case.

One discrepancy was much larger: Somalia fit better with EA6 (F_{ST}=1.5%) than with the nominal EA3 (F_{ST}=2.2%); Somalia was subsequently included with IC6 rather than IC3. Although Somalia borders Kenya (EA3), it is also geographically close to the Arab world, and there have historically been many links. Note that Kenya also fits EA6 better than EA3, suggesting genetic links reaching through Somalia

IC1	n	Direct			Indirect		
		2.5	50	97.5	2.5	50	97.5
Eire	1949	0.1	0.2	0.2	0.0	0.0	0.1
Great Britain	1416	0.1	0.1	0.1	0.0	0.0	0.0
Eastern Europe	61	0.2	0.5	1.0	0.1	0.3	0.7
Northern Europe	45	0.0	0.3	0.8	0.0	0.2	0.5
Southern Europe	60	0.0	0.2	0.5	0.0	0.1	0.3
Western Europe	13	0.1	0.7	2.1	0.0	0.5	1.8
Anglo New World	13	0.1	0.5	1.7	0.0	0.3	1.4
Latin America	25	0.5	1.3	2.4	0.6	1.3	2.4

Table 3.3: 2.5, 50 and 97.5 posterior percentiles of F_{ST} (expressed as %) for EA1. Subpopulations were compared both individually with the reference population EA1 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). **n** denotes the sample size (number of individuals).

and extending as far as Kenya, but the difference was smaller, so the nominal classification for Kenya was retained. Both mitochondrial [Mikkelsen et al., 2012] and Y-chromosome [Sanchez et al., 2005] studies have suggested a strong Arab influence in Somali genetics, although their highest similarity is usually with neighbouring Eastern Ethiopians and Northern Kenyans. HLA typing [Mohamoud, 2006] suggests that Somalis are more similar to Arabs than to Sub-Saharan Africans, while admixture mapping estimates the Eurasian ancestry of Somalis at roughly 38% [Pickrell et al., 2014], supporting the low F_{ST} estimate for Somalia with the EA6 database.

3.4 EA1 F_{ST} estimates

Compared to the EA1 reference population, all European subpopulations, except Western Europe, have a posterior 97.5 percentile F_{ST} estimate $< 1\%$ (Table 3.3). All subpopulations, including Europe, have a posterior median F_{ST} estimate $< 1\%$, which suggests that the high 97.5 percentile estimate for Western Europe is due to a small sample size rather than true genetic dissimilarity. Anglo New World has posterior 97.5 percentile F_{ST} estimates slightly lower than Western Europe, but the small sample size of each, along with low median F_{ST} estimates suggest that each fits the IC1 population well. Southern Europe has a low F_{ST} estimate, supporting a merger of European-origin IC2 individuals with the IC1 population. Conversely, Latin America has both the highest median and 97.5 percentile F_{ST} estimates; Latin Americans are known to be admixed between native Amerindians, Europeans and Africans [Ruiz-Linares et al., 2014, Salzano and Sans, 2014, Moreno-Estrada et al., 2013], with African ancestry largely localised to the Caribbean islands [Moreno-Estrada et al., 2013] and some coastal regions of South America [Ruiz-Linares et al., 2014]. Subsequently, IC2 might reasonably be redefined as a Latin American population with predominantly European ancestry,

IC3	n	Direct			Indirect		
		2.5	50	97.5	2.5	50	97.5
Ghana	214	0.8	1.1	1.6	0.2	0.3	0.5
Jamaica	166	0.5	0.7	1.0	0.0	0.1	0.2
Kenya	51	0.7	1.2	1.9	0.8	1.3	1.9
Nigeria	444	0.9	1.2	1.5	0.2	0.3	0.3
Sierra Leone	41	0.7	1.3	2.2	0.1	0.3	0.8
Uganda	63	0.3	0.5	1.0	0.0	0.2	0.4
Unknown IC3	864	0.4	0.5	0.7	0.0	0.0	0.0
Other Caribbean	20	0.5	1.5	2.9	0.1	0.4	1.3
Other C/S Africa	55	0.3	0.6	1.1	0.0	0.1	0.3
Other E Africa	66	0.3	0.7	1.1	0.0	0.1	0.4
Other W Africa	48	0.1	0.5	1.0	0.0	0.1	0.3

Table 3.4: 2.5, 50 and 97.5 posterior percentiles of F_{ST} (expressed as %) for EA3. Subpopulations were compared both individually with the reference population EA3 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). **n** denotes the sample size (number of individuals).

as opposed to individuals with primarily Amerindian ancestry who would require a separate database that would be rarely utilised in the UK, or individuals with primarily African ancestry that may fit better in the IC3 population.

Lower F_{ST} estimates are obtained with the indirect method than with the direct method for the majority of subpopulations, which is due to inferred ancestral allele probabilities being towards the centre of the subpopulation values, while the direct method compares to the original EA1 database compiled by the FSS which is likely biased towards individuals of British ancestry. Conversely, Latin American F_{ST} estimates remain almost unchanged, as the ancestral allele probability inferences are dominated by the European subpopulations, which comprise 99.3% of the total sample size in the IC1 population.

3.5 EA3 F_{ST} estimates

Using the direct method, the African national subpopulations of Ghana, Kenya, Nigeria, and Sierra Leone have higher F_{ST} estimates than the African mixed subpopulations of Unknown IC3, East, West and Central-Southern Africa. Conversely, the F_{ST} estimate for the Caribbean mixed subpopulation Other Caribbean is much higher than for the Caribbean national subpopulation Jamaica, while simultaneously being high in relation to all other subpopulations. Jamaicans have a predominantly African origin [Caribbean Community Capacity Development Programme, 2009], and there are approximately 800 000 people of Jamaican descent living in the UK [International Organisation for Migration, 2007], which is close to half the UK population categorised as black [Office for National Statistics, 2011], therefore the EA3 database may be expected to

IC4	n	Direct			Indirect		
		2.5	50	97.5	2.5	50	97.5
Afghanistan	47	0.1	0.3	0.9	0.1	0.4	0.9
Bangladesh	53	0.1	0.4	0.9	0.0	0.1	0.4
India	49	0.0	0.3	0.8	0.0	0.1	0.4
Pakistan	60	0.0	0.2	0.5	0.0	0.2	0.5
Unknown IC4	76	0.0	0.2	0.5	0.0	0.1	0.2

Table 3.5: 2.5, 50 and 97.5 posterior percentiles of F_{ST} (expressed as %) for EA4. Subpopulations were compared both individually with the reference population EA4 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). **n** denotes the sample size (number of individuals).

include a large number of Jamaicans, explaining the low F_{ST} estimate for Jamaica relative to the EA3 database.

Indirect estimation (Table 3.4b) gives results divergent to the direct method. For the majority of subpopulations the F_{ST} estimate is greatly reduced, except for Kenya which is geographically remote from the majority of subpopulations, which are predominantly West African or Caribbean. In Section 3.3.4 it was noted that Kenya fits almost equally well with both EA3 and EA6 using direct estimation, and in fact fits slightly better with EA6, suggesting some genetic influence from the Arab world, which is supported by Somalia, a neighbour of Kenya, being included with IC6 rather than IC3. Jamaica has a much lower F_{ST} estimate with indirect estimation, which is supported by results from Benn-Torres et al. [2008] who estimated West African admixture of Jamaicans at 84.4%. The large 97.5 percentile F_{ST} estimate for Other Caribbean may be due to small sample size as the median estimate is more in line with other subpopulations. This is once again supported by West African admixture estimates in individuals from Barbados and St Thomas of 89.6% and 86.8% respectively [Benn-Torres et al., 2008] and individuals from Haiti being estimated as having majority African ancestry [Moreno-Estrada et al., 2013]. However, individuals from the Dominican Republic and Puerto Rico are estimated to have majority European ancestry [Moreno-Estrada et al., 2013], which may suggest that the elevated F_{ST} estimate for the mixed Caribbean subpopulation compared to the ancestral African population investigated here may be due to the conflicting ancestries of different Caribbean islands, rather than an artefact of small sample size; the Other Caribbean subpopulation comprises nine separate Caribbean islands.

3.6 EA4, EA5 and EA6 F_{ST} estimates

For IC4, the F_{ST} estimates are all low for both direct and indirect methods, with no outliers (Table 3.5). The F_{ST} estimates for India and Bangladesh are much lower for the indirect than the direct method, suggesting

IC5	n	Direct			Indirect		
		2.5	50	97.5	2.5	50	97.5
NE Asia	260	0.1	0.2	0.3	0.1	0.4	0.8
SE Asia	44	0.0	0.2	0.7	0.0	0.1	0.4

Table 3.6: 2.5, 50 and 97.5 posterior percentiles of F_{ST} (expressed as %) for EA5. Subpopulations were compared both individually with the reference population EA5 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). **n** denotes the sample size (number of individuals).

IC6	n	Direct			Indirect		
		2.5	50	97.5	2.5	50	97.5
Iran	12	0.1	0.9	2.4	0.1	0.9	2.7
Iraq	28	0.0	0.2	0.7	0.0	0.2	0.7
Somalia	494	1.1	1.3	1.7	1.2	1.6	2.1
Turkey	20	0.1	0.5	1.6	0.2	0.9	2.1
Middle East	24	0.1	0.7	1.8	0.1	0.5	1.6
N Africa	26	0.2	0.7	1.7	0.1	0.6	1.5

Table 3.7: 2.5, 50 and 97.5 posterior percentiles of F_{ST} (expressed as %) for EA6. Subpopulations were compared both individually with the reference population EA6 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). **n** denotes the sample size (number of individuals).

that the EA4 database is skewed towards Pakistani individuals; Pakistanis are the second most populous South Asian subpopulation in the UK [Office for National Statistics, 2011], behind Indians and ahead of Bangladeshis, making this plausible. This is supported by increasing direct method F_{ST} estimates with increasing geographic distance from Pakistan. Conversely, F_{ST} estimates for IC4 using the indirect method increase with increasing distance from India/Bangladesh.

Similar to IC4, the F_{ST} estimates for IC5 are low for both methods (Table 3.6). The F_{ST} estimate for NE Asia is higher than that for SE Asia using the direct method, but lower using the direct method. This suggests the EA5 database largely consists of individuals from NE Asia, with the NE Asian sample in this study being majority Chinese. This is likely as the 2011 UK Census lists Chinese as a separate ethnic group, but all other East Asian subpopulations are combined in "Other Asia" [Office for National Statistics, 2011] suggesting that the Chinese population of the UK is considerably larger than that of any other East Asian country.

Most IC6 subpopulations have low sample sizes, so posterior 97.5 percentiles will be elevated and show a high variance. Therefore the posterior median will be discussed here rather than the posterior 97.5 percentile. Iraq has low F_{ST} estimates, lower than its geographic neighbour Iran (Table 3.7). Somalia has the largest F_{ST} estimates of all subpopulations using both methods, unsurprisingly. The direct and indirect methods give similar estimates for most subpopulations, however, Turkey has a noticeably larger F_{ST} estimate using the indirect method, perhaps indicating that Turkish individuals are well represented in

Fringe	Reference				
	EA1	EA3	EA4	EA5	EA6
Afghanistan	1.17	2.90	0.78	1.87	0.78
Kenya	2.32	1.39	2.51	2.32	1.36
Southern Europe	0.30	2.99	1.20	2.03	0.34
Unknown IC4	1.68	2.80	0.62	1.17	0.72

Table 3.8: Posterior median F_{ST} (%) for fringe subpopulations. Fringe subpopulations are those for which another reference population gives a median F_{ST} estimate using the direct method within 0.001 of the lowest (best fit) value.

the EA6 database. The indirect method F_{ST} estimates show increasing F_{ST} with increasing distance away from Iraq.

3.7 Fringe regions

"Fringe" subpopulations are those that have similar affinity to two populations (difference in median F_{ST} < 0.001). These fringes are found at the boundaries of the defined continental-scale populations (Table 3.8) reflecting a genetic cline; an overall smooth change in allele frequencies with geography [Ramachandran et al., 2005]. Logically, individuals from Bangladesh and Turkey will be more genetically distinct than individuals from Afghanistan and Iran; the first pair may only fit well in their nominal population, while the latter two may fit almost equally well in each other's nominal population. In the DNA17 dataset presented here, Afghanistan fits in IC4 and IC6 similarly well, S Europe fits IC1 and IC6 similarly well, and Kenya fits IC3 and IC6 similarly well, with all three being on the geographic boundary between the two population designations. These results suggest a relatively low differentiation between IC6 and its surrounding populations, IC1, IC3 and IC4; a global study of admixture and migration between well-defined ethnic groups [Pickrell and Pritchard, 2012] supports the low differentiation between IC6 and both IC1 and IC3 through complex signals of admixture and migration between IC6 and the other two populations, while low differentiation between IC4, IC6 and IC1 is directly inferred through low drift separating the three populations. Only IC5 is not linked to other populations through a fringe subpopulation, perhaps due to the mountains separating China from South Asia, and its geographical remoteness from IC1 and IC3. This agrees with a previous report that East Asian populations are distinct from those of South Asia, but are close to South East Asian populations [Consortium et al., 2009], and IC5 ethnic groups being significantly drifted away from IC4 ethnic groups [Pickrell and Pritchard, 2012] with little migration between the two.

Global	n	Reference					
		EA1	EA3	EA4	EA5	EA6	Indirect
IC1	3582	0.4	3.1	1.9	1.9	0.9	2.7
IC3	2032	1.7	0.7	1.7	1.4	1.1	1.0
IC4	285	1.4	3.1	0.7	1.3	0.8	2.3
IC5	304	3.1	4.2	2.4	0.5	2.0	3.3
IC6	604	1.8	1.7	1.9	1.7	0.9	1.4

Table 3.9: Posterior median F_{ST} (%) for inter-population comparisons. Populations IC1-6 were compared to each reference population in turn using the direct method. The indirect method was used to compare each population to a hypothetical global ancestral population.

3.8 Inter-population comparisons

Continental-scale sample populations are compared to continental-scale reference populations (direct) or a global-scale inferred ancestral population (indirect) here, as opposed to the largely national-scale subpopulation comparisons with continental-scale reference populations in Sections 3.4 to 3.7. Each column of Table 3.9 shows a different F_{ST} analysis of the five IC populations, using an EA database as the reference database in the direct method (columns 3-7), or using the indirect method (column 8).

For the direct method, each IC database gives the lowest F_{ST} estimate with its corresponding EA database, supporting a reasonable consistency of definitions between IC and EA databases. The highest F_{ST} value for IC1, IC4 and IC5 are all obtained relative to EA3, suggesting higher divergence from EA3 than other populations. This reflects the increased genetic diversity seen in Africa compared to other populations due to sequential founder effects as humans spread from Africa, and is echoed in [Pickrell and Pritchard, 2012], who found the substantially larger drift between African ethnic groups and non-African ethnic groups than within the non-African ethnic groups. Conversely, looking down the columns of Table 3.9, IC5 shows the highest F_{ST} value for each EA database except EA5, indicating that IC5 is genetically distinct from all other populations, as seen in Pickrell and Pritchard [2012]. The IC6 database shows similar F_{ST} values with respect to all four EA databases other than EA6, with a large influence from the Somalian subpopulation. Note, the IC5 database shows the highest F_{ST} estimates compared to each EA database, providing further evidence that the IC5 population is distinct from all other tested populations.

Using indirect estimation, IC3 and IC6 show the lowest F_{ST} values, increasing through IC4, IC1 and IC5 in turn, corresponding to an inferred ancestral human population similar to that of modern North-East Africa [Pemberton et al., 2013]. This recapitulates the pattern seen in Pickrell and Pritchard [2012], with the genetic similarity of IC3 and IC6 being likely due to recent migrations, which STR data is well placed to determine.

3.9 Precision

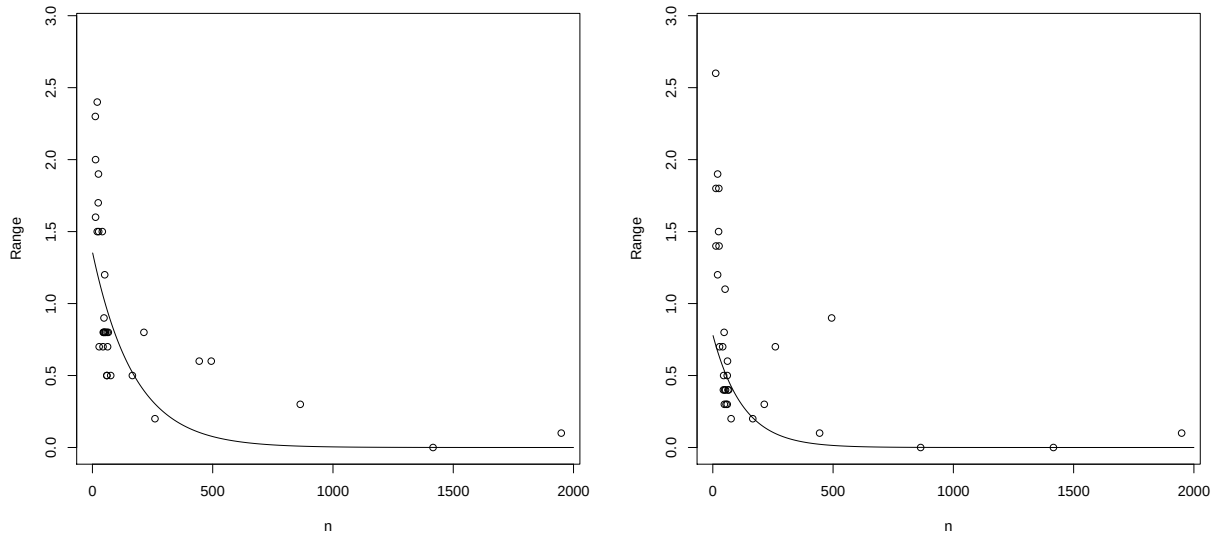


Figure 3.5: F_{ST} posterior 95% intervals (range) against log subpopulation sample size (n) for the direct method (left) and the indirect method (right). The fitted line is a linear regression with n as a predictor and log posterior 95% interval as a response.

Good precision (tight 95% posterior intervals) of F_{ST} estimates has been obtained here, despite only examining 10 or 15 STR loci; this was possible due to the multi-allelic nature of STRs, and large sample sizes for many of the investigated subpopulations. This is demonstrated by the fact that the precision of F_{ST} estimates increases as the sample size of a subpopulation increases (Figure 3.5). However, F_{ST} estimates depend sensitively on the choice of reference population, in particular the choice of a hypothetical ancestral population or a population database, which are usual practice in forensic genetics and population genetics respectively.

3.10 Comparison with published estimates

Silva et al. [2012] estimated global F_{ST} separately from a collection of worldwide forensic STR databases and from the non-forensic Human Genome Diversity Project (HGDP) dataset, with F_{ST} estimates of 2.3% and 5.3% from the forensic and non-forensic datasets respectively. The forensic estimate is similar to the inter-population estimates presented here (Table 3.9), however, the non-forensic estimate is considerably larger. Silva et al. suggest that this discrepancy is because STR markers are chosen for forensic use in part because

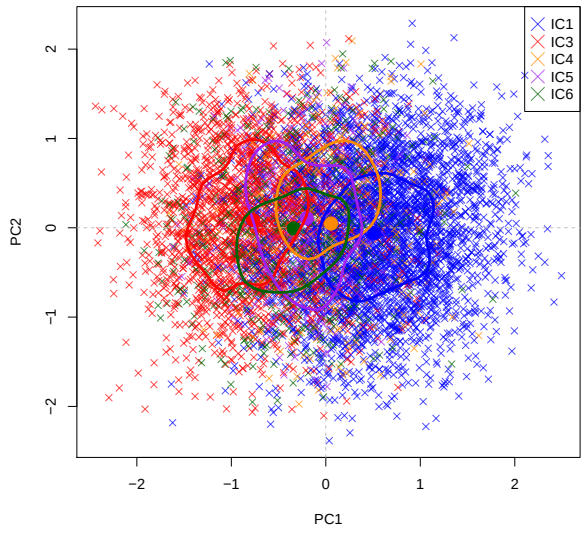


Figure 3.6: First and second principal components (PCs) of a principal components analysis of the individual genotypes of the 7 121 individuals that comprise the DNA17 dataset, at the 10 SGM+ loci. Crosses indicate individuals, filled circles indicate mean principal components for each population, coloured lines indicate density contours for each population, while colours indicate the population.

they have low differentiation between populations. This can be seen in a principal components analysis (PCA), Figure 3.6, where there is little differentiation between the different populations investigated. The largest differentiation is between IC1 and IC3, which may be due to these two populations having the largest sample size. Silva et al. also demonstrate, however, that R_{ST} [Slatkin, 1995], an F_{ST} analogue for STRs that assumes a stepwise mutation model, is decreased by selecting high heterozygosity markers; forensic markers that are currently in use have also been selected in part to maximise heterozygosity. Forensic surveys for generating STR databases tend to sample individuals from large ethnically diverse populations, while the HGDP dataset, and population genetics datasets in general, tend to sample individuals from small ethnically distinct populations; these different strategies and aims may explain the different F_{ST} estimates reported by Silva et al..

Both the F_{ST} estimates of Silva et al., and the F_{ST} estimates presented in this chapter, are considerably lower than those presented by Nelis et al. [2009], who estimate continental genetic distance between Africa, Asia and Europe using the HapMap SNP database (HapMap 2), and obtained F_{ST} estimates ranging from 11% (European/Asian comparison) to 19% (African/Asian comparison). STR mutations are not uniform; expansion mutations are favoured in short alleles [Xu et al., 2000], while contraction mutations are favoured in long alleles [Sibly et al., 2003, Dupuy et al., 2004, Lu et al., 2012], leading to relatively stable

allele fractions across populations due to the high mutation rate seen in STRs [Weber and Wong, 1993]. This may explain the higher F_{ST} estimates obtained from SNP data than from STR data. Excoffier and Hamilton [2003] demonstrated that this discrepancy can be removed by modelling the stepwise mutation seen in STRs; the probability of an allele mutating to repeat unit lengths other than x-1 or x+1 is negligible. However, the broad pattern of variation is similar for both STRs and SNPs [Ramachandran et al., 2005, Pemberton et al., 2013].

3.11 Guidelines for forensic practice

From the results presented here, an $F_{ST} \leq 3\%$ should be appropriate for a wide variety of forensic calculations, involving any of the populations investigated here. The 97.5 posterior percentile for F_{ST} is $\leq 3\%$ relative to the best fit population for all subpopulations tested here; this agrees with more limited results that have been previously published [Balding and Nichols, 1997, Gill et al., 2003]. Low values may be applicable in some situations, e.g. $F_{ST} = 1\%$ may be acceptable for Asians (both South and East), however, $F_{ST} = 3\%$ is less susceptible to the incorrect assignment of the reference population for an unknown contributor, which may include X under H_d; this is discussed further in Chapter 4. With appropriate case circumstances, it may be possible to tailor the F_{ST} value used to the case at hand, based on the subpopulation estimates presented here. For example, a lower F_{ST} value may be appropriate for Jamaican alternative contributors than would be appropriate for an alternative contributor from another Caribbean island.

3.11.1 Future work

Cowell [2016] have recently published work incorporating both an F_{ST} adjustment and uncertainty in p_j to the calculation of forensic likelihoods, giving p_j a Dirichlet prior. Such uncertainty in p_j may have an effect on the F_{ST} estimates under the direct method investigated here, so more accurate estimates of F_{ST} may be obtained by incorporating a Dirichlet prior on the p_j, similar to Cowell [2016], rather than assuming the p_j as known here.

Chapter 4

Choice of population database for forensic DNA profile analysis

Work in this chapter has been published in Steele and Balding [2014a], see Appendix B. A model for forensic likelihoods that allows for different contributors to come from different populations without dropout or dropin was developed and coded by me, and was adapted from the likeLTD discrete model developed by Prof. David Balding. All simulations and analyses were performed by me. The dataset was collected and provided by Dr. Denise Syndercombe Court (see Chapter 3).

4.1 Effect of database choice on the WoE

As discussed in Chapters 1 and 3, the misassignment of population database for any unknown contributor to a CSP can have an important impact on the WoE. The rarity of a CSP allele is directly linked to the WoE against a Q who possesses the allele. Returning to the scenario that generated (1.6), where $\mathcal{C}$ =ABC, $\mathcal{G}_Q$ =AB, $\mathcal{G}_K$ =AC, both at good template, with LR=$1/(2p_Ap_B + p_B^2 + 2p_Bp_C)$ if shared ancestry between Q and X is ignored. If a population allele probability database shown in row 1 of Table 4.1 is used, the WoE is 0.7 bans, and the most likely $\mathcal{G}_X$ is BC. However, if a different database is used, in which allele B is now rare in the population, $p_B = 0.01$, then the WoE against Q increases considerably. The WoE increases because Q matches a CSP allele that is rare in the chosen population, so observing that allele in X under H_d becomes *a priori* unlikely; all three $\mathcal{G}_X$ probabilities drop substantially as they all require X to have a B allele (Table 4.1, second row). The BC genotype is still the most likely for X, but the AB genotype is now more likely than the BB genotype. Note that the B allele here cannot be explained by K, and therefore must be explained by X (all possible genotypes for X include the B allele), so any reduction in p_B will increase the WoE against Q. If instead K is able to explain the rare allele, the WoE remains similar to that evaluated

F_{ST}	p_A	p_B	p_C	Pr($\mathcal{G}_X$=AB)	Pr($\mathcal{G}_X$=BB)	Pr($\mathcal{G}_X$=BC)	WoE
0.00	0.10	0.20	0.30	4.0e-2	4.0e-2	1.2e-1	0.7
	0.10	0.01	0.30	2.0e-3	1.0e-4	6.0e-3	2.1
	0.01	0.20	0.30	4.0e-3	4.0e-2	1.2e-1	0.8
0.03	0.10	0.20	0.30	5.4e-2	4.7e-2	1.2e-1	0.7
	0.10	0.01	0.30	9.5e-3	1.5e-3	2.2e-2	1.5
	0.01	0.20	0.30	1.7e-2	4.7e-2	1.2e-1	0.7

Table 4.1: WoE and genotype probabilities for X ($\mathcal{G}_X$) for a good-template CSP of observed alleles ABC, with a queried contributor with genotype AB and a known contributor with genotype AC.

using the database with no rare allele (Table 4.1, third row). The probability of the $\mathcal{G}_X$ that includes the rare allele is once again decreased, however, the remaining two genotype probabilities remain unchanged. In this situation X is most likely to be BC, with K explaining the remaining rare A allele in the CSP.

When possible distant relatedness between Q and X is taken into account by setting F_{ST}=0.03, the WoE is unaffected when both of the alleles of Q are common (Table 4.1, fourth row); the probabilities of genotypes $\mathcal{G}_X$=AB and $\mathcal{G}_X$=BB have increased slightly, but the probability of $\mathcal{G}_X$=BC remains unchanged. However, when one of the alleles of Q is rare the WoE is reduced significantly (fifth row); the probability of all three genotypes under H_d have been increased, as p_B is increased through the F_{ST} adjustment. This allows for both Q and X to match the CSP, even if Q has not contributed to the CSP, and hence the WoE against Q should be reduced accordingly. When the rare allele can be explained by K, the WoE is reduced by just a tenth of a ban (deciban, sixth row); once again only the probabilities of $\mathcal{G}_X$=AB and $\mathcal{G}_X$=BB have increased.

The choice of database also affects the WoE through Pr($\mathcal{G}_U$), since both X and U are drawn from the population. When evaluating the WoE the most relevant database will be that which is most appropriate for the population of X, as X is assumed to be the true source of DNA. Without a U, this database choice is unnecessary under H_p, as X is assumed to be Q and $\mathcal{G}_Q$ is known. This leaves the question of what database is most appropriate for X; Balding and Nichols [1994] argue for using the database of Q even if the ancestry of X is unknown, while many other authors have stated that the most appropriate database for Q may not be the most appropriate for X [National Research Council, 1996, Foreman et al., 1998]. Balding and Nichols support their argument through a size-bias effect; once $\mathcal{G}_Q$ has been observed then Pr($\mathcal{G}_X = \mathcal{G}_Q$) increases if X is assumed to come from the same population as Q but not if X is assumed to come from a different population.

# populations	# U	# permutations (H_d)
3	0	3
	1	9
	2	27
5	0	5
	1	25
	2	125
7	0	7
	1	49
	2	343

Table 4.2: Number of database permutations for unknown contributors to a CSP under H_d. The number of permutations is given by n^u where n is the number of population databases, and u is the number of unknown contributors including X.

4.2 Evaluation with all possible databases

Currently, common practice when the ancestry of X is unknown is to evaluate the WoE with multiple population databases for X, and to choose the database that returns the minimum WoE in the interest of being conservative. However, it should not be necessary to favour defendants in such an arbitrary way, that may have no bearing on the realities of the case. As an illustrative example, Q is Caucasian and the lowest WoE is obtained with a database of Vietnamese individuals. It may be reasonable to report the Vietnamese WoE if the population in the area local to the crime includes many Vietnamese individuals, however, if the local area contains few, if any, Vietnamese individuals then it may be unhelpful to present the Vietnamese WoE to the court. For instance reporting a Vietnamese WoE may be appropriate if the crime was committed in London, a cosmopolitan city, but would not be appropriate if the crime was committed in rural Devon. The world's population can be categorised in a large number of ways; it is neither possible nor desirable to enumerate the WoE for all population designations in order to report the smallest WoE. However, reasonable population choices should be considered based on the available knowledge about the nature and location of the crime. Approximations that favour the defence may be desirable to simplify the necessary analyses and to avoid courtroom challenges, as the number of reasonable population choices can still be large. One such approximation, investigated in this chapter, is to assume that X and all other hypothesised Us come from the population that is most appropriate for Q.

The number of permutations of population databases for unknown contributors is not prohibitive for few U or for few possible databases (Table 4.2), however, as either increases the number of possible permutations quickly becomes prohibitive. Also, the computational effort of a single evaluation increases with the number of unknown contributors, meaning that the overall computational effort to consider multiple

databases for each unprofiled contributor can be restrictive. Therefore, the possibility of evaluating a single WoE, assuming the database of Q for all unknowns, becomes attractive given that it can be demonstrated that it does not favour the prosecution. Additionally, an F_{ST} adjustment, as discussed in Chapters 1 and 3, is necessary if Q and X are assumed to come from the same population [Balding and Nichols, 1994]. Chapter 3 demonstrated that an F_{ST} value of 3% is greater than the F_{ST} estimates for all tested subpopulations [Steele et al., 2014b], and should therefore be conservative in all scenarios covered in that study. As discussed in Chapter 1, the F_{ST} adjustment increases the population probability of alleles observed in Q, and therefore decreases the population probability of alleles not observed in Q, essentially implementing the size-bias effect into the calculation of the WoE.

4.3 Effect of F_{ST} on mixtures

For contributors to a CSP other than Q/X, utilising an F_{ST}=0.03 should be conservative. When a K is hypothesised the problem simplifies to assigning the correct database for X, because the alleles of K are not drawn from a population. Conversely, when a U is assumed in a mixture, the population allele probability is important for both $\Pr(\mathcal{G}_X)$ and for $\Pr(\mathcal{G}_U)$, so it is necessary to check that assigning a population database most relevant to Q does not adversely affect the WoE when such an assignment is in fact incorrect for any U and/or X. Note that the F_{ST} adjustment utilised in the heuristic only increases the population allele probability for any alleles of Q, increasing the probability that both X and U share any allele with Q, which supports the defence case more than with no F_{ST} adjustment, meaning that the F_{ST} adjustment is conservative (errs on the side of supporting H_d).

While favouring the defence through an F_{ST} adjustment is desirable, it is not possible to guarantee that a proposed WoE evaluation will be conservative compared to all possible alternative evaluations. Instead, it should be sufficient to demonstrate that the majority of evaluations are conservative with the heuristic compared to a set of alternative evaluations. When a database for an unknown contributor's population allele probabilities is used that differs from the database most relevant to Q, due to some other evidence about the ethnicity of X, perhaps colour CCTV footage of some exposed skin or ancestry inference from a SNP panel [Yang et al., 2005, Phillips et al., 2007, Halder et al., 2008, Jia et al., 2014], assumed coancestry between Q and X is no longer applicable, so F_{ST}=0 may be most appropriate. However, distant populations share some genetic background, so a small F_{ST} of 0.01 may be more appropriate to account for this very distant shared ancestry across populations [Balding, 2005]. Such an F_{ST} value will introduce a small bias in favour of the defence proposition, and will allow for the database population to differ from the population

that X originates from slightly. However, here an F_{ST}=0 will be used in such a case, as the study design already incorporates a slight bias towards the defence, through comparing to the minimum WoE over a set of alternative database choices.

4.4 Unknown ancestry of Q

The true ancestry of Q may be unknown, for example if Q does not report their own ancestry or if Q is adopted and does not know his ancestry. Similarly, the true ancestry of Q may be misassigned, for example if Q is impersonating another individual, his ancestry is ambiguous such as through admixture, or his ancestry has been misreported by the reporting police officer based on physical appearance. Q may also be of some ancestry that is poorly represented in any available database, such as a native Amerindian individual arrested in the UK. In such cases, observing an allele in Q does not make it more likely to observe that allele in the database that has been assigned to Q. This does not match the adjustment that would be incorporated through F_{ST}, so the WoE would be adversely impacted, however, the F_{ST} value used is already generous, so the impact of population misassignment should be small, and within the range of the F_{ST} adjustment allowing for misassignment of databases.

4.5 Databases and modelling choices

Frequency data was used at 16 STR loci for five UK populations, which are identical to those used in Chapter 3: Caucasian (IC1), African and African Caribbean (IC3), South Asian (IC4), East Asian (IC5) and Middle Eastern (IC6), Table 4.3. See Chapter 3 for a full description of the dataset. These data were used to simulate 16-locus profiles which were simulated at both Hardy Weinberg equilibrium and linkage equilibrium. Dropout and dropin were not simulated here, and were not modelled when evaluating the WoE.

The WoE was evaluated using the likelihood ratio framework [Gill et al., 2006], see Chapter 1 for details. All hypothesis pairs include Q as a contributor under H_p, and replace Q with an unrelated X under H_d. An F_{ST} adjustment [Balding, 2005] to the population allele probabilities that match Q was implemented whenever the most appropriate database for Q was assumed for X as well; $F_{ST} = 0.03$ is used when the adjustment is applied, and $F_{ST} = 0$ is used otherwise. A sampling adjustment of one was added to the database counts of alleles of Q (see Chapter 1), which avoids underestimating the population probabilities of rare alleles [Balding, 1995].

Allele counts	IC1	IC3	IC4	IC5	IC6
D3S1358	6878	3941	520	599	1202
TH01	6816	3918	514	598	1202
D21S11	6870	3941	520	599	1199
D18S51	6808	3930	520	600	1195
D16S539	6818	3927	514	600	1199
VWA	6877	3936	520	600	1201
D8S1179	6871	3941	520	600	1202
FGA	6853	3938	516	600	1201
D19S433	6702	3868	507	595	1197
D2S1338	6443	3758	491	594	1176
D22S1045	1816	2482	421	498	954
D1S1656	1827	2508	426	504	959
D10S1248	1815	2499	416	500	912
D2S441	1800	2473	420	493	943
D12S391	1857	2543	437	499	945
SE33	368	872	237	394	268

Table 4.3: Number of allele observations at each locus for each population database: Caucasian (IC1), Afro-Caribbean (IC3), South Asian (IC4), East Asian (IC5) and Middle Eastern (IC6)

4.6 Single-contributor CSPs

4.6.1 Matching database

Initially 10,000 single contributor CSPs were simulated from each database in turn, leading to 50,000 profiles in total. The WoE for every simulated CSP was calculated using each database in turn for the population allele probabilities for X, using hypotheses of the form:

$$H_p: \quad Q$$
$$H_d: \quad X.$$

The minimum WoE over the four incorrect database assignments was subtracted from the heuristic WoE, which correctly assumes the database of Q as that for X using $F_{ST} = 0.03$; a negative result indicates that the heuristic is conservative compared to all alternatives tested, regardless of the true ancestry of X.

With probability > 0.999, the heuristic returns a lower WoE than any of the four alternative database assignments for X(Figure 4.1), indicating that the heuristic is conservative in the vast majority of cases. Two observations of non-conservative WoEs using the heuristic were seen in the 50 000 simulated profiles. The mean difference between the heuristic WoE and the minimum WoE from all four alternative WoEs was 0.3 bans per locus (Table 4.4, column 1).

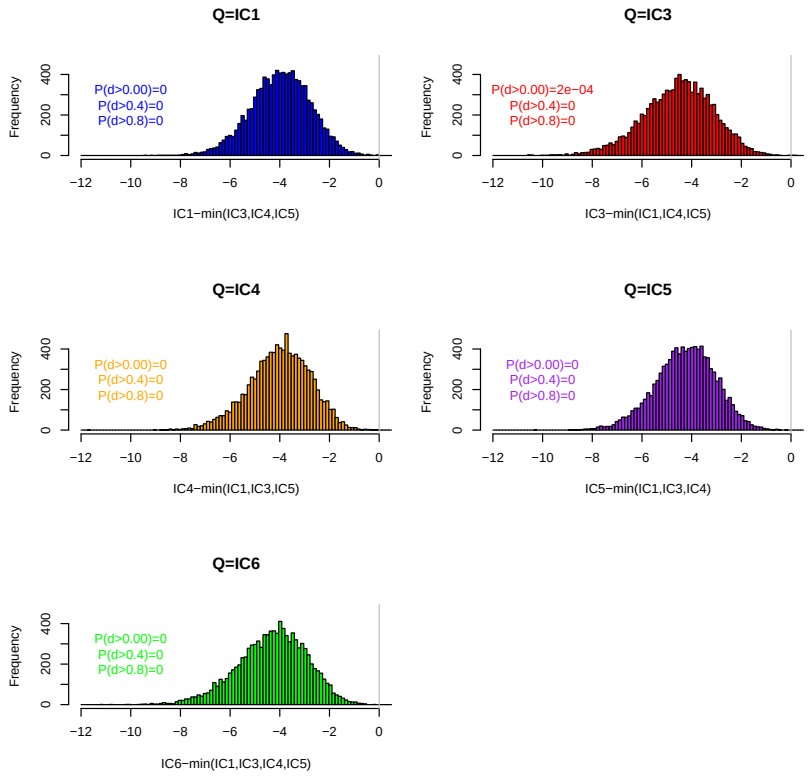


Figure 4.1: The effect of database on Weight of Evidence (WoE) calculations for a one-contributor CSP. The databases are described in Table 4.3. The x-axis shows the WoE computed using the database from which the contributor Q was simulated (indicated in the subplot title) with $F_{ST} = 0.03$, minus the lowest WoE computed using each of the four alternative databases and $F_{ST} = 0$. P(d>x) indicates the proportion of differences that are > x.

4.6.2 Imperfect database

CSPs were simulated with a varying value of F_{ST} (0.01, 0.02, or 0.03) from the IC1 database, mimicking the situation where no available database matches the ancestry of Q exactly. 30 000 CSPs were simulated, with 10 000 at each possible F_{ST} value. The hypotheses compared were identical to those in Section 4.6.1, and the WoE was evaluated assuming IC1 as the database for X (the heuristic scenario), and assuming all other databases in turn for the database of X (all incorrect assignments of the database). Once again the minimum of the incorrect assignment WoEs was subtracted from the heuristic WoE.

The maximum number of non-conservative evaluations over the three F_{ST} values was 3 out of 10 000 evaluations, which is higher than that when the database of Q exactly matches that of X (F_{ST}=0, IC1, Section 4.6.1, 0 out of 10 000 non-conservative) as expected, but the difference is both small and non-significant (Fisher's exact test, H_0: odds ratio=1, H_A: odds ratio $\neq$ 1, p=0.25). This demonstrates that the heuristic remains conservative compared to all tested alternative calculations even when the database assumed for Q does not match the database of X exactly, which may occur if no database matches Q exactly (which should be true for all databases) or if Q and X are in fact different individuals but share some coancestry.

4.7 Two-contributor CSPs

Next 25 000 two-contributor profiles were simulated, with 1 000 simulations from each possible permutation of database choices for the two contributors. The WoE was evaluated for these CSPs assuming the second contributor as either known (Section 4.7.1) or unknown (Section 4.7.2).

4.7.1 Known second contributor

Initially, the second contributor was assumed to be a known contributor, rather than an unknown contributor. This means that there is no database assigned to K when evaluating the WoE, so only the database choice for X need be investigated. The WoE for the 25 000 profiles was evaluated assuming the correct database for X (the heuristic calculation) and assuming each of the four incorrect database assignments, with hypotheses of the form:

$$H_p: \quad Q + K$$
$$H_d: \quad X + K$$

and once again the minimum of the four incorrect assignment WoEs was subtracted from the heuristic WoE.

The WoE for all evaluations (both correct and incorrect) was reduced by approximately 3 bans compared to the single-contributor CSPs when a known contributor was introduced into the CSP (Table 4.4, column 2). There was little change in the difference between the heuristic evaluation and the minimum of the incorrect assignment evaluations (Table 4.4, column 1). Across the five databases used to simulate Q, the probability of a conservative heuristic evaluation ranged from 0.994 to 0.999, slightly wider than the range for the single contributor tests.

4.7.2 Unknown second contributor

The WoE for the same 25 000 two-contributor CSPs was then evaluated assuming that second contributor was unknown, giving hypotheses of the form:

$$H_p: \quad Q + U$$
$$H_d: \quad X + U$$

For these evaluations the heuristic WoE was evaluated (assuming the database of Q for both X and U), as well as the WoE for three separate alternative calculations:

1. Correct database assignment for both X and U.
2. Correct database for U, minimum WoE over all possible assignments of the database for X.
3. Database of X and U are assumed the same, minimum WoE over the four possible alternative databases.

Alternative 1 is only applicable in the 20 datasets where the true database of X does not match that of U, while alternatives 2 and 3 are applicable in all 25 datasets.

Alternative 1 can be thought of as the most appropriate WoE, the heuristic should be conservative because the probability that Q and U share alleles is increased through the action of the F_{ST} adjustment. When the second contributor to the CSP is unknown, on average 0.997 of the simulations were conservative using the heuristic compared to using alternative 1, with the minimum fraction of conservative simulations = 0.993 over the 20 database choices (Figure 4.2). The heuristic is conservative because the defence likelihood is maximised when the probability of U and Q having matching alleles is maximised, which is achieved by using the same database for U that Q was sampled from (as in the heuristic) together with a high value of F_{ST} (0.03 here). The prosecution and defence likelihoods are similarly affected by the probabilities of alleles

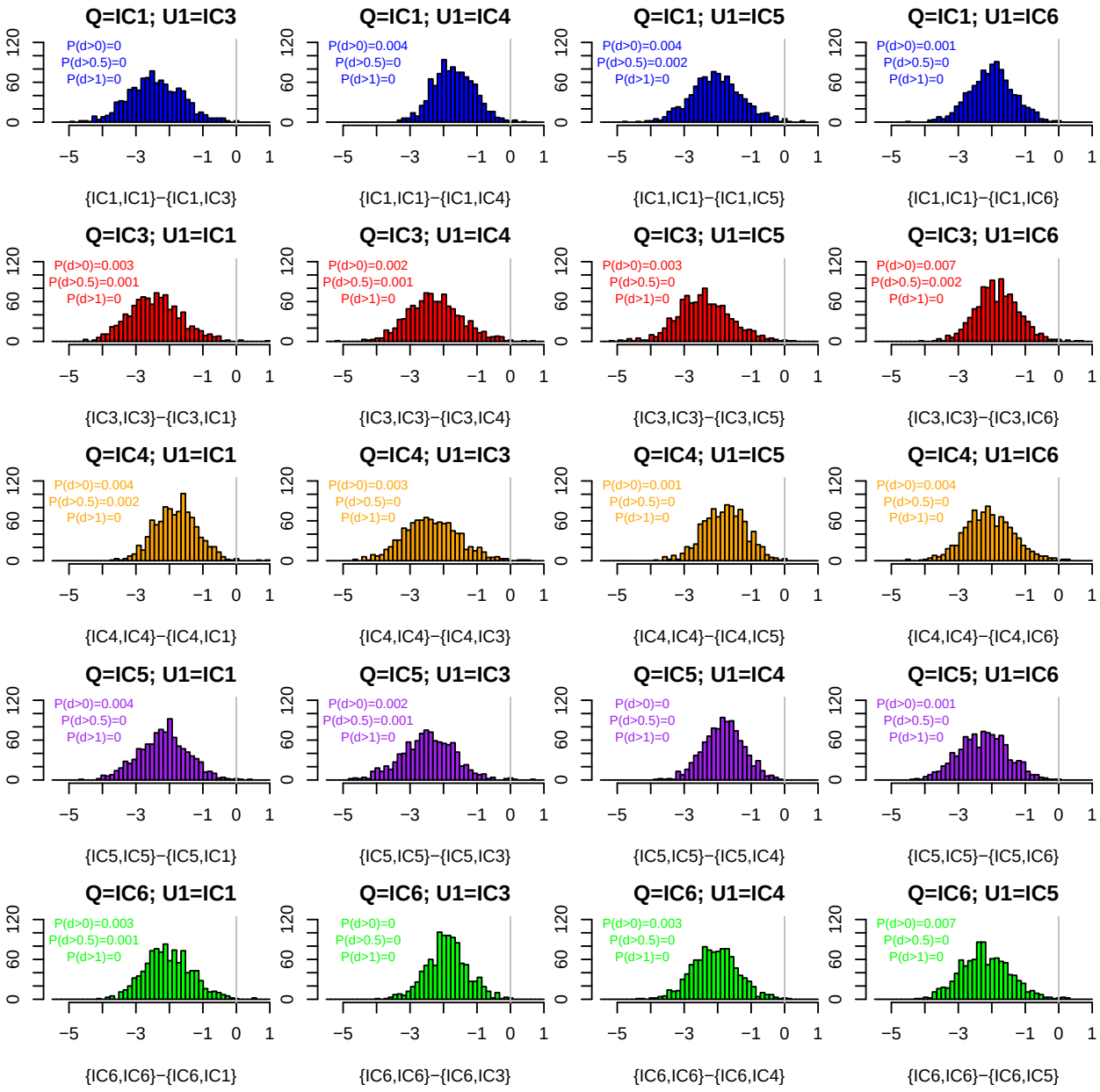


Figure 4.2: The effect of database on Weight of Evidence (WoE) for two-contributor CSPs with alternative 1. The databases are described in Table 4.3. The x-axis shows the WoE computed using the database of Q for both contributors minus that obtained using the correct databases for X and U. The title of each subplot indicates the databases from which each contributor was simulated, where Q is the queried contributor and U is an unknown contributor. The x-axis labels indicate the databases used for each contributor in the analysis. P(d>x) indicates the proportion of differences that are > x. Colour indicates the database of Q.

Contributors under Hd	**X**	**X+K**	**X+U**		
			True both	True U	Same dbase
Heuristic (bans)	20.3	17.8	10.7	10.7	10.7
Alternative (bans)	24.5	20.7	12.8	14.1	14.0
Difference (bans)	4.2	3.0	2.1	3.4	3.2
Difference (%)	18.8	15.6	17.9	27.4	25.9

Table 4.4: Mean Weight of evidence (WoE) for the heuristic rule and the alternatives discussed in the text. The mean of the differences between the heuristic and alternative scenarios is also shown. The % Difference row shows the mean difference as a percentage of the average of the heuristic and alternative means.

that U and Q do not share, so these are less important. The heuristic gives P(WoE>9) = 0.90295, so the LR is in excess of one billion in the majority of cases, which is the maximum LR reported to court in the UK.

Alternative 2 compares the heuristic WoE when the database of Q has been correctly ascertained to the WoE in a situation where the database for U has been correctly ascertained, but the database of Q has been incorrectly ascertained, so the database used for X is incorrect. This mimics a situation where there may be information on the ancestry of U, who perhaps refused to be genotyped as he was not under suspicion, but the ancestry of X is unknown. For alternative 2, the heuristic WoE is greater than the minimum of alternative WoEs with probability >0.994 (Figure 4.3), with an average difference between the heuristic and minimum alternative WoEs of 3.4 bans (Table 4.4). Note that the largest discrepancies are observed when the heuristic incorrectly assigns the database of U (plots off the diagonal).

Alternative 3 compares the heuristic WoE when the database of Q has been assigned correctly to X with the WoE when the heuristic has been used, but the database of Q has been misassigned so incorrectly assigned to X, taking the minimum over all possible misassignments. This mimics a situation where the heuristic is in use, but perhaps Q has lied about his ancestry, or the arresting officer has incorrectly assigned his ancestry. For alternative 3, the heuristic WoE is greater than the minimum of alternative WoEs with probability >0.996 (Figure 4.4), with an average difference between the heuristic and minimum alternative WoEs of 3.2 bans (Table 4.4).

4.8 Heuristic vs. alternatives

Summary results for the simulation experiments presented in this Chapter are given in Table 4.4. In absolute terms the mean difference between the heuristic WoE and alternative WoEs decreases as the problem difficulty increases (with a corresponding decrease in WoE), but increases in relative terms compared to the alternative

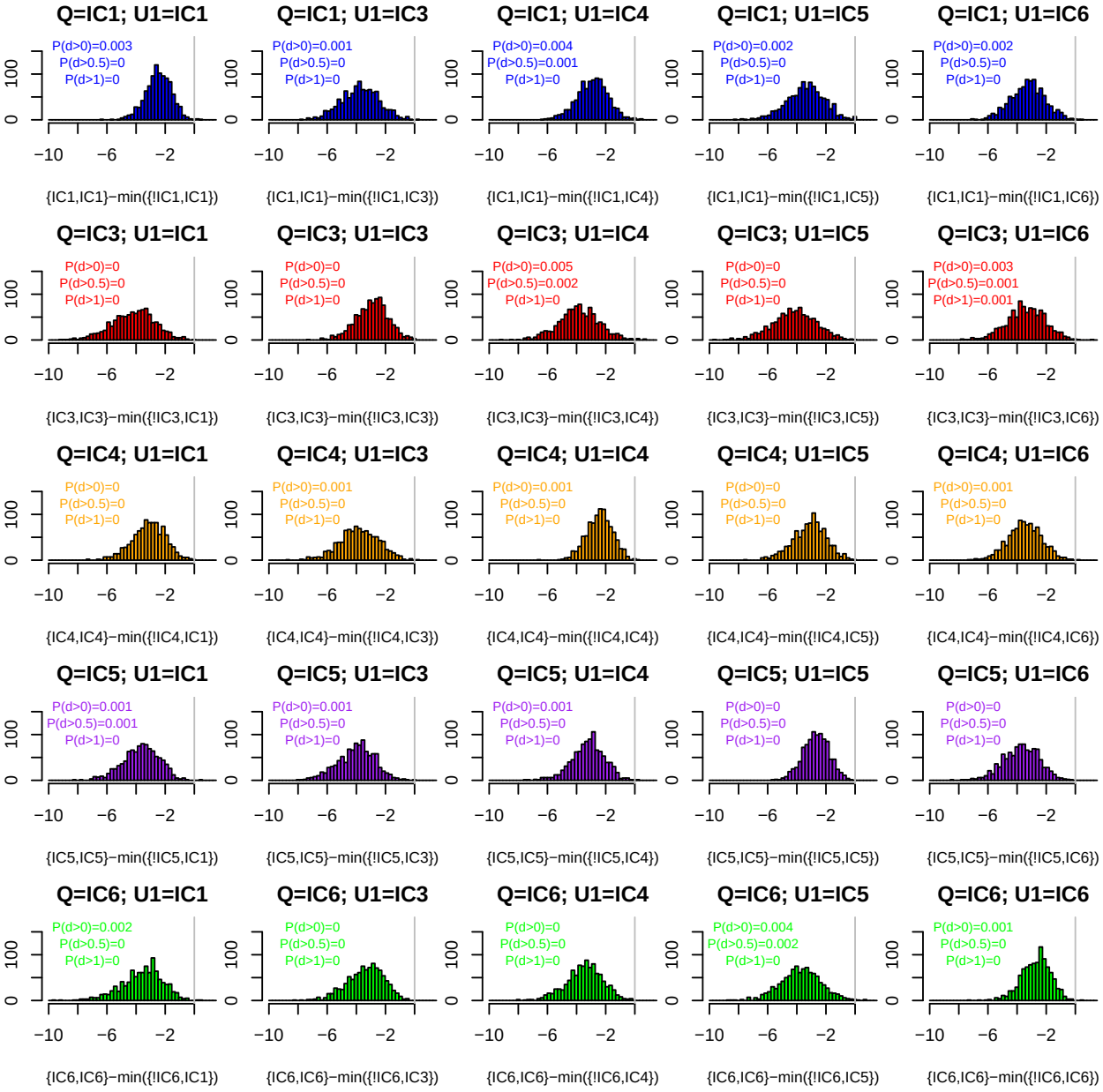


Figure 4.3: The effect of database on Weight of Evidence (WoE) for two-contributor CSPs with alternative 2. The databases are described in Table 4.3. The x-axis shows the WoE computed using the database of Q for both contributors minus the minimum WoE obtained over all other choices of databases for X, always using the correct database for U. The title of each subplot indicates the databases from which each contributor was simulated. The x-axis labels indicate the databases used for each contributor in the analysis (!IC1 indicates all databases other than IC1). P(d>x) indicates the proportion of differences that are > x. Colour indicates the database of Q.

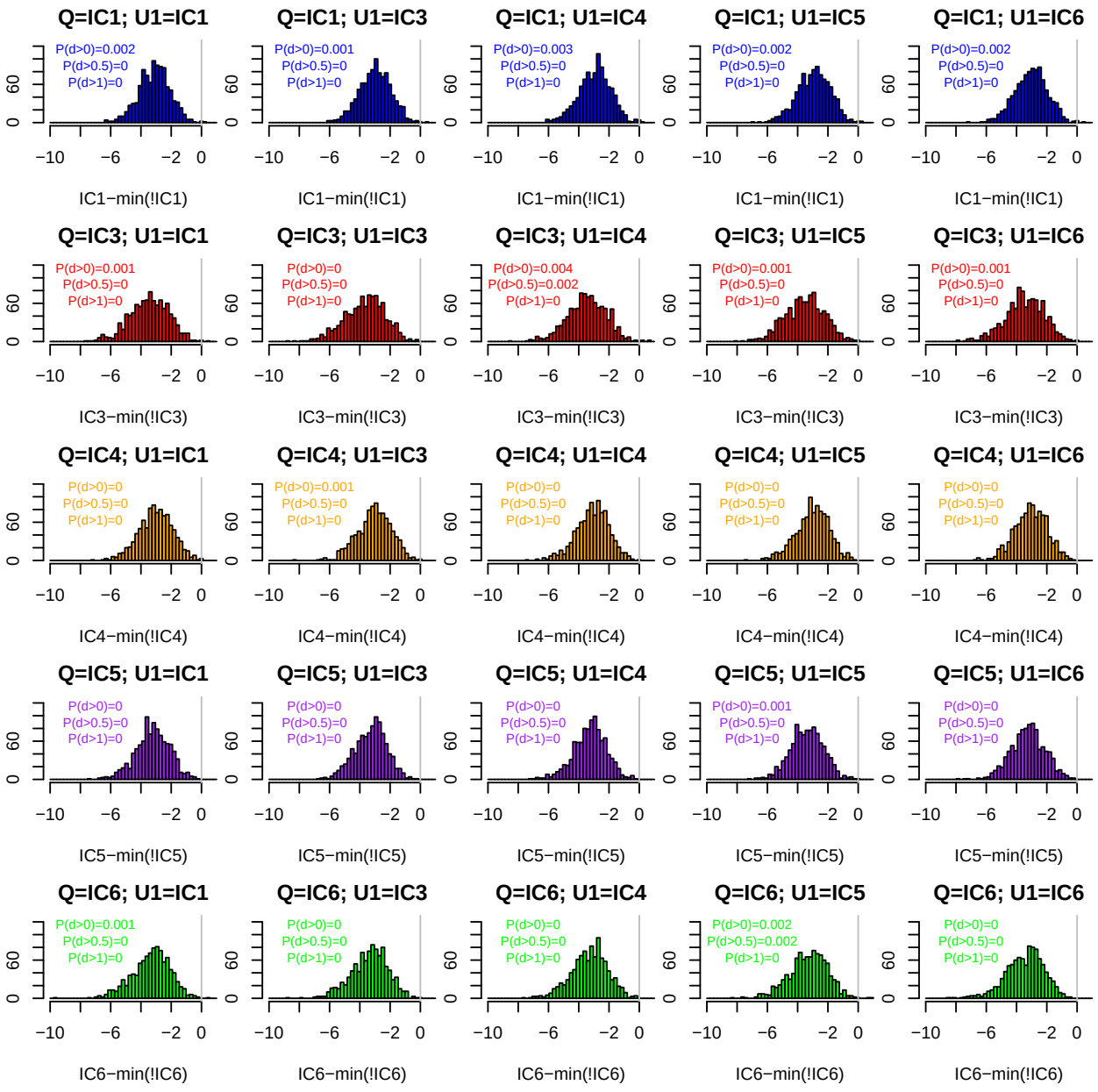


Figure 4.4: The effect of database on Weight of Evidence (WoE) for two-contributor CSPs with alternative 3. The databases are described in Table 4.3. The x-axis shows the WoE computed using the database of Q for both contributors minus the minimum WoE obtained using each other database in turn for both X and U. The title of each subplot indicates the databases from which each contributor was simulated. The x-axis labels indicate the database used for both contributors in the analysis. P(d>x) indicates the proportion of differences that are > x. Colour indicates the ancestry of Q.

WoE. This suggests that if three or more contributors were to be considered the mean difference between the heuristic WoE and alternative WoEs would be a larger percentage of the alternative WoEs. However, more than two contributors were not considered here due to the computational demands of such a large simulation study.

These results demonstrate that the heuristic WoE calculation proposed here is almost always conservative (at least 99.3% of simulations) compared to multiple alternatives for single contributor situations (Figure 4.1), two-contributor situations where the second contributor is an uncontested known contributor, and two-contributor situations where the second contributor is instead unknown when the alternative is the ground truth (Figure 4.2), when the alternative is the correct database for U but the most favourable database for X (Figure 4.3) and when the alternative is the most favourable database for X applied to both X and U (Figure 4.4). When the heuristic WoE was not conservative, the difference between the heuristic WoE and alternative WoE was always < 1.5 bans.

4.9 World populations

There are many ways to divide the world's populations into different subpopulations. Chapter 3 along with [Ramachandran et al., 2005] demonstrated that allele probabilities generally change smoothly with geographic distance. As an example from Chapter 3, whether or not to define Afghanistanis as Middle Eastern or South Asian becomes a somewhat subjective choice, depending on the criteria that are most important to the current study; they will be genetically very similar to both Iranians and Pakistanis, their Middle Eastern and South Asian neighbours respectively, but will be distinct from Moroccans and Bangladeshis, the Middle Eastern and South Asian nations with the greatest geographical distance from Afghanistan. As a consequence, in any particular case there is no "correct" choice of alternative subpopulations to evaluate the WoE for. Instead, the desired statistic would be a WoE averaged over each possible database allocation for X, where each database WoE is weighted by its plausibility given the specifics of the case in question; CCTV footage showing some exposed Caucasian skin may weight the average WoE so that it is almost identical to the IC1 WoE, whereas the WoE in a case where each population is equally likely (e.g. a crime in a cosmopolitan city for which there is no evidence indicating the ethnicity of the true culprit) would essentially be an unweighted average WoE across the possible database choices.

4.10 Casework recommendations

The heuristic presented here will return results that are almost certainly conservative, and therefore favours defendants. Section 4.6.2 verified that this behaviour does not result from Q being sampled from the same database that is subsequently used in the analysis phase, analogous to ensuring that a model is not being tested on the same data that it was trained on.

Altering the value of F_{ST} alters the extent to which the heuristic WoE is conservative, so the choice of an F_{ST} value can be used to tailor the WoE analysis to the case at hand. Chapter 3 demonstrated that $F_{ST} = 0.03$ is greater than the majority of median F_{ST} estimates from global comparisons of subpopulations with continental populations. This chapter has further demonstrated that $F_{ST} = 0.03$ is also sufficient for the heuristic WoE to be conservative compared to various alternative calculations for the majority of cases. Therefore, the recommendations of Chapter 3 to utilise an $F_{ST} = 0.03$ in routine casework is reiterated here. While the results in Chapter 3 suggests that a lower value of F_{ST} may be used for e.g. Caucasians, the results presented in this chapter demonstrate that utilising a sufficiently large value of F_{ST} ensures that the WoE is conservative compared to a range of possible alternative calculations. As a result, the recommendation stemming from this chapter is to utilise an identical sufficiently large value of F_{ST} regardless of the population of Q (here $F_{ST} = 0.03$), despite the demonstration in Chapter 3 that within-population F_{ST} values differ across populations. With such a recommendation, the WoE should remain conservative even if the population of Q has been misassigned.

4.11 Misassigning the database of Q

Here, using the database most appropriate for Q for all unknowns has been demonstrated to be favourable to defendants, given a sufficiently large value of F_{ST}. Misassignment of the ancestry of Q and subsequent use an incorrect database, or if no database is appropriate to Q, will overstate the evidence against Q through an inflated WoE. However, this is mainly accounted for here through the use of a large value for F_{ST}, which decreases the WoE, ensuring that the WoE remains conservative even when the database of Q is misassigned, and reducing the detrimental impact of the misassignment. A smaller value of F_{ST} may not be sufficient to counteract the effect of database misassignment, while a larger value of F_{ST} may favour the defendant unnecessarily.

Chapter 5

Developing a peak height (PH) model for the evaluation of forensic likelihoods

Some of the work in this chapter has been published in Steele et al. [2016], see Appendix B for the accepted manuscript. All work was performed by me.

5.1 Information available from PHs

All results presented in previous chapters use a discrete model to calculate the probability of a CSP given some hypothesis, where the CSP peaks are classified as "certain", "uncertain" or "non-allelic". This classification of peaks uses the PH information, so the discrete model indirectly incorporates PHs. However, much information available in the CSP is lost by not modelling epg PHs explicitly. In this chapter, an implementation of a PH model is described that aims to fully utilise the PH information available in a CSP; the results from validation tests of the implemented model will be presented in Chapter 6. This model has been published as v6.0 of the CRAN package likeLTD.

As a motivating example for the utility of a model that incorporates PHs explicitly, suppose a single-locus CSP as shown in Figure 5.1 has been observed. Assume that using the discrete model the peaks at positions 4, 6, 8 and 10 would be called by the forensic practitioner as non-allelic, while peaks at positions 5, 7, 9 and 11 would be called as allelic. With these allele calls, a reasonable set of hypotheses would be $Q/X+$ U. Ignoring the ordering of contributors and the effects of degradation, under H_d the discrete model described in Chapter 1 would give equal weights to the genotypes {5,7}&{9,11}, {5,9}&{7,11}, and {5,11}&{7,9}, each with probability $4p_5p_7p_9p_{11}(1-D_{1,1})^2(1-D_{1,2})^2$, where $D_{1,1}$ is the heterozygote dropout probability

for the first contributor, and $D_{1,2}$ is the corresponding probability for the second contributor. However, information from PHs, as seen in Figure 5.1, allows for an inference that the genotype pair {5,9}&{7,11} is the most likely with average DNA contributions in RFU of 500 and 1000 RFU respectively, assuming a low PH variability. Therefore, a suitable PH model should return this genotype pair as the most likely, and should provide minimal weight to the genotype pairs {5,7}&{9,11} and {5,11}&{7,9}, as these would require a high variability in PHs to explain the observed PHs; in each case of an incorrect genotype assignment the best fit to the data would be two equal contribution contributors at 750 RFU, with a corresponding high PH variance. A high PH variability may negate any inference of true genotypes, so that all three genotype pairs are reasonable. The utilisation of PH information to enhance inference has previously been described in Pascali and Merigioli [2012], Perlin and Sinelnikov [2009], Perlin and Szabady [2001].

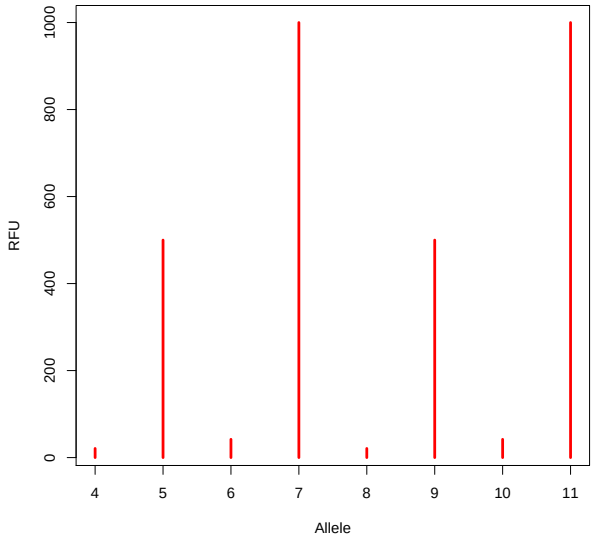


Figure 5.1: An example two-contributor single-locus CSP for which the incorporation of PH information will give useful insight as to the true genotypes of the contributors. Using a PH model, peaks at alleles 4, 6, 8 and 10 will be determined to be non-allelic, peaks at alleles 5 and 9 will be determined to originate from a minor contributor, while peaks at alleles 7 and 11 will be determined to originate from a major contributor.

A benefit of using a PH model, that is not highlighted by this example, is lack of having to classify peaks as allelic/uncertain/non-allelic. For complex mixtures it can be difficult to decide if a peak in a stutter position of a major peak has some allelic contribution, especially with high peak height variability. Furthermore, because peaks do not have to be classified, the detection threshold employed to generate the CSP can be lowered, giving extra information regarding peaks in stutter positions, and low-level peaks. A high detection threshold has the potential to remove allelic peaks of a minor contributor of interest, in an

effort to remove non-allelic peaks of a major contributor.

To allow for inter-Q comparisons, throughout this chapter results will be presented as WoE/$\log_{10}$(IMP), which will be termed the Information Gain Ratio (IGR). The IGR has a maximum at 1.0, and an IGR<0 supports H_d.

5.2 The model

A CSP with replicates R and loci L, a hypothesis with contributors C and a database with alleles I are all provided. Each element of G is one set of genotype allocations to the hypothesised contributors that can explain the CSP. Q and K have known genotypes, so the elements of G vary according to the genotypes allocated to any U. A single locus is considered here, $l \in L$, for which the database alleles at locus l are available, I_l, as well as joint genotype allocations that are being considered at locus l, G_l. From these the effective dose, E, at replicate $r \in R$, joint genotype allocation $g \in G_l$ and contributor $c \in C$ for a specific allele $i \in I_l$ is calculated as:

$$E_{l,r,g,c,i} = n_{g,c,i}\rho_r\chi_c(1+\delta_c)^{-f_i}, \tag{5.1}$$

where ρ is a multiplicative adjustment to the expected dose across replicates, χ is the expected DNA contribution of an individual in the first replicate in RFU, δ is the degradation parameter that adjusts the expected dose based on the mean adjusted length of an allele in base pairs, f is the mean adjusted fragment length and $n_{g,c,i} \in \{0,1,2\}$ indicating how many copies of allele i contributor c possesses in joint genotype allocation g. n can be thought of as a three-dimensional array with rows G_l, columns I_l and layers C, with each cell taking values 0, 1 or 2, and the rows sum to 2. Equation (5.1) gives the replicate and degradation-adjusted dose for a single individual at allele i.

The expected dose that remains at position i (A), stutters to position $i-1$ (S), double-stutters to position $i-2$ (D) or over-stutters to position $i+1$ (O) for each $E_{l,r,g,c,i}$ is calculated as:

$$\begin{aligned}
S_{l,r,g,c,i} &= \beta\alpha_l u_i E_{l,r,g,c,i},\\
D_{l,r,g,c,i} &= \eta E_{l,r,g,c,i},\\
O_{l,r,g,c,i} &= \theta E_{l,r,g,c,i},\\
A_{l,r,g,c,i} &= E_{l,r,g,c,i} - (S_{l,r,g,c,i} + D_{l,r,g,c,i} + O_{l,r,g,c,i})
\end{aligned}$$

where β is the mean gradient for the linear relationship between longest uninterrupted sequence (LUS, u) and stutter fraction [Brookes et al., 2012, Bright et al., 2013c, Kelly et al., 2014] across loci, α_l is a locus

adjustment parameter for the stutter gradient for $l \in L$, η is the mean double-stutter fraction across loci and θ is the mean over-stutter fraction across loci. In reality the over- and double-stutter rates (η and θ) also have a linear relationship with LUS, however, the expected number of observed over- and double-stutters in a single CSP is too low to estimate the gradient.

The A values for each i are then summed with all of the S values from position $i+1$, D values from position $i+2$ and O values from position $i-1$ across all individuals c to give the expected PH at each position i:

$$P_{l,r,g,i} = p_i\lambda(1+\epsilon)^{-f_i} + \sum_{c \in C} A_{l,r,g,c,i} + S_{l,r,g,c,i+1} + D_{l,r,g,c,i+2} + O_{l,r,g,c,i-1}. \tag{5.2}$$

where $p_i\lambda$ is the unadjusted dropin dose for allele i, with λ being the dropin parameter, which gives a dropin dose in RFU, which is then adjusted for degradation, where ϵ is the degradation rate for dropin peaks. Dropin of a given allele can reasonably be expected to occur in proportion to the frequency of that allele in the population, so if there is a given environmental DNA load in RFU, λ, then for each allele, i, in the population database we expect $p_i\lambda$ dose of that allele as a dropin dose in RFU in a given replicate of a CSP. Note that this dropin dose is subject to degradation at a separate rate to that of non-dropin doses.

The PHs at positions i, $h_{l,r,i}$, are then assumed to be gamma distributed:

$$h_{l,r,i} \sim \Gamma(P_{l,r,g,i}, \sigma P_{l,r,g,i}), \tag{5.3}$$

where σ is the scale parameter for the gamma distribution which is constant across joint genotype allocations and loci. Here the parameterisation of the gamma distribution is using the mean ($P_{l,r,g,i}$) and variance ($\sigma P_{l,r,g,i}$). For observed peaks, a discrete approximation to the probability mass function is computed as:

$$F(h_{l,r,i} + 0.5 | P_{l,r,g,i}, \sigma P_{l,r,g,i}) - F(h_{l,r,i} - 0.5 | P_{l,r,g,i}, \sigma P_{l,r,g,i}), \tag{5.4}$$

where F is the cumulative distribution function for the gamma distribution. This approximation partitions the continuous probability for PHs into discrete bins, where each RFU value of PH incorporates the probability of all PHs that would be rounded to that RFU value.

Any expected peak from the model for the current joint genotype allocation, g, that was unobserved in the CSP (dropout) is given a probability mass of:

$$F(t_l - 0.5 | P_{l,r,g,i}, \sigma P_{l,r,g,i}), \tag{5.5}$$

where t_l is the detection threshold used when analysing the epg at locus l. This is the probability of a PH having been sub-threshold, given the mean and variance of the expected peak.

Unobserved alleles with a low probability in the population are classified as "rare", and combined into a single allele. When the genotypes of the unprofiled contributors (X and U) include > 1 alleles classified as rare, these are assumed to be distinct i.e. no shared alleles among them. Note that rare combined alleles are necessarily unobserved, therefore non-dropout probabilities do not need to be adjusted.

5.2.1 Penalties and constraints

Parameters are penalised as shown in Table 5.1. The δ, ϵ and σ penalties are designed to constrain the parameters so that overly large values are penalised. The α penalty constrains each locus to have a gradient close to the mean gradient. The β, η and θ penalties are intended to support a wide range of plausible values. The incorporation of a penalty to the likelihood function is the maximisation equivalent to defining a prior distribution on parameters in an integration scenario such as that used in Markov chain Monte Carlo (MCMC) methods.

Parameter	Distribution	Mean	SD
β	N	0.013	0.010
η	Γ	0.02	0.019
θ	Γ	0.02	0.019
$\log_{10}\alpha_l$	N	0	0.300
δ_c	e	0.02	0.020
ϵ	e	0.02	0.020
σ	e	100	0.010

Table 5.1: Penalties applied to the parameters of the PH model. Distribution gives the penalty distribution; N=normal, Γ=gamma, e=exponential.

5.2.2 Combining probabilities and maximisation

At allele i the probability of an observed peak with height $h_{l,r,i}$ is given in (5.4), while the probability of no peak observation ($h_{l,r,i} < t_l$) is given in (5.5). If this conditional probability is represented as a function $a(h_{l,r,i}, P_{l,r,g,i}, \sigma, t_l)$, then individual peak probabilities are combined to form a likelihood as:

$$\prod_{l\in L}\left[\sum_{g\in G_l}\left[\prod_{r\in R}\left[\prod_{i\in I_l} a(h_{l,r,i}, P_{l,r,g,i}, \sigma, t_l)\right]\prod_{c\in C} Pr(\mathcal{G}_{g,c})\right]\pi_l\right] \tag{5.6}$$

where π_l is the combined penalty on the likelihood at locus l given the parameter values for β, η, θ, α_l, δ, ϵ

and σ. Note that (5.6) is essentially a restating of (1.2), but with a focus on PHs and an incorporation of a penalty.

The likelihood is then maximised using a genetic algorithm that simulates "mutation", "recombination" and "selection" on sets of randomly generated parameter values to obtain the set of parameters that gives the highest penalised likelihood [Mullen et al., 2011]. Note that the prosecution and defence likelihoods are maximised separately.

5.3 Theoretical predictions

The genotype probabilities for a number of artificial single-locus CSPs were evaluated using both a simplified version of the PH model, and the dropout model. CSPs were generated from the genotypes of each contributor, and a heterozygote dose for each contributor, from which PHs of the CSP can be calculated. The stutter fraction used to generate the CSPs was constant across alleles i.e. does not vary with LUS, and was set at 0.1. CSPs were not subject to degradation. All peaks above t_l were included in the CSP, where $t_l = 50$ RFU.

For the PH model, $P_{l,r,g,i}$ were calculated assuming no double-stutter or over-stutter, a constant stutter rate across alleles, no degradation and no dropin. Contributor doses, χ_c, were assumed equal to the heterozygote dose used to generate the CSP. To calculate the genotype probabilities a number of parameters were fixed; $\sigma = 10$, $t_l = 50$. Probabilities of hypothesised dropouts were calculated using (5.5), while probabilities of observed peaks were calculated using (5.4).

For the discrete model, all allelic peaks were correctly labelled as allelic, while all stutter peaks were labelled as non-allelic. Dropout probabilities were calculated using the scheme of Tvedebrink et al. [2009] with $\beta_1 = -4.35$, as estimated by Tvedebrink et al. [2009], and $\beta_0 = 18.556$ which is an average over locus estimates in Tvedebrink et al. [2009].

5.3.1 Single contributor

Using both the discrete model and PH model, a breakdown of the likelihood ratios for a number of single-contributor CSPs is given in Table 5.2. Under $\mathcal{C}$ bold alleles indicate allelic peaks, while non-bold alleles indicate stutter peaks, so the first four CSPs are low-template, and the last three are good-template. Focussing on $\Pr(R|\mathcal{G})$, any mismatch between a hypothesised peak and its observation will have a low probability, leading to a low $\Pr(\mathcal{G}_X)$ e.g. under $\mathcal{C}$=**5**,**8**, $\mathcal{G}_X$=6,8 the $D(A_6)$ indicates that a hypothesised allelic dose

Label	$\mathcal{C}$	$\mathcal{G}_X/$ $\mathcal{G}_Q$	$\Pr(R\|\mathcal{G})$		P(G)	
			Discrete	PH	H_p	H_d
(a)	**5**,**8**	5,8	$(1-D)^2$	$O(h_5,A_5)O(h_8,A_8)D(S_5)D(S_8)$	1	$2p_5p_8$
		6,8	-	$D(A_6)O(h_8,A_8)O(h_5,S_6)D(S_8)$	-	$2p_6p_8$
		5,9	-	$O(h_5,A_5)D(A_9)D(S_5)O(h_8,S_9)$	-	$2p_5p_9$
		6,9	-	$D(A_6)D(A_9)O(h_5,S_6)O(h_8,S_9)$	-	$2p_6p_9$
(b)	**5**,**6**	5,6	$(1-D)^2$	$O(h_5,A_5+S_6)O(h_6,A_6)D(S_5)$	1	$2p_5p_6$
		6,6	-	$O(h_5,2S_6)O(h_6,2A_6)$	-	p_6^2
		5,7	-	$O(h_5,A_5)O(h_6,S_7)D(A_7)$	-	$2p_5p_7$
		6,7	-	$O(h_6,A_6+S_7)D(A_7)O(h_5,S_6)$	-	$2p_6p_7$
(c)	**5**	5,5	$1-D_2$	$O(h_5,2A_5)D(2S_5)$	1	p_5^2
		5,6	-	$O(h_5,A_5+S_6)D(A_6)D(S_5)$	-	$2p_5p_6$
		6,6	-	$D(2A_6)O(h_5,2S_6)$	-	p_6^2
		4,5	-	$D(A_4+S_5)O(h_5,A_5)D(S_4)$	-	$2p_4p_5$
		5,Z	$D(1-D)$	$O(h_5,A_5)D(A_Z)D(S_5)D(S_Z)$	-	$2p_5p_Z$
		6,Z	-	$D(A_6)D(A_Z)O(h_5,S_6)D(S_Z)$	-	$2p_6p_Z$
(d)	**5**	5,5	$(1-D_2)$	$O(h_5,2A_5)D(2S_5)$	-	p_5^2
		5,6	-	$O(h_5,A_5+S_6)D(A_6)D(S_5)$	-	$2p_5p_6$
		6,6	-	$D(2A_6)O(h_5,2S_6)$	-	p_6^2
		4,5	-	$D(A_4+S_5)O(h_5,A_5)D(S_4)$	-	$2p_4p_5$
		5,Z	$D(1-D)$	$O(h_5,A_5)D(A_Z)D(\mu_S,5)D(S_Z)$	-	$2p_5p_Z$
		6,Z	-	$D(A_6)D(A_Z)O(h_5,S_6)D(S_Z)$	-	$2p_6p_Z$
		5,8	$D(1-D)$	$O(h_5,A_5)D(A_8)D(S_5)D(S_8)$	1	$2p_5p_8$
		5,9	-	$O(h_5,A_5)D(A_9)D(S_5)D(S_9)$	-	$2p_5p_9$
		6,8	-	$D(A_6)D(A_8)O(h_5,S_6)D(S_8)$	-	$2p_6p_8$
		6,9	-	$D(A_6)D(A_9)O(h_5,S_6)D(S_9)$	-	$2p_6p_9$
(e)	4,**5**,7,**8**	5,8	$(1-D)^2$	$O(h_5,A_5)O(h_8,A_8)O(h_4,S_5)O(h_7,S_8)$	1	$2p_5p_8$
(f)	4,**5**,**6**	5,6	$(1-D)^2$	$O(h_5,A_5+S_6)O(h_6,A_6)O(h_4,S_5)$	1	$2p_5p_8$
		4,6	-	$O(h_4,A_4)O(h_6,A_6)D(S_4)O(h_5,S_6)$	-	$2p_4p_6$
		5,7	-	$O(h_5,A_5)D(A_7)O(h_4,S_5)O(h_6,S_7)$	-	$2p_5p_7$
(g)	4,**5**	5,5	$1-D_2$	$O(h_5,2A_5)O(h_4,2S_5)$	1	p_5^2
		5,6	-	$O(h_5,A_5+S_6)D(A_6)O(h_4,S_5)$	-	$2p_5p_6$
		4,5	-	$O(h_4,A_4+S_5)O(h_5,A_5)D(S_4)$	-	$2p_4p_5$
		4,6	-	$O(h_4,A_4)D(A_6)D(S_4)O(h_5,S_6)$	-	$2p_4p_6$
		5,Z	$D(1-D)$	$O(h_5,A_5)D(A_Z)O(h_4,S_5)D(S_Z)$	-	$2p_5p_Z$

Table 5.2: Likelihoods for a number of single contributor CSPs ($\mathcal{C}$) split into $Pr(R|G)$ and $P(G)$ using both the discrete model and the PH model. The genotype of Q or X for each term of the likelihood is given in $\mathcal{G}_X/\mathcal{G}_Q$. CSP alleles in bold are allelic peaks, non-bold are non-allelic. Only single-stutter is considered here. CSPs (a-d) simulate low-template profiles while CSPs (e-g) simulate good-template profiles. D is the dropout probability, which under the PH model is a function of expected PH, D_2 gives the dropout probability for a homozygous allele. O is the probability of an observed peak as a function of both the observed PH, h, and the expected PH. A_x indicates the allelic contribution to expected PH at allele x, S_x indicates the stutter contribution to expected PH at allele $x-n$ that stuttered from allele x.

allele has dropped out, which has low probability if the allelic dose is sufficiently large, just as $O(h_5, S_6)$ indicates that an observed truly allelic peak is hypothesised as having only stutter dose which is likewise improbable if the allelic dose is sufficiently large. Taking into account these mismatches, most CSPs are left with a single $\mathcal{G}_X$ having a high $\Pr(R|\mathcal{G})$. To verify this, the $\Pr(R|\mathcal{G})$s were calculated for a range of expected heterozygous peak doses (51-151 RFU for low-template CSPs, 300-1500 for good-template CSPs).

Calculating $Pr(R|\mathcal{G})$ for all X in Table 5.2 gives support for a single genotype under H_d for most CSPs under both models over the range of heterozygote doses tested (see Figure 5.2), so discrete and PH models are broadly analogous for single-contributor CSPs, regardless of DNA template, and should give similar LRs. CSPs (c) and (d) depart from this behaviour, as in both cases each model starts to support a $\mathcal{G}_X$ that includes dropout at low RFU. In CSP (c) the true contributor is homozygous and the support for dropout genotypes is $\approx$ 20% at 51 RFU, whereas in CSP (d) Q is heterozygous and the support for dropout genotypes is $\approx$ 90% using the PH model at 51 RFU. Similarly, the discrete model has 20% support for dropout genotypes at low RFU when the true contributor is homozygous and 40% support for dropout genotypes when the true contributor is heterozygous. What appears to be a significant difference between the two models is in fact a difference of degree rather than a difference of kind; both models have increasing support for a dropout genotype at low RFU, and both models support a dropout genotype more when the true contributor is heterozygous than when he is homozygous. Note that the discrete model has greatest support for dropout genotypes when D=0.5, as this is the value of D that maximises $D(1-D)$, which occurs at 71 RFU, which corresponds to a heterozygote dose of 79 RFU taking into account that H_1 is estimated after stutter; this matches the highest probability of dropout genotypes seen at $\approx$80 RFU in Figure 5.2.

5.3.2 Two contributors

No shared alleles

For a mixed CSP the PH model is expected to utilise the extra information in the PH data to return a more extreme WoE in favour of the true hypothesis compared to the discrete model when the contributors have unequal contributions and do not share any alleles. When the contributors have approximately equal contributions the two models should perform similarly, as genotype deconvolution is difficult at equal contributions and there are no shared alleles in this CSP. Once again this can be demonstrated by breaking down the likelihood ratios (Table 5.3) for various CSPs. The same improbable observed/expected PH pairings that are encountered in single contributor profiles apply; hypothesised non-allelic positions paired with observed truly allelic peaks and *vice versa* will have a low probability e.g $D(A_{8,X})$ in the second CSP. In

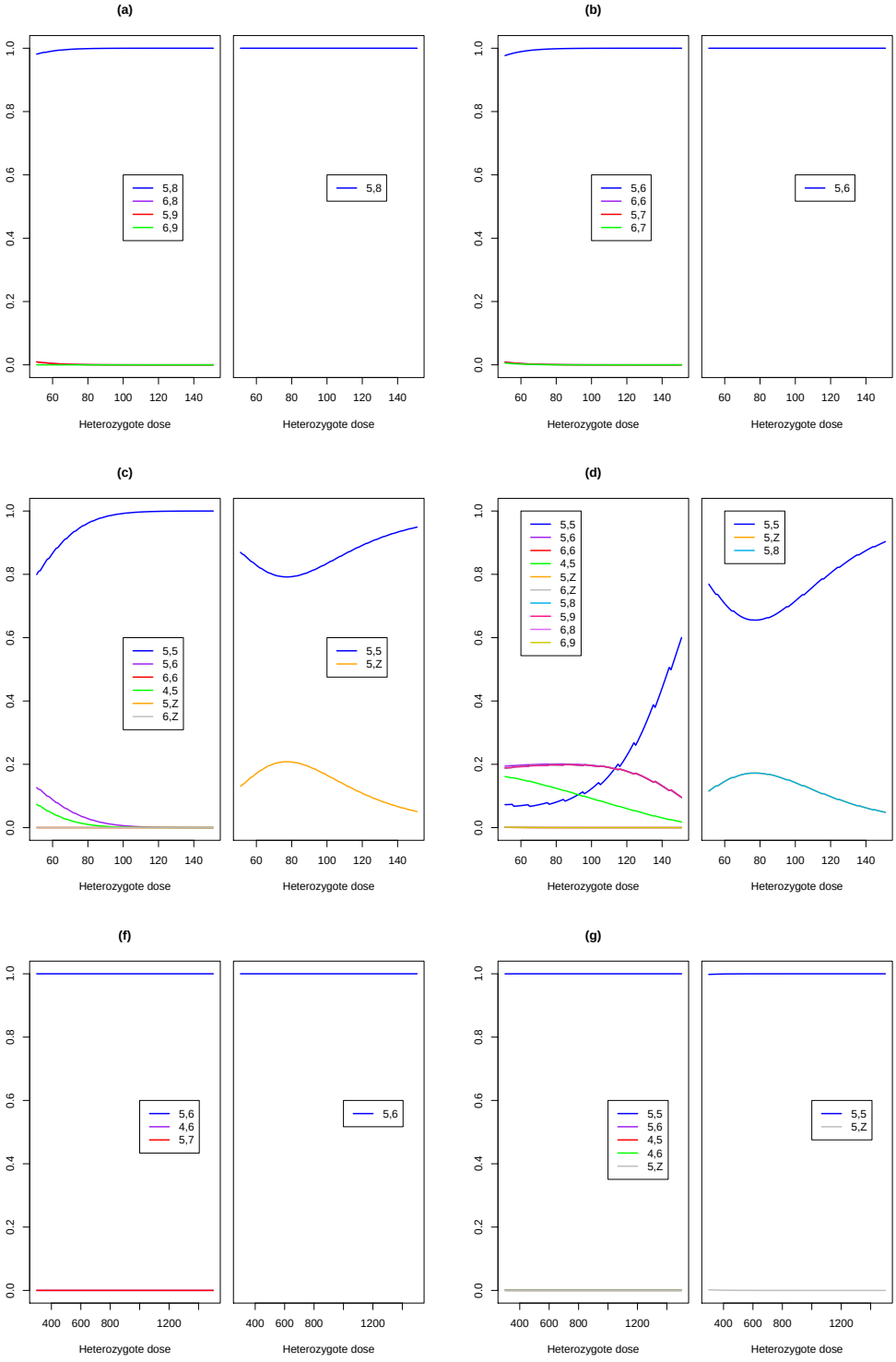


Figure 5.2: $Pr(R|\mathcal{G}_X)$ under H_d for CSPs in Table 5.2. CSP (e) is omitted as there is only a single genotype considered. Y-axis shows the normalised $Pr(R|\mathcal{G}_X)$, x-axis shows the heterozygous dose used to generate the CSP and calculate $Pr(R|\mathcal{G}_X)$. Legend displays $\mathcal{G}_X$.

$\mathcal{C}$	$\mathcal{G}_X$/ $\mathcal{G}_Q$	$\mathcal{G}_U$/ $\mathcal{G}_K$	$\Pr(R\|G)$	
			Discrete model	PH model
2,**3**,4,**5**, 6,**7**,8,**9**	3,5	7,9	$(1-D_U)^2(1-D_X)^2$	$O(h_2,S_{3,X})O(h_3,A_{3,X})O(h_4,S_{5,X})O(h_5,A_{5,X})O(h_6,S_{7,U})O(h_7,A_{7,U})O(h_8,S_{9,U})O(h_9,A_{9,U})$
	3,7	5,9		$O(h_2,S_{3,X})O(h_3,A_{3,X})O(h_4,S_{5,U})O(h_5,A_{5,U})O(h_6,S_{7,X})O(h_7,A_{7,X})O(h_8,S_{9,U})O(h_9,A_{9,U})$
	3,9	5,7		$O(h_2,S_{3,X})O(h_3,A_{3,X})O(h_4,S_{5,U})O(h_5,A_{5,U})O(h_6,S_{7,U})O(h_7,A_{7,U})O(h_8,S_{9,X})O(h_9,A_{9,X})$
	7,9	3,5		$O(h_2,S_{3,U})O(h_3,A_{3,U})O(h_4,S_{5,U})O(h_5,A_{5,U})O(h_6,S_{7,X})O(h_7,A_{7,X})O(h_8,S_{9,X})O(h_9,A_{9,X})$
	5,9	3,7		$O(h_2,S_{3,U})O(h_3,A_{3,U})O(h_4,S_{5,X})O(h_5,A_{5,X})O(h_6,S_{7,U})O(h_7,A_{7,U})O(h_8,S_{9,X})O(h_9,A_{9,X})$
	5,7	3,9		$O(h_2,S_{3,U})O(h_3,A_{3,U})O(h_4,S_{5,X})O(h_5,A_{5,X})O(h_6,S_{7,X})O(h_7,A_{7,X})O(h_8,S_{9,U})O(h_9,A_{9,U})$
4,**5**,**6**,**7**	4,7	5,6	-	$D(S_{4,X})O(h_4, A_{4,X}+S_{5,K})O(h_5, A_{5,K}+S_{6,K})O(h_6, A_{6,K}+S_{7,X})O(h_7, A_{7,X})$
	4,8		-	$D(S_{4,X})O(h_4, A_{4,X}+S_{5,K})O(h_5, A_{5,K}+S_{6,K})O(h_6, A_{6,K})O(h_7, S_{8,X})D(A_{8,X})$
	5,7		$(1-D_{X+K})(1-D_K)(1-D_X)$	$O(h_4, S_{5,K}+S_{5,X})O(h_5, A_{5,K}+A_{5,X}+S_{6,K})O(h_6, A_{6,K}+S_{7,X})O(h_7, A_{7,X})$
	5,8		-	$O(h_4, S_{5,K}+S_{5,X})O(h_5, A_{5,K}+A_{5,X}+S_{6,K})O(h_6, A_{6,K})O(h_7, S_{8,X})D(A_{8,X})$
	6,7		$(1-D_{X+K})(1-D_K)(1-D_X)$	$O(h_4, S_{5,K})O(h_5, A_{5,K}+S_{6,K}+S_{6,X})O(h_6, A_{6,K}+A_{6,X}+S_{7,X})O(h_7, A_{7,X})$
	6,8		-	$O(h_4, S_{5,K})O(h_5, A_{5,K}+S_{6,K}+S_{6,X})O(h_6, A_{6,K}+A_{6,X})O(h_7, S_{8,X})D(A_{8,X})$
	7,7		$(1-D_{2X})(1-D_K)^2$	$O(h_4, S_{5,K})O(h_5, A_{5,K}+S_{6,K})O(h_6, A_{6,K}+2S_{7,X})O(h_7, 2A_{7,X})$
	7,8		-	$O(h_4, S_{5,K})O(h_5, A_{5,K}+S_{6,K})O(h_6, A_{6,K}+S_{7,X})O(h_7, A_{7,X}+S_{8,X})D(A_{8,X})$
	7,Z		$D_X(1-D_X)(1-D_K)^2$	$O(h_4, S_{5,K})O(h_5, A_{5,K}+S_{6,K})O(h_6, A_{6,K}+S_{7,X})O(h_7, A_{7,X})D(S_{Z,X})D(A_{Z,X})$
	8,8		-	$O(h_4, S_{5,K})O(h_5, A_{5,K}+S_{6,K})O(h_6, A_{6,K})O(h_7, 2S_{8,X})D(2A_{8,X})$
	8,Z		-	$O(h_4, S_{5,K})O(h_5, A_{5,K}+S_{6,K})O(h_6, A_{6,K})O(h_7, S_{8,X})D(A_{8,X})D(S_{Z,X})D(A_{Z,X})$

Table 5.3: Likelihoods for two contributor CSPs ($\mathcal{C}$), only showing $Pr(R|G)$, using both the discrete model and the PH model. The genotype of Q or X for each term of the likelihood is given in $\mathcal{G}_X/\mathcal{G}_Q$, whereas $\mathcal{G}_U/\mathcal{G}_K$ gives the genotype for U in the first CSP and the genotype for K in the second CSP. CSP alleles in bold are allelic peaks, non-bold are non-allelic. Only single-stutter is considered here. Genotype combinations that are congruent with H_p are highlighted in red. D is the dropout probability, which under the PH model is a function of expected PH, and under the discrete model is subscripted by X, U or K for heterozygous dropout probabilities for each contributor, an additional subscript of 2 for a homozygous dropout probability. A subscript of $X+K$ indicates the dropout probability of a hypothesised shared allele between X and K. O is the probability of an observed peak as a function of both the observed PH, h, and the expected PH. $A_{x,y}$ indicates the allelic expected PH at allele x from contributor y, $S_{x,y}$ indicates the stutter expected PH at allele $x-n$ that stuttered from allele x.

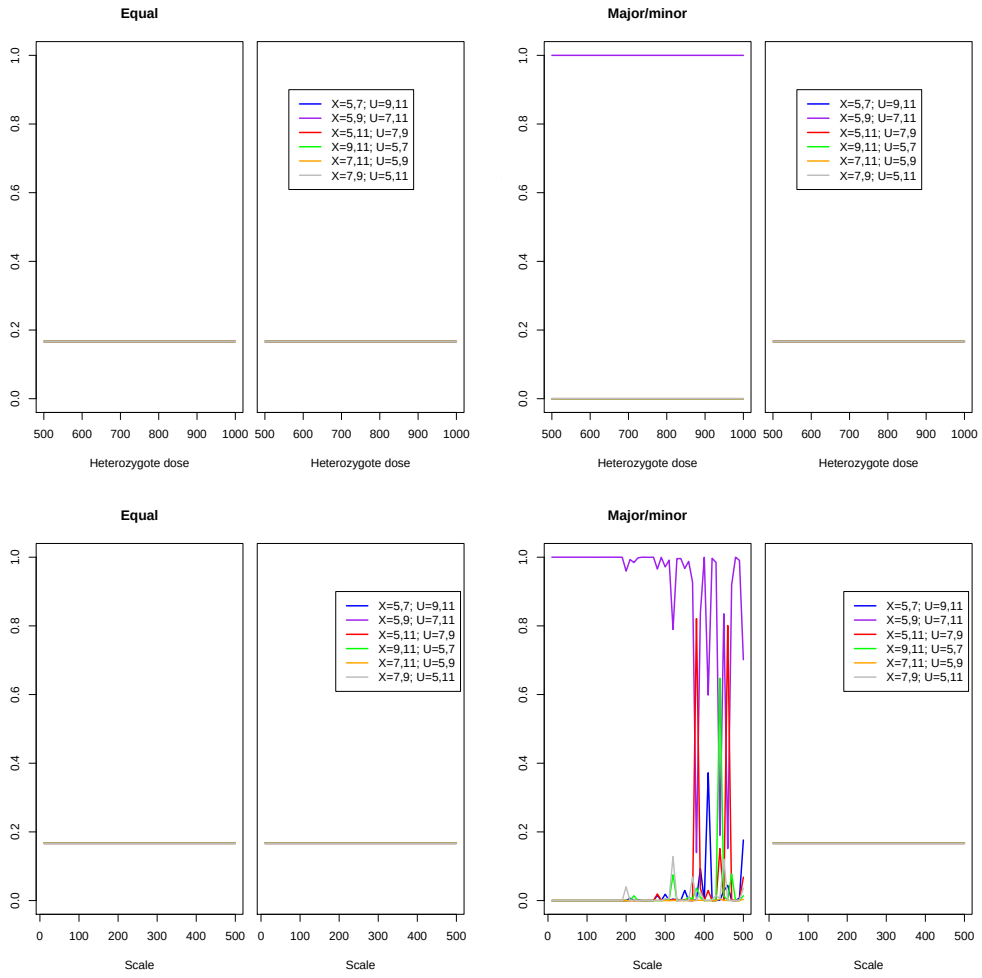


Figure 5.3: $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$ under H_d for the first two-contributor CSP in Table 5.3 with varying heterozygote dose (top, scale fixed at 10) or scale (bottom, heterozygote dose fixed at 1000), and either a major/minor (left, 1:10 ratio) or equal contributions (right, 1:1 ratio) design. Y-axis shows the normalised $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$, x-axis shows the heterozygous dose used to generate the CSP and calculate $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$ or the scale used to generate the CSP and calculate $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$. Legend displays $\mathcal{G}_X$ and $\mathcal{G}_U$.

addition to those improbable pairings, if the two contributors have dissimilar DNA contributions a mismatch of contributor dose and observed peak will be improbable e.g. $O(h_5, A_{5,X})$ in the first CSP will have a low probability if the peak at position 5 is large but X is hypothesised to be the minor contributor or *vice versa*. In contrast, if the contributors have approximately equal contributions no such combination is possible, so there will be no improbable allelic peak/hypothesised contributor pairings: $O(h_5, A_{5,X}) \approx O(h_5, A_{5,U})$ in the first CSP of Table 5.3. So for CSPs with equal contributions, multiple genotype combinations will be supported under H_d, which is similar to the discrete model, whereas with unequal contributions the PH model will only support a single, or a few, genotype combinations, and the discrete model will again support many genotypes equally. This is seen in Figure 5.3 (top row), where the $\Pr(R|G)$ have been evaluated for the first CSP of Table 5.3 with varying heterozygote doses, and with either the first and second contributor having equal contributions (Equal), or the second contributor having $10\times$ the dose of the first contributor (Maj/min).

The theoretical predictions so far have assumed a small value for the gamma distribution scale, $\sigma = 10$, however, it is expected that the utility of the PH model breaks down with a very large PH variability. To demonstrate this behaviour, the first CSP from Table 5.3 was simulated again, but with a fixed heterozygote dose for X of 1000 RFU, a varying value of σ, where now the observed PHs were drawn from a gamma distribution with mean equal to the expected PH and scale=σ. The probability of peaks under $Pr(\text{R}|\text{G})$ were dynamically calculated, so that an "observed" peak $< t$ was given the dropout probability rather than the observed peak probability. No allelic peaks dropped out during the simulation, but dropouts of stutter peaks were observed. As σ becomes very large the PH model starts to support some incorrect $\mathcal{G}_X$ for major/minor mixtures, while the discrete model continues to support all genotypes equally (Figure 5.3, bottom row, Maj/min). Because an incorrect genotype fits the data best for some CSPs, a greater σ will be required to explain the CSP under H_p, which will be penalised by the PH model, so for some CSPs with high PH variability the discrete model will perform better than the PH model. When the contributors have equal contributions, the PH model continues to support all genotypes equally under H_d despite a high σ (Figure 5.3, bottom row, Equal); the H_p explanation may still require a greater σ than the average over all possible genotypes under H_d, so may similarly be penalised in some high variability CSPs.

Shared alleles

Now consider a CSP with a shared allele, where the true genotypes of the contributors are $\mathcal{G}_1$=5,7 and $\mathcal{G}_2$=5,6, and an additional stutter peak was observed at allele 4 (see the second CSP in Table 5.3).

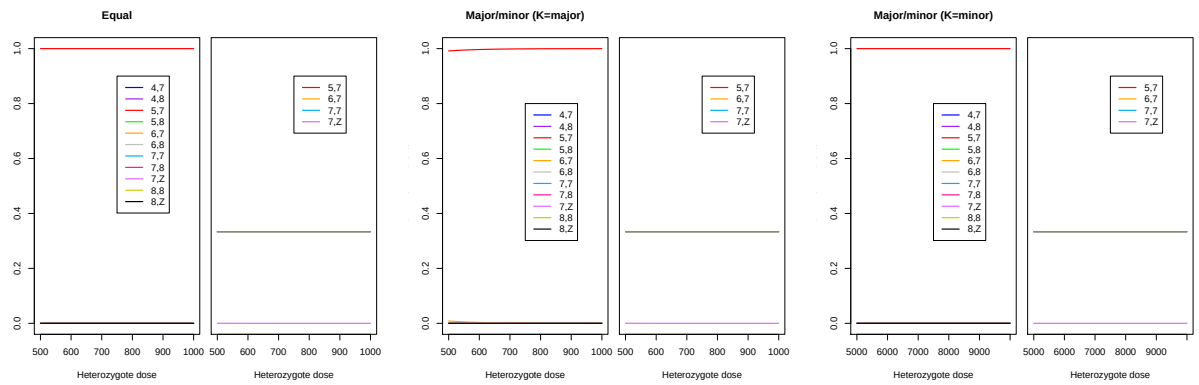


Figure 5.4: $Pr(R|\mathcal{G}_X, \mathcal{G}_K)$ under H_d for the second CSP in Table 5.3, which is a two-contributor CSP with one known contributor (K), with varying heterozygote dose. The unprofiled contributor (Q/X) has the same as (left), 0.1× (middle) or 10× (right) the heterozygote dose of K. Y-axis shows the normalised $Pr(R|\mathcal{G}_X, \mathcal{G}_K)$, x-axis shows the heterozygous dose for X used to generate the CSP and calculate $Pr(R|\mathcal{G}_X, \mathcal{G}_K)$. Legend displays $\mathcal{G}_X$.

When one of the contributors to the two person mixture is assumed known, the PH model is expected to be able to utilise the extra PH information in the CSP to return a more extreme WoE in favour of the true hypothesis compared to the discrete model. Regardless of the relative DNA contributions, the discrete model is unable to distinguish between the three possible non-dropout genotypes for X (Figure 5.4), as they all have a simplified probability of $(1-D)^4$ (Table 5.3). In contrast, the PH model supports the true genotype of X over all other possibilities in each scenario, as once again mismatches between a hypothesised contribution and observed PH are improbable. For example, under equal contributions at 500 RFU heterozygote dose the CSP heights are 100, 950, 500 and 450 RFU for alleles 4, 5, 6 and 7 respectively, and a hypothesised $\mathcal{G}_X$=7,7 will have expected PHs 50, 500, 550 and 900 RFU; alleles 5 and 6 will have a low probability, so $\mathcal{G}_X$=7,7 will have a low probability. Note that this is different to a $Q/X{+}U$ scenario, where the PH model would support genotypes $\mathcal{G}_X{=}\{5,6\}, \mathcal{G}_U{=}\{5,7\}$, $\mathcal{G}_X{=}\{5,7\}, \mathcal{G}_U{=}\{5,6\}$, $\mathcal{G}_X{=}\{5,5\}, \mathcal{G}_U{=}\{6,7\}$ and $\mathcal{G}_X{=}\{6,7\}, \mathcal{G}_U{=}\{5,5\}$ equally. In this way, including a known contributor can enable the PH model to utilise extra information over the discrete model, even when the two contributors have equal contributions.

If instead the second contributor is assumed unknown, the PH model is expected to gain information over the discrete model for both unequal and equal contribution mixtures. Probabilities of genotypes are not shown here, as there are 157 and 18 possible genotype combinations for the PH and discrete models respectively. When the two contributors have equal contributions, the PH model supports four genotype allocations, those where two 5 alleles are allocated across the two genotypes, whereas the discrete model supports 12 separate genotype allocations, those where no dropout is hypothesised in either contributor

(Figure 5.5, top left). When the mixture is instead a major/minor mixture, where the first contributor is present at the heterozygote dose, and the second contributor is present at a tenth of the heterozygote dose, the PH model gives most support to the true genotype allocation, $\mathcal{G}_1$=5,7 and $\mathcal{G}_2$=5,6, but supports multiple genotypes for the minor contributor, $\mathcal{G}_2$, as their genotype cannot be entirely deconvoluted at their low dose of between 60 and 100 RFU (Figure 5.5, top right). Conversely the discrete model largely continues to support the same 12 genotypes as the equal contributions condition. If instead the second contributor contributes 10× the heterozygote dose, the PH model supports a single genotype combination, which is the ground truth, while the discrete model continues to support the same 12 genotype combinations (Figure 5.5, bottom left). If now the heterozygote dose is fixed at 600 RFU, but the mixture ratio is varied from 0.1 to 10, a similar pattern is seen, where multiple genotypes are supported at a low mixture ratio, because the minor contributes a small amount so is difficult to deconvolute, before supporting only the ground truth genotypes between approximately mixture ratios 0.3 and 0.6 (Figure 5.5, bottom right). Around mixture ratio 1, four genotype combinations that include a double dose of allele 5 are supported, while above approximate mixture ratio 2.5 the ground truth genotype allocation is supported again.

5.4 Other continuous models

There are currently six continuous models, including likeLTD, that have been developed (summarised in Table 5.4), each with differing modelling choices and assumptions.

5.4.1 DNAmixtures

DNAmixtures is based on a Bayesian network implementation of a PH model published by Cowell et al. [2015]. Similar to likeLTD, DNAmixtures assumes that PHs are gamma distributed. The stutter model of DNAmixtures is simpler than that of likeLTD, with a single parameter for the mean stutter proportion, β, so that for every allelic peak, $(1 - \beta)$ of the dose remains at the parent peak position, while β of the dose stutters to a position one repeat unit shorter. Note that the stutter proportion does not vary by locus, and is not affected by the LUS or any other characteristic of an allele. DNAmixtures does not model OS or DS. Neither does DNAmixtures model degradation, although the authors note that implementation of a non-contributor-specific degradation model would be simple. DNAmixtures explicitly models silent alleles, rather than treating them as dropout alleles. DNAmixtures implements a Bayesian network incorporating known and unknown genotypes, observed PHs, and model parameters [Graversen and Lauritzen, 2014], which can

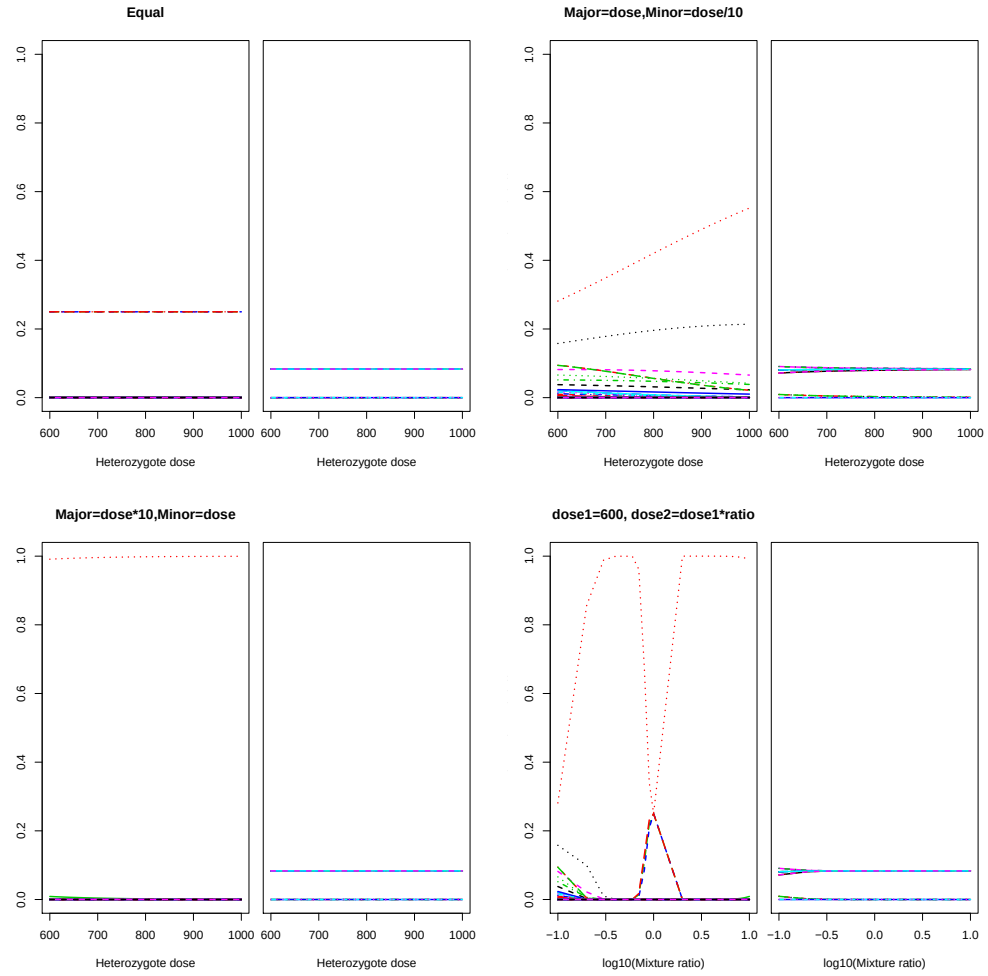


Figure 5.5: $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$ under H_d for the second two-contributor CSP in Table 5.3 with varying heterozygote dose (top and bottom left, scale fixed at 10) or mixture ratio (bottom right, heterozygote dose fixed at 600). Y-axis shows the normalised $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$, x-axis shows the heterozygous dose used to generate the CSP and calculate $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$ or the mixture ratio used to generate the CSP and calculate $Pr(R|\mathcal{G}_X, \mathcal{G}_U)$.

Program	PH distribution	Parameter elimination	Stutter model	Dropin model	DS	OS	Degradation model
DNAmixtures	$\Gamma(O_i \mid E_i, E_i\sigma)$	Maximisation	α	Extra U	×	×	
EuroForMix	$\Gamma(O_i \mid E_i, E_i\sigma)$	Both	α	$Cp_ik(h_i - t\mid\zeta)$	×	×	$\delta^{\frac{F_{l,i}-125}{100}}$
LiRa	$\mathrm{Dir}(\frac{O_i}{\sum \mathbf{O}} \mid E_i, E_i\sigma)$	Maximisation	$\frac{\beta_l \sum \mathbf{O}_l}{F_{i,l}}$	$g(n\mid\pi)\prod_{m=1}^{n} p_m j(h_m\mid\psi,\omega)$	×	×	
likeLTD	$\Gamma(O_i \mid E_i, E_i\sigma)$	Maximisation	$\beta_l u_{i,l}$	$p_i\lambda(1+\epsilon)^{-f_i}$	✓	✓	$(1+\delta_c)^{-f_i}$
STRmix	$\log N(\frac{O_i}{E_i} \mid 0, \frac{c^2}{E_i})$	Integration	$\beta_l \sum_j max(s_{j,l} - x_l, 0) + \alpha_l$	ae^{-bh_i}	×	✓	$\Delta \times e^{\delta \times F_a}$
TrueAllele	$\mathcal{N}(\mathbf{O}\mid E_i, \psi^2 \cdot \mathrm{diag}(\mathbf{O}) + \xi^2)$	Integration	✓	✓	?	?	✓

Table 5.4: Summary of the characteristics of the currently available continuous models for the evaluation of low-template DNA mixtures. Under "PH distribution", $\mathbf{O}$ indicates a vector of observed PHs, E_i indicates an expected PH for a position in $\mathbf{O}$, σ indicates the scale parameter for the gamma distribution, c indicates a PH variance constant for STRmix, ψ indicates an amplification variance parameter and ξ indicates a baseline variation parameter. Under "Stutter model" α is an intercept parameter, and β is a gradient parameter, locus dependency is indicated by subscript l, while allele dependency is indicate by subscript i and $F_{l,i}$ is the allele length in base pairs. For STRmix j is the number of uninterrupted repeat section in allele i and x_l can be thought of as a number of repeats below which the repeated section does not contribute to stutter. Under "Dropin model" C indicates a probability of dropin, k indicates the exponential pdf, ζ is the exponential rate parameter which is fixed, g indicates the Poisson pmf, n indicates the number of hypothesised dropin alleles, π indicates the mean dropin rate per locus (experimentally determined), j indicates the gamma pdf, ψ and ω are the shape and scale parameters of the gamma distribution, while a & b are constants for the STRmix dropin model. Under "Degradation model" δ is degradation parameter (exponent for STRmix), f is the mean adjusted allele length in base pairs, and Δ is the constant of proportionality.

be used to compute the likelihood efficiently using the HUGIN software. This allows DNAmixtures to model dropin as originating from an extra unknown contributor of low-level. Each locus is treated as independent, and so is assigned a separate Bayesian network, which may cause problems modelling locus dependency such as linkage for closely related individuals, or distant relatedness as accounted for using F_{ST}. However, this problem was solved for distant relatedness by Tvedebrink et al. [2015] who incorporated an F_{ST} correction into the Bayesian network using a multivariate Dirichlet-multinomial distribution implementation. The use of the proprietary HUGIN software means that DNAmixtures is not fully open source.

5.4.2 EuroForMix

Similar to DNAmixtures, EuroForMix is built on the model of Cowell et al. [2015], but has been extended to incorporate degradation, dropin and distant relatedness through an F_{ST} adjustment [Bleka et al., 2016], which are all missing in DNAmixtures, and are important phenomena to model (see Chapter 1). However, the implementation of EuroForMix does not depend on HUGIN. It is the only fully open-source software for computation of forensic WoE using PH information, other than likeLTD. Both Bayesian (non-MCMC integration) and Frequentist (maximisation) methods of inference are available in the package. EuroForMix requires a laboratory specific parameter when modelling dropin that may make portability across laboratories more difficult than likeLTD; the probability of a dropin allele is dependent on the rate parameter for the exponential distribution, which is fixed by the user and is likely to vary across genotyping machines (the authors use the Applied Biosystems 3500xl Genetic Analyzer), kits and laboratories. Dropin is handled differently from likeLTD, which includes dropin dose at each peak position and calculates the probability of all peaks as gamma distributed. EuroForMix instead treats unexplained peaks in a hypothesis differently, classifying them as dropins that are exponentially distributed above the detection threshold, which is modulated by a probability of dropin, C, while the probability of all non-dropin peaks are then modulated by $1 - C$. This appears to be a hybrid between the discrete dropin model, C and $1 - C$, and the continuous model, $k(h_i - t|\zeta)$.

5.4.3 LiRa

LiRa models the proportion of PH over sum PHs for a locus as Dirichlet distributed [Puch-Solis et al., 2013], which is equivalent to modelling the individual PHs as gamma distributed with a common scale parameter, similar to DNAmixtures, EuroForMix and likeLTD. The inclusion of dropout alleles alters the computation, as the DNA quantity proxy used by Puch-Solis et al. [2013], $\sum(\mathbf{O})$, needs to be upweighted to account for

lost quantity due to PHs that are lower than the detection threshold. This is done by sequentially estimating the expected PH at each unobserved position, and adding that expected PH to the total PH at the locus, before estimating the next unobserved expected PH. The stutter model of LiRa is more complicated than that of the two Cowell et al. based models [Cowell et al., 2015, Bleka et al., 2016], as the stutter proportion has a linear relationship with the length of allele in base pairs, F, so is dependent on both the marker, l, and specific allele, i. Similar to likeLTD, LiRa assumes that the linear relationship of the stutter proportion goes through the origin, as there is no intercept in the model. The model also differs between CSPs, as the stutter ratio is actually linear in relation to $\sum(\mathbf{O}/F_{i,l})$, so the relationship will vary if the CSP is good-template or low-template. The LiRa dropin model multiplies the Poisson-distributed probability of observing the n hypothesised dropin peaks by the product of the population allele probability and the gamma pdfs for the dropin peak for each of the n hypothesised dropin alleles [Puch-Solis, 2014], as shown in Table 5.4. If no dropin alleles are hypothesised this simplifies to the Poisson probability of observing zero dropins, given some mean dropin rate, which was experimentally determined by Puch-Solis [2014]. Note that this means that dropin cannot have occurred at any position where an explained peak has been observed, similar to EuroForMix and in contrast to likeLTD. Treating both dropin peaks and non-dropin peaks as gamma distributed seems preferable to the EuroForMix implementation of gamma-distributed non-dropin peaks and exponentially-distributed dropin peaks. Note that LiRa was developed in house by LGC Forensics, and is not available for use externally.

5.4.4 STRmix

STRmix has the most sophisticated stutter model, with stutter ratio being proportional to the size of all uninterrupted repeat sections in a given allele, which was specifically designed to deal with SE33 which has two long repeat sections, so has a stutter behaviour that cannot be correctly modelled in terms of LUS [Taylor et al., 2016]. The STRmix degradation model is based on the work of Tvedebrink et al. [2009, 2012], which found that the effect of degradation on the dropout probability is exponentially distributed with the molecular weight of an allele. The authors of STRmix subsequently confirmed that the effect of degradation on PHs was best modelled with such an exponential relationship, rather than with a linear relationship [Bright et al., 2013b]. STRmix also models the PH of dropin events as exponentially distributed [Taylor et al., 2013], similar to EuroForMix. Once again this means that dropin cannot be modelled as having occurred at a position where a non-dropin peak has been observed. The probability of an allele dropping out is modelled as the CDF of the gamma distribution, using the same parameters as those for non-dropout

alleles. Mixture deconvolution is performed by applying these degradation and stutter models to a CSP, with parameter elimination being through the use of Markov chain Monte Carlo (MCMC) [Taylor et al., 2013]. Perhaps the most important difference between all of the models previously described and STRmix is that STRmix models the ratio of observed PHs to expected PHs obtained from their model as lognormal distributed, whereas the previous models all assume that PHs are gamma distributed. STRmix includes locus-specific parameters that allow the amplification efficiency to differ between loci, meaning different loci can have different mean PHs, whereas likeLTD incorporates this information into an increased variability over the whole profile, rather than adding extra parameters.

5.4.5 TrueAllele

TrueAllele differs from the other PH models, as it fully models the EPG trace, rather than modelling PHs above the detection threshold and dropouts below [Perlin et al., 2011]. Each time point of the epg will have a trace height, which will be referred to here as PH for simplicity. TrueAllele models PHs with a truncated (>0) multivariate normal distribution, where the distribution parameters are a vector of expected PHs for each position, and a covariance matrix between amplification variation and a diagonal matrix of PHs at each position, with baseline variance incorporated. The expected PH at each position is the sum of genotypic contributions over all individuals. The mixture weights used to determine genotypic contribution are locus specific, drawn from a multivariate normal distribution with mean equal to the mean mixture weights, and covariance matrix being a mixture weight variance parameter multiplied by an identity matrix. All non-allelic positions of the epg will have expected PH 0, as there is no contribution from baseline to the expected PH, so all baseline positions are explained through variance away from 0. Both the model variables, and the genotypes of contributors, are integrated over using MCMC, giving the posterior probabilities for the genotypes of hypothesised contributors. Perlin et al. [2011] do not state whether, or how, they model stutter, dropin or degradation, but note that it is possible to do so.

5.4.6 Comparison of models

Distribution

In Kelly et al. [2012] the authors state that modelling peak heights as gamma distributed captures a skewness in heterozygote balance that the lognormal distribution is unable to capture. However, Kelly et al. [2012] chose to implement the lognormal distribution in STRmix due to the relative simplicity of the lognormal distribution and because the normal distribution is more familiar to biologists and forensic practitioners.

Therefore, modelling peak heights as gamma distributed gives a better fit to peak height data than the lognormal, and is desirable without taking into account issues of complexity and familiarity.

While many papers have been published validating TrueAllele [Perlin et al., 2011, Kadash et al., 2004, Perlin et al., 2013, 2014], there has been no consideration of the suitability of the truncated normal distribution to PH data compared to other choices, such as the gamma or lognormal. Without such comparisons, it is difficult to comment on the suitability of the truncated normal.

Stutter models

The stutter models employed range from a simple constant stutter rate across the profile (DNAmixtures, EuroForMix) to a full linear relationship between LUS and stutter ratio (STRmix). During development, likeLTD had a full linear stutter model, however, the intercept was estimated at 0 for all CSPs tested, so the intercept was subsequently removed from the model. Additionally, likeLTD is the only model that incorporates double-stutter and over-stutter, which were both observed many times during validation of the likeLTD model. Specifically, over-stutter is relatively common at the trinucleotide repeat locus D22. Other models may choose to reduce the incidence of double- and over-stutter peaks by increasing the detection threshold, however, this risks losing minor allelic peaks of interest.

Dropin models

Dropin models employed can be split into those that model dropin as exponential (EuroForMix, STRmix), and those that model it as gamma (LiRa, likeLTD). Because dropin peaks are generated through the same processes as non-dropin peaks it seems consistent to model both types of peak as originating from the same distribution, which only LiRa and likeLTD do. Additionally, both the EuroForMix and STRmix dropin models depend on laboratory-specific parameters, reducing the applicability of their dropin models in different laboratories. This may be seen as a positive, by ensuring that a model is tailored to the laboratory it is being used in, or a negative, in requiring extra validation before a model can be employed in a new laboratory. likeLTD employs the only dropin model that allows peaks to share dropin and non-dropin doses; if dropin is thought of as sporadic observations of DNA peaks, rather than as unexplained alleles given a hypothesis, then there is no reason why a dropin event could not occur at an allelic position, which becomes increasingly likely if the allele is common in the population. A dropin event that overlaps an allelic peak will alter the peak height at that position, which is unable to be accounted for by any other model. Additionally, the likeLTD dropin model is subject to degradation, which no other model accounts for. Dropin peaks

may be expected to experience more degradation than non-dropin peaks, so this is a realistic modelling assumption. It also allows for minor contributors to be explained as dropin with degradation, which other models are unable to do.

Availability

EuroForMix and likeLTD are open-source, DNAmixtures requires proprietary back-end software, but is itself freely available, STRmix is proprietary but is available for use, while LiRa is proprietary and unavailable for free use, as it is in-house software. Open source software is important for easy testing and altering of modelling assumptions.

Runtime

For the Meredith Kercher bra clasp (see Section 6.6), likeLTD took between 16 and 17 minutes to query Raffaele Sollecito, and between 25 and 30 minutes to query Amanda Knox. In contrast, both EuroForMix and STRmix took less than a minute to run for each Q. This is a slight deficiency of likeLTD, but the runtimes are not prohibitive.

Uptake

EuroForMix was developed, and is used extensively, in Europe as part of EuroForGen. STRmix was developed as a joint venture between the New Zealand and Australian forensic services, and is used extensively throughout Australia and New Zealand. STRmix is additionally gaining some traction within UK forensic providers. LiRa is used in-house at LGC-forensics in the UK, but is not available for use externally. TrueAllele has been used mainly in the US, but has seen use elsewhere.

Conclusions

Based on the available information, EuroForMix and STRmix seem to be the strongest models based on availability, model complexity, runtime and uptake. While likeLTD is slightly slower than EuroForMix and STRmix, it models phenomena that no other model has incorporated that are important for some CSPs, so may be preferable in some situations. It is difficult to comment on the suitability of TrueAllele due to a lack of information regarding its stutter, dropin and degradation models. DNAmixtures does not seem sufficiently developed to be of use in casework due to a lack of a degradation model, dropin model and adjustment for

distant relatedness; as discussed in Chapter 3, failing to account for distant shared ancestry between Q and X can be unfairly biased against the suspect.

5.5 Further considerations and modelling choices

Now we return to considering the modelling choices that were made while developing the likeLTD PH model.

5.5.1 Stutter model

There have been some suggestions that the stutter ratio of a given allele depends on the AT content of the amplified fragment [Brookes et al., 2012], which could have been modelled with a parameter similar to LUS, u, but for the average AT content of a specified allele in the population. However, since there is contradictory evidence on this supposed relationship [Gill et al., 2015], no such effect has been modelled here.

Stutters are possible in whole repeat positions other than x-n, x-2n and x+n. Furthermore, stutters are possible in partial repeat positions away from the parent peak e.g. Gill et al. [2015] give examples of stutters observed in positions two basepairs shorter than the parent peak at the tetranucleotide loci D1S1656 and SE33. However, these classes of stutters are expected to be both rare and have a small stutter ratio (and hence small PH), and so are expected to be relatively unimportant to the modelling of PHs. The explanatory power that modelling these rare stutter classes would provide to the model does not warrant the increased computational complexity that would be introduced. Any observations of these classes of stutters can be explained as dropins by the existing model.

5.5.2 Heterozygote balance

Heterozygote imbalance is believed to predominantly originate from pipetting variability rather than PCR variability [Gill et al., 2005]. As a result, no relationship between locus and heterozygote balance has been observed [Kelly et al., 2012], as pipetting has no locus-specific effects. Therefore, there is no need for the variance of PHs to behave differently at different loci, so the model contains only a single σ parameter for the whole profile.

5.5.3 Relative DNA contribution

The model specified here does not allow the relative contributions from individuals to vary across replicates. Some authors instead choose to allow the relative contributions of DNA to vary [Perlin et al., 2011], as it is

expected that through pipetting variability there will be some noise in the amount DNA from contributors taken forward for PCR and subsequent analysis across replicates. However, this effect is not expected to be systematic, so should be accounted for by a relatively high value of σ for CSPs where the relative contributions vary considerably across replicates.

5.6 Towards a model with no detection threshold

As described in Section 5.4, the currently available continuous models are split between those that employ a detection threshold, below which an allele has dropped out (DNAmixtures, EuroForMix, likeLTD, LiRa and STRmix), and one that does not employ a detection threshold, but rather models baseline noise explicitly (TrueAllele). The removal of a detection threshold has the advantage of utilising all of the information available in a CSP, however, it introduces some modelling difficulties.

5.6.1 Baseline noise

Perlin et al. [2011] explicitly model baseline variance in an epg, to allow for automatic determination of the genotypes present in a particular CSP. Mönich et al. [2015] show that dye-specific effects on baseline noise are not sufficient to describe reality, but instead locus-specific effects should be modelled. In their study of 946 single-contributor CSPs they observed that if allelic peaks, stutter peaks, and double-stutter peaks are removed from the data, the gamma distribution gives the closest fit to the baseline data for Identifiler Plus, but a lognormal distribution gives the closest fit for PowerPlex HS. While baseline noise is associated with the quantity of DNA input if double-stutter peaks are retained, it appears that this relationship does not exist once double-stutter peaks are accounted for. The authors describe a method for employing a detection threshold based on the specific CSPs baseline characteristics, however, the eventual removal of a detection threshold is the ultimate goal. For consistency with the current likeLTD PH model, a possible implementation of baseline noise would assume that baseline dose would be gamma distributed, with Equation (5.2) having an additional term ν_l, the expected dose from baseline noise in RFU, and the baseline peaks would be given the same σ as all other peaks in the CSP. This is different from the Perlin et al. [2011] method of modelling baseline noise, which assumes an expected PH at pure baseline positions as 0, because both the shape and scale parameters of the gamma distribution must be greater than 0. In practice the majority of baseline observations have a height of 0 [Mönich et al., 2015], so ν_l would be constrained to be small.

5.6.2 Pull-up

When there is a large peak in an epg lane it can cause small artefactual peaks in the other lanes at the same position, termed pull-up peaks. In current practice pull-up peaks are removed from the dataset manually, however, to minimise manual interpretation for a model with no detection threshold the effects of pull-up should be accounted for in the model. A possible implementation for a given position, i, that has some non-baseline dose would be to add $P_{l,r,g,i}\upsilon$ to all $P_{l,r,g,x}$ where υ is the pull-up proportion that denotes how tall a pull-up peak is relative to its parent peak, and x are all positions with $f_i = f_x$ other than i itself, where f is the length of an allele in base pairs. It may be necessary for υ to be specific for every combination of dye colours, so that e.g. peaks in the yellow lane cause larger pull-up peaks in the blue lane than in the green lane.

Chapter 6

Validation of the PH model

Some of the work in this chapter has been published in Steele et al. [2016], see Appendix B for the accepted manuscript. All work, other than the Bright et al. [2015] results for models other than likeLTD and the STRmix results for the Meredith Kercher case, was performed by me.

6.1 Motivation

I designed a set of tests to validate the model described in Section 5.2. Firstly 72 one- to three-contributor CSPs were generated in the laboratory using DNA samples from 36 donors, for which the WoE was evaluated using both the discrete and PH models (see Section 6.2), to ensure that the two models provide broadly similar results. Secondly, I queried one of the laboratory generated CSPs, with altered model assumptions to ensure the model behaves as expected e.g. not modelling double-stutter (see Section 6.3). Next, I altered one of the laboratory generated CSPs one peak at a time to ensure that the resulting behaviour of the WoE is congruent with the change applied to the input data (see Section 6.4). I then used the proposed PH model to evaluate the WoE for a set of CSPs for which the WoE has previously been published with multiple DNA analysis softwares [Bright et al., 2015]. Results were compared across the published evaluations and the PH model evaluations (see Section 6.5). Lastly, I evaluated the WoE for real-world crime sample with multiple models to compare how they perform in a real-world scenario (see Section 6.6).

6.2 Laboratory validation

A series of laboratory-generated CSPs were evaluated with the likeLTD PH and discrete models to compare the behaviour of the PH model in relation to the discrete model, and in relation to the theoretical predictions set out in Section 5.3. In particular the previous predictions implied that the PH model should have improved

# Cont	# Samples	DNA mass (pg)	Approx. cell equiv.
1	9	250	42
	9	62	10
	9	16	3
	9	4	1
2	12	266 (250:16)	44 (42:3)
	12	62 (31:31)	10 (5:5)
3	6	328 (250:62:16)	55 (42:10:3)
	6	93 (31:31:31)	16 (5:5:5)

Table 6.1: Laboratory protocol for generation of single contributor and multiple contributor CSPs from 36 donated DNA samples. DNA masses and approximate cellular equivalents are rounded and given as a total contribution, with individual contributions in brackets.

genotype inference over the discrete model in both equal- and unequal-contributions CSPs with low PH variability, but that genotype inference may break down with high PH variability. The tests here will give an idea of whether the PH variability for laboratory-generated mixtures is large enough to reduce the genotype inference to the level of the discrete model. These tests should also reveal the types of CSPs for which evaluation with a PH model is desirable, and the types for which the simpler discrete model is adequate.

6.2.1 Laboratory protocol

Cheek swab samples were collected from 36 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 samples were created from the 36 DNA samples as shown in Table 6.1. These created samples were amplified using the AmpFℓSTR® NGMSelect® PCR kit as per 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 3kV voltage. The results were analysed using GeneMapper® ID v3.2 with a detection threshold of 20 RFU, and no stutter threshold, so that both allelic and non-allelic (stutter, over-stutter or double-stutter) peaks were recorded. Mixtures were generate from the extracted DNA samples. Approximate DNA masses are given for pre-amplification samples.

PH CSPs were converted to discrete CSPs using the rules set out in Table 6.2. Designations defaulted to the lowest confidence of calling a peak if it had multiple possible designations e.g. $\mathcal{C}$=13,14,15 and h_l=800,35,600, the 14 allele would be called as non-allelic if assumed to be an OS of the 13 allele ($x = 0.044$),

Designation	S	DS and OS
Non-allelic	$x < 0.05$	$x < 0.05$
Uncertain	$0.05 \leq x < 0.15$	$0.05 \leq x < 0.1$
Allelic	$x \geq 0.15$	$x \geq 0.1$

Table 6.2: Rules for classification of peaks as stutter (S), double-stutter (DS) or over-stutter (OS) of a parent peak when converting a PH CSP to a discrete CSP. x indicates the ratio of the stutter position PH to the parent PH.

but uncertain if assumed to be a S of the 15 allele ($x = 0.058$). In this situation the allelic call defaults to non-allelic due to the non-allelic call from the 13 parent peak. For multi-contributor CSPs, each contributor was queried in turn, leading to 36, 48 and 36 evaluations for the single-, two- and three-contributor CSPs respectively.

6.2.2 Single contributor

The hypotheses compared for the 36 single-contributor laboratory generated CSPs were of the form:

$$H_p\text{: } Q+ \text{ dropin},$$
$$H_d\text{: } X+ \text{ dropin}.$$

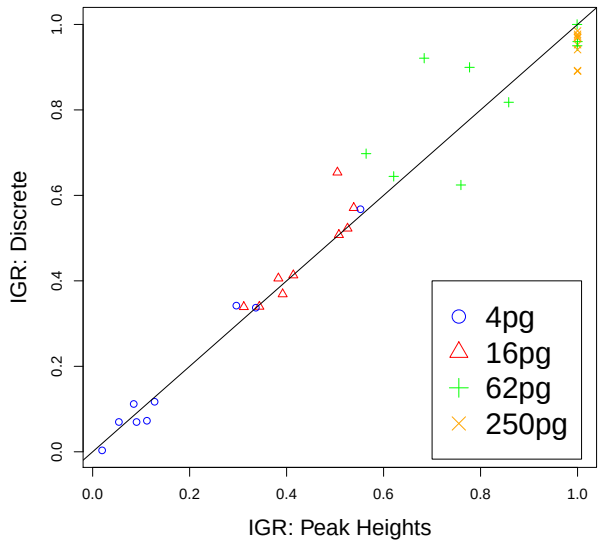


Figure 6.1: Information gain ratio (WoE/IMP) for 36 single-contributor CSPs using both the PH (x-axis) and discrete (y-axis) models. Legend indicates the approximate DNA mass used to generate the CSPs.

IGR increases as the DNA mass increases, for both the PH and discrete models (Figure 6.1). IGR is approximately equal between the two models for the majority of CSPs, as predicted by theory (see Table 5.2 and Figure 5.2). For a single-contributor CSP, PHs can give little information beyond presence/absence; PHs may allow for a homozygote peak to be distinguished from a heterozygote peak with a dropout, however, the results here indicate that this has little impact on the WoE in practice.

At 16pg there is one exception to the parity between the two models, in which the discrete model returns a larger WoE than the PH model (CSP shown in Appendix Figure A.1(a)). This CSP shows high variability through dropouts of alleles at vWA, D16, D18, D22 and SE33, all of which have corresponding non-dropout peaks ranging from 155 to 205 RFU, making these dropouts unlikely given the height of the non-dropout allele using the PH model, due to the penalty on σ. Under H_p the PH likelihood will be penalised due to this high σ, whereas under H_d the peaks can be hypothesised to be homozygous, leading to a lower penalty on σ, and a correspondingly lowered IGR. Since the discrete model does not take into account PHs, these observations can be incorporated into a high dropout rate, with no penalty for the fact that the corresponding non-dropout peaks are relatively tall. The PH model here is utilising the PH information to propose an explanation that fits the data well under H_d, but which the discrete model is unable to support as it does not have access to the PH information.

At 62pg there is greater variability from the x=y line (Figure 6.1). Appendix Figure A.1(b) shows the CSP which gave highest discrete IGR relative to the PH IGR. This result is largely driven by locus D1, at which an 11 allele has dropped out while the remaining non-dropout allele is observed at 282 RFU. Once again, the penalty on σ in the PH model prevents the PH variability being large enough to adequately explain this dropout under H_p, so the discrete model returns a greater IGR than the PH model. Appendix Figure A.1(c) shows the CSP for which the PH model returns the highest IGR relative to the discrete model. This is largely driven by the misclassification of the 28.2 allele at SE33 as allelic for the discrete CSP, which must explain this as a dropin under H_p, whereas the PH model is able to explain it as a stutter peak.

At 250pg the PH model results all obtain $\sim$1.0 IGR, whereas the discrete model IGR is between 0.89 and 0.98. The reduction from IGR=1.0 for the discrete model is largely driven by misclassification of stutter peaks as allelic or uncertain. Appendix Figure A.1(d) shows the CSP for which the discrete model returns the lowest IGR at 250pg. Alleles 25, 25 and 20 at loci FGA, D12 and SE33 respectively are all called as allelic for the discrete CSP as they are > 0.15 of the parent peak, so must be explained as dropins, whereas the PH model can explain them as large stutters. While in a single-contributor scenario the forensic practitioner is unlikely to mistake a stutter peak for an allelic peak, this reflects the difficulty in classifying

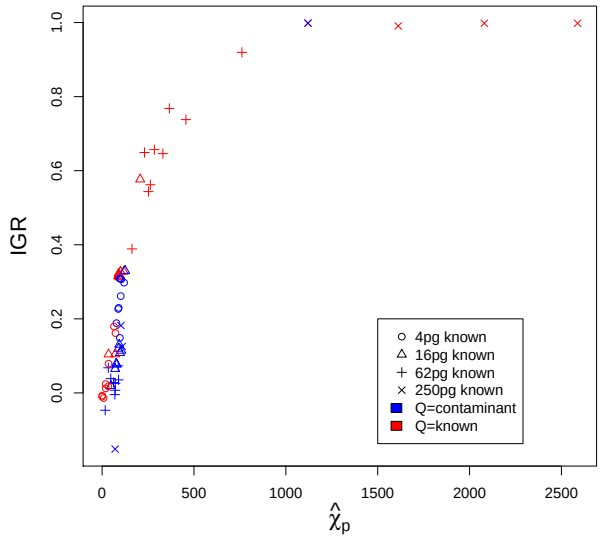


Figure 6.2: Information gain ratio (WoE/IMP) for 36 CSPs with one contributor of known DNA contribution (red), and one contributor that originates from contamination (blue), with DNA contribution estimate under the prosecution hypothesis (x-axis).

peaks in more complex scenarios.

6.2.3 One-contributor contamination

While generating the single-contributor CSPs, a plate of the same design became contaminated with DNA from one of the 36 donors. This provides the opportunity to investigate 35 two-contributor CSPs where the DNA contribution of one contributor is known, and one single-contributor CSP with an unknown DNA contribution. All 36 CSPs were queried for both contributors using the PH model with the hypotheses:

$$H_p\text{: } Q+\ U+ \text{ dropin},$$
$$H_d\text{: } X+\ U+ \text{ dropin}.$$

Similar to the single-contributor results, when the known-contribution individual is queried the IGR largely segregates with known DNA contribution (Figure 6.2, red). As would be expected, the IGR increases with increasing estimated DNA contribution under H_p, $\widehat{\chi}_p$. In contrast to the single contributor results, here four of the 4 pg CSPs return an LR in favour of H_d, which may be a result of either low DNA contribution (< 1 cell worth of DNA) or the reduced confidence of deconvolution for a very low-template mixture.

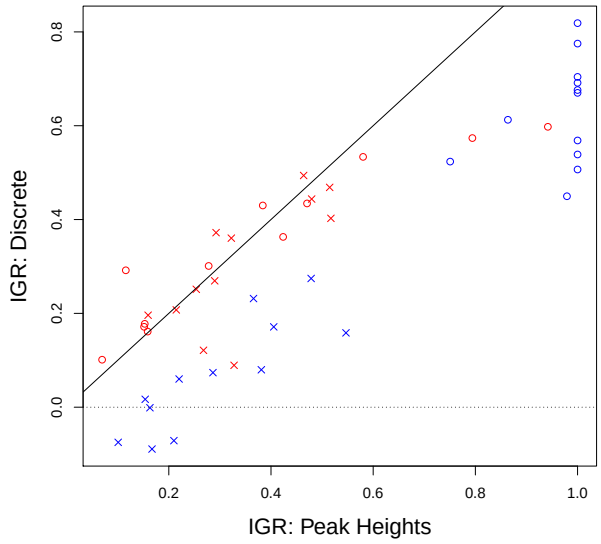


Figure 6.3: Information gain ratio (WoE/IMP) for 12 two-contributor equal-contribution CSPs (red) and 12 two-contributor major/minor CSPs (blue) using both the PH (x-axis) and discrete (y-axis) models. Both contributors to each CSP were queried, with circles and crosses indicating the first- and second-contributor respectively.

When the individual that is the source of the contamination is queried, $0 < \widehat{\chi}_p < 200$, with -0.2 <IGR< 0.4 (Figure 6.2, blue), ignoring the result for the single-contributor CSP, where the known contributor is the same as the contaminant contributor. These results suggest less than 16 pg of DNA contaminant per CSP, or less than three cells worth of DNA. The only CSP with IGR>0.4 is that for the donor who contaminated the other samples, with an IGR≈1.0, which suggests that the PH model is not affected by specifying more unknown contributors than is necessary to explain the CSP: it returns $\widehat{\chi} = \{1e^{-6}, 1121\}$ under both H_p and H_d.

6.2.4 Two contributors

The WoE for the 24 laboratory-generated two-contributor CSPs was evaluated with each contributor in turn as Q, and hypotheses of the form:

$$H_p\text{: } Q+\ U+ \text{ dropin},$$
$$H_d\text{: } X+\ U+ \text{ dropin}.$$

The IGR is approximately equal using the PH and discrete models when the equal-contributions CSPs are queried (Figure 6.3, red). Two of the equal-contributions CSPs localise with the major/minor

CSPs, which suggests that these two CSPs had a large discrepancy in contributions due to pipetting errors. This is supported by $\widehat{\chi_c}$ which give an approximate estimated contribution ratio of 1:4 and 1:3 for each CSP, considerably different from the expected 1:1 ratio. This can be confirmed through a visual inspection of the CSPs, given in Appendix Figures A.2(a) and A.2(c). One equal-contributions CSP returns an IGR with the discrete model that is noticeably larger than that returned by the PH model. This CSP exhibits a large amount of variability in PHs (Appendix Figure A.2(b)), which is exemplified at SE33 where the two contributors have peak pairs at 98 and 327 RFU for the first contributor and 493 and 111 RFU for the second contributor. As described in Section 5.2.1 the PH model penalises σ, meaning L_p is penalised as the allele pairs require large variability, whereas under H_d the PH model pairs the 327 and 493 RFU peaks together as a major contributor and the 98 and 111 RFU peaks together as a minor contributor, reducing the required variability and so incurring less of a penalty to L_d. Using the discrete model, there is no information regarding PHs, and therefore no concept of PH variability to penalise under H_p, leading to the higher IGR with the discrete model in this case. Note that with no PH information, the discrete model supports all six possible genotype pairings of the four observed certain alleles under H_d with equal weight.

All of the major/minor CSPs return an IGR that is larger with the PH model than with the discrete model (Figure 6.3, blue), and cluster according to whether the major (circles) or the minor (crosses) contributor was queried. When the minor is queried, the PH model returns an IGR$>$0 for all CSPs, while the discrete model returns an IGR$\leq$0 for four CSPs, supporting the false H_d. The CSPs for two of these H_d supporting cases are given in Appendix Figures A.3(d) and A.3(b). In CSP 3 only two unmasked peaks of the minor are called as allelic (at vWA and FGA), while in CSP 12 only five unmasked peaks of the minor are called as allelic (two at D21, one at D18, TH01 and SE33). Determining whether or not a minor contributor has an allelic peak in a masking position of the major contributor (either masked by stutter or masked by allelic), is one of the more challenging tasks a forensic scientist has to deal with when calling alleles for a discrete model. This can be seen in both of these CSPs where a number of allelic peaks of the minor have been called as non-allelic, being below the stutter (one of x-n, x-2n or x+n) threshold of one of the peaks of the major contributor. Utilising a continuous model removes this difficulty, as there is no need to call any peaks; the PH model instead estimates what is the most likely genotype of each contributor, given the observed CSP peaks and estimated model parameters. When querying the major contributor the majority of PH IGRs are $\sim$1.0, so the PH model has correctly identified the genotype of Q/X under H_d, while the discrete IGRs range from $\sim 0.4 - 0.8$, so the discrete model has been unable to correctly identify the major genotype. The CSP that returns an IGR close to 1.0 using the PH model, but the lowest IGR

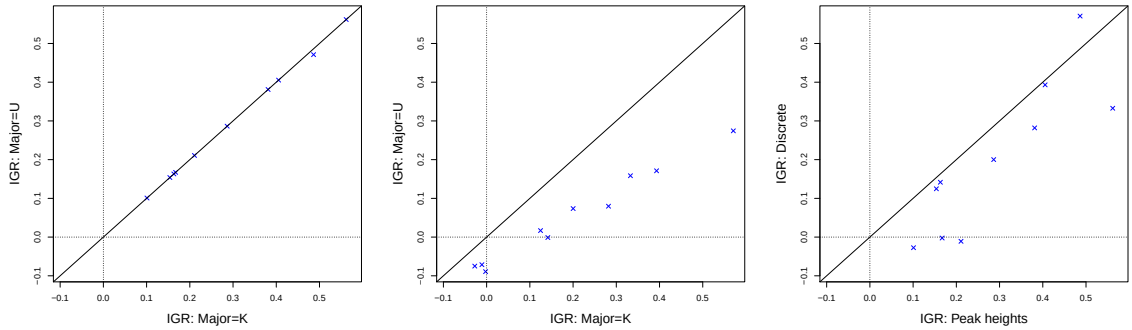


Figure 6.4: Information gain ratio (WoE/IMP) for 10 two-contributor major/minor contribution CSPs comparing the major as unknown (y-axis) with major as known (x-axis) using the PH model (left) or the discrete model (middle), or comparing the discrete model (y-axis) with the PH model (x-axis) when the major is known (right).

with the discrete model is given in Appendix Figure A.3(a), where at loci with < 4 peaks called as allelic H_p will support multiple genotypes for the minor contributor using the discrete model, whereas the PH model will sometimes support a single genotype for the minor contributor, dependent on PH variability. Likewise, under H_d the discrete model will give equal weight to the possible genotypes (see Figure 5.3), whereas the PH model will give the majority of the weight to a single or a few, but fewer than the discrete model, genotype combinations. Two CSPs gave a WoE considerably below the maximum with both the discrete and PH models, the most extreme of which is shown in Appendix Figure A.3(c). The maximum RFU for this CSP is 572 RFU at a homozygous allele of the major contributor, in contrast to RFUs between 1500 and 3000 in the other major/minor CSPs. This suggests that less DNA was introduced into the reaction than intended, resulting in the correspondingly lower WoE with both models.

6.2.5 Major as a known contributor

For the major/minor mixtures in the previous section, the difference in peak contributions is often enough to be able to manually deconvolute the major contributor, at which point they can be included in the hypotheses as a known contributor. This was done for all major contributors that gave an IGR > 0.95 in the two-contributor IGR evaluations. This closely matches forensic practice when a CSP includes a clear unknown major contributor, giving the hypotheses:

$$H_p\text{: } Q+\ K+ \text{ dropin},$$
$$H_d\text{: } X+\ K+ \text{ dropin}.$$

Using the PH model, the IGR assuming the major contributor as known is approximately equal to that assuming major contributor as unknown (Figure 6.4, left) indicating that the PH model is able to fully distinguish the unknown genotype of the major contributor in these cases under both H_p and H_d.

Conversely, when the discrete model is used, the IGR assuming the major contributor as known is always greater than that assuming the major contributor as unknown (Figure 6.4, middle). The discrete model supports many genotype combinations under H_d when the major contributor is unknown, but these genotypes are restricted to those for which $\mathcal{G}_{U1} = \mathcal{G}_K$ when the major contributor is assumed known. Thus the discrete model supports fewer genotypes under H_d with the major contributor as known than unknown, increasing the IGR. The same effect is seen under H_p, where the discrete model supports multiple genotypes for U when the major contributor is unknown, but is restricted to a single genotype when the major contributor is known.

When, instead, the IGR from the PH and discrete models is compared assuming the major contributor as known, the PH model returns a greater IGR than the discrete model in all cases but one (Figure 6.4, right), as predicted by theory. Notably, three cases provide support for H_d using the discrete model, but provide support for H_p using the PH model, which is known to be true here. In all three of these cases some allelic peaks have been called as non-allelic for the discrete CSP (2 peaks in two cases, 5 peaks in the third case). While this is a direct result of the allele calling used for the discrete CSPs, it does mimic the difficulties faced when a forensic practitioner is calling peaks that are in masking positions. Moreover, a number of these mis-called peaks are below the detection threshold routinely employed by forensic labs (50 RFU), and so would not have been available to analyse in normal practice, regardless of having been called as non-allelic due to stutter ratio.

Also of note is the CSP that returns a greater IGR with the discrete model than the PH model (Figure 6.4, right, discrete IGR ≈ 0.55). This discrepancy is largely driven by a high PH variability in this CSP (Appendix Figure A.3(a)) which leads to incorrect genotype inference for X under H_d using the PH model, but in ways that are consistent with the PH information. At D22 ($\mathcal{G}_Q$=11,11) the PH model supports heterozygous genotypes due to the low PH of the 11 peak (251 RFU). At D12 ($\mathcal{G}_Q$=18,23) the PH model assigns the 18 peak as stutter of the major 19 peak. At D1 ($\mathcal{G}_Q$=12,14) the PH model estimates that the minor and major share an allele at the highest peak, whereas the discrete model estimates that they share an allele at the lower of the two major peaks. At D2S441 ($\mathcal{G}_Q$=10,11) a shared major/minor peak is lower than an unshared major peak, so the PH model instead assigns the taller peak as shared and the lower peak as unshared. At each of these loci the prosecution explanation is contrary to the data when PHs are taken

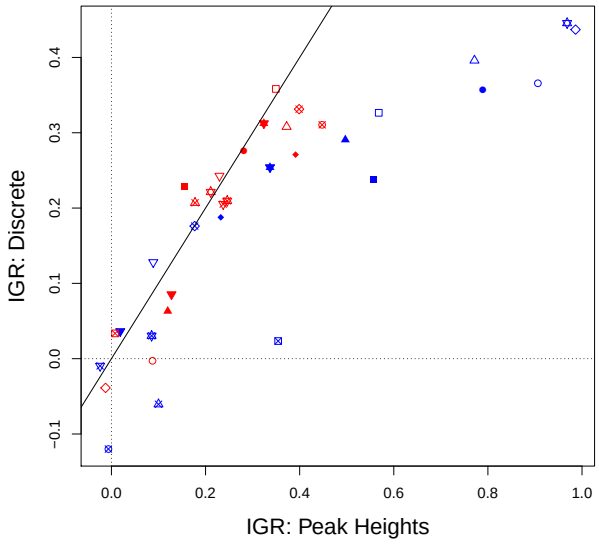


Figure 6.5: Information gain ratio (WoE/IMP) for 6 three-contributor equal contribution CSPs (red, 31:31:31pg) and 6 three-contributor unequal contribution CSPs (blue, 250:64:16pg) using both the PH (x-axis) and discrete (y-axis) models. The six cases of each condition are represented by square, circle, up-triangle, down-triangle, diamond and star symbols. Each contributor to the CSPs was queried in turn, with empty, filled and crossed symbols indicating the first-, second and third-contributor respectively as Q.

into account due to high PH variability, but fits the data better when PHs are not included, so discrete IGR > PH IGR. This once again highlights the difficulty of dealing with PH variability in PH models.

6.2.6 Three contributors

The hypotheses for the three-contributor CSPs were of the form:

$$H_p\text{: } Q\text{+ } U1 + U2,$$
$$H_d\text{: } X\text{+ } U1 + U2.$$

Dropin was not included for the three-contributor CSPs to reduce computation time for the PH model; many dropin alleles will be able to be explained as a stutter/double-stutter/over-stutter of one of the allelic peaks.

When three-contributor equal-contribution CSPs are queried (Figure 6.5, red) the IGR is close to equal between the discrete and PH models for the 18 evaluations. One WoE evaluation (blank diamond) supports H_d with both models (Appendix Figure A.4(b)). Many of the alleles of Q have dropped out in this CSP with both models, and many more are masked. When the first contributor of CSP 4 (Appendix Figure

A.4(a)) is queried, the PH model supports H_p, while the discrete model supports H_d (blank circle). Seven of the alleles of Q have dropped out, which still supports H_p with the PH model due to a low estimated DNA contribution ($\widehat{\chi_p} = 41.9$ RFU).

When unequal-contribution CSPs are queried (Figure 6.5, blue) seven evaluations are at parity between the two models, while 11 evaluations return a larger IGR with the PH model. All six CSPs obtain an IGR ordering of IGR_{250} >IGR_{62} >IGR_{16} for both models, as would be expected. Of the seven evaluations that are at parity between the two models, three correspond to CSP 5 (downward-triangle symbols). This CSP (Appendix Figure A.5(d)) consists of 13 loci with no peaks observed, one locus with a single dropin peak observed (D8), a locus where peaks are observed from each contributor (D2S1338), and a locus where peaks are only observed for the 62 pg and 250 pg contributors (SE33). Querying this CSP gives limited support for H_p when the 250 pg (blank down-triangle) and 62 pg (filled down-triangle) contributors are queried and limited support for H_d when the 16 pg contributor is queried (crossed down-triangle) under both models. Querying the minor contributor of CSP 2 (crossed circle) provides support for H_d with both models, but stronger support with the discrete model. Six alleles of Q have dropped out (Appendix Figure A.5(b)). For the discrete model two and seven alleles of Q have been called as non-allelic and uncertain respectively, of which one of the uncertain alleles is homozygous. For the PH model nine allelic peaks can be explained well as stutters from major contributors, reducing the likelihood that Q contributes. When the minor contributor of CSP 1 is queried (crossed square) the PH IGR is approximately 0.3 greater than the discrete IGR, because of two uncertain and two non-allelic calls for alleles of Q (Appendix Figure A.5(a)). When the major contributor of CSP 4 is queried (blank diamond) the PH IGR is approximately 0.6 greater than the discrete IGR, because the PHs support a single genotype at each locus, while the discrete model supports multiple genotypes.

6.2.7 Runtime

Mean runtime for the laboratory validation CSPs was 5, 10 and 92 minutes for single-, two- and three-contributor CSPs respectively, with all evaluations completing in under four hours (Figure 6.6).

6.2.8 Laboratory conclusions

The results presented here demonstrate that the PH variability in these laboratory-generated CSPs is not large enough to break down genotype inference for unequal-contributions CSPs with the PH model, leading to higher IGRs with the PH model. Conversely, the PH variability is large enough to reduce the ability of

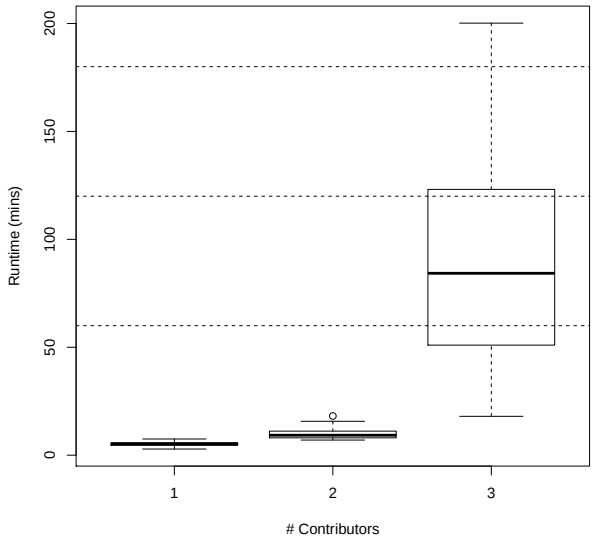


Figure 6.6: Runtime for the laboratory validation evaluations. Hypotheses included dropin for the single- and two-contributor evaluations, three-contributor evaluations did not include dropin. Horizontal dashed lines indicate whole hours.

the PH model to infer contributor genotypes to that of the discrete model when contributors are represented equally, leading to approximately equal IGRs between the PH and discrete models.

The problem of PH variability is highlighted for CSPs where the discrete model returns greater IGRs than the PH model; the PH model utilises PH information to form incorrect inferences about the genotypes of the contributors, but these inferences are those that best fit the data.

These results suggest that evaluation with a PH model is desirable for unequal-contributions mixtures, especially for a queried minor contributor where a more extreme WoE in favour of the true hypothesis is desirable. Conversely, for CSPs that appear to have contributors at equal representations, a discrete model should be adequate to utilise all of the information available in the CSP, as PH variability diminishes the information available in PHs for genotype inference.

Throughout the laboratory validation tests, the detection threshold used to generate the discrete CSPs was 20 RFU. However, forensic laboratories normally use a detection threshold of 50 RFU for discrete analyses, so many low-level peaks are observed in these discrete CSPs that would not be observed in normal practice. This means that these results overstate the power of the discrete model in usual practice, especially for very low-level contributors. Over the CSPs tested, there are a total of 31 peaks of the various queried contributors that are <50 RFU in the single-contributor cases, 63 and 100 such peaks in the

two-contributor major/minor and equal contributions cases respectively, and 19 and 51 such peaks in the three-contributor unequal and equal contributions cases respectively; this information supporting H_p would usually be unavailable to the discrete model.

6.3 Altering the model assumptions

One of the laboratory CSPs was evaluated, with differing modelling assumptions; all combinations of modelling double- and over-stutter, modelling dropin, and removing the locus-dependency of the stutter gradient. This verifies that the WoE for the chosen CSP behaves as expected with the various model assumptions. If a modelling assumption is not important for explaining any feature of the given CSP, it should have little impact on the WoE. Conversely, if a CSP cannot be well explained without a given modelling assumption, altering that assumption should have a large effect on the WoE.

6.3.1 Protocol

One of the three-contributor CSPs was evaluated using just the PH model with hypotheses of the form:

$$H_p\text{: } Q(16\text{pg}) + K1\ (250\text{pg}) + U1,$$
$$H_d\text{: } X + K1\ (250\text{pg}) + U1.$$

The CSP was evaluated with varying assumptions of the model; whether or not to model dropin, double-stutter, over-stutter or a locus-specific stutter gradient.

6.3.2 Results

Modelling dropin does not change the WoE for this CSP (Table 6.3), as dropin is not necessary to explain the CSP when double- and over-stutter are both modelled, as evidenced by the dropin estimates of 5 and 5 RFU under H_p and H_d respectively, equal to the minimum dropin value of 5.0. Similarly, removing double-stutter from the model does not change the WoE as there are no peaks in the CSP that can only be explained through double-stutter. Conversely, removing over-stutter from the model reduces the WoE, particularly because the 17 peak at D22 can no longer be explained by over-stutter (D22 WoE decreases from -0.5 bans with SDO to -0.8 and -0.7 bans with SD and S respectively), so must be assumed to be allelic by the program. D22 is subject to over-stutter more commonly than any other locus in the NGM Select™ kit due to being the only locus with repeat units that are three base pairs long, rather than the standard

Alterations	**Basic**	**Stutter model**				
	SDO	SDO+dropin	SD+dropin	SO+dropin	S+dropin	$\alpha_l = 1$
Parameters						
Dropin	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
D	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE
O	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE
WoE						
D10S1248	0.6	0.6	0.6	0.6	0.6	0.6
vWA	1	1	1.1	1.1	1.1	1.1
D16S539	0.5	0.5	0.5	0.5	0.5	0.5
D2S1338	0.5	0.5	0.4	0.5	0.4	0.6
D8S1179	1.1	1.1	0.9	1.1	0.9	1.1
D21S11	1.7	1.7	1.7	1.7	1.7	1.7
D18S51	1	1	1	1	1	1.1
D22S1045	-0.5	-0.5	-0.8	-0.5	-0.7	-0.5
D19S433	-2.8	-2.8	-2.8	-2.8	-2.8	-2.8
TH01	0.5	0.5	0.5	0.5	0.5	0.5
FGA	0.6	0.6	0.5	0.6	0.5	0.5
D2S441	1.2	1.2	1.2	1.2	1.2	1.2
D3S1358	0.7	0.7	0.7	0.7	0.7	0.7
D1S1656	0.7	0.7	0.7	0.7	0.7	0.7
D12S391	1.4	1.3	1.3	1.4	1.3	1.3
SE33	0.1	0.1	0.1	0.1	0.1	0.1
Overall	8.2	8.2	7.6	8.2	7.7	8.4

Table 6.3: Locus and overall WoE for a chosen three-contributor laboratory-generated CSP, with altered assumptions of the model. Column three models dropin, columns four to six alter whether double or over stutter are being modelled while column seven removes the locus dependency on the stutter gradient.

four base pairs. In the PH model the stutter ratio is assumed linear with the LUS of the allele, with the gradient of the linear relationship allowed to differ between loci through a locus adjustment parameter (α) that is a multiplicative adjustment to the mean gradient over loci. When the stutter gradient is instead assumed to not vary between loci (α_l=1) the WoE increases to 8.4 bans. This change in WoE is driven by the defence likelihood at D2S1338; at this locus $\mathcal{G}_Q$=17,22 but the most likely $\mathcal{G}_X$=17,18 meaning that the truly allelic peak at 22 is estimated to be stutter from one of the majors under H_d ($\mathcal{G}_{K1}$=18,23), requiring a large stutter gradient which is not possible when the stutter gradient cannot vary by locus. This means that the defence hypothesis has a higher likelihood at D2S1338 when the stutter gradient is allowed to vary by locus, leading to a lower locus LR with a locus variant gradient (0.46) than with a fixed gradient (0.61).

6.3.3 Conclusions

As expected, the WoE is only altered when a modelling assumption that is important in explaining the features of the CSP is removed. The features range from as simple as a peak in a position that can or cannot

be explained by the modelling assumptions, to more complicated features such as the variance of stutter gradients between loci. From these, it is clear that a simplistic model that assumes a constant stutter rate across the epg, or that cannot model over- or double-stutter, is not sufficient to adequately explain all of the features of this complex CSP, justifying the complexity of the PH model.

6.4 Artificially altering the input data

One of the single-contributor laboratory-generated CSPs was altered one peak at a time, either introducing peaks that had dropped out, altering the PH of observed peaks, or introducing dropin peaks. Introducing a dropped out peak of Q is expected to increase the WoE against Q, removing a peak of Q is expected to decrease the WoE against Q and introducing a dropin peak is expected to decrease the WoE against Q. These are all intuitive expectations, which, if violated, would call into question the validity of the model. Additionally, changes in peak heights that require greater variance of peak heights to explain the CSP under H_p should decrease the WoE and *vice versa*; this expectation stems from the results seen in Section 6.2, where high PH variability reduced the WoE against Q due to impaired genotype inference.

6.4.1 Protocol

The single-contributor CSP from donor 26 (16 pg DNA, approximately equivalent to 3 cells) was chosen to investigate the behaviour of the PH model when altering the CSP, as it had a mixture of locus dropouts (both heterozygote and homozygote), single dropouts (heterozygote) and non-dropouts (both heterozygote and homozygote), so that the behaviour could be investigated at each of these classes of observation. See Table 6.4 for a summary of the alterations made to the CSP.

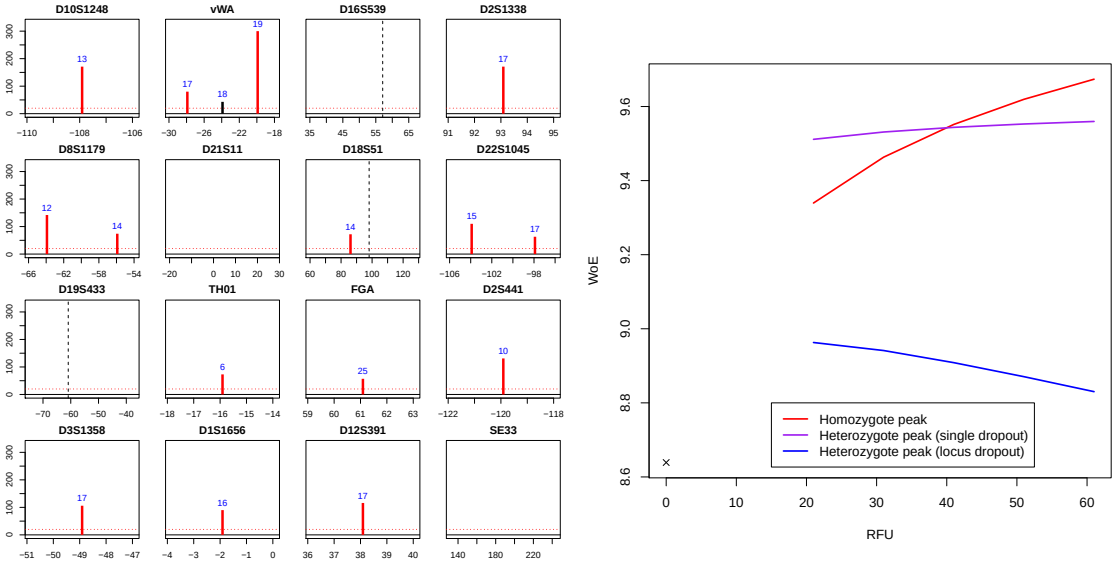
6.4.2 Insertion of a missing peak

A peak at the location of a single allele of Q which had dropped out was added to the CSP with varying PH. This was done at three separate loci with:

1. No observed peaks, Q is homozygous (D16): homozygous locus dropout.
2. No observed peaks, Q is heterozygous (D19): heterozygous locus dropout.
3. One observed peak, Q is heterozygous (D18): heterozygous single dropout.

Locus	$\mathcal{G}_Q$	CSP	Observation	Alteration
D16	13,13	$\emptyset$	Dropout of homozygous 13 allele	Reintroduction of 13 allele
				Introduction of 11 or 15 dropin peak
D18	14,17	14	Dropout of heterozygous 17 allele	Reintroduction of 17 allele
				Introduction of 8 or 12 dropin peak
D22	15,17	15,17	Fully observed heterozygote	Alteration of peak height at allele 17
				Introduction of 16 or 19 dropin peak
D19	13,14	$\emptyset$	Full heterozygous dropout	Reintroduction 13 allele
				Introduction of 15 or 18 dropin peak
TH01	6,6	6	Observed homozygote allele	Alteration of peak height at 6
				Introduction of 8.3 or 9.3 dropin peak
FGA	23,25	25	Dropout of heterozygous 23 allele	Alteration of peak height at 25
				Introduction of 21 or 22.1 dropin peak

Table 6.4: Alterations applied to a single-contributor 16pg CSP at six loci. $\mathcal{G}_Q$ indicates the genotype of Q, the true contributor. $\emptyset$ under CSP indicates no observed peaks above the detection threshold at that locus. Observation gives the true effect seen at the locus. Alteration gives the two changes that were made at each locus. Reintroductions of dropped-out alleles ranged from 0 to 61 RFU, introductions of dropin peaks ranged from 0 to 61 RFU and alterations of observed peaks ranged from 0 to 151 RFU.



(a) CSP (b) Results

Figure 6.7: (a) The single-contributor CSP for which PHs were altered. Vertical dashed lines indicate the position of dropped-out alleles that were inserted. (b) WoE for a single CSP when a dropped out allele is artificially inserted at differing RFUs.

Inserting a homozygous dropout peak of Q increases the WoE from ∼8.6 bans to ∼9.3 bans, which increases to >9.6 bans at 61 RFU (Figure 6.7, red).

Inserting a heterozygous dropout peak of Q for which the corresponding allele was observed increases the WoE from ∼8.6 bans to ∼9.5 bans (Figure 6.7, purple); a larger increase than when a homozygous allele was inserted. However, the WoE increases by a smaller amount as the RFU of the inserted peak increases, reaching ∼9.6 bans at 61 RFU. Below 40 RFU, an insertion of a homozygous dropout gives a lower WoE against Q than an insertion of a heterozygote single dropout. This reverses above 40 RFU. This is intuitive, as a small heterozygous peak is more likely than a small homozygous peak, leading to a greater WoE for the heterozygous peak at small RFUs. Similarly, a large heterozygous peak is less likely than a large homozygous peak, leading to a greater WoE for the homozygous peak at large RFUs.

Inserting a heterozygous dropout peak of Q for which the corresponding allele also dropped out increases the WoE initially from ∼8.6 bans to ∼9.0 bans (Figure 6.7, purple). The WoE decreases with increasing RFU, reaching ∼8.8 bans at 61 RFU. The variability in PHs required to explain the remaining dropout allele with the observed inserted allele under H_p increases with increasing RFU of the introduced peak, increasing the penalty on σ, so reducing the WoE.

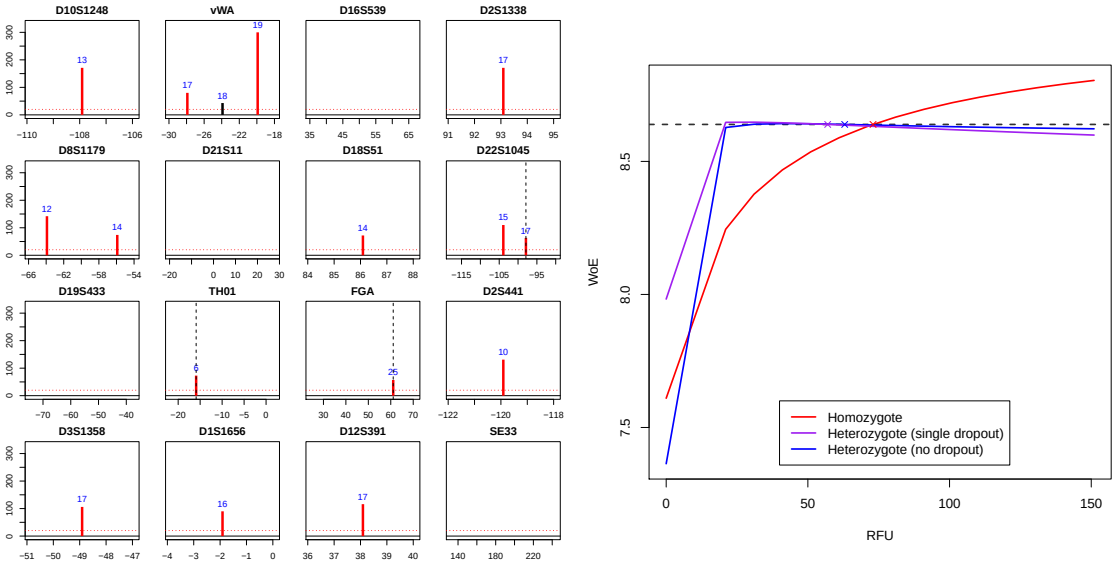
6.4.3 Altering observed CSP peaks

A single peak in the CSP was given an altered RFU, from below the detection threshold (shown as 0 RFU here, analogous to dropout) to 150 RFU. This was performed for peaks at three separate loci with:

1. One observed peak, Q is homozygous (TH01): homozygous peak.
2. One observed peak, Q is heterozygous (FGA): heterozygous peak with dropout.
3. Two observed peaks, Q is heterozygous (D22): heterozygous peak.

When the PH of a homozygous peak of Q is altered, the WoE has a strong positive relationship with the RFU of the peak (Figure 6.8, red), and decreases from ∼8.3 bans at 21 RFU to 7.6 bans when the peak is removed. The WoE increases between 21 and 151 RFU, because the normalised posterior $\Pr(\mathcal{G}_X = \mathcal{G}_Q)$ under H_d increases as the RFU of the peak increases. Below 91 RFU the most likely genotype is heterozygous, including the 6 allele. At 91 RFU and above, the most likely $\mathcal{G}_X$ is a 6,6 homozygote, which is the true genotype of X.

When the PH of a heterozygous peak of Q for which the corresponding allele dropped out is altered, the WoE has a weak negative relationship with the RFU of the peak (Figure 6.8, purple), and decreases from



(a) CSP (b) Results

Figure 6.8: (a) The single-contributor CSP for which PHs were altered. Vertical dashed lines indicate the position of peaks that were altered. (b) WoE for a single CSP when the PHs of an observed peak is artificially altered, from 0 RFU to 151 RFU. Crosses and the dashed horizontal line indicate the WoE and RFU when no peak is altered.

∼8.6 bans at 21 RFU to ∼8.0 bans when the peak is removed. The WoE decreases with increasing RFU because a homozygous genotype is given increasing weight under H_d as the RFU of the peak increases, so H_d fits the data better as the PH increases relative to H_p.

When the PH of a heterozygous peak of Q for which the corresponding allele was also observed is altered, the WoE decreases slightly as the RFU of the peak deviates from that observed in the unaltered CSP (Figure 6.8, blue). The WoE decreases from ∼8.6 bans at 21 RFU to ∼7.4 bans when the peak is removed. The WoE decreases as the RFU moves away from the unaltered CSP because the increased variability increases σ under H_p, which is penalised. However, the reduction in WoE is small because σ is penalised under both H_p and H_d due to increasing PH variability.

The WoE against Q is greater when full locus heterozygous dropout was introduced than when a single heterozygous dropout was introduced (Figure 6.8, RFU=0, blue and purple); this is counter-intuitive, as an extra observed allele may be thought to increase the WoE against Q. This is due to the increased PH variability under H_p introduced by a single heterozygous dropout. Dropout of a homozygous allele gives a greater WoE against Q than a single heterozygous dropout (Figure 6.8, RFU=0, red and blue), because the increased PH variability under H_p with a single heterozygous dropout increases the penalty on σ. Dropout of a homozygous allele gives a lower WoE against Q than full locus heterozygous dropout (Figure 6.8, RFU=0,

Locus	Common		Rare	
	Allele	Probability	Allele	Probability
D16S539	11	0.317	15	7.84e-4
D18S51	12	0.149	8	3.92e-4
D22S1045	16	0.369	19	1.18e-3
D19S433	15	0.179	18	3.92e-4
TH01	9.3	0.334	8.3	1.17e-3
FGA	21	0.179	22.1	3.92e-4

Table 6.5: Dropin alleles that were inserted into the donor 26 16 pg DNA CSP. Common alleles were chosen as the highest frequency allele in the DNA17 NDU1 database not-shared with Q. Rare alleles were chosen as the lowest frequency allele in the database.

red and purple); homozygous dropout is less likely than locus dropout due to the increased expected PH for a homozygous allele.

6.4.4 Insertion of a dropin (non-Q) peak

A single peak was inserted into the CSP at the six previously altered loci, with the newly inserted peak being at a non-Q allele, simulating a dropin event. At each of the six loci both the highest frequency non-Q allele and lowest frequency allele in the DNA17 NDU1 database (Caucasian) were inserted separately. Inserted alleles and population probabilities (without sampling or F_{ST} adjustment) are given in Table 6.5.

At all loci, introducing a dropin peak decreases the WoE from the non-dropin WoE of 8.6 bans to between 7.0 and 8.5 bans (Figure 6.9). For all conditions the WoE is further reduced as the PH of the dropin peak increases from 21 RFU to 61 RFU. The reduction in WoE varies substantially between loci, ranging from 0.05 bans at D22 with a 21 RFU dropin of a common allele to 1.6 bans at D19 with a 21 RFU dropin of a rare allele.

At D22 (red) both of the alleles of Q are observed in the CSP, plus the third introduced dropin peak. The WoE with introduction of a common (solid line) or rare (dashed line) allele diverges as the RFU of the introduced peak increases. Under H_p the dropin peak must be assigned as a dropin, which is more plausible for a common allele than for a rare allele, and becomes increasingly implausible as the RFU of the dropin peak increases because the non-degradation-adjusted dropin dose at each allele is linear with p_x. Under H_d the model correctly assigns the peak as dropin when it is a common allele, but incorrectly assigns an allele of Q as dropin at 61 RFU when the true dropin peak is rare. This incorrect assignment explains the data better than the true circumstances can, because an allelic assignment has no effect of p_x on the expected dose, so H_d has a relatively better fit than H_p when the dropin is rare compared to when it is common, which results in the WoE dropping more with increasing RFU for the rare dropin than for the common dropin.

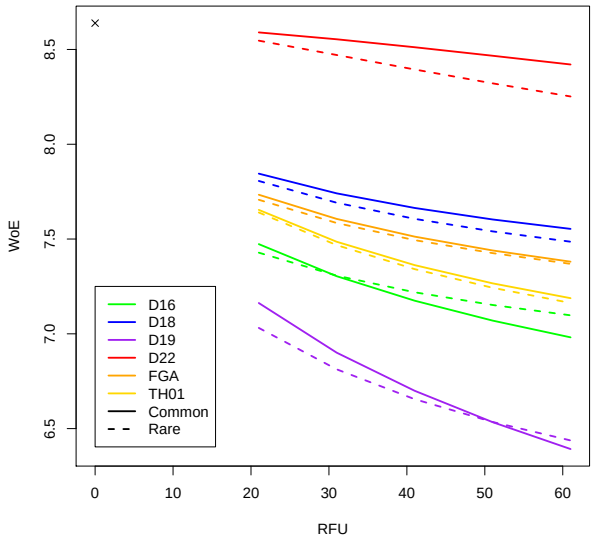


Figure 6.9: Weight-of-evidence for a single-contributor 16 pg DNA CSP when a single rare or common dropin peak is inserted at one of six loci. See Table 6.5 for inserted alleles and their associated population probabilities.

At TH01 (yellow), FGA (orange) and D18 (blue), a single allelic peak (homozygous, heterozygous and heterozygous respectively) was observed in the CSP, plus the introduced dropin peak. At these loci, H_d explains the CSP as a heterozygous genotype composed of the observed true-allelic peak and the introduced dropin peak. The H_p explanation of a dropout and a dropin fits poorly when the dropin peak is rare, while the H_d explanation fits well for both common and rare dropins, so the WoE is lower for a rare dropin than for a common dropin.

At D16 (green) and D19 (purple) no peaks were observed in the original CSP, so the CSPs here consist of just the dropin peak. When the dropin peak is common in the population, under H_d the model explains the observed peak as heterozygous at low RFU, but switches to explaining it as homozygous at high RFU. Conversely, when the dropin peak is rare a homozygote is *a priori* unlikely, as under Hardy-Weinberg assumptions the probability of a homozygote is p_Z^2, which is 6.1e-7 and 1.5e-7 for the rare dropin allele at D16 and D19 respectively. For a common dropin the H_d explanation of a common homozygous allelic peak has an increasingly better likelihood compared to the H_p explanation of a common dropin as the RFU increases, reducing the WoE. However, for a rare dropin, the H_d explanation of a rare heterozygous peak does not increase its likelihood as much when the RFU increases, while the H_p explanation also performs less well as the RFU increases, so there is less discrepancy between the H_p and H_d explanations, leading to

the lower reduction in WoE.

6.4.5 Conclusions

The PH model adheres to all expected behaviours when the input data are altered; the WoE against Q decreases when an allele of Q is dropped out, a non-Q dropin allele is introduced or when PH variability is increased, while the WoE increases when a dropped out allele of Q is introduced or the PH variability is decreased.

These results highlight the importance of PH variability to the behaviour of the model, with some seemingly counter-intuitive results making sense in the light of required PH variability under each hypothesis. For example dropout of homozygous allele of Q should intuitively decrease the WoE against Q by more than a dropout of a heterozygous allele. However, the increased PH variability introduced when a heterozygous allele is dropped out for which the corresponding allele was observed leads to a greater reduction in WoE, going against intuition.

6.5 Published results comparison

A set of artificial CSPs for which the evaluated WoEs with a number of models has been published was evaluated with the likeLTD PH model. These evaluations are useful to benchmark the likeLTD PH model against both discrete and continuous models, on CSPs for which the ground truth is known. Due to the design of the CSPs, these evaluations should also reiterate the results seen in Section 6.2, as these CSPs also consider both major/minor and equal-contributions mixtures.

6.5.1 Published data

Bright et al. [2015] recommended a series of tests for validating probabilistic software, for which they published results using STRmix, and two discrete models, LRmix and LabRetriever. Here the PH model described in Section 5.2 will be referred to as likeLTD, as it has been published in a CRAN package of the same name. Bright et al. [2015] assigned genotypes for two individuals, from which they generated single- and two-contributor CSPs. Parameter and modelling choices were selected to make the evaluation with likeLTD compatible with both the data generation process and the STRmix evaluation; $t_l = 50$, no DS or OS, F_{ST}=0.01, sampling adjustment=0.

There are assumptions in the STRmix model that may have been included when generating the

CSPs, that may differ from the assumptions of likeLTD. These differences are unlikely to affect the results to a large extent, however, it is useful to highlight them:

Stutter: The STRmix stutter model [Bright et al., 2013c] is analogous that of likeLTD, with a locus-specific linear relationship with LUS, however, likeLTD assumes an intercept of 0, while STRmix models the intercept fully. For alleles with an unknown LUS value, STRmix assigns a LUS value equal to the allele designation, while likeLTD extrapolates or interpolates the unknown LUS value.

Locus effects: STRmix allows different loci to have different amplification efficiencies, resulting in different expected RFU values at each locus, while likeLTD does not model this. This is expected to lead to increased amplification efficiency at low molecular weight loci, and higher RFU values at those loci, so can partially be accounted for by degradation with likeLTD.

6.5.2 Results

Case	Q	Hp	nU	Dropin	likeLTD	LRmix	LabRetriever	STRmix
SS1	1	TRUE	0	FALSE	18.8	18.8	18.8	18.8
SS1	2	FALSE	0	TRUE	-61.0	NA	NA	NA
SS2	1	FALSE	0	TRUE	-48.8	NA	NA	NA
SS2	2	TRUE	0	FALSE	19.6	19.6	19.6	19.6
Bal	1	TRUE	1	FALSE	10.1	9.0	9.0	9.8
Bal	2	TRUE	1	FALSE	11.0	10.0	10.0	10.6
MM	1	TRUE	1	FALSE	18.8	18.5	18.5	18.5
MM	2	TRUE	1	FALSE	19.6	16.3	16.2	19.3
Stochastic	1	TRUE	1	FALSE	18.6	NA	NA	18.5
Stochastic	2	TRUE	1	FALSE	19.3	11.4	12.0	15.7

Table 6.6: [WoE for Bright et al. [2015] cases using the likeLTD peak height model, LRmix, LabRetriever and STRmix.] Q indicates the reference profile used as the queried contributor, Hp indicates whether or not Q is a contributor to the CSP, while nU indicates the number of unknown contributors assumed under H_p. Dropin indicates whether dropin was modelled for likeLTD, and was only used when a non-contributor Q was queried for a single-contributor case. The IMP for reference profiles 1 and 2 are 18.8 and 19.6 bans respectively. Blank cells are present for SS1 and SS2 because Bright et al. did not query the non-contributor for single-contributor CSPs. Blank cells are present for Stochastic because Bright et al. did not perform a calculation with LRmix or LabRetriever for the Stochastic CSP.

For a single contributor CSP all four programs return WoE = IMP for a true Q (Table 6.6, SS1 and SS2). These CSPs have no stutter peaks, dropout or dropin, so normalised $P(E|\mathcal{G}_X = \mathcal{G}_Q) = 1$, and

the LR simplifies to $1/P(\mathcal{G}_Q) = \text{IMP}$ for all programs. This is achieved under continuous models (likeLTD and STRmix) by $\prod_{i\in I} P(h_i|E(h_i),\phi)$ being equal under H_p and H_d for $\mathcal{G}_Q$. For discrete models this is achieved by estimating $D = 0$. Bright et al. [2015] did not query the corresponding non-contributor for each single-contributor CSP, but likeLTD gives a WoE that supports H_d for non-contributor evaluations.

For a Major/minor mixture (MM, Table 6.6) the likeLTD WoE is equal to the IMP for both contributors. The minor contributor (Q=2) contributes enough DNA ($\widehat{\chi_c} \approx 590$ under H_p and H_d) to obtain full information about their genotype. Estimated DNA contributions are approximately 1770 and 590 RFU for the major and minor respectively, so a shared allele has expected PH of approximately 2360 RFU, which likeLTD can distinguish from the expected PH of a heterozygous unshared major allele. As an example from the CSP, at D21 $\mathcal{C}_{D21}$=28,29,30,31 with h_{D21}=103, 2046,1487,482 RFU; the shared 29 allele is clearly distinguishable from the unshared major allele at 30. This may not be possible for CSPs with higher variability in PHs or a lower contribution of the minor. STRmix returns WoEs that are slightly lower than likeLTD WoEs for both contributors, whereas LabRetriever and LRmix return significantly reduced WoEs when querying the minor contributor.

When the average RFU of the minor contributor is decreased to below the stochastic threshold (Stochastic, Table 6.6) the likeLTD WoE for the minor contributor (Q=2) falls to three decibans below the IMP, due to reduced DNA contribution ($\widehat{\chi_c} \approx 190$ under both H_p and H_d, close to the published average profile PH of 180 RFU). The estimated DNA contributions are approximately 190 and 3000 RFU for the major and minor contributors respectively, so a shared major/minor peak may not be distinguishable from an unshared heterozygous major peak, reducing the WoE for the minor contributor. The WoE of the major contributor (Q=1) remains at the IMP. The STRmix WoE is unchanged for the major contributor, but is reduced by 3.6 bans for the minor contributor. The WoE for the minor contributor from LRmix and LabRetriever is reduced by 4.9 and 4.2 bans respectively.

When the mixture has equal contributions (Bal, Table 6.6) the WoE for both contributors drops significantly. Genotype deconvolution is more difficult, so multiple genotypes will be supported under H_d (see Table 5.3 and Figure 5.3). The likeLTD WoE falls by 8.6 and 8.3 bans for the first and second reference respectively when compared to the Stochastic case. The STRmix WoE falls by 8.7 and 5.1 bans respectively. The WoE for the second reference falls by 1.4 and 2.0 bans for LRmix and LabRetriever respectively. When two contributors have equal contributions, and an allele is observed that appears to have a double dose, then it is not possible to determine whether the peak is homozygous for contributor A, homozygous for contributor B or shared heterozygous between the two contributors using PHs, just as it is not possible to

determine whether a single dose peak originates from contributor A or B; the same is true of discrete models, so all four programs return similar WoEs.

6.5.3 Conclusions

These results mirror those seen in Section 6.2 with the continuous models, likeLTD and STRmix, providing greater WoEs for a true Q than the discrete models, LRmix and LabRetriever, for unequal-contributions mixtures, especially for a minor contributor. The continuous and discrete models provide similar WoEs for equal-contributions mixtures, or for single-contributor CSPs.

6.6 Real case comparison: Meredith Kercher

A CSP from a real-world crime was evaluated with three continuous models. This comparison benchmarks the models against each other, but in a real-world scenario rather than for the artificial CSPs evaluated in Section 6.5. Assuming that all of the models are valid, the results obtained with each should be similar.

6.6.1 Case circumstances

In November 2007, Meredith Kercher was murdered in her flat in Perugia, Italy. While Rudy Guede was tried and convicted for the crime in under a year with little controversy, the accusation that Raffaele Sollecito and Amanda Knox were involved in the murder was much more controversial. The two were found guilty in December 2009, acquitted in October 2011, found guilty again in January 2014, and finally ruled innocent by the highest court in Italy in March 2015. One of the key, and controversial, pieces of evidence in the case against Knox and Sollecito was Meredith Kercher's bra clasp, item 165B, found on the floor of the room Meredith was murdered in, over a month after the murder occurred. Here, the WoE of the epg arising from the bra clasp for both Knox and Sollecito to be a contributor will be evaluated using the likeLTD, STRmix and EuroForMix PH models. The hypotheses compared are of the form:

$$H_p^S\text{: } Q \text{ (Raffaele Sollecito)} + K1 \text{ (Meredith Kercher)} + U1,$$
$$H_d^S\text{: } X + K1 \text{ (Meredith Kercher)} + U1,$$

and:

$$H_p^K\text{: } Q \text{ (Amanda Knox)} + K1 \text{ (Meredith Kercher)} + U1,$$
$$H_d^K\text{: } X + K1 \text{ (Meredith Kercher)} + U1.$$

Program	likeLTD				STRmix		EuroForMix			
Q	Sollecito		Knox		Sollecito	Knox	Sollecito		Knox	
t	20	50	20	50	50	50	20	50	20	50
D8	0.6	0.7	0.3	-0.2	0.2	1.0	0.5	0.8	0.4	-0.2
D21	0.7	0.5	0.0	-0.1	0.8	0.1	0.9	0.8	0.2	0.1
D7	0.5	0.4	-0.1	-0.2	0.5	-0.4	0.5	0.5	-0.4	-0.1
CSF	0.7	0.7	-0.1	-0.1	0.3	0.0	0.4	0.6	0.0	0.0
D3	1.0	0.8	0.1	-0.4	0.8	0.4	1.0	1.0	-0.1	-0.2
TH01	1.1	1.1	-0.1	-0.3	0.8	-0.6	0.9	1.2	-0.3	-0.3
D13	0.8	0.7	0.1	0.0	0.7	-0.3	0.8	0.8	0.0	-0.1
D16	0.7	0.8	0.0	0.0	0.9	0.0	0.6	0.7	0.2	0.1
D2	2.5	2.4	-0.1	0.1	1.6	0.5	2.0	2.1	0.6	0.0
D19	1.4	1.3	-1.1	-0.9	1.4	-1.4	1.8	1.6	-1.5	-1.4
vWA	1.6	1.5	0.0	-0.2	1.9	-0.7	1.7	1.8	-0.8	-0.4
TPOX	0.9	0.9	0.0	-0.1	0.7	0.1	0.5	0.7	-0.2	-0.1
D18	1.3	1.4	0.1	0.2	1.2	0.3	1.2	1.4	0.4	0.3
D5	-0.3	-0.4	0.0	0.0	0.0	0.3	-0.3	-0.5	0.1	0.1
FGA	-0.9	-1.2	0.0	0.0	0.0	0.1	-0.4	-0.5	0.4	0.0
Overall	12.5	11.5	-0.9	-2.3	11.8	-0.7	12.0	13.0	-1.1	-2.1

Table 6.7: Locus and overall weight of evidence (WoE) for the epg generated from item 165B (bra clasp) in the Kercher case. WoE was evaluated against Raffaele Sollecito or Amanda Knox with a detection threshold (t) of 20 or 50, and with three continuous models; likeLTD, STRmix and EuroForMix. In all evaluations Meredith Kercher was assumed to be a contributor, with another unknown individual and Q/X. The IMP for Sollecito is 18.5 bans.

6.6.2 Results

Raffaele Sollecito

When Raffaele Sollecito is queried with detection threshold, $t = 50$, all three programs return a WoE$\geq$11.5 bans (Table 6.7). At t=50 likeLTD and STRmix have similar WoEs ($\Delta = 0.3$ bans) but EuroForMix has a WoE > 1 ban larger (Δ=1.5 and 1.2 bans for likeLTD and EuroForMix respectively). The three programs have largely good correlation between locus WoEs, with two exceptions:

D5: likeLTD and EuroForMix have similar WoEs supporting H_d, STRmix supports neither hypothesis. Sollecito is homozygous and masked by a heterozygous peak of Kercher.

FGA: likeLTD and EuroForMix support H_d, STRmix supports neither hypothesis, the likeLTD and EuroForMix WoEs are now considerably different. Sollecito is heterozygous and both alleles are masked by alleles of Kercher.

likeLTD reduces the WoE against Sollecito when t is changed from 20 RFU to 50 RFU by 1.0 bans, while EuroForMix increases the WoE by 1.0 bans, giving a more similar WoE when $t = 20$ (Δ=0.5 bans).

At $t = 20$ there are no obvious discrepancies between likeLTD and EuroForMix.

The runtime for likeLTD was between 16 and 17 minutes, while EuroForMix and STRmix took less than a minute to run.

Amanda Knox

When Amanda Knox is queried, all three programs support H_d at $t = 50$ (Table 6.7), with likeLTD and EuroForMix having similar WoEs ($\approx$-2 bans), while STRmix has a noticeably larger WoE (-0.7 bans). There are some notable locus differences between the programs:

D8: EuroForMix and likeLTD support H_d, STRmix has the strongest support for H_p of any program and any locus when querying Knox. Knox has one observed allele in a stutter position of Kercher allele, and a dropout allele in the double-stutter position of the same Kercher allele. The dropped out allele was observed when t=20 explaining the support for H_p by both likeLTD and EuroForMix when t=20.

D3: EuroForMix and likeLTD support H_d, STRmix supports H_p. Knox has one allele masked by Kercher, and another allele that has dropped out. The dropout allele was observed when t=20, with which likeLTD supports H_p, but EuroForMix continues to support H_d.

D13: EuroForMix and STRmix support H_d, likeLTD supports neither hypothesis. One allele of Knox is masked by Kercher, while the other has dropped out. The dropped out allele was observed when t=20, with which likeLTD supports H_p but EuroForMix continues to support H_d.

Both EuroForMix and likeLTD show an increase in WoE when t is changed from 50 RFU to 20 RFU; three peaks that match a heterozygous allele of Knox, and a peak that matches a homozygous allele of Knox are introduced into the CSP when decreasing t, out of six total peaks introduced. The two programs continue to have a similar overall WoE with $t = 20$, however, there are some notable discrepancies in locus WoEs:

D2: likeLTD supports H_d, EuroForMix supports H_p. Knox has one allele masked by Kercher, and one observed allele in the double-stutter position of the same Kercher allele. EuroForMix does not model double-stutter, so must explain the Knox allele as allelic under H_d, supporting H_p, while likeLTD is free to explain the peak as a double-stutter of the Kercher peak under H_d, and so supports H_d.

vWA: likeLTD support neither hypothesis, EuroForMix supports H_d. A homozygous allele of Knox is unobserved at this locus.

FGA: likeLTD supports neither hypothesis, EuroForMix supports H_p. A peak has been observed that matches a homozygous allele of Knox. EuroForMix must explain this as allelic, therefore supporting H_p, however, likeLTD is able to explain it as an over-stutter of a Kercher peak, supporting neither hypothesis.

The runtime for likeLTD was between 25 and 30 minutes, while EuroForMix and STRmix once again required less than a minute for computation.

6.6.3 Conclusions

The three models return similar results for all evaluations, with the largest difference between the models being 1.6 bans for the Knox evaluation with t=50, indicating that all three models are likely to be valid, with some differences due to divergent modelling choices.

Chapter 7

Utilising the PH model to inform casework practice

Some of the work in this chapter has been published in Steele et al. [2016], see Appendix B for the accepted manuscript. All work was performed by me.

7.1 Motivation

The PH model can be used to inform strategies that might be useful in a casework setting. Here, the PH model will be used to investigate whether or not splitting a sample into multiple replicates enhances the WoE (an extension to Chapter 2), whether explaining away the peaks of a minor unknown contributor as dropin is a valid strategy for reducing computational complexity, and whether assuming a clear major unknown contributor as a known contributor is a valid strategy for reducing computational complexity.

7.2 Efficacy of multi-replicate CSPs for LTDNA samples

In Chapter 2 it was demonstrated that the addition of extra replicates in a CSP increases the WoE against Q if H_p is true, up to the IMP. The methods used in that chapter mimic the protocol of some laboratories. DNA extraction produces a fixed volume of extract, at varying concentrations depending on the case circumstances. PCR has a maximum volume of input solution, which is lower than the volume produced by extraction, so multiple PCR reactions can be performed from the same DNA extract. However, other laboratories perform pre-extraction replicates, in which the evidence item is sampled multiple times, essentially splitting the DNA sample before extraction, reducing the DNA mass in each replicate post extraction. To simulate this second protocol, PCR was performed on the extract of two-contributor and three-contributor validation mixtures

# Cont	Condition	Unsplit DNA mass (pg)	# Samples	# Reps	Split DNA mass (pg)	Approx. cell equiv.
2	Equal	62 (31:31)	4	2	31 (16:16)	5 (3:3)
			4	3	21 (10:10)	4 (2:2)
			4	4	16 (8:8)	3 (1:1)
	Maj/min	266 (250:16)	4	2	133 (125:8)	22 (21:1)
			4	3	89 (83:5)	15 (14:1)
			4	4	67 (63:4)	11 (11:1)
3	Equal	93 (31:31:31)	2	2	47 (16:16:16)	8 (3:3:3)
			2	3	31 (10:10:10)	5 (2:2:2)
			2	4	23 (8:8:8)	4 (1:1:1)
	Unequal	328 (250:62:16)	2	2	164 (125:31:8)	27 (21:5:1)
			2	3	109 (83:21:5)	18 (14:4:1)
			2	4	82 (63:16:4)	14 (11:3:1)

Table 7.1: Experimental design for investigating whether replicates provide extra information over an unsplit DNA sample. DNA masses and cellular equivalents are rounded, and are given as a total contribution, with individual contributions in brackets.

(see Table 6.1), but the extract was split into $2 \leq n \leq 4$ replicates, which simulates a situation of splitting a DNA sample with x pg input DNA into multiple samples with on average x/n pg DNA. The rest of the laboratory protocol was kept consistent with that of the validation data set (see Section 6.2).

Using a definition of low template as DNA mass < 200 pg, all total contributions in each replicate are low-template, and all individual contributions per replicate are low template (Table 7.1). In contrast, the total unsplit DNA mass (both as a single sample, and summed over replicates) is low template only for the equal-contributions CSPs, whereas the unequal-contributions mixtures are good template for the major contributor, and low template for subsequent contributors.

The WoE for the replicate CSPs was evaluated using the PH model, and compared to the WoE obtained for the unreplicated CSPs in Chapter 6. There may be little gain from running multiple replicates, as the PH model should be able to utilise close to the total information available in a single replicate, and because the total DNA contribution is the same between the replicated and unreplicated CSPs. Alternatively, the variability in contributions between replicates may give extra information regarding their genotypes, or replication may provide useful information to overcome peak height variability.

7.2.1 Two-contributor CSPs

Unsplit CSP vs. replicates CSP

The majority of two-contributor cases obtain roughly equal information with a single replicate of x pg DNA (x-axis) or n replicates of x/n pg DNA (y-axis), regardless of the relative contributions of each contributor

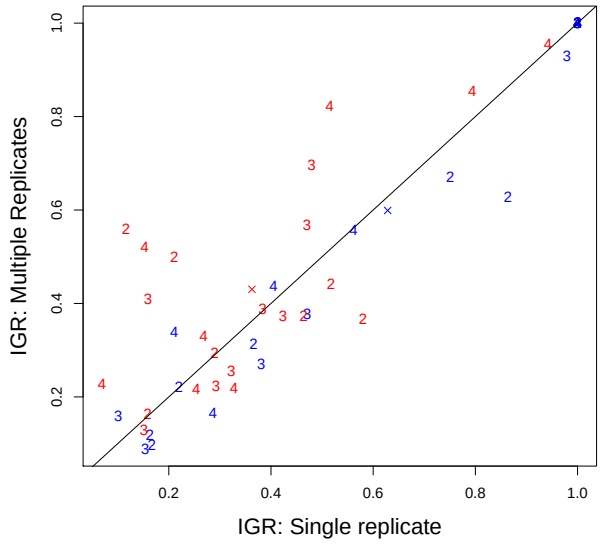


Figure 7.1: Information gain ratio (IGR) for 12 equal-contributions two-contributor CSPs (red) and 12 major/minor two-contributor CSPs (blue) using a single replicate (x-axis) or splitting the sample into n replicates (y-axis).The plotted point numbers indicate how many replicates were run for that CSP. Crosses indicate mean IGRs. Each of the two contributors were queried in turn, leading to the 48 total data points.

(red/blue) or DNA template. In contrast, the multi-replicate CSP for CSP 8 obtains ~ 0.5 greater IGR than the single replicate CSP. This is due to the stochastic nature of peak observation in a particular run; three alleles of Q are observed in the multi-replicate CSP that are not observed in the single-replicate CSP, while two alleles of Q are observed in the single-replicate CSP but not observed in the multi-replicate CSP (Appendix Figure A.6). Of those alleles that are observed only in the multi-replicate CSP, two are rare (p_x=0.03 and 0.05 respectively), so increase the WoE against Q substantially, while the last is relatively common (p_x=0.16). Conversely, the two alleles that are only observed in the single-replicate CSP are common (p_x=0.32 and 0.30 respectively).

The noise seen around the x=y line is likely due to the stochastic sampling of alleles at low template. If mixture generation was performed for the same mixture multiple times, and each mixture was subsequently amplified and analysed, the number of alleles observed in each analysis could reasonably be modelled as being drawn from a truncated Poisson, with some mean proportional to the total DNA contribution in the input mixture, and variance equal to the mean. The noise around the x=y line is analogous to the sampling variance from this hypothetical Poisson distribution; points above the x=y line are those for which more alleles of Q were "sampled" using multiple replicates, while points below the x=y line are those for which more alleles of Q were "sampled" using a single replicate. The effect of this sampling on the WoE is further enhanced

by the population allele probability for the sampled alleles; rare sampled alleles of Q greatly increase WoE, while common sampled alleles of Q only increase the WoE slightly.

Sequential information gain

Here the WoE was evaluated for each replicated CSP with each replicate sequentially added into the CSP, similar to the analysis in Chapter 2. It is expected that with addition of extra replicates, for a true H_p the WoE against Q should increase, mirroring the results in Chapter 2, as individual replicates are very low template here, so subsequent replicates provide information that may be missing in the initial replicate.

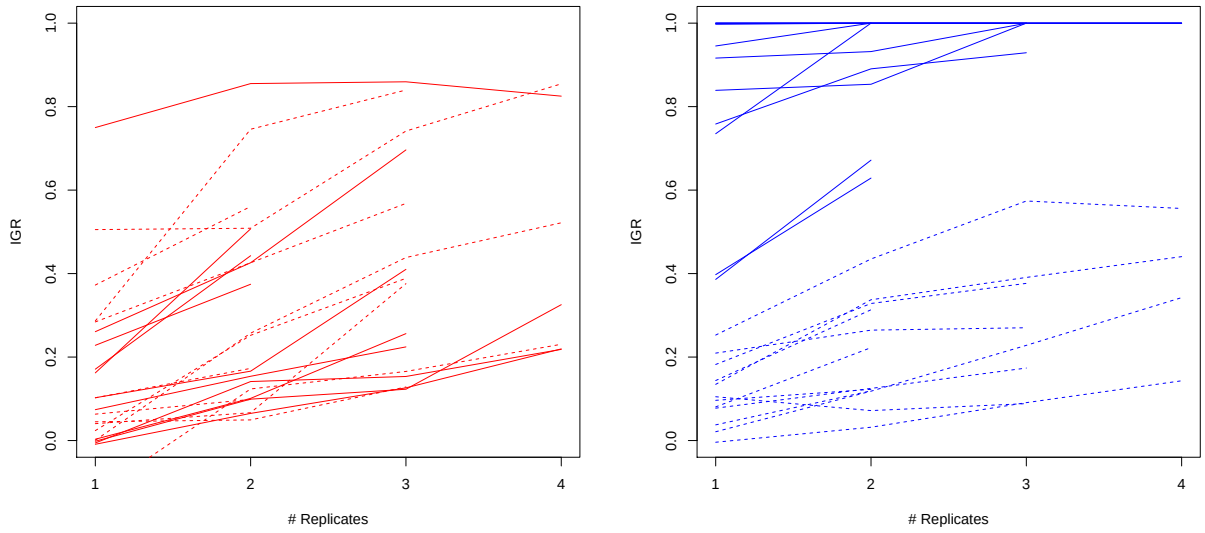


Figure 7.2: Information gain ratio (IGR) for twelve equal contributions (left, red) and twelve major/minor (right, blue) two contributor CSPs with sequential addition of replicates. Dashed or solid lines indicate the queried contributor.

When equal-contributions CSPs are evaluated, the IGR increases with increasing number of replicates, with no effect of the queried contributor (Figure 7.2, left). One exception is seen where the addition of a fourth replicate decreases the IGR. The IGR never reaches 1.0, due to the difficulty of deconvoluting mixtures when the mixture ratio is ≈0.5 (see Section 5.3). A number of CSPs support H_d with a single replicate, but support H_p with additional replicates, which is the ground truth.

When major/minor CSPs are evaluated, the IGR increases with increasing number of replicates (Figure 7.2, right). When querying the major contributor (solid lines), IGR ≈ 1.0 with a single replicate, and IGR=1.0 with a small number of replicates. One exception has IGR≈0.4 at one replicate, and remains

far from IGR=1.0 at two replicates. When querying the minor contributor (dashed lines), the IGR is initially low, increasing towards IGR=1.0 with additional replicates, but never exceeds IGR≈0.4. One CSP supports H_d for the minor Q with a single replicate (IGR<0), but supports H_p with two or more replicates.

7.2.2 Three contributors

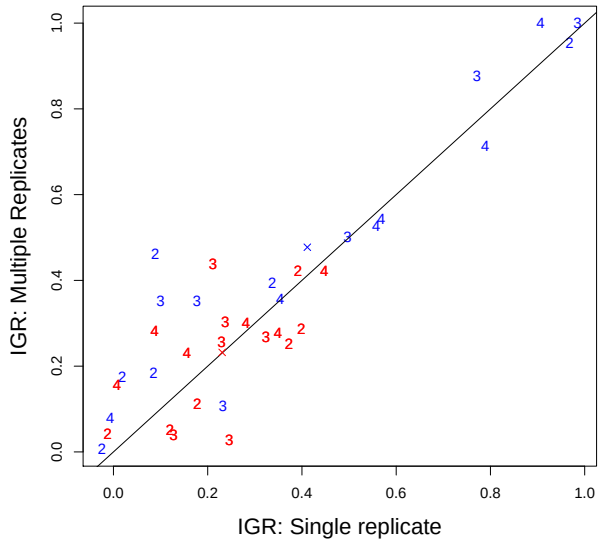


Figure 7.3: Information gain ratio (IGR) for six equal-contributions (red) and six unequal-contributions (blue) three-contributor CSPs using a single replicate (x-axis) or splitting the sample into n replicates (y-axis). The plotted point numbers indicate how many replicates were run for that CSP. Crosses indicate mean IGRs. Each of the three contributors were queried in turn, giving 36 total evaluations.

Similar to the two-contributor replicate results, the three-contributor IGRs are spread around the x=y line for both equal- and unequal-contributions mixtures (Figure 7.3), confirming that splitting an x pg DNA sample into n x/n pg DNA replicates provides no extra information over running the sample as a single replicate x pg CSP. The largest deviation from the x=y line is for a two-replicate CSP that corresponds to the single-replicate unequal-contributions CSP that had 13 whole locus dropouts, and returned a WoE of 1.9, 0.4, and -0.5 bans for the 250, 62 and 16 pg contributors respectively for the single-replicate CSP, which all increased to 10.0, 3.6, and 0.1 bans respectively for the multiple-replicates CSP.

7.2.3 Implications for casework

The results presented here (Figures 7.1 and 7.3) show little to no effect of pre-extraction replication on the WoE using a continuous model, with mean WoEs lying close to the x=y line. This suggests that pre-extraction replication may not be worth the cost. Post-extraction replication increases the WoE towards the IMP with additional replicates (Figure 7.2), so is desirable.

The difference between the IGRs for the three-contributor CSP that had 13 whole locus dropout in the single replicate scenario appears to provide the strongest support for performing multiple replicates; variation is inherent in LTDNA analysis, and includes situations of total failure. Replication minimises this risk as a total failure of a single replicate CSP retains 0% of the potential information, whereas a total failure of one replicate out of n retains some proportion of the potential information in the CSP.

There may be a DNA mass cutoff at which replication cannot be advised, but the DNA mass is unknown before DNA extraction, so in practice the decision to split a sample into replicates cannot be an informed one. The decision is a risk reward analysis, where if splitting a sample is expected to increase the chance of a per-peak dropout probability by less than $1/n$ then the sample should be split.

Using a discrete model, LRmix, Benschop et al. [2015] found that splitting a sample (100:200-600 pg mixture) into four PCR replicates (4 $\times$ 25:50-150 pg mixture) considerably decreased the WoE for the minor contributor, supporting H_d in 90% of evaluations where previously H_p was supported in 100% of evaluations. The WoE for the major contributor was increased for the majority of evaluations. This suggests that discrete models have usable information to gain through pre-extraction replication, as seen for the major contributors. This may be a result of differential dropout rates for contributors across replicates, enhancing deconvolution. However, discrete models lose much information about minor contributors when replicating, as many minor peaks dropout. The PH model, in contrast, gains little to no information for either contributor through replication; full information about the genotype of the major is available from a single replicate, and a low detection threshold enabled by a PH model reduces the information loss for a minor contributor when splitting into replicates. The mixtures used in Benschop et al. had larger DNA contributions for the minor than the mixtures presented here, so may be expected to perform better both with and without splitting. However, direct comparison is made difficult because each typing kit has a different sensitivity.

Condition	Ground Truth	H: Minor=dropin	H: Minor=U
Two contributor	250pg + 16pg	Q/X (250pg) + dropin	Q/X (250pg) + U1
Three contributor	250pg + 62pg + 16pg	Q/X (250pg) + U1 + dropin	Q/X (250pg) + U1 + U2
		Q/X (62pg) + U1 + dropin	Q/X (62pg) + U1 + U2
Contamination	250pg + U (contaminant)	Q/X (250pg) + dropin	Q/X (250pg) + U1
	62pg + U (contaminant)	Q/X (62pg) + dropin	Q/X (62pg) + U1

Table 7.2: Ground truth and hypothesis pairs evaluated when assuming the minor contributor as dropin, or as an unknown contributor.

7.3 Modelling minor contributors as dropin

To reduce the computational complexity of running the PH model, it may be a valid strategy to model minor unknown contributors as dropin. If this strategy is employed, many of the smallest peaks in the CSP will be assigned by the model as dropin peaks, while large peaks will be assigned to one of the assumed contributors. This strategy is not available using a discrete model, as the program has no information available on which peaks to treat as allelic and which to treat as dropins. To test this possibility, the two-contributor (see Section 6.2.4), three-contributor (see Section 6.2.6) and contamination (see Section 6.2.3) CSPs were evaluated with hypothesis pairs given in Table 7.3.

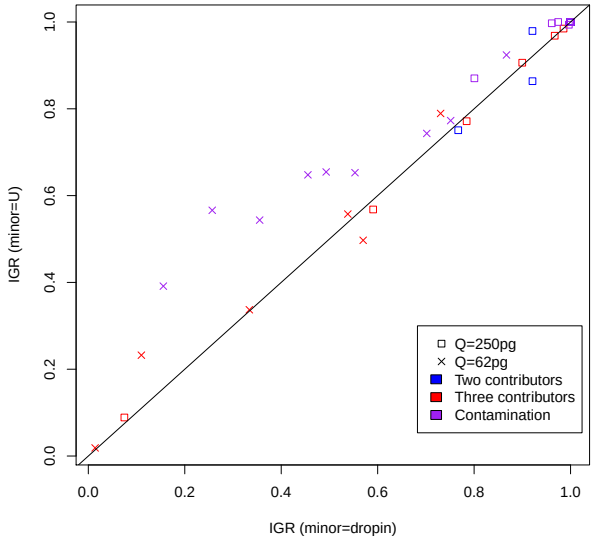


Figure 7.4: Information gain ratio (IGR) for 12 major/minor two contributor CSPs (blue), six unequal contribution three contributor CSPs (red) and 18 contaminated "single contributor" CSPs (blue) assuming the minor as dropin (x-axis) or an unknown contributor (y-axis). Symbols give the DNA contribution of the queried individual.

Assuming the minor contributor as dropin gives approximately equal IGRs with an unknown minor or dropin minor for the majority of CSPs evaluated (Figure 7.4), with exceptions for the contamination cases when the major contributor contributes approximately 62 pg DNA. For these contamination cases, the minor contributor has an unknown amount of DNA, and may not be entirely distinguishable from the major contributor. In this case some of the peaks of Q may be assigned as dropins under H_d if they have a lower PH than some of the contaminant peaks; H_d then fits the data better than H_p is able to, reducing the IGR (Figure 7.4). This has been confirmed by visual inspection, and can be seen in the estimated DNA contributions of the hypothesised contributors when using the full hypotheses of $Q/X + U$, with mean $\widehat{\chi_Q} - \widehat{\chi_U}$ of 91 RFU for the 62 pg cases with dropin IGR<0.6 and 344 RFU for the 62 pg cases with dropin IGR>0.6. Two cases estimate $\widehat{\chi_Q} < \widehat{\chi_U}$ indicating that the contaminant may contribute more DNA to the CSP than Q, so many alleles of Q will be designated as dropins under H_d.

These results suggest that any contributors to a CSP that are represented with substantially less DNA than Q can be treated as dropins. If Q is a minor contributor it is not possible to assume other contributors as dropins, because under H_d many alleles of Q will be assigned as dropin, erroneously reducing the WoE against Q.

7.4 Assuming a major contributor as known

A further potential strategy to reduce computational complexity is to manually deconvolute the genotype of a major unknown contributor, and to assume the deconvoluted genotype as a known contributor. Here the major contributor will be assumed as a known contributor, even if they cannot be clearly distinguished.

When the two contributors can be clearly distinguished, assuming the major contributor as known returns approximately the same IGR as assuming the major as an unknown contributor (Figure 7.5, points along x=y line), so including a clearly-distinguishable major contributor as known is a valid strategy for reducing computational complexity.

When the two contributors are not clearly distinguishable, assuming the "major" contributor as known here simulates a situation where the "major" contributor is believed to be a contributor due to case circumstances, rather than by deconvoluting the major genotype from the CSP. In this situation, assuming a contributor as known increases the WoE against Q if both Q and K are true contributors to the CSP (Figure 7.5, points below x=y line).

Two of the contamination cases give counter-intuitive results; a 62 pg contamination case lies close to the x=y line, while a 250 pg contamination case is clustered with the remaining 62 pg cases. This

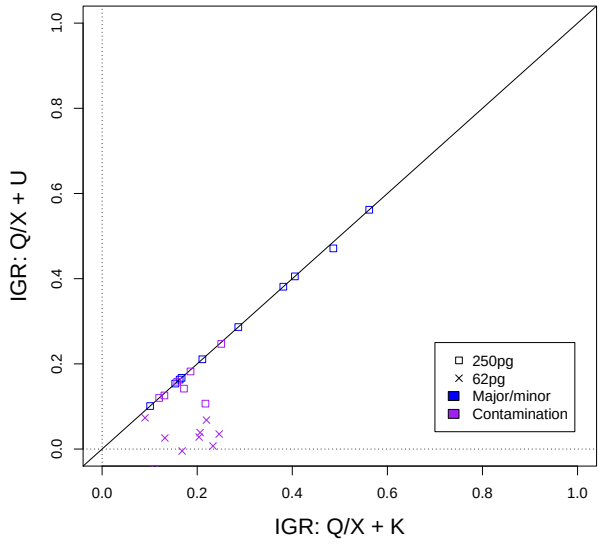


Figure 7.5: Information gain ratio (IGR) for 12 major/minor two contributor CSPs (blue) and 18 contaminated "single contributor" CSPs (purple) assuming the major contributor as a known contributor (x-axis) or as an unknown contributor (y-axis). Symbols give the DNA contribution of the major contributor. The minor contributor was queried.

suggests that the two contributors can be clearly distinguished for the 62 pg case, but cannot for the 250 pg case. This is supported by the DNA contribution estimates of each case; the 62 pg case gives $\widehat{\chi_p} = 865, 98$, $\widehat{\chi_d} = 772, 184$, and the 250 pg case gives $\widehat{\chi_p} = 654, 110$, $\widehat{\chi_d} = 597, 162$. Thus the 250 pg case has closer DNA contribution estimates than the 62 pg case. This is likely due to pipetting errors, where the 250 pg contributor was sampled at lower than 250 pg, and the 62 pg contributor was sampled at greater than 62 pg.

These results highlight that assuming a clearly distinguishable contributor as known is a valid strategy for reducing computational complexity that does not affect the WoE. However, assuming a non-clearly distinguishable unknown contributor as known will unduly favour the prosecution, but including a non-clearly distinguishable contributor as known due to case circumstances will enhance the WoE against a true Q.

Chapter 8

Conclusions

8.1 F_{ST} recommendations

Chapter 3 provides estimates of F_{ST} for worldwide subpopulations relative to continental populations, with recommendations of an F_{ST} value for forensic use of between 2% and 3% for common populations, and perhaps as high as 5% for isolated populations. Chapter 4 further clarifies that the choice of an appropriate F_{ST} value can account for the fact that the appropriate database for an unknown contributor may be unknown, and so may be misassigned, or for the fact that there may not be an appropriate database for a given Q. This leads to the recommendation of 3% F_{ST} for all common populations, to allow for these effects.

8.2 Population genetics

In the process of generating estimates of F_{ST} (Chapter 3) some interesting inferences regarding the population genetics of both global populations and forensic databases were obtained (Chapter 3). The observation of "fringe" subpopulations, that fit almost equally well in multiple populations, reiterates that human allele probabilities change smoothly with geographical distance [Ramachandran et al., 2005], and that designation of subpopulations into larger populations is subjective and somewhat arbitrary. Continental-scale estimates of F_{ST} recapitulate the "out of Africa" theory of the origin of humans due to increasing indirect estimates of F_{ST} with increasing distance from Africa, as well as demonstrating, along with a lack of fringe subpopulations, that East Asia is more genetically distinct from other Old World populations than is typical. Estimates of subpopulation F_{ST} relative to forensic databases suggests that some subpopulations are highly represented in some databases; Jamaicans in the FSS EA3 database, Pakistanis in the EA4 database and Chinese in the EA5 database.

8.3 Forensic databases

Chapter 4 demonstrated that, with an appropriate value of F_{ST}, assuming that all unknown contributors are drawn from the population most relevant to Q is a valid heuristic rule for simplifying the computation of forensic LRs, which does not unduly favour the prosecution, even if the population of Q has been misassigned or Q does not fit their most appropriate database well.

8.4 Use of replicates

8.4.1 Pre-extraction

Chapter 7 demonstrated that pre-extraction splitting of a sample into multiple replicates does not increase the WoE against a true Q, because the total amount of DNA in a single replicate with x pg of DNA, and n replicates with x/n pg DNA are approximately equal, with some noise originating from the stochastic nature of allele observation at low template. This work, together with the post-extraction investigation, suggests that primary samples should be taken with a view to maximising the amount of DNA present in the sample, therefore maximising the information available in the subsequent CSPs, and any replication should be performed post-extraction to reduce the risk of large scale genotyping failure at low DNA levels.

8.4.2 Post-extraction

Chapters 2 and 7 further demonstrated that the sequential inclusion of extra post-extraction replicates in a CSP increases the WoE towards the IMP when Q is a contributor, for both a discrete model (Chapter 2) and a continuous model (Chapter 7). This work suggests that post-extraction replicates should be performed, particularly to guard against the possibility of whole-profile failure, or extensive locus dropout, that are both risks encountered at very low template. In addition, the discrete model is able to exceed the mixLR, the LR from a high quality mixed sample, with just a few replicates, indicating that multiple low-template replicates can provide more information than a single good-template sample through differential dropout rates.

8.4.3 As validation

Using multiple replicates to validate the behaviour of a forensic likelihood implementation was explicitly demonstrated for the discrete model in Chapter 2, and implicitly demonstrated for the PH model in Chapter 7, where it was observed that with increasing numbers of replicates the WoE tends towards the IMP but does

not exceed the IMP, as predicted by theory. Therefore both the PH and discrete models have been validated through their behaviour with multiple replicates, suggesting that this behaviour is useful for validating any implementation of this class of model for evaluation of forensic STR LRs.

8.5 PH model

A PH model was developed in Chapter 5, which was subsequently validated in Chapter 6, and utilised in Chapter 7 to investigate some modelling assumptions that can be employed to decrease computational complexity in some situations.

8.5.1 Validation

Validation of the model showed that it behaves as expected in relation to the discrete model for laboratory-generated CSPs ranging from one to three contributors, that the WoE behaves as expected when the input data is altered artificially and when the model assumptions are altered. Additionally the model returns similar WoEs to other continuous models. The breadth and depth of tests presented throughout Chapter 6 constitute an extensive and thorough validation of the PH model for use in forensic casework.

8.5.2 Uses

Including a major unknown contributor as a known contributor, and modelling minor unknown non-Q peaks as dropin are both strategies for decreasing computational complexity that are valid when a major and/or minor contributor can be clearly distinguished from any other contributor in the CSP. Currently, modelling a major unknown contributor as a known contributor is standard practice when they are clearly distinguishable. Modelling minor peaks as dropin is not currently practiced, but is analogous to employing a high detection threshold to remove minor peaks that are not of interest to the court.

8.6 Limitations

The first limitation in the thesis is the low sample sizes of some populations and subpopulations in Chapter 3. This is most notable for IC6 which had many subpopulations with sample size < 30. However, it was a limitation for all populations, as ideally all national subpopulations would have a sufficient sample size to avoid combining into regional subpopulations e.g. Germany into Western Europe.

Dropout probabilities when simulating profiles in Chapter 2 did not vary with the length of an allele in base pairs, which would simulate the effects of degradation. This may have contributed to the slightly different behaviour observed for the laboratory and simulated CSPs. Instead, it would be desirable to run multiple simulations, with varying degrees of degradation.

A slight limitation of the proposed PH model is the runtime, taking longer to run than other available software. However, some of the modelled phenomena in the PH model are not modelled in other packages, and are important for fully explaining common observations in CSPs.

8.7 Further work

8.7.1 Baseline

The PH model outlined in Chapter 5 employs a detection threshold, below which an allele is deemed to have dropped out, as do all other available models except TrueAllele. To be able to utilise the full information in a CSP that is below the current detection threshold it may be desirable to remove the detection threshold entirely. Currently, this has not been implemented in the PH model, and would require extensive work on determining which genotype allocations to consider so that silent alleles or total dropouts (RFU=0, rather than RFU$< t$) are able to be handled by the model. As is the case for all forensics software, this would require extensive validation tests, similar to those in Chapter 6, but designed specifically to challenge a baseline model.

8.7.2 Single-nucleotide polymorphism (SNP) WoE

SNP panels for ancestry prediction [Yang et al., 2005, Phillips et al., 2007, Halder et al., 2008, Jia et al., 2014], or trait prediction such as eye [Walsh et al., 2011, 2013], hair [Branicki et al., 2011] and skin [Myles et al., 2007, Beleza et al., 2013] pigmentation, are beginning to be utilised to inform forensic investigation while no suspect is known to the police. Currently these data are not then utilised to generate a WoE against any subsequent suspect. A model to combine the WoE from both STRs and SNPs would utilise the full genetic data for identification in such cases, without necessarily utilising the trait prediction based on the SNP data. While some models have been created to generated a WoE from SNP panels, they are not incorporated into models for STR WoEs, so this would be an avenue for further extensions to likeLTD.

8.7.3 Sequencing WoE

In conjunction with generating SNP data, sequencing chips are now available that sequence both the SNP panels and the forensic STR sets together. This provides the possibility of being able to distinguish between microvariants of STRs, so $[ATTC]_5$ could be distinguished from $[ATTC]_2ATGC[ATTC]_2$, where current STR typing through capillary electrophoresis would be unable to distinguish such a difference. Sequencing STRs is still in its infancy, with no database of population frequencies of different microvariants, and indeed no agreement on a naming convention for such microvariants. However, as the sequencing of STRs matures, a model for WoE of sequenced STRs would be necessary to present the evidence in court, and would be a considerable extension to likeLTD.

Bibliography

D. J. Balding. Estimating products in forensic identification using DNA profiles. *Journal of the American Statistical Association*, 90(431):839–844, 1995.

D. J. Balding. Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, 63(3):221 – 230, 2003.

D. J. Balding. *Weight-of-evidence for Forensic DNA Profiles*. Wiley, 2005.

D. J. Balding. Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proceedings of the National Academy of Sciences*, 110(30):12241–12246, 2013.

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

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

D. J. Balding and R. A. Nichols. Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity*, 78(6):583–589, 1997.

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

J. Ballantyne, E. K. Hanson, and M. W. Perlin. DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information. *Science & Justice*, 53(2):103–14, 2013. ISSN 1355-0306. doi: 10.1016/j.scijus.2012.04.004.

M. A. Beaumont and D. J. Balding. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13(4):969–980, 2004.

S. Beleza, N. A. Johnson, S. I. Candille, D. M. Absher, and e. a. Coram M. A. Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *PLoS Genetics*, 9(3):e1003372, 2013.

J. Benn-Torres, C. Bonilla, C. M. Robbins, L. Waterman, T. Y. Moses, W. Hernandez, E. R. Santos, F. Bennett, W. Aiken, T. Tullock, et al. Admixture and population stratification in African Caribbean populations. *Annals of Human Genetics*, 72(1):90–98, 2008.

C. C. G. Benschop, S. Y. Yoo, and T. Sijen. Split DNA over replicates or perform one amplification? *Forensic Science International: Genetics Supplement Series*, 5:e532–e533, 2015.

D. Bentley and P. Lownds. Low template DNA. *Archbold Review*, 1, 2011.

G. Bhatia, N. Patterson, S. Sankararaman, and A. L. Price. Estimating and interpreting Fst: The impact of rare variants. *Genome Research*, 23:1514–1521, 2013.

Ø. Bleka, G. Storvik, and P. Gill. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. *Forensic Science International: Genetics*, 21:35–44, 2016.

W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pośpiech, S. Walsh, T. Kupiec, A. Wojas-Pelc, and M. Kayser. Model-based prediction of human hair color using DNA variants. *Human Genetics*, 129(4): 443–454, 2011.

J.-A. Bright, J. M. Curran, and J. S. Buckleton. Relatedness calculations for linked loci incorporating subpopulation effects. *Forensic Science International: Genetics*, 7(3):380–383, 2013a.

J.-A. Bright, D. Taylor, J. M. Curran, and J. S. Buckleton. Degradation of forensic DNA profiles. *Australian Journal of Forensic Sciences*, 45(4):445–449, 2013b.

J.-A. Bright, D. Taylor, J. M. Curran, and J. S. Buckleton. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Science International: Genetics*, 7(2):296–304, 2013c.

J.-A. Bright, I. W. Evett, D. Taylor, J. M. Curran, and J. Buckleton. A series of recommended tests when validating probabilistic DNA profile interpretation software. *Forensic Science International: Genetics*, 14: 125–131, 2015.

C. Brookes, J.-A. Bright, S. Harbison, and J. Buckleton. Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, 6(1):58–63, 2012.

J. Buckleton and J. Curran. A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Science International: Genetics*, 2(4):343–348, 2008.

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

B. Budowle, A. J. Eisenberg, and A. van Daal. Validity of low copy number typing and applications to forensic science. *Croatian Medical Journal*, 50(3):207–217, 2009. doi: 10.3325/cmj.2009.50.207.

J. M. Butler, E. Buel, F. Crivellente, and B. R. McCord. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis*, 25(1011):1397–1412, 2004.

B. Caddy, G. R. Taylor, and A. M. T. Linacre. A review of the science of low template DNA analysis. *Office of the Forensic Regulator*, pages 23–24, 2008.

Caribbean Community Capacity Development Programme. National Census Report 2001, Jamaica. Online, 2009. URL `http://www.caricomstats.org/Files/Publications/NCR%20Reports/Jamaica.pdf`.

H. P.-A. S. Consortium et al. Mapping human genetic diversity in Asia. *Science*, 326(5959):1541–1545, 2009.

E. A. Cotton, R. F. Allsop, J. L. Guest, R. R. E. Frazier, P. Koumi, I. P. Callow, A. Seager, and R. L. Sparkes. Validation of the AMPFlSTR® SGM Plus system for use in forensic casework. *Forensic Science International*, 112(2):151–161, 2000.

R. Cowell. Combining allele frequency uncertainty and population substructure corrections in forensic DNA calculations. *Forensic Science International: Genetics*, 23:210–216, 2016.

R. G. Cowell, T. Graversen, S. L. Lauritzen, and J. Mortera. Analysis of forensic DNA mixtures with artefacts. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 64(1):1–48, 2015.

J. M. Curran, P. Gill, and M. R. Bill. Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure. *Forensic Science International*, 148(1):47, 2005.

T. M. Diegoli, M. Farr, C. Cromartie, M. D. Coble, and T. W. Bille. An optimized protocol for forensic application of the PreCR Repair Mix to multiplex STR amplification of UV-damaged DNA. *Forensic Science International: Genetics*, 6(4):498–503, 2012.

B. M. Dupuy, M. Stenersen, T. Egeland, and B. Olaisen. Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Human Mutation*, 23(2):117–124, 2004.

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

I. W. Evett, C. Buffery, G. Willott, and D. Stoney. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. *Journal of the Forensic Science Society*, 31(1):41–47, 1991.

EWCA. R v. Reed and Reed; R v. Garmson, 2009.

EWCA. R v. Broughton, 2010.

L. Excoffier and G. Hamilton. Comment on Genetic structure of human populations.. *Science*, 300(5627): 1877, 2003.

L. A. Foreman and I. W. Evett. Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system. *International Journal of Legal Medicine*, 114(3):147–155, 2001.

L. A. Foreman, J. A. Lambert, and I. W. Evett. Regional genetic variation in Caucasians. *Forensic Science International*, 95(1):27–37, 1998.

L. Forster, J. Thomson, and S. Kutranov. Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle low copy number(LCN) method for analysis of trace forensic DNA samples. *Forensic Science International: Genetics*, 2(4):318–328, 2008. doi: 10.1016/j.fsigen.2008.04.005.

W. K. Fung and Y.-Q. Hu. *Statistical DNA forensics: theory, methods and computation*. John Wiley & Sons, 2008.

P. Gill and H. Haned. A new methodological framework to interpret complex DNA profiles using likelihood ratios. *Forensic Science International: Genetics*, 7(2):251–263, 2013.

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

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

P. Gill, J. Curran, and K. Elliot. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research*, 33(2):632–643, 2005.

P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayr, N. Morling, M. Prinz, P. M. Schneider, and B. S. Weir. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, 160(2):90–101, 2006.

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

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

P. Gill, L. Gusmão, H. Haned, W. R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. M. Schneider, and B. S. Weir. 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 Science International: Genetics*, 2012. doi: 10.1016/j.fsigen.2012.06.002.

P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, and T. Egeland. Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches - Twenty years of research and development. *Forensic Science International: Genetics*, 2015.

I. J. Good. Probability and the Weighing of Evidence, 1950.

I. J. Good. Studies in the history of probability and statistics. XXXVII AM Turing's statistical work in World War II. *Biometrika*, pages 393–396, 1979.

T. Graversen and S. Lauritzen. Computational aspects of DNA mixture analysis. *Statistics and Computing*, 25(3):527–541, 2014.

K. S. Grisedale and A. van Daal. Comparison of STR profiling from low template DNA extracts with and without the consensus profiling method. *Investigative Genetics*, 3:14, 2012. doi: 10.1186/2041-2223-3-14.

I. Halder, M. Shriver, M. Thomas, J. R. Fernandez, and T. Frudakis. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Human Mutation*, 29(5):648–658, 2008.

International Organisation for Migration. Jamaica Mapping Exercise. Online, July 2007. URL `http://www.iomlondon.org/doc/mapping/IOM_JAMAICA.pdf`.

J. Jia, Y.-L. Wei, C.-J. Qin, L. Hu, L.-H. Wan, and C.-X. Li. Developing a novel panel of genome-wide ancestry informative markers for bio-geographical ancestry estimates. *Forensic Science International: Genetics*, 8(1):187–194, 2014.

K. Kadash, B. E. Kozlowski, L. A. Biega, and B. W. Duceman. Validation study of the TrueAllele automated data review system. *Journal of Forensic Sciences*, 49(4):660–667, 2004.

H. Kelly, J.-A. Bright, J. M. Curran, and J. Buckleton. Modelling heterozygote balance in forensic DNA profiles. *Forensic Science International: Genetics*, 6(6):729–734, 2012.

H. Kelly, J.-A. Bright, J. S. Buckleton, and J. M. Curran. Identifying and modelling the drivers of stutter in forensic DNA profiles. *Australian Journal of Forensic Sciences*, 46(2):194–203, 2014.

D. Lu, Q. Liu, W. Wu, and H. Zhao. Mutation analysis of 24 short tandem repeats in Chinese Han population. *International Journal of Legal Medicine*, 126(2):331–335, 2012.

B. McCord, K. Opel, M. Funes, S. Zoppis, and L. Meadows Jantz. An investigation of the effect of DNA degradation and inhibition on PCR amplification of single source and mixed forensic samples. *US Department of Justice*, pages 1–66, 2011.

B. R. McCord, J. M. Jung, and E. A. Holleran. High resolution capillary electrophoresis of forensic DNA using a non-gel sieving buffer. *Journal of Liquid Chromatography & Related Technologies*, 16(9-10):1963–1981, 1993.

G. Meakin and A. Jamieson. DNA transfer: review and implications for casework. *Forensic Science International: Genetics*, 7(4):434–443, 2013.

M. Mikkelsen, L. Fendt, A. W. Röck, B. Zimmermann, E. Rockenbauer, A. J. Hansen, W. Parson, and N. Morling. Forensic and phylogeographic characterisation of mtDNA lineages from Somalia. *International Journal of Legal Medicine*, 126(4):573–579, 2012.

A. M. Mohamoud. P52 Characteristics of HLA Class I and Class II Antigens of the Somali Population. *Transfusion Medicine*, 16(s1):47–47, 2006.

U. J. Mönich, K. Duffy, M. Médard, V. Cadambe, L. E. Alfonse, and C. Grgicak. Probabilistic characterisation of baseline noise in STR profiles. *Forensic Science International: Genetics*, 19:107–122, 2015.

A. Moreno-Estrada, S. Gravel, F. Zakharia, J. L. McCauley, J. K. Byrnes, C. R. Gignoux, P. A. Ortiz-Tello, R. J. Martínez, D. J. Hedges, R. W. Morris, et al. Reconstructing the population genetic history of the Caribbean. *PLoS Genetics*, 9(11):e1003925–e1003925, 2013.

J. E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1-2):65–82, 1962.

K. M. Mullen, D. Ardia, D. L. Gil, D. Windover, J. Cline, et al. DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software*, 40(6):1–26, 2011.

S. Myles, M. Somel, K. Tang, J. Kelso, and M. Stoneking. Identifying genes underlying skin pigmentation differences among human populations. *Human Genetics*, 120(5):613–621, 2007.

S. Nathakarnkitkool, P. J. Oefner, G. Bartsch, M. A. Chin, and G. K. Bonn. High-resolution capillary electrophoretic analysis of DNA in free solution. *Electrophoresis*, 13(1):18–31, 1992.

National Research Council. *The evaluation of forensic DNA evidence*. Washington DC: National Academies Press, 1996.

M. Naughton and G. Tan. The need for caution in the use of DNA evidence to avoid convicting the innocent. *The International Journal of Evidence & Proof*, 15(3):245–257, 2011.

M. Nelis, T. Esko, R. Mägi, F. Zimprich, A. Zimprich, D. Toncheva, S. Karachanak, T. Piskáčková, I. Balaščák, L. Peltonen, E. Jakkula, K. Rehnström, M. Lathrop, S. Heath, P. Galan, S. Schreiber, T. Meitinger, A. Pfeufer, H. Wichmann, B. Melegh, N. Polgár, D. Toniolo, P. Gasparini, P. D'Adamo, J. Klovins, L. Nikitina-Zake, V. Kučinskas, J. Kasnauskienė, J. Lubinski, T. Debniak, S. Limborska, A. Khrunin, X. Estivill, R. Rabionet, S. Marsal, A. Julià, S. Antonarakis, S. Deutsch, C. Borel, H. Attar, M. Gagnebin, M. Macek, M. Krawczak, M. Remm, and M. Metspalu. Genetic structure of Europeans: a view from the North–East. *PloS ONE*, 4(5):e5472, 2009.

NICC. The Queen v. Hoey, 2007. URL `http://www.bailii.org/nie/cases/NICC/2007/49.html`.

Office for National Statistics. 2011 census data (England and Wales) [computer file], 2011.

V. L. Pascali and S. Merigioli. Joint Bayesian analysis of forensic mixtures. *Forensic Science International: Genetics*, 6(6):735–748, 2012.

T. J. Pemberton, M. DeGiorgio, and N. A. Rosenberg. Population structure in a comprehensive genomic data set on human microsatellite variation. *G3: Genes— Genomes— Genetics*, 3(5):891–907, 2013.

M. W. Perlin and A. Sinelnikov. An information gap in DNA evidence interpretation. *PLoS One*, 4(12): e8327, 2009.

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

M. W. Perlin, M. M. Legler, C. E. Spencer, J. L. Smith, W. P. Allan, J. L. Belrose, and B. W. Duceman. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*, 56(6):1430–1447, 2011.

M. W. Perlin, J. L. Belrose, and B. W. Duceman. New York State TrueAllele® Casework validation study. *Journal of Forensic Sciences*, 58(6):1458–1466, 2013.

M. W. Perlin, K. Dormer, J. Hornyak, L. Schiermeier-Wood, and S. Greenspoon. TrueAllele casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. *PloS One*, 9(3):e92837, 2014.

C. M. Pfeifer, R. Klein-Unseld, M. Klintschar, and P. Wiegand. Comparison of different interpretation strategies for low template DNA mixtures. *Forensic Science International: Genetics*, 6(6):716–722, 2012.

C. Phillips, A. Salas, J. J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, M. Calaza, M. C. de Cal, D. Ballard, M. V. Lareu, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics*, 1(3):273–280, 2007.

J. K. Pickrell and J. K. Pritchard. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8(11):e1002967, 2012.

J. K. Pickrell, N. Patterson, P.-R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, and D. Reich. Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 111(7):2632–2637, 2014.

R. Puch-Solis. A dropin peak height model. *Forensic Science International: Genetics*, 11:80–84, 2014.

R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I. Evett, J. Curran, and D. Balding. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Science International: Genetics*, 7(5):555–563, 2013.

S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial

founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–15947, 2005.

A. D. Roeder, P. Elsmore, M. Greenhalgh, and A. McDonald. Maximizing DNA profiling success from sub-optimal quantities of DNA: A staged approach. *Forensic Science International: Genetics*, 3(2):128–137, 2009. doi: 10.1016/j.fsigen.2008.12.004.

A. Ruiz-Linares, K. Adhikari, V. Acuña-Alonzo, M. Quinto-Sanchez, C. Jaramillo, W. Arias, M. Fuentes, M. Pizarro, P. Everardo, F. de Avila, et al. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genetics*, 10(9):e1004572–e1004572, 2014.

M. C. Ruiz-Martinez, O. Salas-Solano, E. Carrilho, L. Kotler, and B. L. Karger. A sample purification method for rugged and high-performance DNA sequencing by capillary electrophoresis using replaceable polymer solutions. A. Development of the cleanup protocol. *Analytical Chemistry*, 70(8):1516–1527, 1998.

F. M. Salzano and M. Sans. Interethnic admixture and the evolution of Latin American populations. *Genetics and Molecular Biology*, 37(1):151–170, 2014.

J. J. Sanchez, C. Hallenberg, C. Børsting, A. Hernandez, and N. Morling. High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. *European Journal of Human Genetics*, 13(7):856–866, 2005.

L. Schneps and C. Colmez. *Math on trial: How numbers get used and abused in the courtroom.* Wiley Online Library, 2013.

R. M. Sibly, A. Meade, N. Boxall, M. J. Wilkinson, D. W. Corne, and J. C. Whittaker. The structure of interrupted human AC microsatellites. *Molecular Biology and Evolution*, 20(3):453–459, 2003.

N. M. Silva, L. Pereira, E. S. Poloni, and M. Currat. Human neutral genetic variation and forensic STR data. *PloS ONE*, 7(11):e49666, 2012.

M. Slatkin. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139 (1):457–462, 1995.

C. D. Steele and D. J. Balding. Choice of population database for forensic DNA profile analysis. *Science & Justice*, 54(6):487–493, 2014a.

C. D. Steele and D. J. Balding. Statistical Evaluation of Forensic DNA Profile Evidence. *Annual Review of Statistics and Its Application*, 1:20–1, 2014b. doi: 10.1146/annurev-statistics-022513-115602.

C. D. Steele, M. Greenhalgh, and D. J. Balding. Verifying likelihoods for low template DNA profiles using multiple replicates. *Forensic Science International: Genetics*, 13:82–89, 2014a.

C. D. Steele, D. Syndercombe Court, and D. J. Balding. Worldwide F_{ST} estimates relative to five continental-scale populations. *Annals of Human Genetics*, 2014b.

C. D. Steele, M. Greenhalgh, and D. J. Balding. Evaluation of low-template DNA profiles using peak heights. *Statistical Applications in Genetics and Molecular Biology*, 2016. doi: 10.1515/sagmb-2016-0038.

P. Taberlet, S. Griffin, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L. P. Waits, and J. Bouvet. Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, 24(16): 3189–3194, 1996.

D. Taylor, J.-A. Bright, and J. Buckleton. The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 7(5):516–528, 2013.

D. Taylor, J. Buckleton, and I. Evett. Testing likelihood ratios produced from complex DNA profiles. *Forensic Science International: Genetics*, 16:165–171, 2015.

D. Taylor, J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut, and J. Buckleton. Validating multiplexes for use in conjunction with modern interpretation strategies. *Forensic Science International: Genetics*, 20: 6–19, 2016.

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

T. Tvedebrink, P. S. Eriksen, M. Asplund, H. S. Mogensen, and N. Morling. Allelic drop-out probabilities estimated by logistic regression - further considerations and practical implementation. *Forensic Science International: Genetics*, 6(2):263–267, 2012.

T. Tvedebrink, P. S. Eriksen, and N. Morling. The multivariate Dirichlet-multinomial distribution and its application in forensic genetics to adjust for subpopulation effects using the θ-correction. *Theoretical Population Biology*, 105:24–32, 2015.

United Nations Statistics Division. Standard country and area codes classifications (m49). Online, Jan. 2014. URL `http://unstats.un.org/unsd/methods/m49/m49regin.htm`.

S. Walsh, F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, and M. Kayser. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Science International: Genetics*, 5(3):170–180, 2011.

S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, and M. Kayser. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics*, 7(1):98–115, 2013.

J. L. Weber and C. Wong. Mutation of human short tandem repeats. *Human Molecular Genetics*, 2(8): 1123–1128, 1993.

B. S. Weir. *Genetic data analysis II. Methods for discrete population genetic data.* Sinauer Associates, Inc. Publishers, 2001.

B. S. Weir. The rarity of DNA profiles. *The Annals of Applied Statistics*, 1(2):358–370, 2007.

B. S. Weir and W. G. Hill. Estimating F–statistics. *Annual Review of Genetics*, 36(1):721–750, 2002.

B. S. Weir, C. M. Triggs, L. Starling, K. A. J. Stowell, and J. Buckleton. Interpreting DNA mixtures. *Journal of Forensic Science*, 42:213–222, 1997.

P. E. Williams, M. A. Marino, S. A. Del Rio, L. A. Turni, and J. M. Devaney. Analysis of DNA restriction fragments and polymerase chain reaction products by capillary electrophoresis. *Journal of Chromatography A*, 680(2):525–540, 1994.

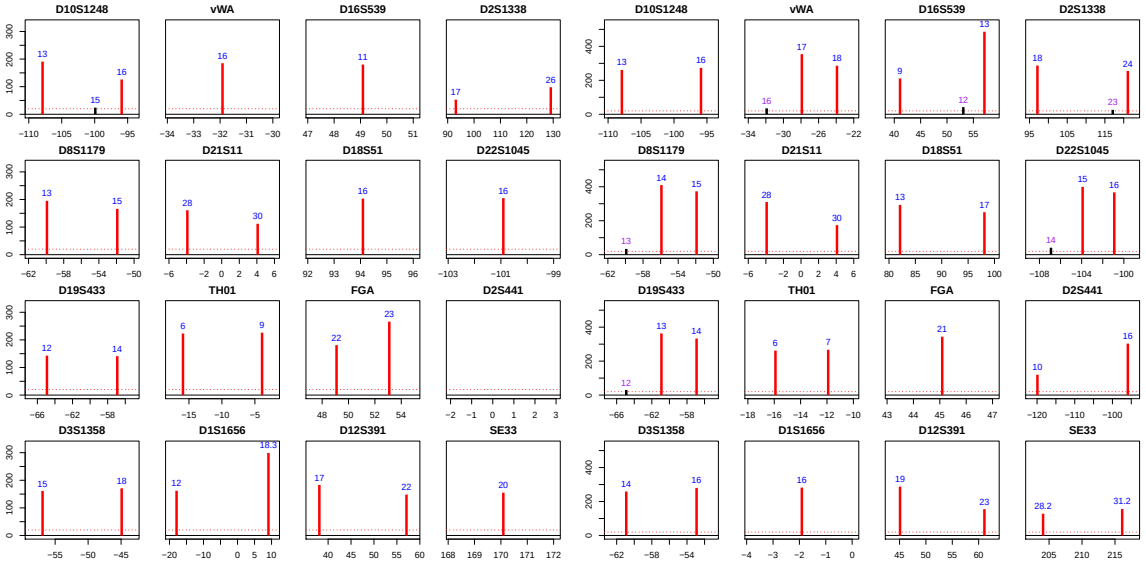
S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, 1949.

X. Xu, M. Peng, Z. Fang, and X. Xu. The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, 24(4):396–399, 2000.

N. Yang, H. Li, L. A. Criswell, P. K. Gregersen, M. E. Alarcon-Riquelme, R. Kittles, R. Shigeta, G. Silva, P. I. Patel, J. W. Belmont, et al. Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Human Genetics*, 118(3-4):382–392, 2005.

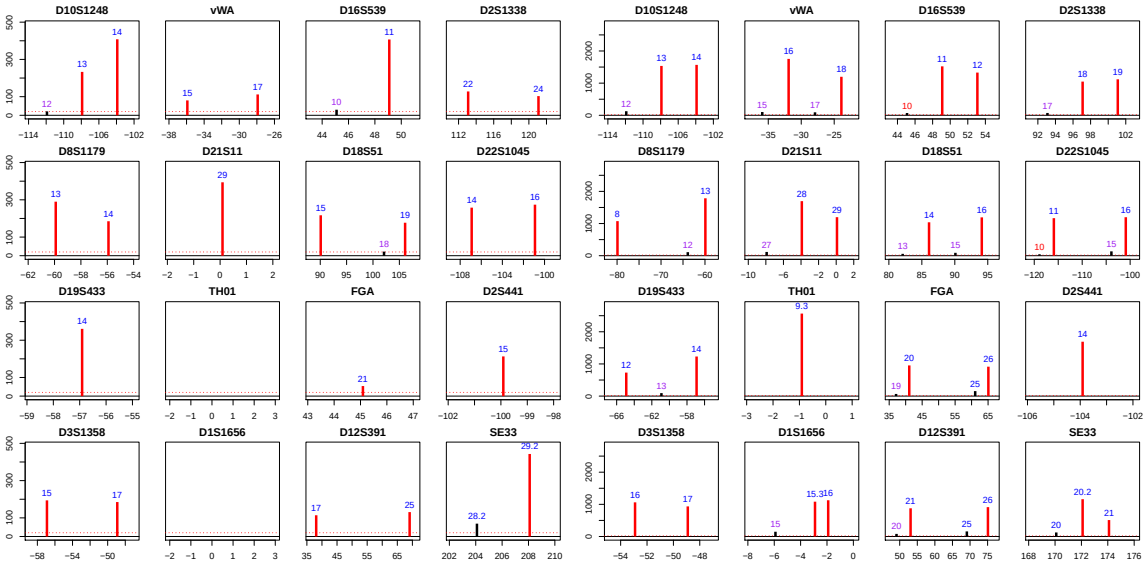
Appendix A

Laboratory CSPs of interest



(a) CSP 22 (16pg)

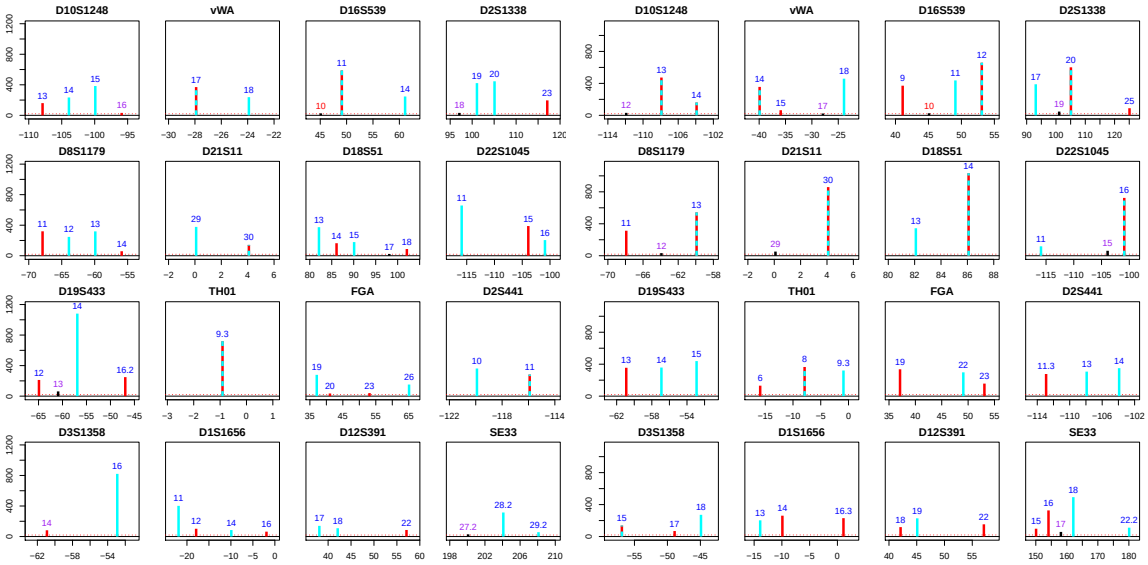
(b) CSP 10 (62pg)



(c) CSP 15 (62pg)

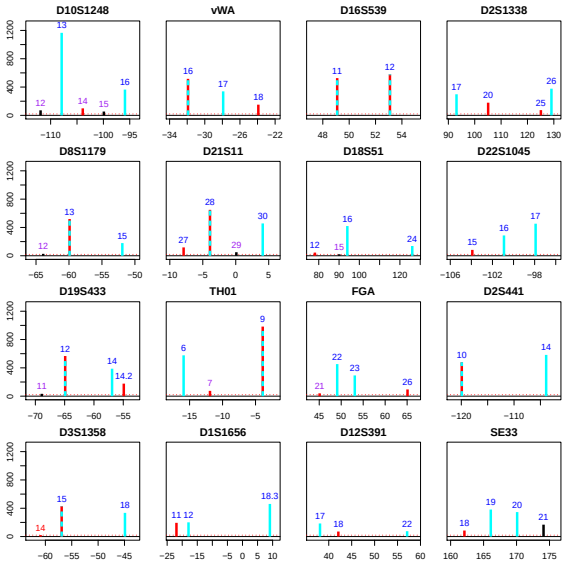
(d) CSP 3 (250pg)

Figure A.1: CSPs for a number of notable single-contributor results; red bars indicate alleles of Q while black bars indicate unattributable peaks. Blue, purple and red labels indicate peaks called as allelic, uncertain and non-allelic respectively for the discrete model CSP.



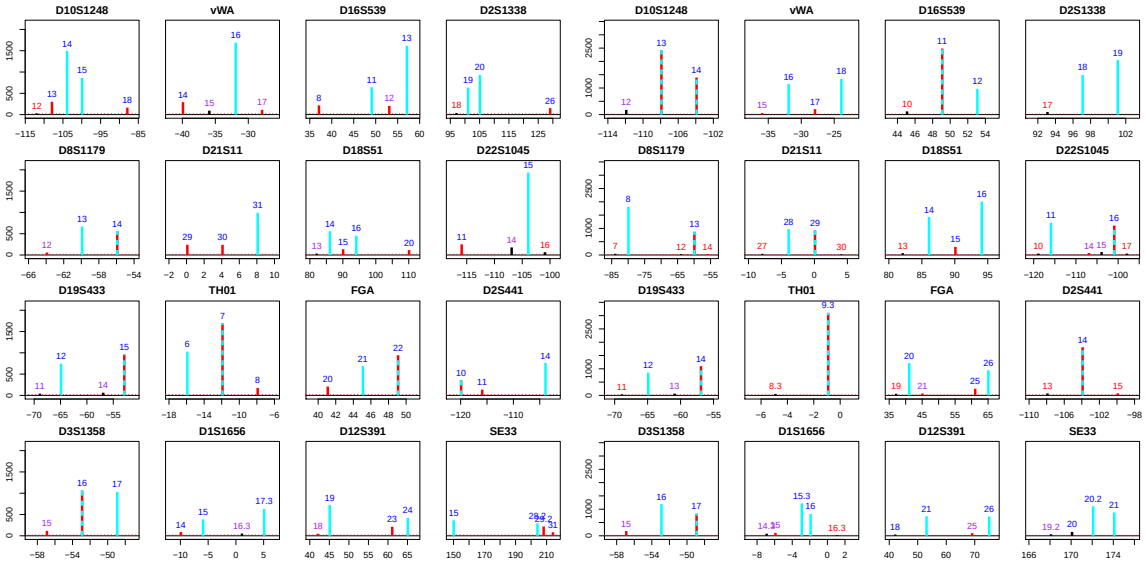
(a) CSP 4

(b) CSP 8



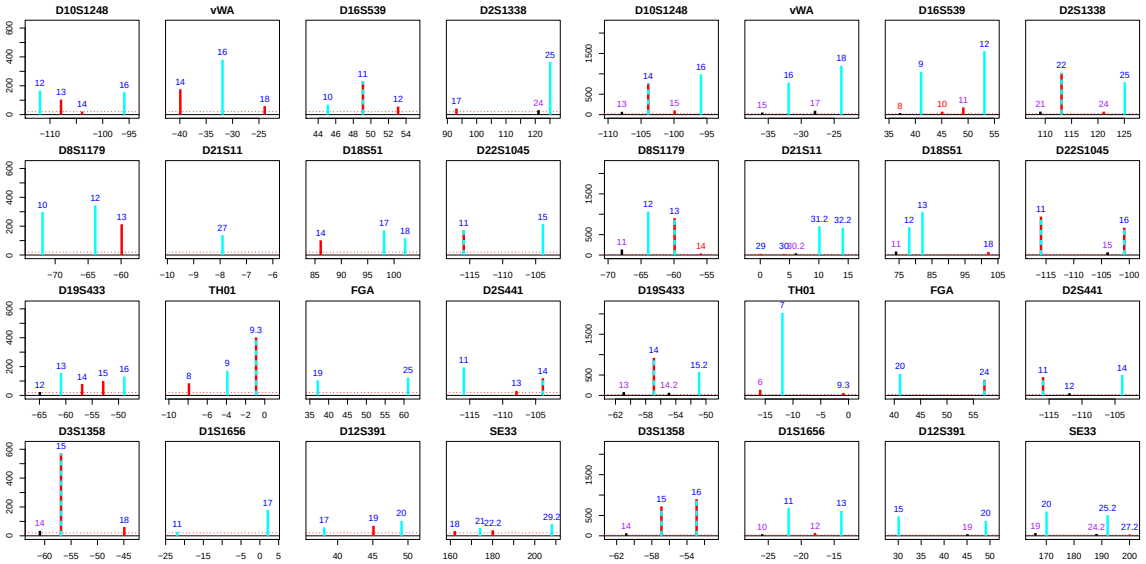
(c) CSP 10

Figure A.2: CSPs for a number of notable two-contributor equal-contribution results; red bars indicate alleles of the first contributor, turquoise bars indicate alleles of the second contributor and black bars indicate unattributable peaks. Blue, purple and red labels indicate peaks called as allelic, uncertain and non-allelic respectively for the discrete model CSP.



(a) CSP 2

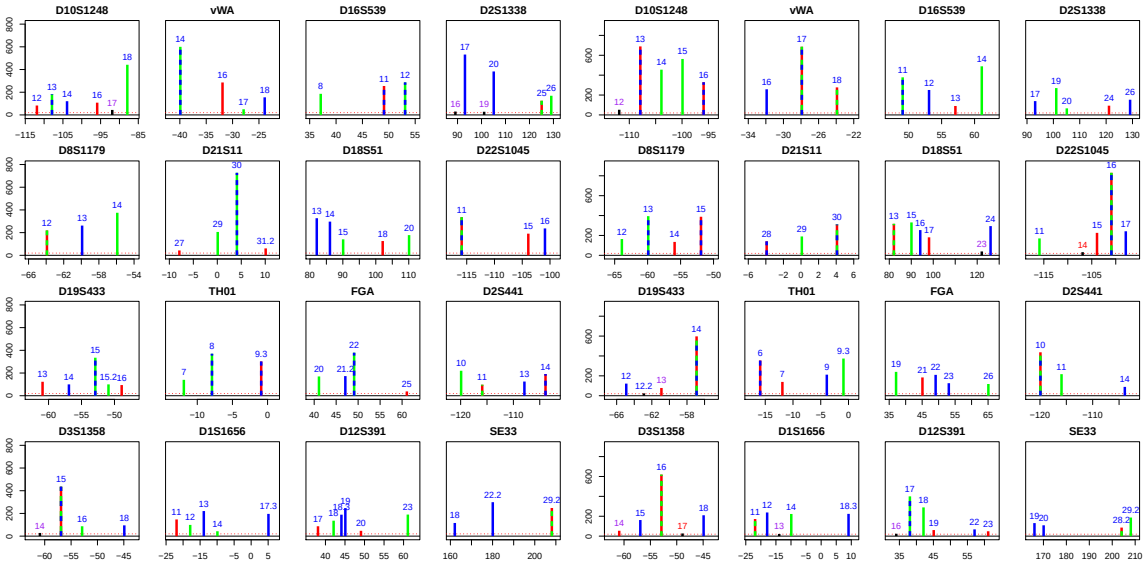
(b) CSP 3



(c) CSP 8

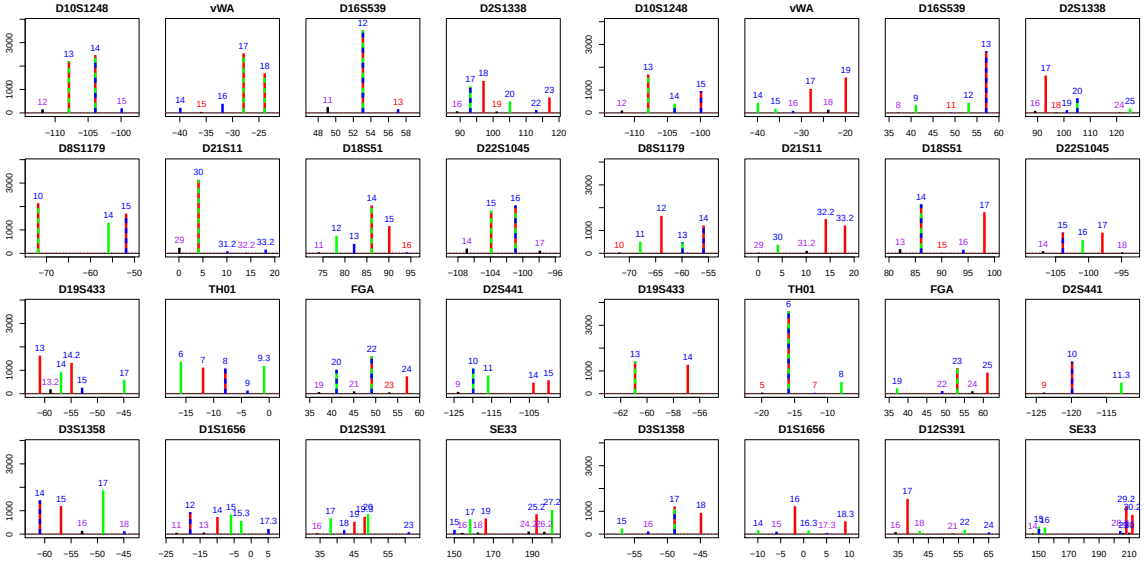
(d) CSP 12

Figure A.3: CSPs for a number of notable two-contributor major/minor results; red bars indicate alleles of the minor contributor, turquoise bars indicate alleles of the major contributor and black bars indicate unattributable peaks. Blue, purple and red labels indicate peaks called as allelic, uncertain and non-allelic respectively for the discrete model CSP.



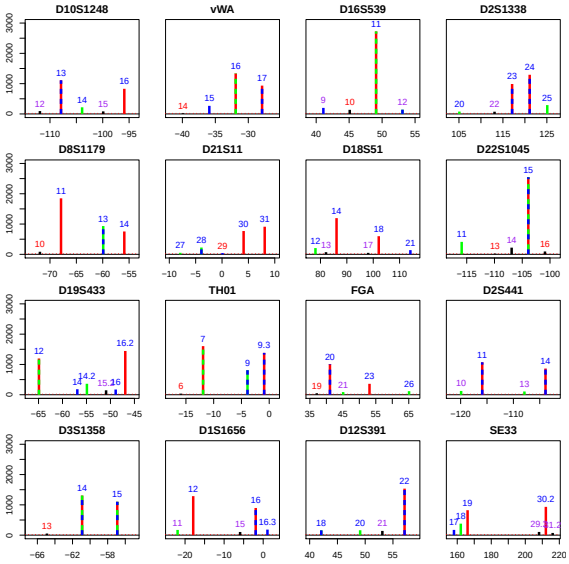
(a) CSP 2 (b) CSP 4

Figure A.4: CSPs for a number of notable three-contributor equal-contribution results; red, blue, green and black bars indicate alleles of the first, second and third contributors and unattributable peaks respectively. Blue, purple and red labels indicate peaks called as allelic, uncertain and non-allelic respectively for the discrete model CSP.

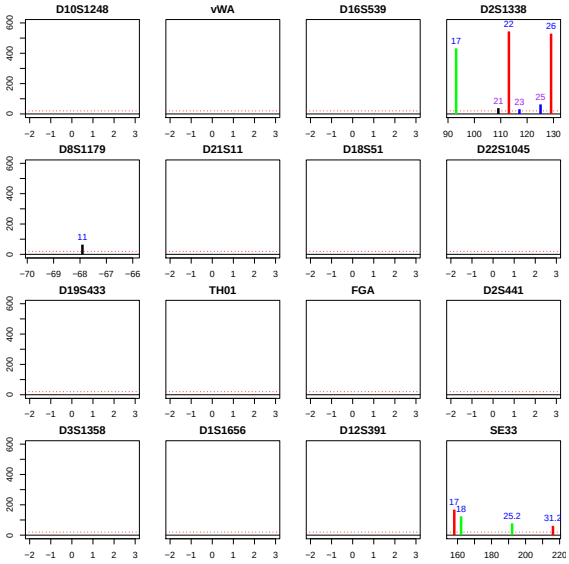


(a) CSP 1

(b) CSP 2

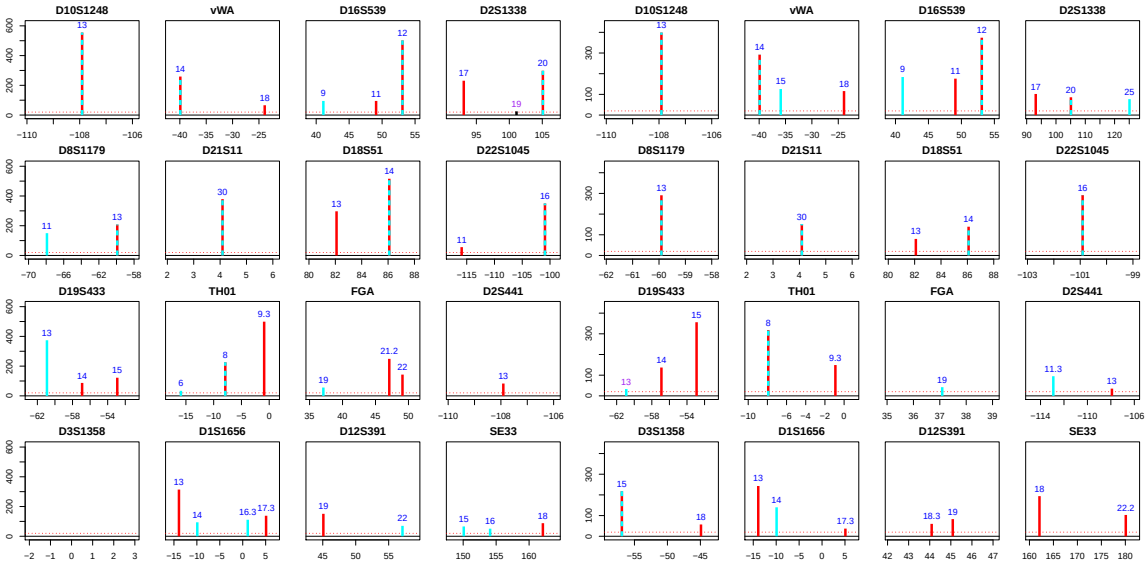


(c) CSP 4



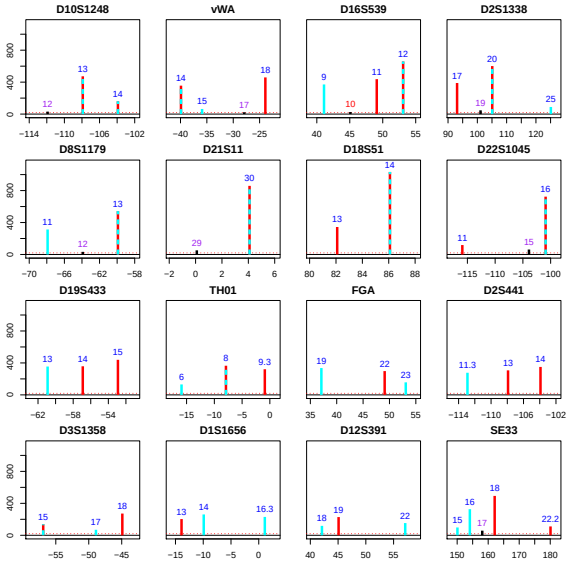
(d) CSP 5

Figure A.5: CSPs for a number of notable three-contributor unequal-contribution results; red, green, blue and black bars indicate alleles of the 250pg, 62pg and 16pg contributors and unattributable peaks respectively. Blue, purple and red labels indicate peaks called as allelic, uncertain and non-allelic respectively for the discrete model CSP.



(a) Multiple replicates: 1

(b) Multiple replicates: 2



(c) Single Replicate

Figure A.6: Two-contributor equal-contributions CSP 8, with (a-b) multiple replicates and (c) single replicate.

Appendix B

Publications

ACKNOWLEDGMENTS

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

LITERATURE CITED

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Forensic Science International: Genetics 13 (2014) 82–89

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Verifying likelihoods for low template DNA profiles using multiple replicates

Christopher D. Steele [a,*], Matthew Greenhalgh [b], David J. Balding [a]

[a] UCL Genetics Institute, Darwin Building, Gower Street, London WC1E 6BT, UK
[b] Orchid Cellmark Ltd., Abingdon Business Park, Blacklands Way, Abingdon OX14 1YX, UK

ARTICLE INFO

Article history:
Received 19 February 2014
Received in revised form 9 June 2014
Accepted 30 June 2014

Keywords:
Low-template DNA
DNA mixtures
Likelihood ratio
Replicates
Forensic
likeLTD

ABSTRACT

To date there is no generally accepted method to test the validity of algorithms used to compute likelihood ratios (LR) evaluating forensic DNA profiles from low-template and/or degraded samples. An upper bound on the LR is provided by the inverse of the match probability, which is the usual measure of weight of evidence for standard DNA profiles not subject to the stochastic effects that are the hallmark of low-template profiles. However, even for low-template profiles the LR in favour of a true prosecution hypothesis should approach this bound as the number of profiling replicates increases, provided that the queried contributor is the major contributor. Moreover, for sufficiently many replicates the standard LR for mixtures is often surpassed by the low-template LR. It follows that multiple LTDNA replicates can provide stronger evidence for a contributor to a mixture than a standard analysis of a good-quality profile. Here, we examine the performance of the `likeLTD` software for up to eight replicate profiling runs. We consider simulated and laboratory-generated replicates as well as resampling replicates from a real crime case. We show that LRs generated by `likeLTD` usually do exceed the mixture LR given sufficient replicates, are bounded above by the inverse match probability and do approach this bound closely when this is expected. We also show good performance of `likeLTD` even when a large majority of alleles are designated as uncertain, and suggest that there can be advantages to using different profiling sensitivities for different replicates. Overall, our results support both the validity of the underlying mathematical model and its correct implementation in the `likeLTD` software.

1. Introduction

In forensic DNA profiling, a likelihood ratio (LR) is calculated to measure the support provided by DNA evidence (E) for a proposition H_p favouring the prosecution case, relative to its support for H_d representing the defence case. The LR can be written as

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

Each of H_p and H_d specifies a number of unprofiled contributors and a list of contributors whose DNA profiles are known (included in E). Typically H_p includes a profiled, queried contributor that we designate Q, who is replaced under H_d by an unprofiled individual X. Q may be an alleged offender, or a victim, while X is an alternative, usually unknown, possible source of the DNA. It usually suffices to limit attention to H_p and H_d that differ only in replacing Q with X, otherwise the LR is difficult to interpret as a measure of the weight of evidence for Q to be a contributor of DNA.

In addition to reference profile(s), of Q and possibly other known contributors, the DNA evidence consists of one or more profiling runs performed on a DNA sample recovered from a crime scene, or from an item thought to have been present when the crime occurred. Each profiling run generates graphical results in an electropherogram (epg), which we assume has been interpreted by a forensic scientist who decides a list of alleles observed at each locus, and also a list of potential alleles about which there is substantial uncertainty, perhaps due to possible stutter. Alleles not on either list are regarded as unobserved in that run.

In low-template DNA (or LTDNA) profiling, each epg can be affected by stochastic effects such as dropin, dropout and stutter [1]. To help assess stochastic effects, it is common to perform multiple profiling runs, possibly varying the laboratory conditions but these are nevertheless referred to as replicates.

* Corresponding author.
E-mail addresses: c.steele.11@ucl.ac.uk (C.D. Steele), mgreenhalgh@cellmark.co.uk (M. Greenhalgh), d.balding@ucl.ac.uk (D.J. Balding).

http://dx.doi.org/10.1016/j.fsigen.2014.06.018

Joint likelihoods for multiple replicates are obtained by assuming that the replicates are independent conditional on the genotypes of all contributors and parameters ϕ such as the amounts and degradation levels of DNA from each contributor [2]. We can write

$$\Pr(E|H) = \sum_j \Pr(\mathcal{G}_j) \prod_i \Pr(R_i|\mathcal{G}_j, \phi), \tag{2}$$

where R_i is the set of allele designations in the ith replicate run of the crime scene profile (CSP), $\mathcal{G}_j$ denotes the jth set of contributor genotypes, and the summation is over all possible sets of contributor genotypes under H. $\Pr(\mathcal{G}_j)$ is computed under a standard population genetics model [1]. The unknown parameters ϕ can be replaced with estimates, or eliminated by maximisation or integration with respect to a prior distribution.

Currently, there are only limited possibilities to check the validity of an algorithm for evaluating an LTDNA LR (henceforth ltLR). One approach is to evaluate the ltLR when Q is repeatedly replaced by a random profile [3]. In that case H_p is false and we expect the majority of computed ltLRs to be small. Here, we propose to investigate a performance indicator for ltLR algorithms when H_p is true. Under H_d, it may occur that $\mathcal{G}_X = \mathcal{G}_Q$, where $\mathcal{G}_X$ and $\mathcal{G}_Q$ denote the genotypes of X and Q. This occurs with probability π_Q, the match probability for Q. Since $\Pr(E|H_d, \mathcal{G}_X = \mathcal{G}_Q) = \Pr(E|H_p)$, it follows that [4]

$$\begin{aligned} \text{ltLR} &= \frac{\Pr(E|H_p)}{\Pr(E|H_d, \mathcal{G}_X = \mathcal{G}_Q)\pi_Q + \Pr(E|H_d, \mathcal{G}_X \neq \mathcal{G}_Q)(1-\pi_Q)} \\ &\leq \frac{1}{\pi_Q}. \end{aligned} \tag{3}$$

We will refer to $1/\pi_Q$ as the inverse match probability (IMP).

Consider first that Q is the major contributor to an LTDNA profile. Intuitively, if E implies that $\mathcal{G}_X = \mathcal{G}_Q$ then equality should be achieved in Eq. (3). The key idea of this paper is that if H_p is true then increasing numbers of LTDNA replicates should provide increasing evidence that $\mathcal{G}_X = \mathcal{G}_Q$, and so the ltLR should converge to the IMP. This holds even for mixtures if Q is the major contributor, since differential dropout rates should allow the alleles of Q to be identified from multiple replicates. However, any inadequacies in the underlying mathematical model or numerical approximations may become more pronounced with increasing numbers of replicates, preventing the ltLR from approaching the IMP. Therefore we propose to consider convergence of the ltLR towards the IMP as the number of replicates increases as an indicator of the validity of an algorithm to compute the ltLR when Q is the major contributor.

If Q is not the major contributor, even for many replicates there may remain ambiguity about the alleles of Q so that there remains a gap between the ltLR and IMP. However, the bound (3) still holds, and there is a useful guide to the appropriate value of the ltLR provided by the mixture LR for good-quality CSPs computed using only presence/absence of alleles [5]. If under H_p the contributors are Q and U, where U denotes an unknown, unprofiled individual, and H_d corresponds to two unknown contributors X and U, an example of a mixture LR is

$$\begin{aligned} \text{mixLR} &= \frac{\Pr(\text{CSP} = \text{ABC}, \mathcal{G}_Q = \text{AB}|\text{Q}, \text{U})}{\Pr(\text{CSP} = \text{ABC}, \mathcal{G}_Q = \text{AB}|\text{X}, \text{U})} \\ &= \frac{\Pr(\mathcal{G}_U \text{ is one of AC, BC, CC})}{\Pr((\mathcal{G}_X, \mathcal{G}_U) \text{ is one of} \begin{array}{l}(\text{AA}, \text{BC}), (\text{AC}, \text{BB}), (\text{AB}, \text{CC}), \\ (\text{AB}, \text{AC}), (\text{AB}, \text{BC}), (\text{AC}, \text{BC}))\end{array}}, \end{aligned} \tag{4}$$

where within-pair ordering is ignored in the denominator. Under the standard population genetics model [6,7] and setting $F_{ST} = 0$, the mixLR for this example is

$$\frac{\Pr(\text{CSP} = \text{ABC}, \mathcal{G}_Q = \text{AB}|\text{Q}, \text{U})}{\Pr(\text{CSP} = \text{ABC}, \mathcal{G}_Q = \text{AB}|\text{X}, \text{U})} = \frac{2p_A + 2p_B + p_C}{6p_A p_B(p_A + p_B + p_C)}, \tag{5}$$

where the p are population allele probabilities. As expected, mixLR $<$ IMP $= 1/2p_Ap_B$. See Ref. [8] for further details and examples. Note that the mixLR does not use peak height information.

Multiple LTDNA replicates should allow identification of all alleles present in any contributor, and hence the ltLR should reach the mixLR. In fact, ltLR will typically exceed mixLR because the alleles of different contributors may be distinguished over the multiple replicates through differential dropout rates. Indeed, Ref. [9] propose subsampling to generate different mixture ratios in low-template replicates as a strategy to assist mixture deconvolution. We cast light on this possibility below by considering a real CSP that has been profiled using multiple replicates at two different levels of sensitivity. More generally, we examine the behaviour of ltLR in relation to mixLR and IMP, and the utility of each of these for verifying the validity of ltLR computations.

`likeLTD` is an open-source R package that computes likelihoods for low-template DNA profiles [10]. `likeLTD` allows for the designation of epg peaks as uncertain in addition to the usual allelic/non-allelic classification, but does not directly use epg peak heights. Uncertain alleles are treated as if they were masked in calculation of the likelihood: the presence/absence of the allele is regarded as unknown. The effect of an uncertain call on calculation of the likelihood is illustrated in Table 1. When B is called as uncertain rather than absent and the hypothesised contributor has a B allele, a dropout term D is removed from the likelihood because the dropout status of B is unknown. We use `likeLTD` here both to confirm its good performance in computing ltLRs, and to illustrate the value of the IMP as a strict upper bound and the mixLR as an approximate lower bound. We apply `likeLTD` to lab-based profiling replicates, simulated replicates, and replicates obtained by re-sampling the five actual replicates of a real CSP.

Throughout this paper, ltLR, mixLR and IMP will be reported in units of bans, which is a base 10 logarithmic scale introduced as a measure of weight of evidence by Alan Turing during his wartime code breaking work [11]. Thus 6 bans corresponds to an LR of 1 million on the natural scale.

2. Materials and methods

2.1. Laboratory replicates

Cheek swab samples were obtained from five volunteers, and 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. The samples were then quantified using the Life Technologies Quantifiler® Human DNA Quantification kit as per the manufacturer's recommendations.

Table 1
Likelihood calculations for a CSP when the queried contributor Q has genotype AB and [] indicates an allele designated as uncertain. L_p is the likelihood under the prosecution hypothesis, and D is the dropout probability. Under H_d are possible genotypes for the alternative contributor X, where Z is any other allele. L_d is the corresponding contribution to the likelihood under the defence hypothesis, where p_x is the probability of allele x, and D_2 is the homozygote dropout probability.

CSP	L_p	H_d	L_d
A	$D(1-D)$	AA	$p_A^2(1-D_2)$
		AZ	$2p_A(1-p_A)D(1-D)$
A[B]	$1-D$	AA	$p_A^2(1-D_2)$
		AB	$2p_Ap_B(1-D)$
		AZ	$2p_A(1-p_A-p_B)D(1-D)$

Table 2
Sample preparation and genotyping protocol for all conditions examined in the lab-based experiments (described in Table 3). Each condition was replicated eight times. The initial DNA concentration (column 3), dilution (column 4) and volume (column 5) generate approximately the DNA mass indicated in column 6. Columns 7 and 8 show the number of PCR cycles and the volume of PCR product added to each well for the genotyping. Columns 9 and 10 show the ratio of Hi-Di™ formamide to GeneSan™ 400HD ROX™ and the volume of the mixture added to each well. Apmr stands for as per manufacturers recommendations.

Condition	Contributor	Init. conc. (ng μL^{-1})	Dilution (%)	Volume (μL)	Mass (pg)	Cycles	Product (μL)	Formamide: ROX	F/ROX mixture (μL)
(i)	B	31.0	1	1.6	500	28	apmr	apmr	apmr
(ii)	B	31.0	0.1	2.0	60				
(iii)	B	31.0	0.01	5.0	15				
(iv)	A	23.0	1	17.6	500	28	apmr	apmr	apmr
	C	18.1	0.1	16	30				
(v)	A	23.0	0.1	22.4	60	28	apmr	apmr	apmr
	C	18.1	1	22.0	500				
(vi)	A	23.0	0.1	2.7	60	28	apmr	apmr	apmr
	B	31.0	0.1	2.0	60				
	C	18.1	0.1	3.5	60				
(vii)	A	23.0	0.1	2.7	60	28	1	600:1	9
	B	31.0	0.1	2.0	60				
	C	18.1	0.1	3.5	60				
(viii)	A	23.0	0.1	2.7	60	28	9	366:1	11
	B	31.0	0.1	2.0	60				
	C	18.1	0.1	3.5	60				
(ix)	A	23.0	0.1	2.7	60	30	apmr	apmr	apmr
	B	31.0	0.1	2.0	60				
	C	18.1	0.1	3.5	60				

Each sample was serially diluted on a $\log_{10}$ scale, and then amplified using the AmpFℓSTR® SGM Plus® PCR kit as per the manufacturer's recommendations on a Veriti® 96-Well Fast Thermal Cycler.

An ABI 3130 Sequencer was used to analyse 1 μL of the PCR products, with 10 second injections at 3 kV; these settings were used for all subsequent analyses. The results returned from the 3130 sequencer were analysed using GeneMapper® ID v3.2 to determine which samples were suitable for further use.

For the one-contributor investigation eight replicates of each of three conditions were created (Table 2). The conditions were created to investigate increasing dropout rate. For the 500 pg and

Table 3
Experimental conditions and hypotheses compared. pg denotes picograms and measures DNA mass; Pr(D) denotes the probability of dropout for a heterozygote allele, while Pr(C) denotes the probability of dropin. Pr(unc) indicates the probability of designating a CSP allele as uncertain. υ indicates the number of uncertain dropins per locus per replicate; see text for further details of "Condition". Q denotes the queried contributor, who is one of A, B or C as indicated in parentheses. X is an unknown alternative to Q under H_d, while U1 and U2 are unknown contributors under both H_p and H_d.

Study	# Contributors	Condition	H_p	H_d
Lab-based	1	500 pg (i)	Q (B)	X
		60 pg (ii)	Q (B)	X
		15 pg (iii)	Q (B) + dropin	X + dropin
	2		Q (A) + dropin	X + dropin
		A=500 pg; C=30 pg (iv)	Q (A) + U1	X + U1
			Q (C) + U1	X + U1
			Q (C) + dropin	X + dropin
		A=60 pg; C=500 pg (v)	Q (C) + U1	X + U1
			Q (A) + U1	X + U1
	3	28 cycles (vi)	Q (A) + U1 + U2	X + U1 + U2
		Phase 1 (vii)	Q (A) + U1 + U2	X + U1 + U2
		Phase 2 (viii)	Q (A) + U1 + U2	X + U1 + U2
		30 cycles (ix)	Q (A) + U1 + U2	X + U1 + U2
Simulation	1	$Pr_B(D) = 0$; $Pr(C) = 0$	Q (B)	X
		$Pr_B(D) = 0.4$; $Pr(C) = 0.05$	Q (B) + dropin	X + dropin
		$Pr_B(D) = 0.8$; $Pr(C) = 0.05$	Q (B) + dropin	X + dropin
		Pr(unc) = 0.8; $\upsilon \sim$ Pois($\lambda = 1$)	Q (B)	X
		Pr(unc) = 0.4; $\upsilon \sim$ Pois($\lambda = 1$)	Q (B)	X
	2		Q (A) + dropin	X + dropin
		$Pr_{A,C}(D) = \{0.2, 0.8\}$; $Pr(C) = 0$	Q (A) + U1	X + U1
			Q (C) + U1	X + U1
			Q (A) + dropin	X + dropin
		$Pr_{A,C}(D) = \{0.2, 0.6\}$; $Pr(C) = 0$	Q (A) + U1	X + U1
			Q (C) + U1	X + U1
	3	$Pr_{A,B,C}(D) = \{0.8,0.5,0.2\}$; $Pr(C) = 0$	Q (A) + U1 + U2	X + U1 + U2
		$Pr_{A,B,C}(D) = \{0.5,0.5,0.5\}$; $Pr(C) = 0$	Q (A) + U1 + U2	X + U1 + U2
		$Pr_{A,B,C}(D) = \{0.2,0.5,0.8\}$; $Pr(C) = 0$	Q (A) + U1 + U2	X + U1 + U2
Real-world	≥3	Standard and sensitive	Q + U1 + U2	X + U1 + U2
		Standard only	Q + U1 + U2	X + U1 + U2
		Sensitive only	Q + U1 + U2	X + U1 + U2

Table 4
Five replicates of a crime scene profile, three from a sensitive LTDNA profiling technique and two from standard DNA profiling. Alleles shown in [] were called as uncertain.

Locus	Sensitive profiling			Standard profiling	
	Run 1	Run 2	Run 3	Run 4	Run 5
D3	16, [15]	16, [15]	16, 18, [15]	16	16
vWA	15, 16, [17]	15, [14]	15, 18, [14]	15	15
D16	9	9	9, 11, [10]	9	9
D2	17, 19, 24	16, 17, 24,[23]	17, [16]	24	24
D8	8, 13, 15, 16	8, 12, 13, 16, [15]	8, 13, 14, 16, [15]	[8]	
D21	30, 32, 33.2	32, 32.2, 33.2	32, 32.2, 33.2, 34, [31]	[32], [32.2]	[33.2]
D18	12, 17	12, 17, 19	12, 17, [11], [16]	[17]	17
D19	14, 21, [13]	11, 14, [13]	14, [13]	14	14
TH01	6, 9.3	6, 9.3	6, 8, 9.3	[6], [9.3]	[6]
FGA	21	21, [20]	21, 20	21	

60 pg conditions, one-contributor hypotheses were compared, B under H_p and X under H_d, while for the 15 pg condition dropin was also modelled under both hypotheses (Table 3).

For the two-contributor investigation eight replicates of each of two conditions were created (Table 2). The major and minor contributors were reversed between conditions, with an increased DNA contribution from the minor. These samples were amplified and analysed as described previously. Two-contributor hypotheses were compared, with each of A and C in turn playing the role of Q, while the other contributor was treated as unknown. Additionally one-contributor-plus-dropin hypotheses were compared, with only the major contributor playing the role of Q (Table 3).

For the three-contributor investigation eight replicates of each of four conditions were created (Table 2). The conditions were created to investigate different profiling protocols. The Phase 1 and Phase 2 conditions are post-PCR purification protocols designed to enhance the sensitivity of detection of the standard protocol [12], and both involve concentrating the post-PCR product using an Amicon® PCR microcon unit according to the manufacturer's recommendations. Phase 1 enhancement increases the amount of formamide in the mixture compared to the manufacturer's recommendations, while Phase 2 enhancement increases the amount of DNA, formamide and ROX compared to Phase 1. For all four conditions (30 cycles, 28 cycles, Phase 1, and Phase 2), three-contributor hypotheses were compared, with A playing the role of Q and the other contributors treated as unknown (Table 3). Dropin was not modelled under either hypothesis, although dropin was included in the simulations. This reflects a realistic challenge for few replicates with multiple contributors, whereby any dropin alleles may be wrongly attributed to one of the contributors. However the incorrect model will lead to deterioration of inferences for larger numbers of replicates.

2.2. Simulated replicates

All of the conditions that we now describe were simulated in eight replicates, with the whole simulation being performed five times. Initially a number of single-contributor CSPs were simulated using the profile of individual B. The first condition investigated was a "perfect match", in which all eight replicates generated exactly the profile of B. Next, we introduced mild dropout ($\Pr(D) = 0.4$) and severe dropout ($\Pr(D) = 0.8$) of the alleles of B, in each case with dropins included at rate $\Pr(C) = 0.05$ (at most one dropin per locus per replicate). The homozygous dropout probability was set equal to $\Pr(D)^2/2$, as suggested by [13]. We then examined the effect of uncertain allele designations by randomly designating some alleles of B as uncertain, first with $\Pr(\text{unc}) = 0.4$ and then $\Pr(\text{unc}) = 0.8$. In both conditions, at each locus and in each replicate a Poisson mean one number of alleles not in the profile of B was also designated as uncertain, with types randomly selected according to frequencies in the UK Caucasian database. For all these simulated profiles, one-contributor hypotheses were compared, B under H_p and X under H_d.

Next two-contributor CSPs were simulated, based on the profiles of A and C. Two conditions were simulated, both used $\Pr_A(D) = 0.2$, while $\Pr_C(D)$ was initially 0.8 and then 0.6. Dropin was not

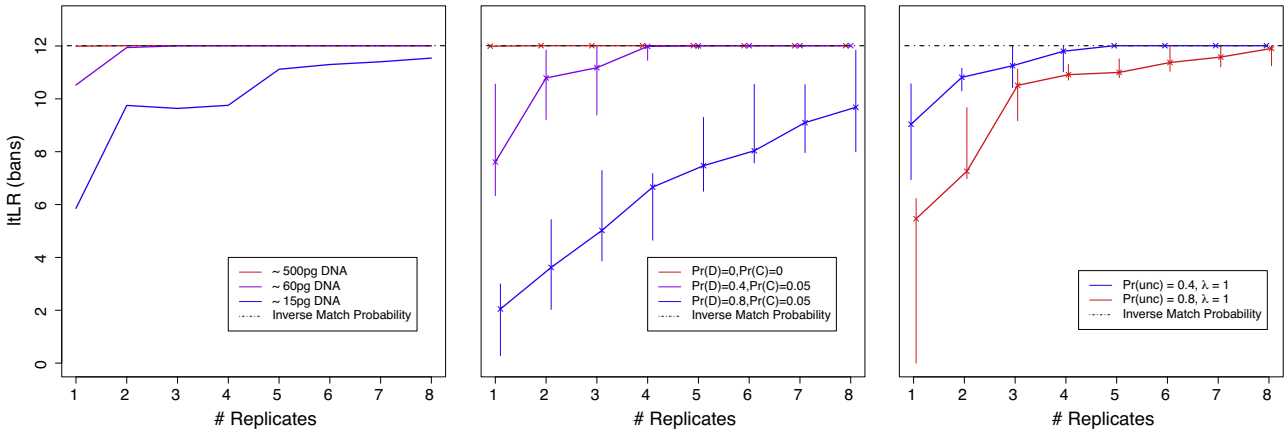


Fig. 1. The ltLR shown on a logarithmic scale (in bans) from one-contributor CSPs evaluated using from one up to eight replicates. Left: lab-based replicates, with DNA template (in pg) as shown in the legend box. Middle: simulated replicates with dropout (probability Pr(D)) and dropin (probability Pr(C)); the plotted points represent the median from five repetitions of the simulation, and the vertical bars show the range. Right: simulated replicates with uncertain allele calls (probability Pr(unc) for a true allele to be uncertain, and a Poisson (rate λ) number of non-alleles labelled as uncertain at each locus.

simulated. For shared alleles the dropout probability was the product of the dropout probabilities for each contributor having that allele. Two-contributor hypotheses were compared, with each of A and C in turn taking the role of Q, while the other was treated as unknown in the analysis. Additionally one-contributor-plus-dropin hypotheses were compared, only for A playing the role of Q (Table 3).

Three-contributor CSPs were then simulated under three conditions, with dropout probabilities for Donors A, B and C as shown in Table 3. Dropin was included as for the one-contributor simulations. Three-contributor hypotheses were compared, with A playing the role of Q and the other two contributors being treated as unknown.

2.3. Crime case replicates

We used a CSP from an actual crime investigation, consisting of five replicates: two using standard SGM+ profiling and three generated using an LCN protocol with 34 PCR cycles (Table 4). This example was submitted to us for `likeLTD` analysis, and as is typical only limited information about the profiling protocol was provided by the profiling lab. These details are not required by `likeLTD` because it estimates the unknown parameters from the CSP allele designations. We re-sampled the five actual replicates to generate simulated profiles with up to eight replicates, consisting of standard replicates only, sensitive replicates only, or both. Six distinct alleles were observed at locus D8, but no more than three

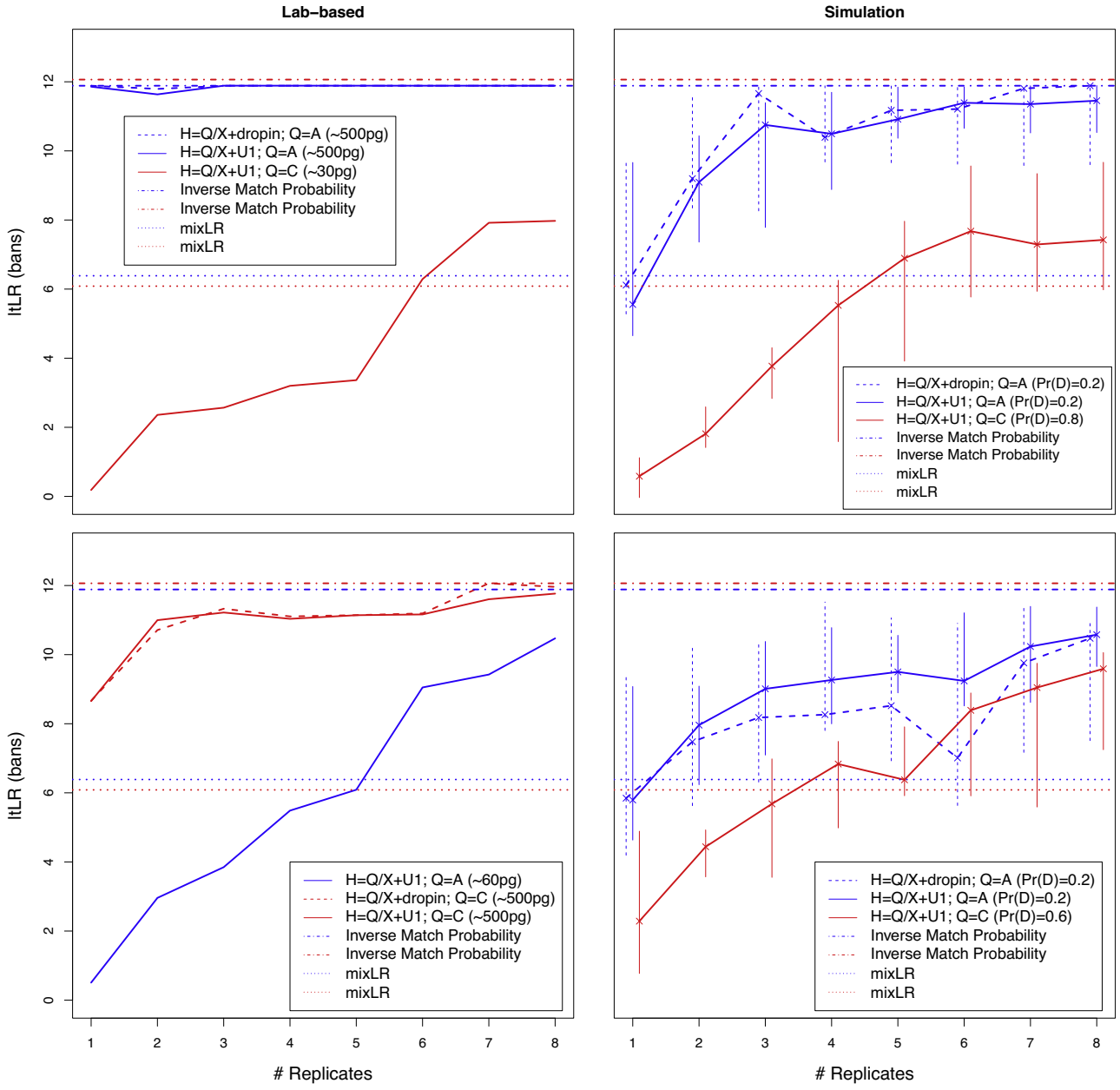


Fig. 2. The low-template likelihood ratio (ltLR) from two-contributor CSPs profiled at up to eight replicates. Left: lab-based replicates, with the DNA template from the minor contributor greater in the lower panel (see legend boxes). Right: simulation-based replicates, with the minor contributor having reduced dropout in the lower panel. The simulated CSPs were generated from the profiles of Donors A and C, and the line colours on the graph indicate whether the queried individual (Q) is A (blue) or C (red). Solid lines indicate a two-contributor analysis, with the non-Q individual regarded as unknown (U1). Dashed lines indicate a one-contributor analysis that also allows for dropin (only for Q the major contributor). The inverse match probability is shown with dot-dash lines, coloured according to Q. The mixLR is shown with dotted lines, coloured according to Q. In the legend boxes, H indicates the hypotheses with X an unknown alternative to Q, and Pr(D) indicates the probability of dropout.

replicated alleles were observed at any locus. Three-contributor hypotheses were compared, with all contributors unknown under H_d, and no dropin (Table 3).

3. Results

3.1. One contributor

3.1.1. Lab-based

For the good-template experiments (500 pg), Fig. 1 (left) shows that the ltLR equals the IMP for all numbers of replicates (one through eight). This is the expected result, and the exercise shows that in this simple setting there is no deterioration in the quality of the computed LR for large numbers of replicates. Low DNA template (60 pg) generates an ltLR about 1.6 bans below the IMP for one replicate, but the gap is very small for two replicates and is negligible for larger numbers of replicates. For very low DNA template (15 pg) the ltLR is just under 6 bans for a single replicate, about 6 bans below the IMP. Replicate profiling substantially narrows the gap, but does not completely close it, with a difference of about 3 decibans remaining at eight replicates.

3.1.2. Simulation

The corresponding simulation studies show broadly similar trends to the lab-based data. For both the perfect match ($\Pr(D) = 0$) and mild dropout ($\Pr(D) = 0.4$) conditions, the median ltLR rapidly reaches the IMP but does not exceed it, while under severe dropout ($\Pr(D) = 0.8$) the median ltLR rises towards the IMP but does not reach it (Fig. 1, middle). For the low and high rates of uncertain calls, the IMP is approximately reached at a five and eight replicates, respectively (Fig. 1, right).

3.2. Two contributors

3.2.1. Lab-based

When the minor contributor provides only 30 pg of DNA (Fig. 2, top left panel), then if Q is the major contributor the ltLR is very close to the IMP for all numbers of replicates, whereas if Q is the minor contributor then there remains a substantial gap between ltLR and IMP even at eight replicates. However, even with this very low template, the ltLR exceeds the mixLR beyond five replicates. When the major and minor contributors are reversed, and the amount of DNA from the minor is doubled (Fig. 2, bottom left), then if Q is the minor contributor the ltLR substantially exceeds mixLR from six replicates and rises to within two bans of the IMP at eight replicates. Under both conditions, the two-contributor analysis gives a very similar result to the one-contributor-with-dropin analysis.

3.2.2. Simulation

When the minor contributor is subject to high dropout (Fig. 2, top right), then if Q is the major contributor the ltLR exceeds the mixLR after one replicate, and rises rapidly to within about 2 bans of the IMP, but the gap narrows only slowly thereafter. The one-contributor-plus-dropin analysis gives an ltLR that is broadly similar to the two contributor analysis, but with a wider range indicating greater variability. If Q is the minor contributor, the median ltLR increases rapidly from a low base, and appears to stabilise after about five replicates, about four bans below the IMP but exceeding the mixLR. The range increases after three replicates, and remains high up to eight replicates.

With reduced dropout for the minor contributor (Fig. 2, bottom right), inferring the presence of a major contributor Q is harder because of additional masking by the minor contributor. The median ltLR in both the two contributor and one-contributor-plus-dropin analyses eventually reaches within 2 bans of the IMP, with the latter showing a greater range. Conversely, the lower dropout rate leads to improved inference for a minor contributor Q, with the median ltLR rising to about three bans below the IMP at eight replicates, and exceeding the mixLR from four replicates. Interestingly, after six replicates the range of the minor contributor ltLR overlaps the range for the major contributor.

3.3. Three contributors

3.3.1. Lab-based

The 30 PCR cycles condition gives the highest ltLR at one replicate but little improvement with additional replicates (Fig. 3, left). The other amplification methods do show an increasing ltLR trend with additional replicates, but in no case did the ltLR reach within four bans of the IMP. As expected, the ltLR for both phase 1 and phase 2 enhancement exceeds that for standard 28 PCR cycles at all numbers of replicates, and phase 2 enhancement ltLR typically gives a small improvement over phase 1 enhancement. For 30 PCR cycles, the ltLR exceeds the mixLR for a single replicate but dips slightly below it at six replicates. For the other conditions, the mixLR is always exceeded from four replicates.

3.3.2. Simulation

All three curves in Fig. 3 (middle) show an increasing trend with number of replicates, with the median ltLR being in the expected order throughout (decreasing ltLR with increasing

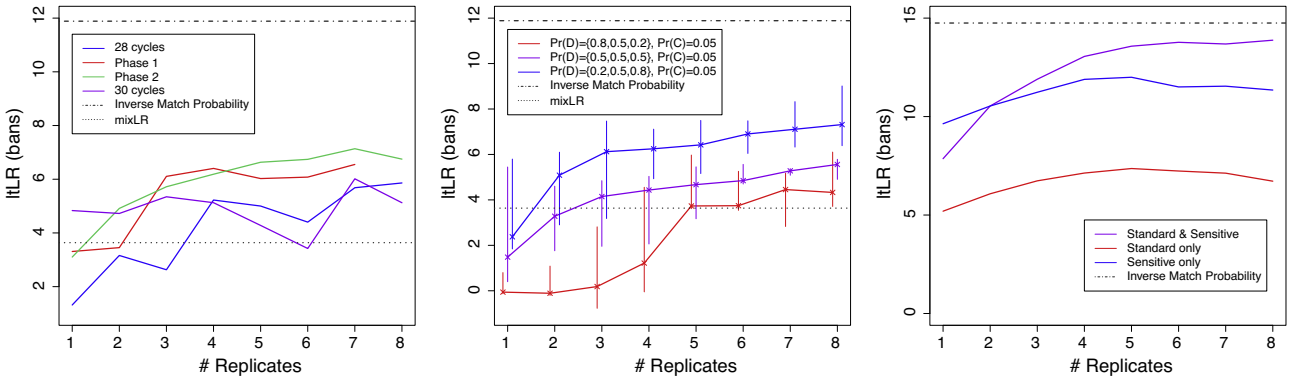


Fig. 3. The low-template likelihood ratio (ltLR) for three-contributor crime stains profiled with one to eight replicates. Left: laboratory replicates using four lab techniques indicated in the legend box and described further in Section 2. Middle: simulated replicates with dropout rates for the three contributors as shown in the legend box against Pr(D), the first value being for the queried contributor. Pr(C) is the dropin probability. Right: re-sampled actual crime-stain replicates; the original data are two standard profiling replicates, and three replicates using enhanced sensitivity. The ltLR returned from a perfect replicate of the contributors (consisting of every allele from each contributor) is shown with dotted lines; this is not possible for the real-world case, as the true contributors are unknown.

dropout for Q). The median ltLR exceeds the mixLR after one replicate (low dropout), after two replicates (medium dropout) and after four replicates (high dropout). The range is often wide, reflecting a strong dependence of the ltLR on the details of the simulation (in particular the number of alleles shared across contributors).

3.3.3. Real-world case

The ltLR returned when only standard or only sensitive replicates are used shows a similar trend, but nearly five bans lower for the standard replicates (Fig. 3, right). For three or more replicates, using mixed types of replicates is superior even to only using sensitive replicates, coming to within two bans of the IMP. This partly reflects the limited pool of replicates used in the actual crime case, but suggests that using different sensitivities in the profiling replicates may convey an advantage due to different contributors being better distinguished.

4. Discussion

We have shown that ltLR computed by `likeLTD` is bounded above by the IMP in every condition considered, as predicted by theory (Eq. (3)). That the bound is often tight when Q is the major contributor (Figs. 1 and 2 (top)) supports the validity of the underlying mathematical model, and its correct implementation in the `likeLTD` software. Our results should help counter any misconception that combining multiple noisy profiling replicates only compounds the noise: in fact, multiple noisy replicates can fully recover the genotype of a contributor [14].

A novel feature of `likeLTD`, is that it can accommodate uncertain allele designations, which diminishes the problem of an all-or-nothing allele call, therefore mitigating the problem highlighted by [15] of choosing a detection threshold. We have shown (Fig. 1 (right)) that introducing many uncertain allele calls leads to ltLRs that satisfy the bound, which is reasonably tight with as few as three replicates even when 80% of true alleles are designated as uncertain and there are also multiple uncertain non-alleles.

We have further shown that mixLR, the LR computed from knowing every allele that is represented in the profile of at least one contributor to the CSP, is often surpassed after only a handful of replicates. Then, multiple LTDNA replicates provide stronger evidence than a single good quality profile correctly representing the alleles of all contributors, which occurs because the alleles of different contributors can to some extent be distinguished through differential dropout rates in multiple replicates. These results lend support in principle to the proposal of [9].

Fig. 2 shows that, for two-person mixtures, the analysis assuming one-contributor-plus-dropin gave a very good approximation for the lab-based replicates (left panels), and a reasonably good approximation for the simulation replicates, but with more variable ltLR values, as indicated by the wider range.

4.1. Choice of profiling technique

We generated three-contributor CSPs in order to compare different LTDNA profiling techniques. We chose the most challenging condition in which all three contribute the same DNA template, making it impossible to deconvolve the mixture into the genotypes of individual contributors. We found that PCR performed with 28 cycles (regardless of enhancement) is preferable to 30 cycle PCR beyond one replicate (Fig. 3). More PCR cycles introduces more stochasticity in the results, as stated in the AmpFℓSTR® SGM Plus® PCR Amplification Kit user guide. We found that enhancement of the post-PCR sample is advantageous, with Phase 2 enhancement providing a small further improvement over Phase 1 (Fig. 3). These results support those of Forster et al. [16], who demonstrated that increasing PCR cycles increases the size of stutter peaks and the incidence of dropin; we observed no improvement in the WoE for 30 PCR cycles, possibly due to these stochastic effects.

The results from the real crime case (Fig. 3, right) suggest that if possible, a mixture of LTDNA replicates with differing sensitivities should be employed, as this allows better discrimination between the alleles of different contributors and hence a higher ltLR than the same number of replicates all using the same sensitivity.

4.2. Use of replicates

Splitting the sample reduces the quality of results expected in each replicate compared with that which would be obtained from a single profiling run using all available DNA. Grisedale and van Daal [17] favour use of a single run, but their comparison was with a consensus sequence obtained from multiple replicates, rather than the more efficient statistical analysis available through analysing individual replicates. Our results show increasing information obtained from additional replicates, which may tilt the argument towards use of multiple replicates but we have not done a comparison directly addressing this question. To fully test the performance of `likeLTD` in relation to mixLR and IMP we have used up to eight replicates. Taberlet et al. [18] suggest seven replicates to generate a quality profile when the amount of DNA is low, but this many replicates is rarely available for low-template crime samples [15].

Acknowledgements

CDS is funded by a PhD studentship from the UK Biotechnology and Biological Sciences Research Council and Cellmark Forensic Services.

References

[1] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Appl. 1 (2014) 20–21. , http://dx.doi.org/10.1146/annurev-statistics-022513-115602.
[2] J.M. Curran, P. Gill, M.R. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (1) (2005) 47.
[3] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. (2012), http://dx.doi.org/10.1016/j.fsigen.2012.11.002.
[4] R.G. Cowell, T. Graversen, S. Lauritzen, J. Mortera, Analysis of DNA mixtures with artefacts, in: ArXiv e-Prints, 2013, February.
[5] B.S. Weir, C.M. Triggs, L. Starling, K.A.J. Stowell, J. Buckleton, Interpreting DNA mixtures, J. Forensic Sci. 42 (1997) 213–222.
[6] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2) (2006) 90–101.
[7] P. Gill, L. Gusm ao, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, B.S. Weir, 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. (2012), http://dx.doi.org/10.1016/j.fsigen.2012.06.002.
[8] D.J. Balding, Weight-of-Evidence for Forensic DNA Profiles, Wiley, 2005.
[9] J. Ballantyne, E.K. Hanson, M.W. Perlin, DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information, Sci. Justice 53 (2) (2013) 103–114. , http://dx.doi.org/10.1016/j.scijus.2012.04.004, ISSN 1355-0306.
[10] D.J. Balding, Evaluation of mixed-source, low-template DNA profiles in forensic science, Proc. Natl. Acad. Sci. U. S. A. (2013), http://dx.doi.org/10.1073/pnas.1219739110.
[11] J. Good Irving, Studies in the history of probability and statistics. xxxvii am Turing's statistical work in world war ii, Biometrika 66 (2) (1979) 393–396.
[12] A.D. Roeder, P. Elsmore, M. Greenhalgh, A. McDonald, Maximizing DNA profiling success from sub-optimal quantities of DNA: a staged approach, Forensic Sci. Int. Genet. 3 (2) (2009) 128–137. , http://dx.doi.org/10.1016/j.fsigen.2008.12.004.
[13] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (1) (2009) 1–10. , http://dx.doi.org/10.1016/j.fsigen.2009.03.003.

[14] L. Schneps, C. Colmez, Math on trial: how numbers get used and abused in the courtroom, Wiley Online Library, 2013.
[15] B. Budowle, A.J. Eisenberg, A. van Daal, Validity of low copy number typing and applications to forensic science, Croat. Med. J. 50 (3) (2009) 207–217. , http://dx.doi.org/10.3325/cmj.2009.50.207.
[16] L. Forster, J. Thomson, S. Kutranov, Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle "low copy number"(LCN) method for analysis of trace forensic DNA samples, Forensic Sci. Int. Genet. 2 (4) (2008) 318–328. , http://dx.doi.org/10.1016/j.fsigen.2008.04.005.
[17] K.S. Grisedale, A. van Daal, Comparison of STR profiling from low template DNA extracts with and without the consensus profiling method, Invest. Genet. 3 (2012) 14, http://dx.doi.org/10.1186/2041-2223-3-14.
[18] P. Taberlet, S. Griffin, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L.P. Waits, J. Bouvet, Reliable genotyping of samples with very low DNA quantities using PCR, Nucleic Acids Res. 24 (16) (1996) 3189–3194.

Annals of human genetics

doi: 10.1111/ahg.12081

Worldwide F_{ST} Estimates Relative to Five Continental-Scale Populations

Christopher D. Steele[1]*, Denise Syndercombe Court[2] and David J. Balding[1]

[1]*UCL Genetics Institute, Darwin Building Gower Street, London, WC1E 6BT, UK*

[2]*Analytical & Environmental Sciences, Kings College London, Stamford Street, London, SE1 9NH, UK*

Summary

We estimate the population genetics parameter F_{ST} (also referred to as the fixation index) from short tandem repeat (STR) allele frequencies, comparing many worldwide human subpopulations at approximately the national level with continental-scale populations. F_{ST} is commonly used to measure population differentiation, and is important in forensic DNA analysis to account for remote shared ancestry between a suspect and an alternative source of the DNA. We estimate F_{ST} comparing subpopulations with a hypothetical ancestral population, which is the approach most widely used in population genetics, and also compare a subpopulation with a sampled reference population, which is more appropriate for forensic applications. Both estimation methods are likelihood-based, in which F_{ST} is related to the variance of the multinomial-Dirichlet distribution for allele counts. Overall, we find low F_{ST} values, with posterior 97.5 percentiles $< 3\%$ when comparing a subpopulation with the most appropriate population, and even for inter-population comparisons we find $F_{ST} < 5\%$. These are much smaller than single nucleotide polymorphism-based inter-continental F_{ST} estimates, and are also about half the magnitude of STR-based estimates from population genetics surveys that focus on distinct ethnic groups rather than a general population. Our findings support the use of F_{ST} up to 3% in forensic calculations, which corresponds to some current practice.

Keywords: Microsatellite, short tandem repeat, F_{ST}, fixation index, forensic

Introduction

We analyse an extensive new data set of the short tandem repeat (STR) profiles of individuals with worldwide origins, to estimate F_{ST} for national-scale subpopulations relative to continental-scale populations. We use two approaches to estimating F_{ST}, which differ according to the choice of reference population: a direct method that is appropriate for forensic applications, and an indirect method that reflects current population genetics practice.

In a forensic setting, F_{ST} is used to account for distant relatedness (coancestry) between the queried contributor (Q) and the unknown individual X that replaces Q in the defence hypothesis (Weir, 2007). Larger values of F_{ST} imply greater coancestry and so a greater probability that the profiles of X and Q are similar. This results in a lower likelihood ratio (LR), meaning that ignoring coancestry between X and Q is unfavourable to the defendant. The difference is unimportant for full-profile matches because even after F_{ST} adjustment the resulting LR is extremely large, and may be rounded down for example to 1 billion for reporting in court. However, F_{ST} adjustments are widely used, and can have a substantial impact, in analyses of mixed and low-template DNA profiles. The use of an F_{ST} adjustment can be regarded as allowing for additional uncertainty arising from the fact that the available database does not fit the circumstances of the case perfectly, which logically reduces confidence in the result, reflected in the reduced LR.

The appropriate value of F_{ST} in forensic work is relative to the reference database used, and may therefore differ substantially from F_{ST} estimates arising in population genetics research. Even if Q and X have a very similar ethnic background, a low F_{ST} value may suffice if the allele frequency database is directly appropriate for both Q and X, whereas the more distant they are from the database population, the larger the F_{ST} value that is required (Steele & Balding, 2014).

*Corresponding author: Christopher D. Steele, UCL Genetics Institute, Darwin Building Gower Street, London, WC1E 6BT, UK. Tel: +44 (0) 20 7679 4392; E-mail: c.steele.11@ucl.ac.uk

It is usually regarded as reasonable to give the defence some benefit of doubt and to apply a generous F_{ST} value to all possible X drawn from the same population as Q. If, on the other hand, Q is Caucasian and we wish to consider an X who is Afro-Caribbean, then the Afro-Caribbean database is appropriate and since little coancestry is expected between Q and X relative to this database, only a low value of F_{ST} would be required. There is always some uncertainty about the appropriate F_{ST} values: there is the usual variation in any statistical estimate but we have additional uncertainty here because F_{ST} is rarely estimated at the scale appropriate for a particular forensic analysis, and also different alternative contributors have different genetic backgrounds.

The origins of our study subjects are recorded at a national level, without reference to subnational ethnic identities. For example, in the analyses below Nigeria is treated as a subpopulation of a broader Afro-Caribbean population, but this ignores the substantial genetic variation among different groups within Nigeria. In forensic applications, it is appropriate to consider a distribution of F_{ST} values over alternative possibilities for X. Because an LR involves in effect a product over loci with an F_{ST} value applied at each locus, a single F_{ST} value for use in computing the LR should come from the upper tail of the F_{ST} distribution. Below, we will report posterior median estimates of F_{ST}, but when discussing forensic applications we will use the posterior 97.5 percentile, thus tending to over-estimate which is favourable to defendants.

We report F_{ST} values that are much lower than have been obtained from single nucleotide polymorphisms (SNPs). This in part reflects the within-nation population mixing described above, but low F_{ST} estimates also suggest a homogenising effect of STR mutation, which has previously been reported (Xu et al., 2000; Lu et al., 2012). It may also reflect that STRs employed in forensics were chosen in part on the basis of limited variation across populations, although many of the loci were chosen when little population data were available.

An extensive survey of worldwide human STR loci (Pemberton et al., 2013) focussed on well-defined ethnic groups, often with small population sizes, rather than the large and often ethnically mixed populations that are expected to be well represented in our database. Another recent study (Silva et al., 2012) has used worldwide forensic STR databases. We go beyond these papers in giving F_{ST} estimates at both within-continent and between-continent scales, and in using both observed and inferred reference populations. Our estimates are likelihood based, thus correctly account for variable sample size and provide posterior quantiles. They are directly relevant for forensic casework, and are also of broader interest in understanding human genetic variation in general populations at national, regional and continental scales.

Table 1 Number of alleles typed per locus and population. IC1-6 correspond to populations; Caucasian (IC1), Black African/Caribbean (IC3), South Asian (IC4), East/South-East Asian (IC5), and Middle Eastern/North African (IC6).

Observations	IC1	IC2	IC3	IC4	IC5	IC6	Total
D3S1358	7013	162	5200	704	625	226	13930
TH01	6953	158	5177	694	624	226	13832
D21S11	7006	162	5198	704	624	225	13919
D18S51	6944	157	5180	704	626	226	13837
D16S539	6951	162	5183	694	626	226	13842
VWA	7013	162	5194	704	626	226	13925
D8S1179	7007	162	5200	704	626	226	13925
FGA	6988	162	5196	700	626	226	13898
D19S433	6836	158	5122	687	621	226	13650
D2S1338	6575	152	4995	667	620	220	13229
D22S1045	1822	56	3478	523	506	162	6547
D1S1656	1835	56	3509	528	511	162	6601
D10S1248	1823	56	3497	516	506	118	6516
D2S441	1808	56	3458	521	501	160	6504
D12S391	1869	56	3531	551	507	162	6676
SE33	376	4	1039	308	396	140	2263

Materials and Methods

Database

Our data set includes the STR profiles of 7 121 individuals living in the UK or Eire, or applying to migrate to the UK on the basis of relatedness to a UK resident. They are all genotyped by the same laboratory at up to 16 STR loci. The individuals are self identified into one of six populations: White (IC1 and IC2, with IC2 including darker-skinned individuals of European origin), Black African/Caribbean (IC3), South Asian (IC4), East/South-East Asian (IC5), or Middle Eastern/North African (IC6). They are further classified into subpopulations, in most cases defined at the national level. Our worldwide coverage is extensive (Fig. 1), but some large populations are not included, such as Japan and Indonesia, and the sample sizes from Latin America are small. Our analyses use only allele counts and not individual genotypes. In a few instances of only one allele observed at a locus, the peak intensity was insufficient to confirm homozygote status, leading to only one allele being recorded at that locus. Thus, total allele counts are not always even integers (Table 1).

Subpopulations with >40 individuals sampled were included in our analyses. Some subpopulations of particular interest were also included despite having sample size <40. We merged or removed other subpopulations with small sample sizes. Study participants self identified both population and subpopulation labels, and in some cases we changed the population classification to better fit the subpopulation, as described below. These decisions require some subjective

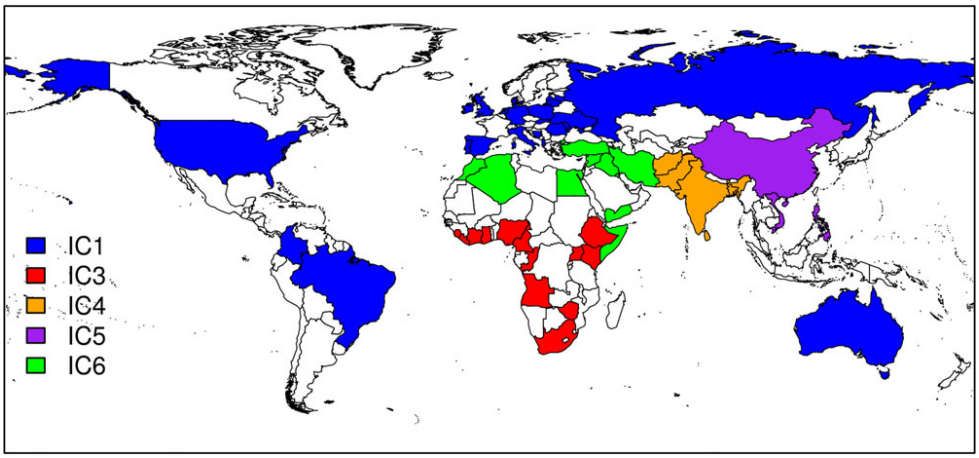


Figure 1 Countries of origin of the individuals included in the study, coloured according to the population that provides the best fit according to the indirect method (see text). White indicates countries represented by fewer than five individuals.

judgement; there is no canonical classification scheme for human populations.

IC1 and IC2

IC2 individuals from Europe were moved to IC1. Two national subpopulations were kept distinct, Eire and Great Britain, while the remaining European subpopulations were merged according to the United Nations geo-scheme for Europe (United Nations Statistics Division, 2014):

Eastern Europe:	Hungary, Moldova, Poland, Romania, Russia, Slovakia, Ukraine.
Northern Europe:	Denmark, Latvia, Lithuania, Sweden.
Southern Europe:	Albania, Bosnia, Croatia, Cyprus, Greece, Italy, Kosovo, Malta, Macedonia, Portugal, Spain, Yugoslavia.
Western Europe:	Belgium, France, Germany, Netherlands.

IC2 individuals from Argentina, Bolivia, Brazil, Columbia, Mexico, and Venezuela were combined ("Latin America"), as were IC1 individuals from Australia, New Zealand, and USA ("Anglo New World"). Those with no subpopulation identified, and those from Jersey, Northern Ireland, or South Africa, were removed.

IC3

Six national subpopulations were kept distinct: Ghana, Jamaica, Kenya, Nigeria, Sierra Leone, and Uganda. The following subpopulations were created from mergers according to the United Nations geo-scheme for Africa (United Nations Statistics Division, 2014), with Middle and Southern Africa combined as Central/Southern Africa:

Other W Africa:	Benin, Gambia, Guinea, Guinea-Bissau, Ivory Coast, Liberia, Mali, Togo.
Other C/S Africa:	Angola, Chad, Congo, Cameroon, South Africa.
Other E Africa:	Burundi, Ethiopia, Eritrea, Malawi, Rwanda, Sudan, Tanzania, Zambia, Zimbabwe.
Other Caribbean:	Barbados, Bermuda, Dominica, Guyana, Grenada, Monserrat, St Lucia, Virgin Islands, Trinidad.

Individuals with missing subpopulation were included as "Unknown IC3." Those with origin not in Africa or the Caribbean were removed (Eire, GB, USA). Algeria, Egypt, Morocco, and Somalia were all included with IC6 (see "Best population fit" below).

IC4

Four national subpopulations were kept distinct: Afghanistan, Bangladesh, India, Pakistan. Individuals with missing subpopulation, or if the subpopulation was Nepal or Sri Lanka, were included as "Unknown IC4." Mauritius was removed.

IC5

SE Asian subpopulations were merged (Cambodia, Indonesia, Philippines, Thailand, Vietnam). Mongolia and South Korea were merged with the much larger China sample to form NE Asia. Fiji was removed.

IC6

Iran, Iraq, Somalia, and Turkey were kept as separate national subpopulations. Other subpopulations were merged

into N Africa (Algeria, Egypt, Morocco) or Middle East (Jordan, Kuwait, Lebanon, Palestine, Qatar, Syria, Yemen, UAE). Those from Georgia or with no subpopulation identified were removed. Afghanistan was moved to IC4.

The UK Forensic Science Service (FSS) previously collated (Foreman & Evett, 2001) databases of STR frequencies at 10 loci, in six populations with similar definitions to our data set: EA1 (Caucasian), EA2 (Mediterranean), EA3 (Afro-Caribbean), EA4 (South Asian), EA5 (East Asian), and EA6 (Middle East/North Africa). These databases are small (<2000 individuals combined) and do not include subpopulation labels. EA5 and EA6 both have sample sizes varying over loci, and the average sample size is reported below. Until recently, these were the reference databases used in most DNA forensics in the UK. Please note that the IC population codes refer to our new 16-locus data set, while the EA codes refer to the historic 10-locus data set.

Filtering Out Possible Relatives

Pairwise allele sharing was measured in all subpopulations, counting only loci for which both individuals were genotyped and including all pairs of individuals that had at least four genotyped loci in common. If >75% of alleles were shared, the individual with the fewest loci typed was removed. For subpopulations with <100 individuals, the threshold for removal was reduced to 50% allele sharing.

Definition and Estimation of F_{ST}

There are various ways to define, estimate and interpret F_{ST} (Bhatia et al., 2013). The original definition (Wright, 1949) compared the variance of an allele fraction over subpopulations (S) to its variance in the total population (T):

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\overline{p}(1-\overline{p})}, \quad (1)$$

where $\overline{p}$ denotes the population allele fraction. The total population used in this formulation is usually a hypothetical ancestral population, from which observed subpopulations are assumed to have descended (Weir, 2001). However, in forensic work it is necessary to compare the subpopulation of a suspect with the population from which the available allele frequency database has been drawn. Thus, the reference population allele fractions are observed rather than inferred (Balding & Nichols, 1997). We will refer to these two approaches to estimation of F_{ST} as the indirect and direct methods, respectively.

Moment-based estimators of F_{ST} are widely used (Bhatia et al., 2013), but we take advantage of the benefits of likelihood-based estimation, which include high precision, correct accounting for sample size and interpretable intervals and quantiles (Balding, 2003, 2005). Weir & Hill (2002) proposed maximum likelihood estimation of F_{ST} using a normal approximation to the multinomial, but the multinomial-Dirichlet (Mosimann, 1962) provides a natural likelihood without a large-sample assumption. Given a locus with k distinct alleles, the multinomial-Dirichlet has $k-1$ parameters specifying the population allele fractions, which are replaced with observed values in the direct method and are unknown parameters in the indirect method. The remaining parameter λ specifies the variance, and $F_{ST} = 1/(1+\lambda)$. Throughout F_{ST} will be reported in percent.

Direct Method

The multinomial-Dirichlet likelihood is used for allele counts in a subpopulation, with reference allele fractions obtained from reference database counts, adjusted by adding a pseudocount of one for each allele in order to avoid zero values. The FSS databases EA1-6 are used as reference databases throughout. The direct analyses below only use the 10 loci in common between our data set and the historic FSS database, which are the loci with total allele counts $> 10^4$ (Table 1).

The likelihood curve for F_{ST} can automatically be interpreted as a posterior density with respect to a uniform prior. To formulate an informative prior, we noted previous work with small sample sizes (Balding & Nichols, 1997) suggesting that F_{ST} typically lies below 4%. Since more diverse subpopulations are considered here, we chose a beta prior distribution for F_{ST}, with median 2.3% and 95% credible interval (CI) from 0.26% to 8.0%.

To illustrate the effects of sample size, we performed direct estimation under both the uniform and beta priors using different sample sizes. Multinomial allele counts were simulated based on allele fractions that were Dirichlet-distributed, with means given by the EA4 allele fractions and $\lambda = 99$ so that $F_{ST} = 1\%$. The 95% CI includes 1% at all sample sizes, and becomes tighter as the sample size is increased (Fig. 2). For small sample sizes, the beta prior leads to slightly smaller posterior interval widths than the uniform, and the posterior median moves towards the prior value.

Figure 3 shows that the choice of prior has a noticeable effect on the posterior for Iran ($n = 13$), and less so for Afghanistan ($n = 42$), in both cases the informative prior shifts the F_{ST} posterior distribution to slightly higher values compared with the uniform prior.

Indirect Method and Locus Dependence

The direct method is the most appropriate for forensic applications because the role of the reference database in F_{ST} estimation matches its role in computing DNA profile likelihoods. The indirect method requires no reference database, so the 10-locus FSS databases are not used in these analyses

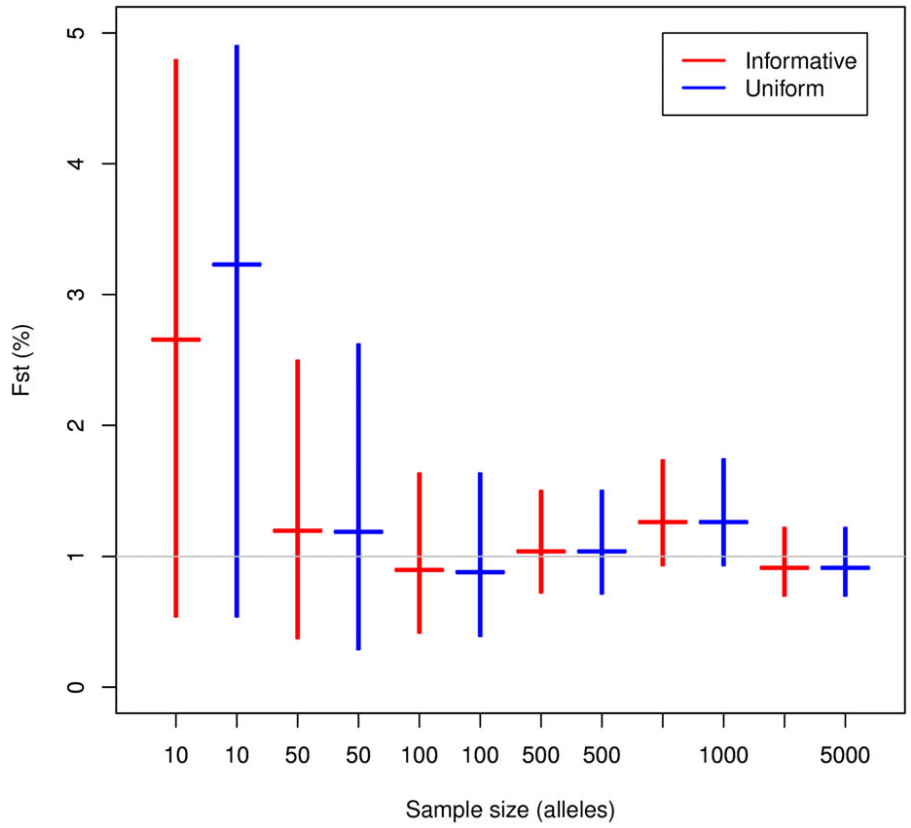


Figure 2 F_{ST} posterior 95% interval using: (red) a beta prior with median 2.3% and 95% CI (0.26%, 8.0%); (blue) the uniform prior. Sample sizes are shown on x-axis. Data were simulated to have $F_{ST} = 1\%$ (horizontal line). The vertical lines indicate the 95% equal-tailed CI, and medians are indicated with horizontal segments.

and we are thus able to utilise 15 of the 16 available loci (SE33 is excluded due to low sample sizes, Table 1).

In the indirect method, the reference population is not observed, but is assumed to be a hypothetical ancestral population from which two or more observed subpopulations have descended independently. We used the BayesFST software (Beaumont & Balding, 2004) which implements a Markov Chain Monte Carlo method to sample from the posterior distribution of F_{ST} in each subpopulation given the allele counts. BayesFST assigns a jointly uniform prior distribution to the ancestral allele fractions at each locus, and uses the model

$$F_{ST}^{i,j} = \frac{e^{a_i + b_j}}{1 + e^{a_i + b_j}}, \qquad (2)$$

where a_i and b_j denote locus and population effects, respectively. All inferences reported below are based on 150 000 posterior values.

We first investigated the variation of F_{ST} estimates across loci, treating IC1 through IC6 as six subpopulations of the hypothetical ancestral population. Each subpopulation parameter b_j was assigned an N(−3, 1.8) prior, while the locus parameters a_i were assigned an N(0,1) prior. The resulting prior distribution for F_{ST} has a prior median 4.7%, with 95% CI from 0.02% to 92%. Table 2 shows that the posterior 95% CI for the a_i include zero for 13 of the 15 loci. In view of this limited evidence for locus heterogeneity, we subsequently set the locus effect parameter to be close to zero in order to estimate an average F_{ST} over loci and hence allow greater comparability across analyses. The implied prior median is then 4.7%, with 95% CI from 0.1% to 63%.

We repeated all analyses with only the 10 loci used in the direct analyses, and confirmed that resulting inferences were similar, but on average more precise with 15 loci (10-locus results not shown). Thus, the differences reported below between direct and indirect F_{ST} values for a subpopulation are almost entirely due to the different reference population, rather than the different number of loci used.

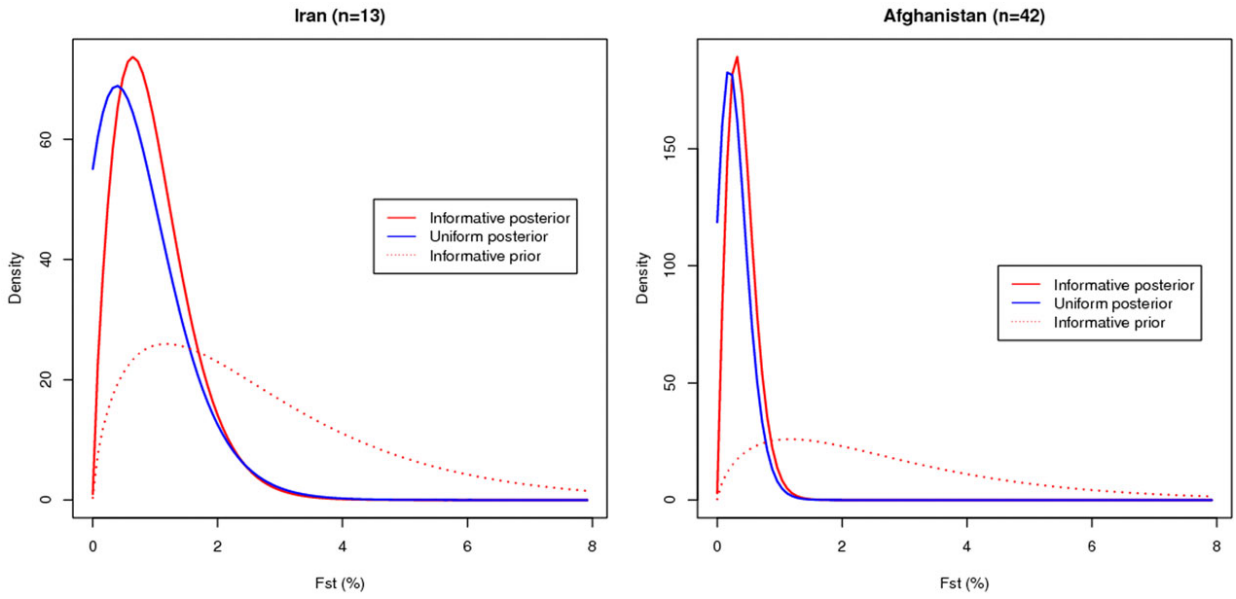


Figure 3 F_{ST} posterior densities (solid lines) using the direct method, given a uniform prior (blue) and an informative beta prior (red). Dotted red lines show the beta prior density. The subpopulations analysed are (left) Iran and (right) Afghanistan, with the reference populations being EA6 (Middle East/North Africa) and EA4 (South Asia), respectively.

Table 2 Posterior 95% intervals for locus effect parameters using the indirect method. The analysis used all 7121 individuals with IC1 through IC6 treated as six subpopulations.

	Percentile			Percentile	
Locus	2.5	97.5	Locus	2.5	97.5
D3	−1.72	−0.2	D19	−0.62	0.62
TH01	0.11	1.58	D2	−0.59	0.62
D21	−0.85	0.45	D22	−0.06	1.32
D18	−0.79	0.38	D1	−0.7	0.52
D16	−1.3	0.15	D10	−0.87	0.6
vWA	−0.93	0.42	D2	−0.21	1.15
D8	−0.73	0.6	D12	−0.71	0.56
FGA	−1.04	0.23			

Best Population Fit

Each subpopulation defined above was assigned to the FSS database giving the "best fit" (lowest median F_{ST} under the direct method), for both direct and indirect method analyses below. The majority of allocations were as expected: most European subpopulations fit best with EA1, most African and Caribbean subpopulations with EA3, all South Asian subpopulations fit best with EA4, both East Asian subpopulations fit best with EA5 and most Arab subpopulations fit best with EA6. Three subpopulations close to the Middle East fit EA6 equally or slightly better than their nominal population: Southern Europe (EA1), Afghanistan (EA4) and Kenya (EA3). The nominal classification was retained in each case.

One discrepancy was much larger: Somalia fit better with EA6 (F_{ST}=1.5%) than with the nominal EA3 (F_{ST}=2.2%), and we subsequently included Somalia with IC6. Although Somalia borders Kenya (EA3), it is also geographically close to the Arab world, and there have historically been many links. Mitochondrial (Mikkelsen et al., 2012) and Y-chromosome (Sanchez et al., 2005) studies have both suggested a strong Arab influence in Somali genetics, although their highest similarity is usually with neighbouring Eastern Ethiopians and Northern Kenyans. HLA typing (Mohamoud, 2006) also suggests that Somalis are more similar to Arabs than to sub-Saharan Africans. Pickrell et al. (2014) estimate the Eurasian ancestry of Somalis at roughly 38% using admixture mapping, supporting the low F_{ST} estimate for Somalia with the EA6 database.

RESULTS

EA1

When comparing subpopulations to the EA1 reference population (Table 3), all the European subpopulations have an F_{ST} estimate (97.5 percentile) under 1%, except Western Europe, which has the smallest sample size. The low F_{ST} estimate for Southern Europe supports the merging of European-origin IC2 individuals with IC1, suggesting that IC2 might usefully be redefined to only include Latin Americans with

Table 3 The 2.5, 50, and 97.5 posterior percentiles of F_{ST} (expressed as %). Subpopulations were compared both individually with the reference population EA1 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). *n* denotes the sample size (number of individuals).

		Direct			Indirect		
IC1	*n*	2.5	50	97.5	2.5	50	97.5
Eire	1949	0.1	0.2	0.2	0.0	0.0	0.1
Great Britain	1416	0.1	0.1	0.1	0.0	0.0	0.0
Eastern Europe	61	0.2	0.5	1.0	0.1	0.3	0.7
Northern Europe	45	0.0	0.3	0.8	0.0	0.2	0.5
Southern Europe	60	0.0	0.2	0.5	0.0	0.1	0.3
Western Europe	13	0.1	0.7	2.1	0.0	0.5	1.8
Anglo New World	13	0.1	0.5	1.7	0.0	0.3	1.4
Latin America	25	0.5	1.3	2.4	0.6	1.3	2.4

predominantly European ancestry. The Anglo New World has slightly lower estimates than Western Europe, but Latin America has a higher F_{ST} estimate, presumably due to admixture with non-European populations.

The indirect method gives lower F_{ST} estimates than the direct method, which is expected because the ancestral allele fractions are inferred to be towards the centre of the subpopulation values. However, the F_{ST} values for Latin America are almost unchanged and are again the highest, because inference of ancestral allele fractions is dominated by the European populations.

EA3

The mixed subpopulations of West, Central-Southern and East Africa, as well as Unknown IC3, have lower F_{ST} estimates under the direct method than the national subpopulations of Ghana, Kenya, Nigeria, and Sierra Leone. The F_{ST} estimate for other Caribbean is high, much higher than for Jamaica. Jamaicans have a predominantly African origin (Caribbean Community Capacity Development Programme, 2009), and there are approximately 800 000 people of Jamaican descent living in the UK (International Organisation for Migration, 2007), which is close to half the UK population categorised as black (Office for National Statistics, (2011)). Therefore the EA3 database may be expected to include a large number of Jamaicans.

Indirect estimation (Table 4b) gives noticeably different results than the direct method. In most cases they are greatly reduced, the exception being Kenya which is geographically remote from the majority of subpopulations, which are in West Africa or the Caribbean. We have noted above that Kenya fits almost equally well with both EA3 and EA6 using direct estimation, suggesting some genetic influence from the Arab world.

Table 4 The 2.5, 50, and 97.5 posterior percentiles of F_{ST} (expressed as %). Subpopulations were compared both individually with the reference population EA3 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). *n* denotes the sample size (number of individuals).

		Direct			Indirect		
IC3	*n*	2.5	50	97.5	2.5	50	97.5
Ghana	214	0.8	1.1	1.6	0.2	0.3	0.5
Jamaica	166	0.5	0.7	1.0	0.0	0.1	0.2
Kenya	51	0.7	1.2	1.9	0.8	1.3	1.9
Nigeria	444	0.9	1.2	1.5	0.2	0.3	0.3
Sierra Leone	41	0.7	1.3	2.2	0.1	0.3	0.8
Uganda	63	0.3	0.5	1.0	0.0	0.2	0.4
Unknown IC3	864	0.4	0.5	0.7	0.0	0.0	0.0
Other Caribbean	20	0.5	1.5	2.9	0.1	0.4	1.3
Other C/S Africa	55	0.3	0.6	1.1	0.0	0.1	0.3
Other E Africa	66	0.3	0.7	1.1	0.0	0.1	0.4
Other W Africa	48	0.1	0.5	1.0	0.0	0.1	0.3

Table 5 The 2.5, 50, and 97.5 posterior percentiles of F_{ST} (expressed as %). Subpopulations were compared both individually with the reference population EA4 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). *n* denotes the sample size (number of individuals).

		Direct			Indirect		
IC4	*n*	2.5	50	97.5	2.5	50	97.5
Afghanistan	47	0.1	0.3	0.9	0.1	0.4	0.9
Bangladesh	53	0.1	0.4	0.9	0.0	0.1	0.4
India	49	0.0	0.3	0.8	0.0	0.1	0.4
Pakistan	60	0.0	0.2	0.5	0.0	0.2	0.5
Unknown IC4	76	0.0	0.2	0.5	0.0	0.1	0.2

EA4, EA5, and EA6

For EA4 and EA5, the F_{ST} estimates are all low for both direct and indirect methods, with no outliers (Tables 5 and 6). The F_{ST} estimates for India and Bangladesh are much lower for the indirect than the direct method. The F_{ST} estimate for NE Asia is higher than that for SE Asia using the direct method, but lower using the direct method. This suggests the EA5 database largely consists of individuals from NE Asia.

Most IC6 subpopulations have low sample sizes, and so we will here discuss the posterior median of F_{ST} rather than the 97.5 percentile. Iraq has low F_{ST} estimates, much lower than its neighbour Iran (Table 7). Unsurprisingly, large F_{ST} estimates were obtained for Somalia. Results are largely congruent between the direct and indirect method, however, Turkey

Table 6 The 2.5, 50, and 97.5 posterior percentiles of F_{ST} (expressed as %). Subpopulations were compared both individually with the reference population EA5 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). *n* denotes the sample size (number of individuals).

		Direct			Indirect		
IC5	*n*	2.5	50	97.5	2.5	50	97.5
NE Asia	260	0.1	0.2	0.3	0.1	0.4	0.8
SE Asia	44	0.0	0.2	0.7	0.0	0.1	0.4

Table 7 The 2.5, 50, and 97.5 posterior percentiles of F_{ST} (expressed as %). Subpopulations were compared both individually with the reference population EA6 (direct method, 10 loci) and analysed jointly to infer ancestral allele fractions (indirect method, 15 loci). *n* denotes the sample size (number of individuals).

		Direct			Indirect		
IC6	*n*	2.5	50	97.5	2.5	50	97.5
Iran	12	0.1	0.9	2.4	0.1	0.9	2.7
Iraq	28	0.0	0.2	0.7	0.0	0.2	0.7
Somalia	494	1.1	1.3	1.7	1.2	1.6	2.1
Turkey	20	0.1	0.5	1.6	0.2	0.9	2.1
Middle East	24	0.1	0.7	1.8	0.1	0.5	1.6
N Africa	26	0.2	0.7	1.7	0.1	0.6	1.5

Table 8 Posterior median F_{ST} (%) for fringe subpopulations: These are subpopulations for which another reference population gives a median F_{ST} estimate using the direct method within 0.001 of the lowest (best fit) value.

	Reference				
Fringe	EA1	EA3	EA4	EA5	EA6
Afghanistan	1.17	2.90	0.78	1.87	0.78
Kenya	2.32	1.39	2.51	2.32	1.36
Southern Europe	0.30	2.99	1.20	2.03	0.34
Unknown IC4	1.68	2.80	0.62	1.17	0.72

Table 9 Posterior median F_{ST} (%):Populations IC1-6 were compared to each reference population in turn using the direct method. The indirect method was used to compare each population to a hypothetical global ancestral population.

		Reference					
Global	*n*	EA1	EA3	EA4	EA5	EA6	Indirect
IC1	3582	0.4	3.1	1.9	1.9	0.9	2.7
IC3	2032	1.7	0.7	1.7	1.4	1.1	1.0
IC4	285	1.4	3.1	0.7	1.3	0.8	2.3
IC5	304	3.1	4.2	2.4	0.5	2.0	3.3
IC6	604	1.8	1.7	1.9	1.7	0.9	1.4

has a larger F_{ST} estimate using the indirect method, which may be due to Turkish individuals being well represented in the EA6 database.

Fringe Regions

We use the term "fringe" for subpopulations that have similar affinity to two populations (difference in median F_{ST} <0.001). Broadly speaking these regions reflect an overall smooth change in allele frequencies with geography, so that the fringe regions are at the boundaries of our continental-scale populations (Table 8). Thus, Afghanistan is near the boundary between IC4 and IC6, and fits them approximately equally well, S Europe is at the boundary between IC1 and IC6, and Kenya is the IC3 country nearest to IC6. These results suggest a relatively low differentiation between IC6 and all three surrounding populations (IC1, IC3, IC4). Only IC5 is not linked to other populations through a fringe subpopulation, perhaps due to the mountains separating China from South Asia, and its geographical remoteness from IC1 and IC3. This agrees with a previous report that East Asian populations are distinct from those of South Asia, but are close to South East Asian populations (HUGO Pan-Asian SNP Consortium, 2009).

Inter-Population Comparisons

Above we have compared subpopulations with continental-scale reference populations, and now we make comparisons among those populations. Each column of Table 9 shows a different F_{ST} analysis of the five IC populations, using an EA database as the reference database in the direct method, or using the indirect method.

For the direct method, each IC database showed the best fit (lowest F_{ST} estimate) with its cognate EA database, reflecting a reasonable consistency of definitions between IC and EA databases. The highest F_{ST} value for IC1, IC4 and IC5 are all obtained relative to EA3. Conversely, looking down the columns of Table 9, IC5 shows the highest F_{ST} value for each EA database except EA5. The IC6 database is influenced by the large sample size from Somalia, and shows similar F_{ST} values with respect to all four EA databases other than EA6.

Using indirect estimation, IC3 and IC6 show the lowest F_{ST} values, while IC5 shows the highest value, corresponding to an inferred ancestral human population similar to that of modern North-East Africa (Pemberton et al., 2013).

Discussion

Although we have only examined 10 or 15 STR loci in this study, their multi-allelic nature and the large sample sizes for many subpopulations means that we have been able to achieve good precision in many of the F_{ST} estimates that we report, as indicated by the 95% posterior intervals. We have shown that F_{ST} estimates depend sensitively on the choice of reference population, and in particular that the use of a population reference database can generate very different F_{ST} estimates from those based on a hypothetical ancestral population, which is the usual practice in population genetic studies.

Silva et al. (2012) collated STR databases worldwide, and reported a global F_{ST} estimate from forensic data sets of 2.3%, comparable with inter-population estimates reported here (Table 9), while the corresponding estimate from the non-forensic Human Genome Diversity Project (HGDP) data set was more than twice as high, at 5.3%. Silva et al. suggest that this discrepancy is due to forensic markers being selected to have low differentiation among populations. However, they also demonstrate that selecting high heterozygosity markers decreases R_{ST}, and current forensic markers were selected in part to achieve high heterozygosity. The difference may also reflect larger and more ethnically mixed populations being included in forensic surveys, while the HGDP data set includes many ethnically distinct populations, often of small size.

Nelis et al. (2009) used the HapMap SNP database (before the upgrade to HapMap 3) to estimate continental genetic distance between Africa, Asia, and Europe. The F_{ST} values ranged from 11% (Europeans compared with Asians) to 19% (Africans compared with Asians), much higher than the STR-based estimates reported here and in Silva et al. (2012). This may be due to the high STR mutation rate (Weber & Wong, 1993) tending to stabilise allele fractions across populations, for example through mutations in short alleles tending to favour expansion, while contractions are favoured in long alleles (Sibly et al., 2003; Dupuy et al., 2004; Lu et al., 2012). Excoffier & Hamilton (2003) demonstrated that the discrepancy between F_{ST} estimates from SNP markers and those from STR markers can be removed by taking into account the stepwise mutation seen at STR markers. However, the broad pattern of variation is similar for STRs as for SNPs (Ramachandran et al., 2005; Pemberton et al., 2013).

One motivation for this research is to guide forensic practice, and overall we find that $F_{ST} \leq 3\%$ should be appropriate for most forensic calculations. The 97.5 posterior percentile for F_{ST} lies under 3% for all subpopulations relative to their best fit population, consistent with more limited previous results (Balding & Nichols, 1997; Gill et al., 2003). Low values can be justified in some settings, for example $F_{ST} =$ 1% appears adequate for Asians (both South and East), but $F_{ST} = 3\%$ would be more robust against incorrect assignment of reference population for an unknown contributor. In some cases it may be possible to tailor the F_{ST} value to specific circumstances, for example a lower F_{ST} value may be appropriate for alternative contributors who are known to be Jamaican, rather than from another Caribbean island.

Acknowledgements

Thanks to Sue Pope for informing us of the EA5 and EA6 databases held in the Forensic Archive. CDS is funded by a PhD studentship from the UK Biotechnology and Biological Sciences Research Council and Cellmark Forensic Services.

References

Balding, D. (2003) Likelihood-based inference for genetic correlation coefficients. *Theor Popul Biol* **63**, 221–230.

Balding, D. (2005) *Weight-of-Evidence for Forensic DNA Profiles*. New York: Wiley.

Balding, D. & Nichols, R. (1997) Significant genetic correlations among caucasians at forensic DNA loci. *Heredity* **78**, 583–589.

Beaumont, M. & Balding, D. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**, 969–980.

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. (2013) Estimating and interpreting *F*st: The impact of rare variants. *Genome Res* **23**, 1514–1521.

Caribbean Community Capacity Development Programme (2009) National census report 2001, Jamaica. Online. http://⟨?PMU ?⟩www.caricomstats.org/Files/Publications/NCR%20Reports/Jamaica.pdf.

Dupuy, B., Stenersen, M., Egeland, T., & Olaisen, B. (2004) Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Hum Mutat* **23**, 117–124.

Excoffier, L. & Hamilton, G. (2003) Comment on genetic structure of human populations. *Science* **300**, 1877.

Foreman, L. & Evett, I. (2001) Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system. *Int J Legal Med* **114**, 147–155.

Gill, P., Foreman, L., Buckleton, J., Triggs, C., & Allen, H. (2003) A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations. *Foren Sci Int* **131**, 184–196.

HUGO Pan-Asian SNP Consortium (2009) Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545.

International Organisation for Migration (2007) Jamaica mapping exercise. Online. http://www.iomlondon.org/doc/mapping/IOM_JAMAICA.pdf.

Lu, D., Liu, Q., Wu, W., & Zhao, H. (2012) Mutation analysis of 24 short tandem repeats in Chinese Han population. *Int J Legal Med* **126**, 331–335.

Mikkelsen, M., Fendt, L., Röck, A.W., Zimmermann, B., Rockenbauer, E., Hansen, A., Parson, W., & Morling, N. (2012) Forensic and phylogeographic characterisation of mtdna lineages from somalia. *Int J Legal Med* **126**, 573–579.

Mohamoud, A. (2006) P52 characteristics of hla class I and class II antigens of the somali population. *Transfus Med* **16**, 47–47.

Mosimann, J. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49**, 65–82.

Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskáčková, T., Balaščák, I., Peltonen, L., Jakkula, E., Rehnström, K., Lathrop, M., Heath, S., Galan, P., Schreiber, S., Meitinger, T., Pfeufer, A., Wichmann, H., Melegh, B., Polgár, N., Toniolo, D., Gasparini, P., D'Adamo, P., Klovins, J., Nikitina-Zake, L., Kučinskas, V., Kasnauskiene, J., Lubinski, J., Debniak, T., Limborska, S., Khrunin, A., Estivill, X., Rabionet, R., Marsal, S., Julià, A., Antonarakis, S., Deutsch, S., Borel, C., Attar, H., Gagnebin, M., Macek, M., Krawczak, M., Remm, M., & Metspalu, M. (2009) Genetic structure of Europeans: A view from the North–East. *PloS ONE* **4**, e5472.

Office for National Statistics (2011) Census: Aggregate data (England and Wales) [computer file]. UK Data Service Census Support. http://infuse.mimas.ac.uk.

Pemberton, T., DeGiorgio, M., & Rosenberg, N. (2013) Population structure in a comprehensive genomic data set on human microsatellite variation. *G3* **3**, 891–907.

Pickrell, J. K., Patterson, N., Loh, P. R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., & Reich, D. (2014) Ancient west eurasian ancestry in southern and eastern africa. *P Natl Acad Sci USA* **111**, 2632–2637.

Ramachandran, S., Deshpande, O., Roseman, C., Rosenberg, N., Feldman, M., & Cavalli-Sforza, L. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *P Natl Acad Sci USA***102**, 15942–15947.

Sanchez, J., Hallenberg, C., Børsting, C., Hernandez, A., & Morling, N. (2005) High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. *Eur J Hum Genet* **13**, 856–866.

Sibly, R., Meade, A., Boxall, N., Wilkinson, M., Corne, D., & Whittaker, J. (2003) The structure of interrupted human AC microsatellites. *Mol Biol Evol* **20**, 453–459.

Silva, N., Pereira, L., Poloni, E., & Currat, M. (2012) Human neutral genetic variation and forensic STR data. *PloS ONE* **7**, e49666.

Steele, C. & Balding, D. (2014) Statistical evaluation of forensic DNA profile evidence. *Annu Rev Stat Appl* **1**, 20–1.

United Nations Statistics Division (2014) Standard country and area codes classifications (m49). Online. http://unstats.un.org/unsd/methods/m49/m49regin.htm.

Weber, J. & Wong, C. (1993) Mutation of human short tandem repeats. *Hum Mol Genet* **2**, 1123–1128.

Weir, B. (2001) *Genetic Data Analysis II. Methods for Discrete Population Genetic Data*. Sunderland, MA: Sinauer Associates, Inc.

Weir, B. (2007) The rarity of DNA profiles. *Ann Appl Stat* **1**, 358–370.

Weir, B. & Hill, W. (2002) Estimating F–statistics. *Annu Rev Genet* **36**, 721–750.

Wright, S. (1949) The genetical structure of populations. *Ann Eugenic* **15**, 323–354.

Xu, X., Peng, M., Fang, Z., & Xu, X. (2000) The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* **24**, 396–399.

Received: 4 February 2014
Accepted: 22 July 2014

Science and Justice 54 (2014) 487–493

Contents lists available at ScienceDirect

Science and Justice

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

Choice of population database for forensic DNA profile analysis

Christopher D. Steele *, David J. Balding

UCL Genetics Institute, Darwin Building, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 17 June 2014
Received in revised form 6 September 2014
Accepted 20 October 2014

Keywords:
Population database
DNA mixtures
Likelihood ratio
Forensic DNA

ABSTRACT

When evaluating the weight of evidence (WoE) for an individual to be a contributor to a DNA sample, an allele frequency database is required. The allele frequencies are needed to inform about genotype probabilities for unknown contributors of DNA to the sample. Typically databases are available from several populations, and a common practice is to evaluate the WoE using each available database for each unknown contributor. Often the most conservative WoE (most favourable to the defence) is the one reported to the court. However the number of human populations that could be considered is essentially unlimited and the number of contributors to a sample can be large, making it impractical to perform every possible WoE calculation, particularly for complex crime scene profiles. We propose instead the use of only the database that best matches the ancestry of the queried contributor, together with a substantial F_{ST} adjustment. To investigate the degree of conservativeness of this approach, we performed extensive simulations of one- and two-contributor crime scene profiles, in the latter case with, and without, the profile of the second contributor available for the analysis. The genotypes were simulated using five population databases, which were also available for the analysis, and evaluations of WoE using our heuristic rule were compared with several alternative calculations using different databases. Using $F_{ST} = 0.03$, we found that our heuristic gave WoE more favourable to the defence than alternative calculations in well over 99% of the comparisons we considered; on average the difference in WoE was just under 0.2 bans (orders of magnitude) per locus. The degree of conservativeness of the heuristic rule can be adjusted through the F_{ST} value. We propose the use of this heuristic for DNA profile WoE calculations, due to its ease of implementation, and efficient use of the evidence while allowing a flexible degree of conservativeness.

1. Introduction

In forensic DNA analysis, unknown contributors to a DNA profile are usually considered to come from one of several populations for which an allele frequency database is available. The choice of database can have an important impact on weight of evidence (WoE): the rarer an allele the stronger the evidence implicating a queried contributor (Q) if he has that allele and it is observed in the crime scene profile (CSP). The most appropriate population is the one that best matches the ancestry of X, the true source of the DNA. Under the prosecution case X is assumed to be Q, but under the defence case there is often little or no information about the ancestry of X. Many authors have noted that the database most appropriate for Q is not necessarily most appropriate for X [6,4]. Conversely, [3] argue for using the database of Q even if the ancestry of X is unknown, in part because the observation of the profile of Q introduces a size-bias effect: an observed profile tends to be more common in the population in which it was observed than in a different population. Thus, having observed the profile of Q, on average the probability for X to have the same profile is higher if X is assumed to come from the same population.

In current forensic practice, when the ancestry of X is unknown, it is common to consider multiple population databases and choose the one that generates the lowest WoE. There should be no requirement to favour defendants in this way. Suppose for example that Q is Caucasian but it is discovered that the lowest WoE is obtained using a database of Vietnamese individuals. If the population local to the crime includes few Vietnamese and there is no evidence to suggest that a Vietnamese person was the source of the DNA, it may not be helpful to the court to report the WoE arising from the Vietnamese database. Similarly, the world's population can be categorised in a vast number of different ways, and it is not possible to investigate them all in order to report the smallest WoE. However, a forensic expert should make reasonable allowance for the different possible ancestries of X, given the available knowledge about the location and nature of the crime. It can be expedient to make approximations that favour the defence in order to permit simplified analyses while avoiding court-room challenges. Here we propose a heuristic for WoE analysis that involves only one calculation, using the database most appropriate for

Abbreviations: WoE, Weight of Evidence; Q, Queried contributor; X, Alternate contributor that replaces Q in defence hypothesis; K, Contributor to the CSP whose reference profile is available; U, Unprofiled contributor to the CSP.

* Corresponding author.
E-mail addresses: c.steele.11@ucl.ac.uk (C.D. Steele), d.balding@ucl.ac.uk (D.J. Balding).

http://dx.doi.org/10.1016/j.scijus.2014.10.004

Q. We show that our heuristic tends to strongly favour defences compared with a range of alternative calculations.

For a one-contributor CSP when there are only, say, five population databases, it is usually easy to compute the WoE for each database and choose the one most favourable to the defence. However, for mixed profiles, the computational effort to consider multiple databases for each unprofiled contributor can be substantial. Thus our heuristic that computes the WoE only using the database of Q would be attractive, provided that it can be established to be conservative (favourable to the defence). If X is from the same population as Q then it becomes relevant to consider that they may also come from the same subpopulation, in which case an F_{ST} adjustment may be required [3]. We have recently published worldwide F_{ST} estimates appropriate for forensic use [7] and concluded that choosing $F_{ST} = 0.03$ is sufficiently large to be almost always conservative. The effect of the F_{ST} adjustment is to increase the probability assigned to the alleles of Q, and consequently decrease the probability for other alleles. Although the rationale for an F_{ST} adjustment is to allow for the possibility that X has ancestry similar to that of Q, we illustrate below that for $F_{ST} = 0.03$ our heuristic calculation is conservative even if X could have come from one of several different populations. It is for this reason that our heuristic uses the same value of F_{ST} whatever the population of Q, even though within-population F_{ST} values differ across populations.

A similar argument applies to other contributors to a mixed CSP. Consider a two-contributor profile, one of the contributors being X, who is alleged to be Q. If the reference profile of the other contributor is known, as is often the case for a victim or bystander, there are no probabilities to assess for the alleles of that individual and so the question of the appropriate population database is essentially the same as for the one-contributor case. When the reference profile of the other contributor, say U, is unavailable, then we show that it is conservative to use for both X and U the database best matching the ancestry of Q, again with $F_{ST} = 0.03$. The F_{ST} adjustment under our heuristic only increases the population allele fraction for the alleles of Q, which is helpful to defences because it increases the probability that X or U share alleles with Q, thus increasing the support for the defence explanation of the observed CSP.

It is not feasible or desirable to guarantee that a proposed WoE calculation is more favourable to the defence than any conceivable alternative calculation. We perform simulation experiments which show that for UK population databases our heuristic WoE calculation is, with probability $\gg 0.99$, more favourable to defendants than a range of reasonable alternative calculations. We first simulate single-contributor CSPs matching the reference profile of the alleged contributor Q. Then the WoE for Q to be a contributor is calculated using the correct database (that used for the simulation) and is compared with the smallest WoE calculated using in turn four other databases. We repeated this exercise for one database using allele fractions that differ from the database values according to each of three values of F_{ST}, and show that our heuristic remains conservative compared to the WoE from the four alternative databases.

We then simulate two-contributor CSPs using all possible choices of two databases from the five available, and compare the WoE computed using the database of Q for both contributors (and $F_{ST} = 0.03$) with (a) the correct assignment of databases, (b) the minimum WoE using each of the four alternative databases for both X and U, and (c) the minimum WoE over the four databases for X, always using the correct database for U. In all our calculations, an adjustment using $F_{ST} = 0.03$ is applied to the alleles of Q when the database of Q is used for X.

When a calculation is performed using a database different from that of Q, perhaps because of evidence about the ethnic background of X, coancestry is not relevant and so it is appropriate to use $F_{ST} = 0$. It has been suggested [2] that even in this setting it would be cautious to use a low value of F_{ST} such as 0.01. This introduces some bias in favour of the defendant in order to allow for the ancestry of X to differ somewhat from the database population. Here we assume that there is no specific suggestion of an alternative population for X, and since a bias in favour of defendants is introduced by taking the minimum WoE over four alternative database choices, we use $F_{ST} = 0$ in calculations using databases different from that of Q.

It is possible that the true ancestry of Q is unknown or misassigned, for example if he impersonates another individual, or an assessment of his physical appearance was incorrect. He may also be of mixed ancestry or some other ancestry not well represented in the available databases. In that case there is no size-bias effect tending to make the observed profile of Q more common in the population to which he is assigned than in other populations. However, although such an error may have an adverse impact on the calculated WoE, the generous value of F_{ST} is the main factor underlying the conservative nature of the WoE analysis that we propose, and so the impact of any population misassignment of Q will be relatively small.

Table 1
Number of allele observations at each locus for each population database: Caucasian (IC1), Afro-Caribbean (IC3), South Asian (IC4), East Asian (IC5) and Middle Eastern (IC6).

Allele counts	IC1	IC3	IC4	IC5	IC6
D3S1358	6878	3941	520	599	1202
TH01	6816	3918	514	598	1202
D21S11	6870	3941	520	599	1199
D18S51	6808	3930	520	600	1195
D16S539	6818	3927	514	600	1199
VWA	6877	3936	520	600	1201
D8S1179	6871	3941	520	600	1202
FGA	6853	3938	516	600	1201
D19S433	6702	3868	507	595	1197
D2S1338	6443	3758	491	594	1176
D22S1045	1816	2482	421	498	954
D1S1656	1827	2508	426	504	959
D10S1248	1815	2499	416	500	912
D2S441	1800	2473	420	493	943
D12S391	1857	2543	437	499	945
SE33	368	872	237	394	268

2. Materials & methods

2.1. Databases

We have used frequency data at 16 STR loci for five UK populations: Caucasian (IC1), African and African Caribbean (IC3), South Asian (IC4), East Asian (IC5) and Middle Eastern (IC6) (Table 1). For further details of the dataset, see [7]. We used these data to simulate 16-locus profiles assuming Hardy–Weinberg and linkage equilibria. Neither dropin nor dropout are included in the simulations, nor are they allowed for in the analyses.

The WoE is computed using the likelihood ratio framework [5], and reported in bans ($= log_{10}$(likelihood ratio)) comparing a hypothesis that includes Q as a contributor with an alternative in which Q is replaced by X, assumed to be unrelated to Q. We implement F_{ST} adjustment [2] to the population fractions of the alleles of Q whenever the database most appropriate for Q is used for X; the adjustment uses

Table 2
Mean weight of evidence (WoE) for the heuristic rule and the alternatives discussed in the text. The mean of the differences between the heuristic and alternative scenarios is also shown. The % Difference row shows the mean difference as a percentage of the average of the heuristic and alternative means.

Contributors under Hd	X	X + K	X + U		
			True both	True U	Same dbase
Heuristic (bans)	20.3	17.8	10.7	10.7	10.7
Alternative (bans)	24.5	20.7	12.8	14.1	14.0
Difference (bans)	4.2	3.0	2.1	3.4	3.2
Difference (%)	18.8	15.6	17.9	27.4	25.9

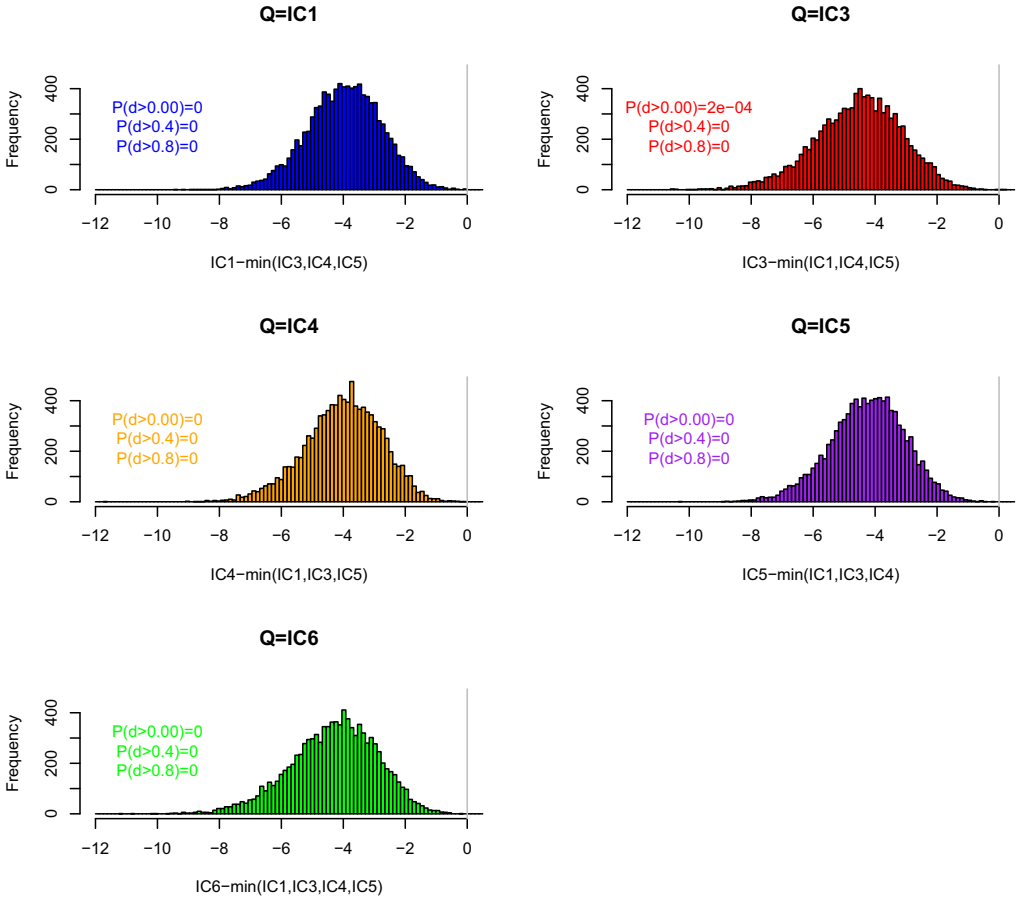


Fig. 1. The effect of database on weight of evidence (WoE) calculations for a one-contributor CSP. The databases are described in Table 1. The x-axis shows the WoE computed using the database from which the contributor Q was simulated (indicated in the subplot title) with $F_{ST} = 0.03$, minus the lowest WoE computed using each of the four alternative databases and $F_{ST} = 0$. P(d > x) indicates the proportion of differences that are > x.

$F_{ST} = 0.03$, and otherwise $F_{ST} = 0$. In all calculations one was added to the database count for each allele of Q, introducing a bias against understating the frequencies of rare alleles [1].

2.2. Simulation experiments

Initially a series of 10 000 one-contributor CSPs were simulated, using in turn allele fractions from each of the five population databases (so 50 000 profiles in total). The WoE for each simulated CSP was calculated five times, each time comparing hypotheses of the form:

H_p : Q
H_d : X

but using a different database. The minimum WoE over the four incorrect databases was then subtracted from the WoE computed using our heuristic (which uses the database of Q and $F_{ST} = 0.03$), so that a negative result indicates that it is favourable to the defence to report our heuristic WoE irrespective of the ancestry of X.

A second set of one-contributor analyses was conducted to investigate the effect of Q having an ancestry that differs from all of the available databases. Simulations were based on the IC1 database but with allele fractions differing from the IC1 values according to three F_{ST} values (0.01, 0.02, and 0.03). Ten thousand CSPs were simulated for each F_{ST} value (30 000 in total). The hypotheses compared were the same as above, and our heuristic was again applied (using the IC1 database) from which was subtracted the minimum WoE using each of the four other databases.

Next, 25 sets of 1000 two-contributor profiles were created, one for each choice of databases for the two contributors. The hypotheses compared were of the form:

H_p : Q + K
H_d : X + K

where K denotes that the second contributor was known (the reference profile was available for the analysis). The WoE computed using our heuristic was compared with the minimum WoE computed using each of the four alternative databases.

We then performed a series of analyses based on the same simulations but now assuming that the uncontested contributor to the two-contributor profiles was unknown, and so the hypotheses compared were of the form:

H_p : Q + U
H_d : X + U

For each dataset we computed the WoE using our heuristic with three alternative WoE calculations.

The first alternative WoE calculation used the correct database for each of X and U, which differs from our heuristic in the 20 datasets with Q and U simulated from different databases. This alternative may be regarded as the most appropriate WoE, while our heuristic WoE is biased in favour of the defence because the F_{ST} adjustment increases the probability for U to share alleles with Q. The second alternative, applicable in all 25 datasets, uses the lowest WoE obtained over all

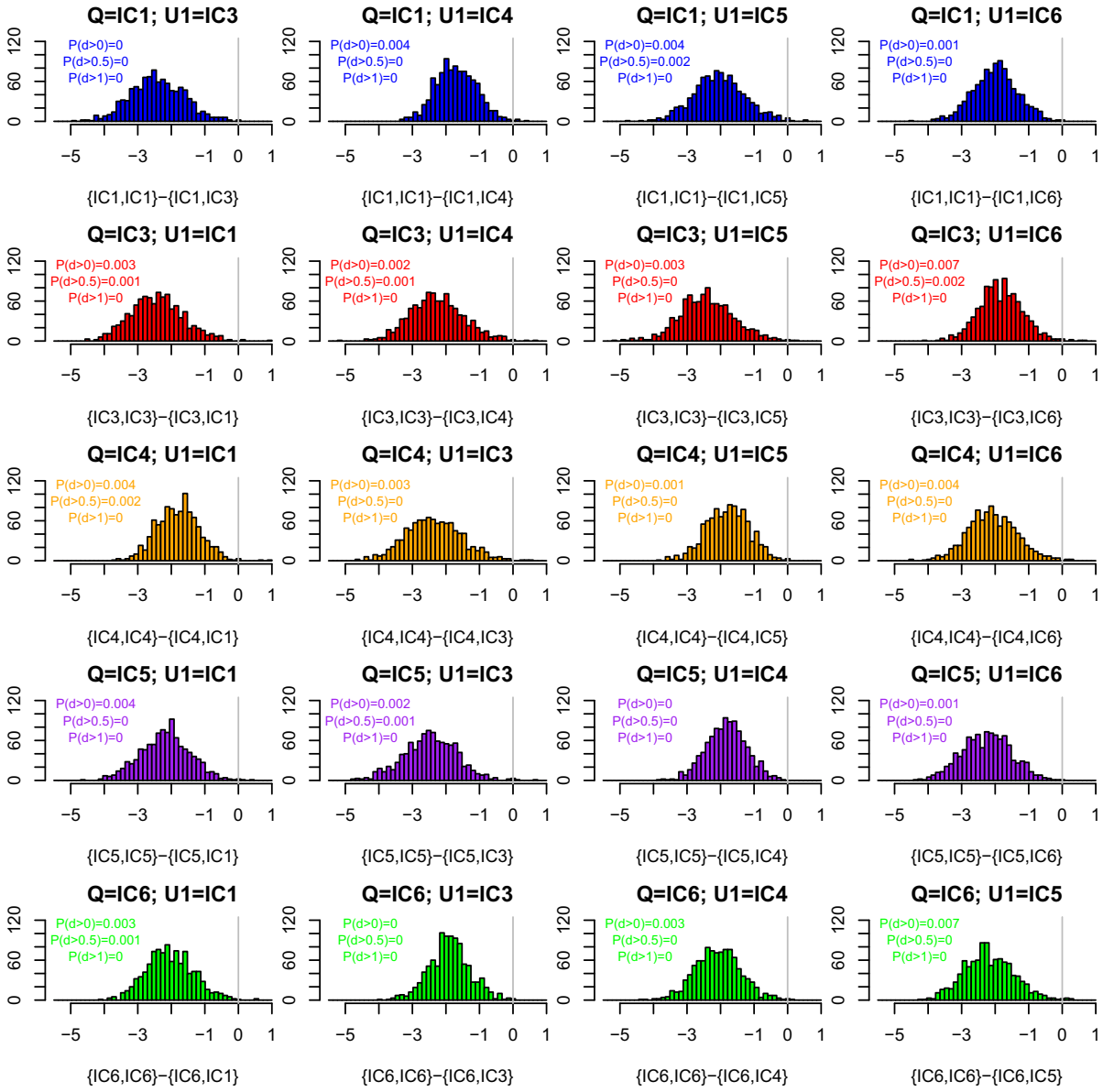


Fig. 2. The effect of database on weight of evidence (WoE) for two-contributor CSPs. The databases are described in Table 1. The x-axis shows the WoE computed using the database of Q for both contributors minus that obtained using the correct databases for X and U. The title of each subplot indicates the databases from which each contributor was simulated, where Q is the queried contributor and U is an unknown contributor. The x-axis labels indicate the databases used for each contributor in the analysis. P(d > x) indicates the proportion of differences that are > x. Colour indicates the database of Q.

possible databases for X, using the correct database for U. The third alternative WoE calculation, also applied in all 25 datasets, uses the lowest WoE over the four alternative databases, the same for both X and U.

3. Results

Table 2 shows summary results for the five simulation experiments. As the difficulty of the inference problem increases and the mean WoE decreases, the mean difference between our heuristic and the alternatives considered also decreases in absolute terms, but increases as a percentage of the alternative WoE. We have not considered three or more contributors because the computational demands of a large simulation study are prohibitive, but this trend suggests that for CSPs with three or more contributors, the mean difference between our heuristic and the alternative WoE would be a large percentage of the latter.

In one-contributor tests, our heuristic gives, with probability > 0.999, a lower WoE than any of the four alternative calculations (Fig. 1). There were two instances in 50 000 simulated profiles of an advantage to the defence from using one of the alternative databases. On average, the WoE obtained using our proposed calculation is lower by 0.3 bans (1 ban = 1 order of magnitude) per locus than the minimum over the four alternative calculations (Table 2, column 1). When the tested individuals are not simulated directly from the database allele frequencies, but differ according to $F_{ST} = 0.01$, 0.02 and 0.03, we found that the number of comparisons that are not conservative is at most 3 out of 10 000, which is as expected higher than when Q is simulated directly from the IC1 database (0 non-conservative out of 10 000) but the difference is small and not significant.

Including a known contributor reduces the WoE for both our heuristic and the minimum of the four alternatives by about 3 bans (Table 2, column 2). The difference between them remains similar to that for

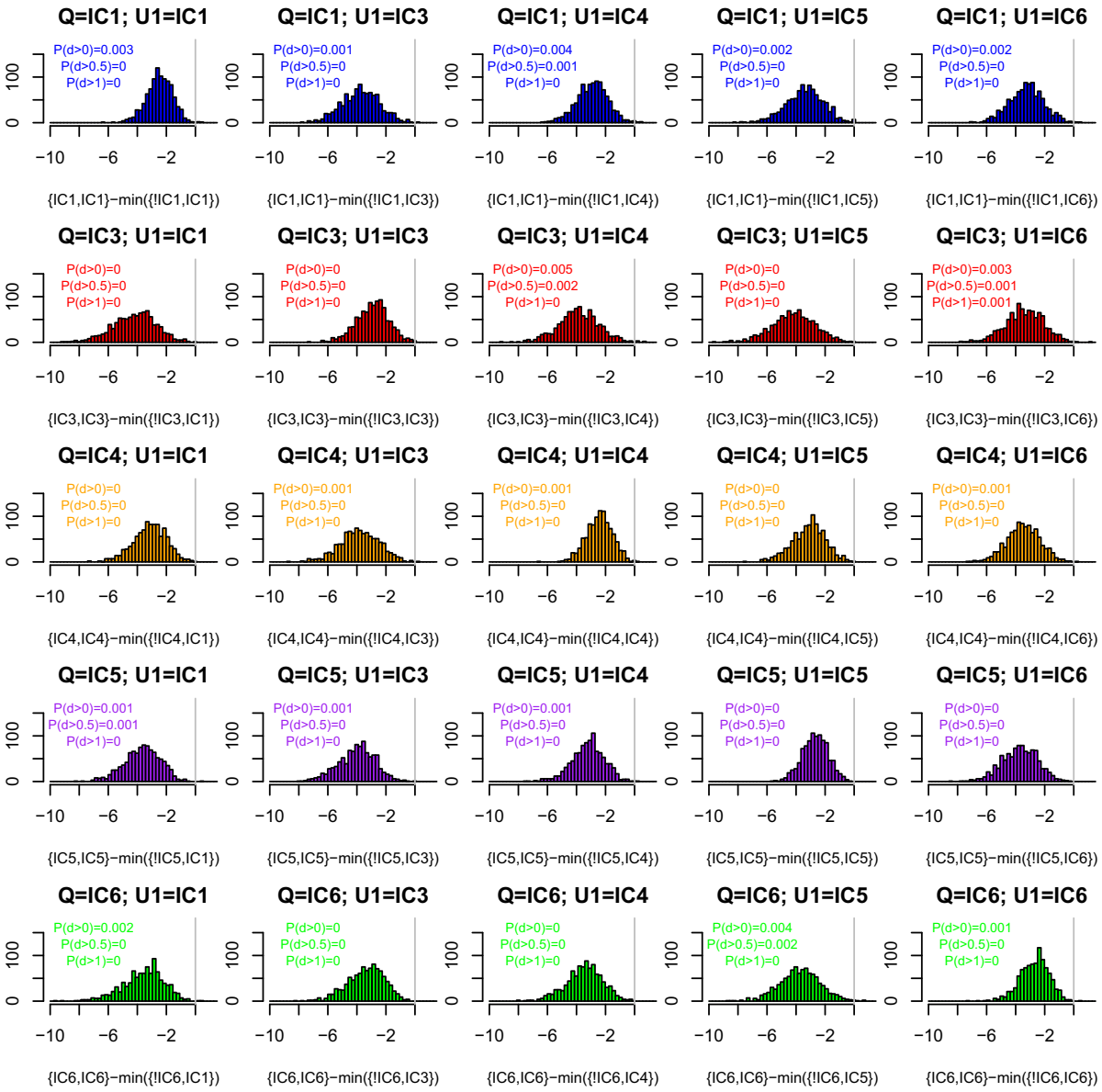


Fig. 3. The effect of database on weight of evidence (WoE) for two-contributor CSPs. The databases are described in Table 1. The x-axis shows the WoE computed using the database of Q for both contributors minus the minimum WoE obtained over all other choices of databases for X, always using the correct database for U. The title of each subplot indicates the databases from which each contributor was simulated. The x-axis labels indicate the databases used for each contributor in the analysis (!IC1 indicates all databases other than IC1). P(d > x) indicates the proportion of differences that are > x. Colour indicates the database of Q.

the one-contributor analyses (column 1). The fraction of simulations in which our heuristic was conservative ranged from 0.994 to 0.999 across the five databases used to simulate Q.

When the additional contributor is unknown (U rather than K), the fraction of simulations in which our heuristic was conservative compared with using the correct databases for each of X and U was on average 0.997, and at least 0.993 over the 20 choices of databases for X and U (Fig. 2). The reason that our heuristic is conservative is that it is helpful to the defence to maximise the probability that U has alleles matching those of Q, and this is achieved in our heuristic using the database of Q together with $F_{ST} = 0.03$. The probabilities assigned to alleles of U not shared with Q are less important because these have a similar effect under both prosecution and defence hypotheses. Using our heuristic, P(WoE > 9) = 0.903, and so the LR is usually but not always in excess of one billion.

Fig. 3 shows that the WoE computed under our heuristic is almost always ($P > 0.995$) less than the minimum value over the four alternative choices of database for X, with U always assigned the correct database. Finally, Fig. 4 shows that if the same database is used for both X and U, it is conservative ($P > 0.996$) to use our heuristic.

4. Discussion

We have shown that for a one-contributor setting, our heuristic WoE calculation that uses only the database of the queried contributor Q is almost always conservative (favours defences) compared with choosing the lowest WoE among four other databases for the alternative contributor X (Fig. 1). Similar results hold when there is additionally an uncontested contributor K with reference profile available. When the additional contributor is unprofiled, using the database of Q for both X and U is almost always conservative compared to (a) using the correct database for each of X and U (Fig. 2), (b) using the correct database for U, and choosing the most favourable alternative database for X (Fig. 3), and (c) choosing the most favourable among alternative

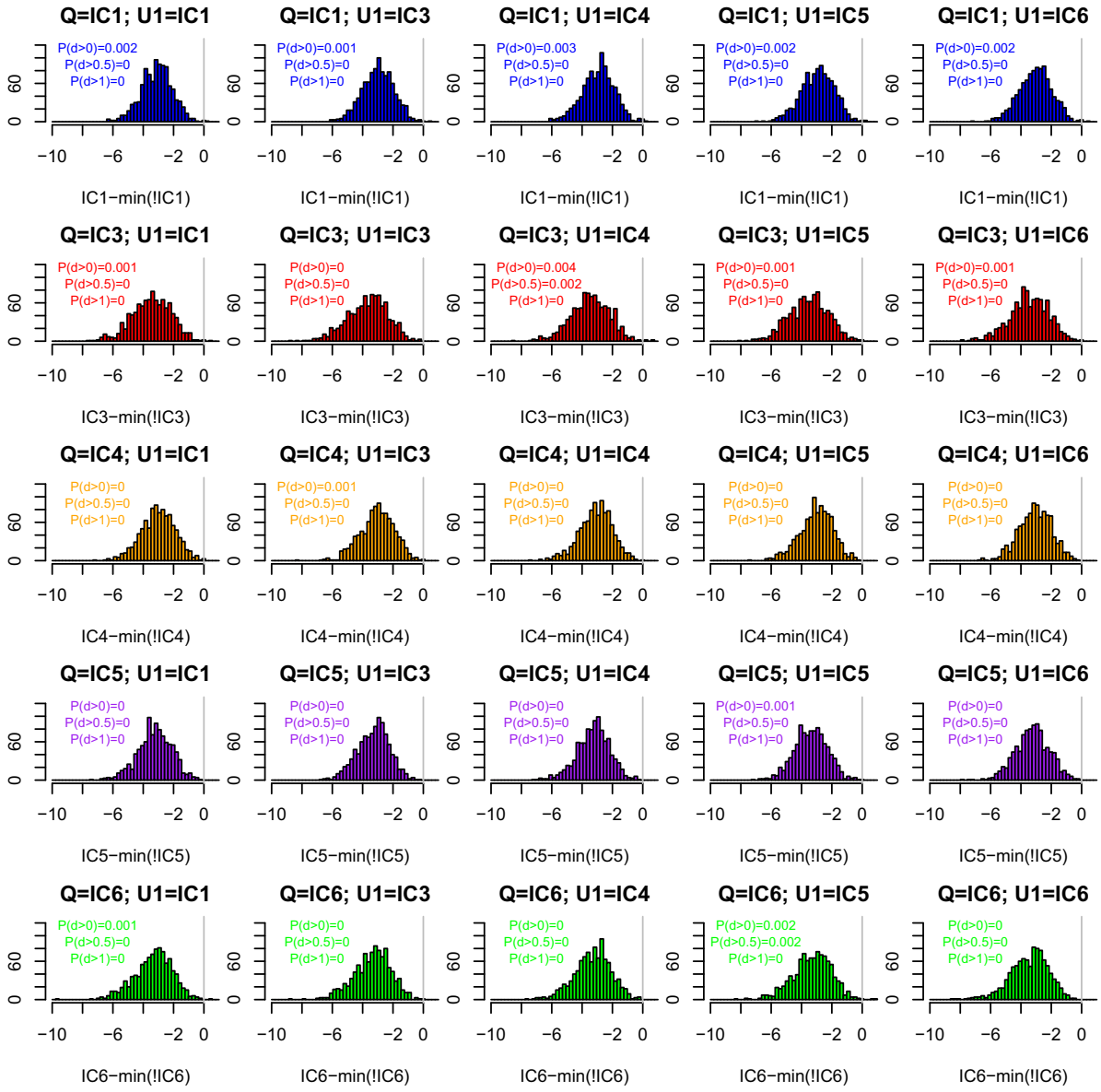


Fig. 4. The effect of database on weight of evidence (WoE) for two-contributor CSPs. The databases are described in Table 1. The x-axis shows the WoE computed using the database of Q for both contributors minus the minimum WoE obtained over using each other database in turn for both X and U. The title of each subplot indicates the databases from which each contributor was simulated. The x-axis labels indicate the database used for both contributors in the analysis. P(d > x) indicates the proportion of differences that are > x. Colour indicates the ancestry of Q.

databases, that database being used for both X and U (Fig. 4). In all calculations we used $F_{ST} = 0.03$ when the database of Q was used, and $F_{ST} = 0$ otherwise. The F_{ST} adjustment increases the population probabilities for the alleles of Q, but not other alleles observed in the CSP nor any other available reference profiles.

In all our simulations, our heuristic is conservative compared with the alternative calculations considered in at least 99.3% of the simulations, and in the few instances that it was not conservative the difference was always < 1.5 bans. The world's population can be divided into an unlimited number of different subpopulations; therefore there can be no precisely correct choice of alternative subpopulations to consider. What is required is an average WoE over each possibility for the ancestry of the alternative contributor X, weighted by its plausibility given the known circumstances of the crime. Our heuristic will almost certainly give a result that is more favourable to defendants. We have verified that our good results are not favourably biased because Q is sampled from the same database used in the analysis.

The degree to which our heuristic favours defences can be controlled by changing the value of F_{ST} from 0.03 used here. [7] found that $F_{ST} = 0.03$ exceeds almost all median F_{ST} values from world-wide comparisons of subpopulations with continental-scale populations, and we have shown here that it also suffices to ensure that using the database of Q with this F_{ST} value almost always returns a lower WoE than a range of alternative calculations.

We have not performed simulations for three or more unknown contributors because of the prohibitive time required for a simulation study, but the same principles apply to ensure that our heuristic will be similarly conservative. The large average reduction in WoE compared with the two-contributor case suggests that the difference in WoE will also be reduced, although it is expected to increase as a fraction of the overall WoE (see Table 2).

We believe that our heuristic offers a good policy for WoE calculation based on DNA evidence that is easy to implement, and almost always favourable to defendants relative to reasonable alternative policies.

Because it is favourable to defendants to use the database most appropriate for Q, it will therefore generally be unfavourable to defendants if the wrong database is used because the ancestry of Q is misassigned, or because there is no appropriate database. However the relatively large value of F_{ST} is the main factor in ensuring that our heuristic tends strongly to favour defendants, and so while misassigning the database of Q will have some detrimental impact, it will usually be small and outweighed by the impact of the F_{ST} value.

Acknowledgments

CDS is funded by a PhD studentship from the UK Biotechnology and Biological Sciences Research Council (BB/J012963/1) and Cellmark Forensic Services (CMD-PHD1).

References

[1] D.J. Balding, Estimating products in forensic identification using DNA profiles, J. Am. Stat. Assoc. 900 (431) (1995) 839–844 (0).
[2] D.J. Balding, Weight-of-evidence for Forensic DNA Profiles, Wiley. com, 2005.
[3] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 640 (2) (1994) 125–140 (0).
[4] L.A. Foreman, J.A. Lambert, I.W. Evett, Regional genetic variation in Caucasians, Forensic Sci. Int. 950 (1) (1998) 27–37 (0).
[5] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 1600 (2) (2006) 90–101 (0).
[6] National Research Council, The Evaluation of Forensic DNA Evidence, National Academies Press, Washington DC, 1996.
[7] C.D. Steele, D. Syndercombe Court, D.J. Balding, Worldwide F_{ST} estimates relative to five continental-scale populations, Ann. Hum. Genet. 78 (6) (November 2014) 468–477.

Evaluation of low-template DNA profiles using peak heights

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

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 modelling 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.

1 Keywords

Low-template DNA, DNA mixtures, likelihood ratio, peak heights, forensic, likeLTD

2 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 (Graversen and Lauritzen, 2014; Cowell et al., 2015; Bleka et al., 2016; Puch-Solis et al., 2013; Bright et al., 2013b; Perlin et al., 2011). 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:

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

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 defence 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 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 artefacts 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

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.	✓	✓	?	?	✓	×

Table 1: Summary of current software for evaluation of complex DNA profiles using peak heights. Distributions: Γ=gamma, $\log N$=lognormal, $\mathcal{N}$=truncated normal. Parameter elimination methods: maximisation (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 modelled 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 modelled but details are unknown, while question marks indicate that we are not aware if the phenomenon is modelled. For all other models ticks and crosses indicate that the phenomenon is or is not modelled. 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).

stutter from the 14 peak, and statistical modelling 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 modelling 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 modelling double-stutter. In addition, `likeLTD` has a more realistic dropin model: dropin is modelled 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).

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 modelling 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 $= \log_{10}(\text{LR})/\log_{10}(\text{IMP})$). IGR allows for easy comparison of LRs across different Q, as max(IGR) = 1.0 for every Q.

3 Materials & Methods

3.1 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 ($\rho_1 = 1$), 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.

Parameter	Distribution	Mean	SD
$\mathrm{E}[\alpha_l]$	N	0.013	0.010
$\log_{10}(\alpha_l/\mathrm{E}[\alpha_l])$	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

Table 2: Penalties applied to the parameters of the peak-height model. Distributions: N=normal, Γ=gamma, e=exponential. The degradation parameters δ have the same penalty for each contributor and for dropin.

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 penalised, 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 maximisation 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 penalised 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:

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

where $\mathcal{G}_{g,c}$ denotes the genotype allocated to c in g, π_l is the combined penalty on the likelihood at locus l given the values for α_l, D, O, σ and the δ, and θ denotes all model parameters. (4) is then maximised over these parameters. `likeLTD` uses a genetic algorithm `DEoptim` that simulates mutation, recombination and selection on parameter vectors to search for the vector that maximises the penalised likelihood (Mullen et al., 2011). Maximisation is performed separately under H_p and H_d and the LR is the ratio of the maximised values.

3.2 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

# 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)

Table 3: Laboratory protocol for generation of single-contributor and multi-contributor CSPs from 36 donated DNA samples. DNA masses are given as a total, with individual contributions in brackets. These are target values, realised values can vary.

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.

3.2.1 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 $\sigma = 10$. 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 $\beta_1 = -4.35$, as estimated by Tvedebrink et al., and $\beta_0 = 18.556$ 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.

3.2.2 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

# Contributors	Condition	Hypotheses	Figure
2	Single Rep	$U1$ + dropin	1(b), 3(a), 4(a), 4(b)
	Multiple Reps	$U1$ + dropin	3(a)
	Minor dropin	dropin	4(b)
3	Data fit	$K1$ (250pg) + $U1$	2
	Single Rep	$U1$ + $U2$	3(b), 4(b)
	Multiple Reps	$U1$ + $U2$	3(b)
	Minor dropin	$U1$ + dropin	4(b)

Table 4: Hypothesis pairs evaluated for the CSPs generated from the mixtures in Tables 3 and 6. 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 16pg contributor. For other conditions, each contributor was queried in turn.

Designation	S	D and O
Non-allelic	$x < 0.05$	$x < 0.05$
Uncertain	$0.05 \leq x < 0.15$	$0.05 \leq x < 0.1$
Allelic	$x \geq 0.15$	$x \geq 0.1$

Table 5: Interpretation rules for epg peaks in positions that could correspond to stutter (S), double-stutter (D) or over-stutter (O). 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.

per 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 3kV 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.

3.2.3 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 5 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.

3.3 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 6). PCR amplification, capillary electrophoresis and genotype analysis were performed for each replicate as described above.

# 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)

Table 6: Experimental design for investigating the relative merits of pre-extraction replication. 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 6) 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).

3.4 Model extensions

3.4.1 Minor contributors modelled 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 hypothesised 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 hypothesised to have some contribution from dropin.

3.4.2 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:

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

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 defence.

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 6 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 4 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 1, and a Caucasian population database for all unknown contributors (Steele and Balding, 2014; Steele et al., 2014b).

4 Results

4.1 Model validation

4.1.1 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 1(a) 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 uses no information that could distinguish the alleles of the two contributors. The right panel of Figure 1(a) 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 recognise 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.

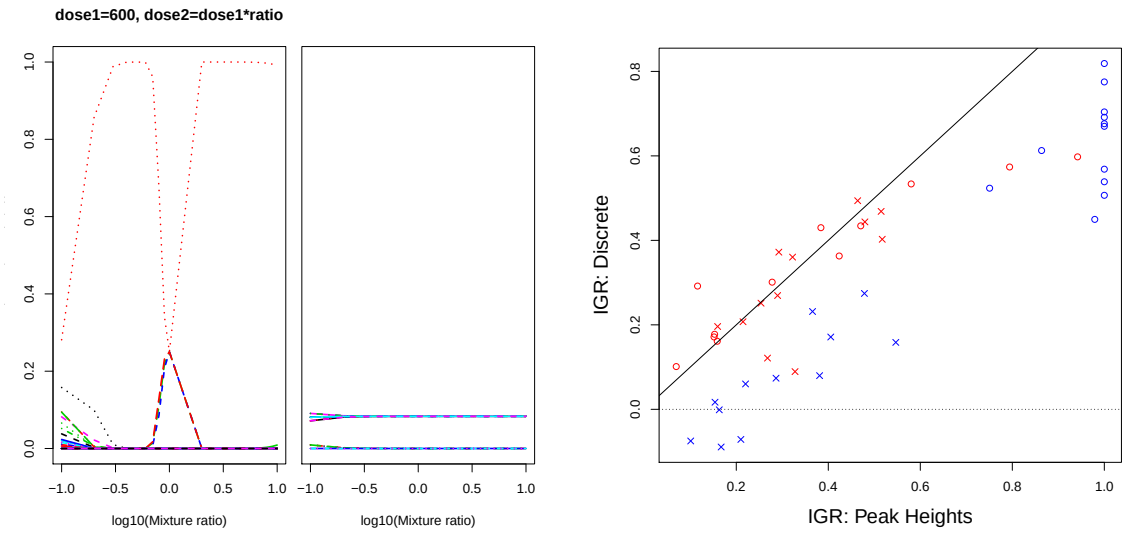
4.1.2 Laboratory data: model fit

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 4, 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.

4.1.3 Comparison with discrete model

Despite the superiority of the peak-height model in a simplified setting with no peak-height variability (Figure 1(a)), 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 1(b), red). In effect, the additional information potentially available from peak heights is lost due to peak-height variability at the low template used here (31 pg per contributor). Even the two red x in 1(b) 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 1(b), blue). In four cases the peak height IGR for the minor (crosses) supports



(a) Theoretical

(b) Laboratory

Figure 1: Simulated two-person mixtures. **(a)** 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 $\times$ 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)** 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.

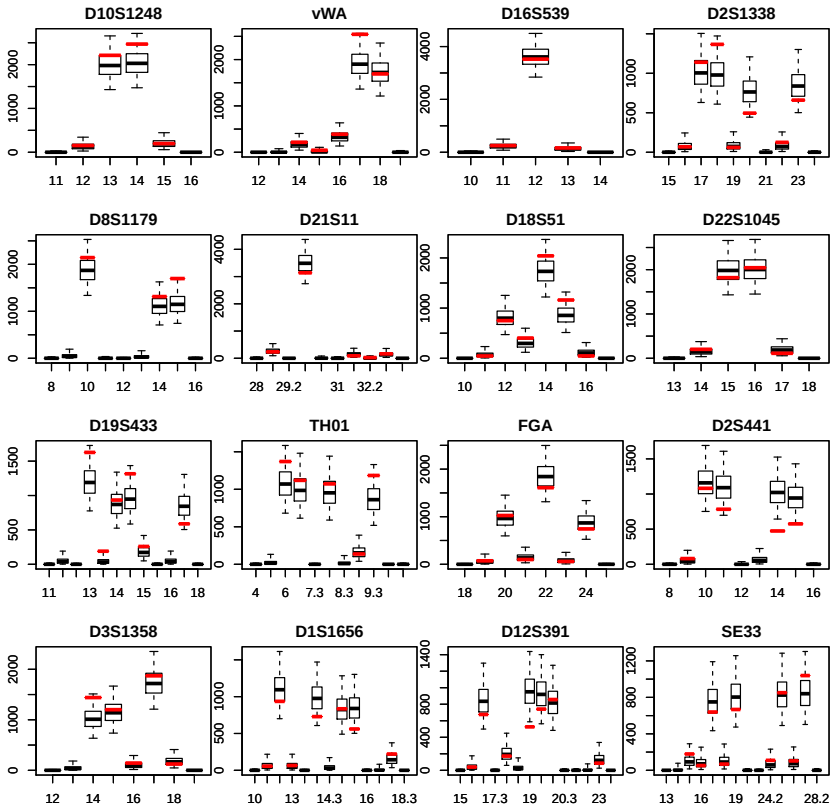


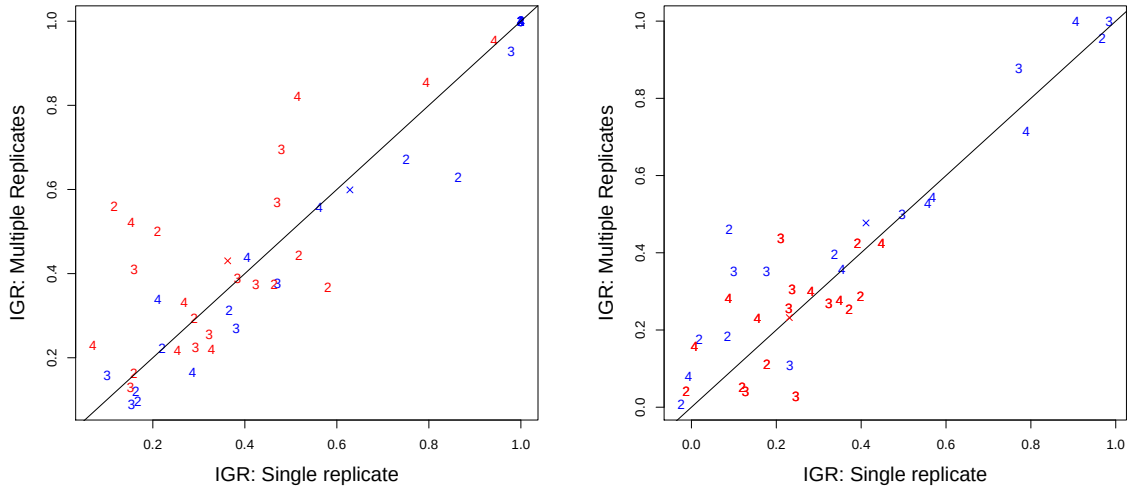
Figure 2: Observed and fitted peak heights under H_d for a CSP assuming a 250 pg K and two unknown contributors. Boxes show the central 50% (inter-quartile range) of the gamma distribution for each hypothesised peak, whiskers represent the 95% equal-tailed probability interval and red bars show observed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

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.

4.2 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 3(a)) and three-contributor CSPs (Figure 3(b)). These results show that with efficient statistical analysis splitting a sample to achieve replication does not lose information. We discuss potential advantages below.

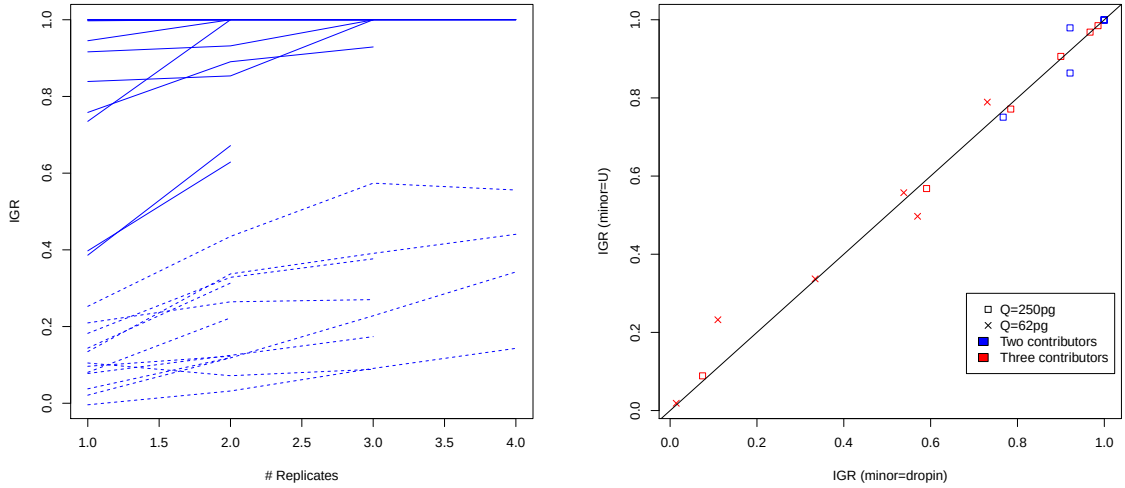
If replication is “free" in the sense of not exhausting the supply of DNA then it is potentially always advantageous. However, there are costs involved and a declining return from additional replicates. Figure



(a) Two contributors

(b) Three contributors

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 colour. Each of the contributors was queried in turn, leading to 48 and 36 data points.



(a) Replicates unequal contributions

(b) Minor as dropin

Figure 4: Information gain ratio (IGR) for **(a)** twelve 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).

4(a) 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.

4.3 Model extensions

4.3.1 Minor contributors modelled 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 modelled as dropin (Figure 4(b)). Because it can be difficult to decide whether additional low-level peaks in an epg should be modelled 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.

4.3.2 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 unfavourable 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.

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

Table 7: WoE (Weight of Evidence = $\log_{10}$(LR)) and $\log_{10}$(IMP) (IMP = Inverse Match Probability) 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. 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).

5 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 modelling 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 modelled, 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 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 minutes, while the 36 three-contributor single-replicate evaluations ranged from 18 to 200 minutes.

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 4(a)), 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 modelling as dropin any low-level contributors not of interest to the investigation (Figure 4(b)). Conceptually, dropin is modelled 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 (Manabe et al., 2013; Haned et al., 2011), 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 modelled as dropin it is unnecessary to specify the number of contributors exactly.

Not adjusting for linked loci tends to favour prosecutions, while the degree that removing one locus from linked pairs favours defences 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., 2015). 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 modelling 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.

6 Bibliography

Balding, D. J. (2013). Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proceedings of the National Academy of Sciences*, 110(30):12241–12246.

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

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

Benschop, C. C. G., Yoo, S. Y., and Sijen, T. (2015). Split DNA over replicates or perform one amplification? *Forensic Science International: Genetics Supplement Series*, 5:e532–e533.

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

Bright, J.-A., Curran, J. M., and Buckleton, J. S. (2013a). Relatedness calculations for linked loci incorporating subpopulation effects. *Forensic Science International: Genetics*, 7(3):380–383.

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

Bright, J.-A., Taylor, D., Curran, J. M., and Buckleton, J. S. (2013b). Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Science International: Genetics*, 7(2):296–304.

Brookes, C., Bright, J.-A., Harbison, S., and Buckleton, J. (2012). Characterising stutter in forensic STR multiplexes. *Forensic Science International: Genetics*, 6(1):58–63.

Buckleton, J. and Curran, J. (2008). A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Science International: Genetics*, 2(4):343–348.

Champod, C. (2013). Dna transfer: informed judgment or mere guesswork? *Frontiers in Genetics*, 4:300.

Cowell, R. G., Graversen, T., Lauritzen, S. L., and Mortera, J. (2015). Analysis of forensic DNA mixtures with artefacts. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 64(1):1–48.

Dørum, G., Kling, D., Tillmar, A., Vigeland, M. D., and Egeland, T. (2015). Mixtures with relatives and linked markers. *International Journal of Legal Medicine*, pages 1–14.

Gill, P., Brenner, C. H., Buckleton, J. S., Carracedo, A., Krawczak, M., Mayr, W. R., Morling, N., Prinz, M., Schneider, P. M., and Weir, B. S. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, 160(2):90–101.

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

Gill, P., Gusmão, L., Haned, H., Mayr, W. R., Morling, N., Parson, W., Prieto, L., Prinz, M., Schneider, H., Schneider, P. M., and Weir, B. S. (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 Science International: Genetics*.

Gill, P. and Haned, H. (2013). A new methodological framework to interpret complex DNA profiles using likelihood ratios. *Forensic Science International: Genetics*, 7(2):251–263.

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

Good, I. J. (1950). Probability and the Weighing of Evidence.

Graversen, T. and Lauritzen, S. (2014). Computational aspects of DNA mixture analysis. *Statistics and Computing*, 25(3):527–541.

Haned, H., Pene, L., Lobry, J. R., Dufour, A. B., and Pontier, D. (2011). Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? *Journal of Forensic Sciences*, 56(1):23–28.

Kelly, H., Bright, J.-A., Buckleton, J. S., and Curran, J. M. (2014). Identifying and modelling the drivers of stutter in forensic DNA profiles. *Australian Journal of Forensic Sciences*, 46(2):194–203.

Manabe, S., Kawai, C., and Tamaki, K. (2013). Simulated approach to estimate the number and combination of known/unknown contributors in mixed DNA samples using 15 short tandem repeat loci. *Forensic Science International: Genetics Supplement Series*, 4(1):e154–e155.

McCord, B. R., Jung, J. M., and Holleran, E. A. (1993). High resolution capillary electrophoresis of forensic DNA using a non-gel sieving buffer. *Journal of Liquid Chromatography & Related Technologies*, 16(9-10):1963–1981.

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., Cline, J., et al. (2011). DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software*, 40(6):1–26.

Nathakarnkitkool, S., Oefner, P. J., Bartsch, G., Chin, M. A., and Bonn, G. K. (1992). High-resolution capillary electrophoretic analysis of DNA in free solution. *Electrophoresis*, 13(1):18–31.

Perlin, M. W., Legler, M. M., Spencer, C. E., Smith, J. L., Allan, W. P., Belrose, J. L., and Duceman, B. W. (2011). Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*, 56(6):1430–1447.

Puch-Solis, R., Rodgers, L., Mazumder, A., Pope, S., Evett, I., Curran, J., and Balding, D. (2013). Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Science International: Genetics*, 7(5):555–563.

Ruiz-Martinez, M. C., Salas-Solano, O., Carrilho, E., Kotler, L., and Karger, B. L. (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. *Analytical Chemistry*, 70(8):1516–1527.

Steele, C. D. and Balding, D. J. (2014). Choice of population database for forensic DNA profile analysis. *Science & Justice*, 54(6):487–493.

Steele, C. D., Greenhalgh, M., and Balding, D. J. (2014a). Verifying likelihoods for low template DNA profiles using multiple replicates. *Forensic Science International: Genetics*, 13:82–89.

Steele, C. D., Syndercombe Court, D., and Balding, D. J. (2014b). Worldwide F_{ST} estimates relative to five continental-scale populations. *Annals of Human Genetics*.

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

Taylor, D., Buckleton, J., and Evett, I. (2015). Testing likelihood ratios produced from complex DNA profiles. *Forensic Science International: Genetics*, 16:165–171.

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

Williams, P. E., Marino, M. A., Del Rio, S. A., Turni, L. A., and Devaney, J. M. (1994). Analysis of DNA restriction fragments and polymerase chain reaction products by capillary electrophoresis. *Journal of Chromatography A*, 680(2):525–540.