Forensic Science International: Genetics 18 (2015) 100–117

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches—Twenty years of research and development

Peter Gill [a,b,*], Hinda Haned [c], Oyvind Bleka [a], Oskar Hansson [a], Guro Dørum [d], Thore Egeland [a,d]

[a] Norwegian Institute of Public Health, Department of Forensic Biology, PO Box 4404 Nydalen, 0403 Oslo, Norway
[b] Department of Forensic Medicine, Sognsvannsveien 20, Rikshospitalet, 0372 Oslo, Norway
[c] Netherlands Forensic Institute, Department of Human Biological Traces, The Hague, The Netherlands
[d] Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Aas, Norway

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form 19 March 2015
Accepted 24 March 2015

Keywords:
Multiplexes
European standard set of loci
National DNA databases
Complex mixtures
Kinship
STRs

ABSTRACT

The introduction of Short Tandem Repeat (STR) DNA was a revolution within a revolution that transformed forensic DNA profiling into a tool that could be used, for the first time, to create National DNA databases. This transformation would not have been possible without the concurrent development of fluorescent automated sequencers, combined with the ability to multiplex several loci together. Use of the polymerase chain reaction (PCR) increased the sensitivity of the method to enable the analysis of a handful of cells. The first multiplexes were simple: 'the quad', introduced by the defunct UK Forensic Science Service (FSS) in 1994, rapidly followed by a more discriminating 'six-plex' (Second Generation Multiplex) in 1995 that was used to create the world's first national DNA database. The success of the database rapidly outgrew the functionality of the original system – by the year 2000 a new multiplex of ten-loci was introduced to reduce the chance of adventitious matches. The technology was adopted world-wide, albeit with different loci. The political requirement to introduce pan-European databases encouraged standardisation – the development of European Standard Set (ESS) of markers comprising twelve-loci is the latest iteration. Although development has been impressive, the methods used to interpret evidence have lagged behind. For example, the theory to interpret complex DNA profiles (low-level mixtures), had been developed fifteen years ago, but only in the past year or so, are the concepts starting to be widely adopted. A plethora of different models (some commercial and others non-commercial) have appeared. This has led to a confusing 'debate' about the 'best' to use. The different models available are described along with their advantages and disadvantages. A section discusses the development of national DNA databases, along with details of an associated controversy to estimate the strength of evidence of matches. Current methodology is limited to searches of complete profiles – another example where the interpretation of matches has not kept pace with development of theory. STRs have also transformed the area of Disaster Victim Identification (DVI) which frequently requires kinship analysis. However, genotyping efficiency is complicated by complex, degraded DNA profiles. Finally, there is now a detailed understanding of the causes of stochastic effects that cause DNA profiles to exhibit the phenomena of drop-out and drop-in, along with artefacts such as stutters. The phenomena discussed include: heterozygote balance; stutter; degradation; the effect of decreasing quantities of DNA; the dilution effect.

1. Introduction

Short tandem repeat (STR) analysis was first introduced into forensic casework 20 years ago. The ability to combine several markers to form multiplexes and to subsequently visualise the results by automated fluorescent sequencing made National DNA databases feasible. The first example was launched in 1995 by the defunct Forensic Science Service (FSS). This 'twentieth anniversary' review is a perspective on the development and the subsequent worldwide adoption of STRs that has taken place over the previous two decades. The review is European-centred and is structured into several sections:

* Corresponding author at: Norwegian Institute of Public Health, Department of Forensic Biology, PO Box 4404 Nydalen, 0403 Oslo, Norway. Tel.: +47 7786126571.
E-mail address: peterd.gill@gmail.com (P. Gill).

http://dx.doi.org/10.1016/j.fsigen.2015.03.014
1872-4973/

The scientific societies in Europe and North America played a crucial role in coordinating standardisation, education and training to facilitate uptake. This work continues to the present day (Section 2). Section 3 begins a historical analysis that describes the key research and developmental advances that went towards designing and standardising multiplexes. In total there have been three iterations of multiplexes, e.g. in many European countries the six-loci SGM system was replaced by the ten-loci SGM-plus in 1999 to improve discriminating power. As databases expand, in conjunction with political initiatives to introduce massive pan-European databases, this in turn drives the requirement for ever more powerful multiplex systems. This continuing need recently led to the implementation of a new European standard set (ESS) markers of twelve-loci (Section 5) that has been adopted by the European Commission following recommendations of the European Network of Forensic Science Institutes (ENFSI). Several commercial companies, who work closely with the scientific societies, now provide new multiplex systems to the community on the basis of these recommendations. Practically speaking there are sixteen loci, since D16S539, D19S433, S2S1338 and SE33 are all included in addition to the ESS markers.

New biochemistry has simultaneously increased the sensitivity of tests, to the extent that the once controversial low-level or low-template (LT-) DNA analysis is considered to be routine (Section 7). However, this is not without challenge. LT-DNA profiles tend to be complex mixtures, with problems of 'missing alleles', known as drop-out. New statistical methods, based on likelihood ratio (LR) estimation have been critical to improve the interpretation (Section 9). A number of different solutions have been proposed and implemented. There is no 'gold standard' or any preferred method, since each has its own advantages and disadvantages that are listed. It is usually assumed that contributors to crime-stains are unrelated. However, this is not always the case. Methods have been developed to analyse mixtures (that may also be low-level) where the contributors may be related e.g. sibs (Section 10).

The new statistical theory also extends to (and improves the efficiency) of national DNA database interrogation (Section 11). It is no longer necessary to think in terms of simple matching or non-matching profiles; since likelihood ratios can be calculated for every single member of a database (illustrated examples are based on the UK database size of 5 million reference profiles). If large databases are searched for potential suspects, there is increased danger of false-positive matches and false-negative non-matches. Forensic scientists are urged to consider the DNA evidence in full context of the non-DNA evidence in court reports.

Finally, there is a comprehensive discussion in Section 16 on the characteristerisation of DNA profiles where the causal reasons for stochastic effects are explored. Stochastic effects are primarily visualised as heterozygote balance (imbalance) and stutters of variable size. Logistic regression is used to measure probability of drop-out relative to allelic peak heights. There is a description of the utilisation of software to predict and to simulate DNA profiles based on DNA quantity, amount pipetted, PCR efficiency and extraction efficiency.

2. The role of the European scientific societies in the evolution of STR-DNA profiling

Within Europe, there are two predominant scientific societies that have a special interest in DNA profiling: the oldest is the International Society of Forensic Genetics (ISFG) which dates from 1968. This society is also the home of the DNA Commission. The DNA commission comprises a peer review body of recognised experts from all over the world (not just Europe) who regularly meet to discuss and to formulate recommendations relating to new techniques or areas that may be controversial. Foundation studies includes Y chromosome STR analysis [1]; the interpretation of mixtures [2]; and the interpretation of low-template STRs [5]. A full list of publications is to be found on the ISFG web site http://www.isfg.org/Publications. These publications are reflective and define the consensus view of the forensic community – consequently, they are an important source of potential court-going documents.

Also under the ISFG umbrella is the European DNA Profiling Group (EDNAP) http://www.isfg.org/ednap/ednap.htm This group came into being in 1988. Currently there are representatives from 17 European countries. The group is very active and practically orientated. It was responsible for originally recognising the potential of short tandem repeat (STR) analysis and was the first to demonstrate uniformity of results across different laboratories. The STRs and methods originally developed by EDNAP have since become acknowledged as worldwide standards.

The DNA working group of the European Network Forensic Science Institutes (ENFSI) http://www.enfsi.eu/about-enfsi/structure/working-groups/dna first met in 1995. This is probably the largest group with more than 30 European countries and close US and Australian/New Zealand links. The group is involved with several different areas: database legislation, development of sampling kits, training, standards for the ENFSI QA programme; methods, analysis and interpretation of evidence; a European population database was developed http://www.str-base.org

3. Historical development of multiplexed systems

Early multiplexes consisted of relatively few loci based on simple STRs. The four locus 'quadruplex' was the first multiplex to be used in casework, and was developed by the Forensic Science Service (FSS) [4]. Because it consisted of just four STRs, there was a high chance of a random match – approximately 1 in 10,000. In 1995, the FSS re-engineered the multiplex, producing a 6 locus STR system combined with the amelogenin sex test [5]. This acquired the name 'second generation multiplex' (SGM). The addition of complex STRs: D21S11 and HUMFIBRA/FGA [6], which have greater variability than simple STRs, decreased the chance of a random match to about 1 in 50 million. In the UK, the introduction of SGM in 1995 facilitated the implementation of the UK national DNA database (NDNAD) [7]. As databases become much larger, the number of pairwise comparisons increases dramatically, so it became necessary to ensure that the match probability of the system was sufficient to minimise the chance of two unrelated individuals matching by chance (otherwise known as an adventitious match). Consequently, as the UK NDNAD grew in its first four years of operation, a new system known as the AmpF*l*STR® SGM Plus® [8], with average match probability of 10^{-13} was introduced in 1999. This system comprised 10 STR loci with amelogenin, replacing the previous SGM system. To ensure continuity of the DNA database, and to enable the new system to match samples that had been collated in previous years, all six loci of the older SGM system were retained in the new AmpF*l*STR® SGM Plus® system.

4. Development and harmonisation of European National DNA databases

Harmonisation of STR loci was achieved by collaboration at the international level. Notably, the European DNA profiling group (EDNAP) carried out a series of successful studies to identify and to recommend STR loci for the forensic community to use. This work began with an evaluation of the simple STRs HUMTH01 and HUMVWFA31 [9]. Subsequently, the group evaluated D21S11 and HUMFIBRA/FGA [10]. Recommendations on the use of STRs were published by the ISFG [11].

Most, if not all, European countries have legislated to implement national DNA databases that are based upon STRs [12]. In Europe, there has been a drive to standardise loci across countries in order to meet the challenge of increasing cross-border crime. In particular, a European Community (EC) funded initiative led by the ENFSI group was responsible for co-ordinating collaborative exercises to validate commercially available multiplexes for general use [13]. National DNA databases were introduced in 1997 in Holland and Austria; 1998 in Germany, France, Slovenia and Cypus; 1999 in Finland, Norway and Belgium; 2000 in Sweden, Denmark, Switzerland, Spain and Italy, Czech Republic; 2002 in Greece and Lithuania; 2003 in Hungary; 2004 in Estonia and Slovakia [14].

A parallel process has occurred in Canada [15,16] and in the US [17]) where standardisation was based on 13 STR loci, known as the Combined DNA Index System (CODIS) core loci. See concurrent review in this series by John Butler [18].

5. Development of the European Set of Standard (ESS) markers

Based on the initial EDNAP exercises and recommendations by ENSFI and the Interpol working party [19], four loci were originally defined as the European standard set (ESS) of loci–HUMTH01, HUMVWFA31, D21S11 and HUMFIBRA/FGA. The identity of these loci was dictated by their universal incorporation into different commercial multiplexes that were utilised by member states. By the same rationale, three further loci were added to this set–D3S1358, D8S1179 and D18S51. These loci are the same as the standard set of loci identified by Interpol for the global exchange of DNA data.

A subsequent expansion of ESS loci was motivated by the Prum treaty of 2005 [20], that was signed by Austria, Germany, France, Spain, Belgium, Luxembourg and the Netherlands (many more states have since signed). This treaty promoted cross-border cooperation by agreement to exchange information, including DNA profiling databases, to be made available for pan-European searches. Clearly, the relatively high combined random match probability of the original ESS loci (approximately 10^{-8}) was not sufficient to enable comparisons to be made without unacceptable risk of chance (adventitious) matches (Section 13.1). In addition, since the development of the original multiplexes, a significant number of new STRs had been discovered and it was shown that 'mini-STRs' had improved potential to analyse compromised (degraded) DNA samples because of their short amplicon size [21]. To meet the challenges, discussions began in Europe in 2005, within the ENFSI organisation. Collaborative experimentation confirmed that 'mini-STRs' showed the expected efficacy [22] to analyse degraded DNA. In consultation with manufacturers of multiplex kits, a list of candidate ESS markers were published [23] and revised [24], so that the final list of five additional loci were: D10S1248, D12S391, D22S1045, D1S1656 and D2S441, making a grand total of 12 ESS loci, with a probability of chance match roughly equal to 10^{-15}. The new loci were officially adopted by the European Commission [25] and Interpol in 2010; this led to development of a series of new multiplexes by the major companies (Promega, Life Technologies™ and Qiagen). See Fig. 1.

Improved platforms and associated biochemistry utilised five-dye technology to create the necessary space for new markers. Using the new multiplex systems, the ENFSI group carried out comparative concordance and population genetics studies between 26 EU laboratories [26]. From 2012 to 2013 there was considerable activity to implement the new multiplex systems throughout Europe. At the time of writing the transition to the new marker systems is almost completed and universal. Concurrent expansion of

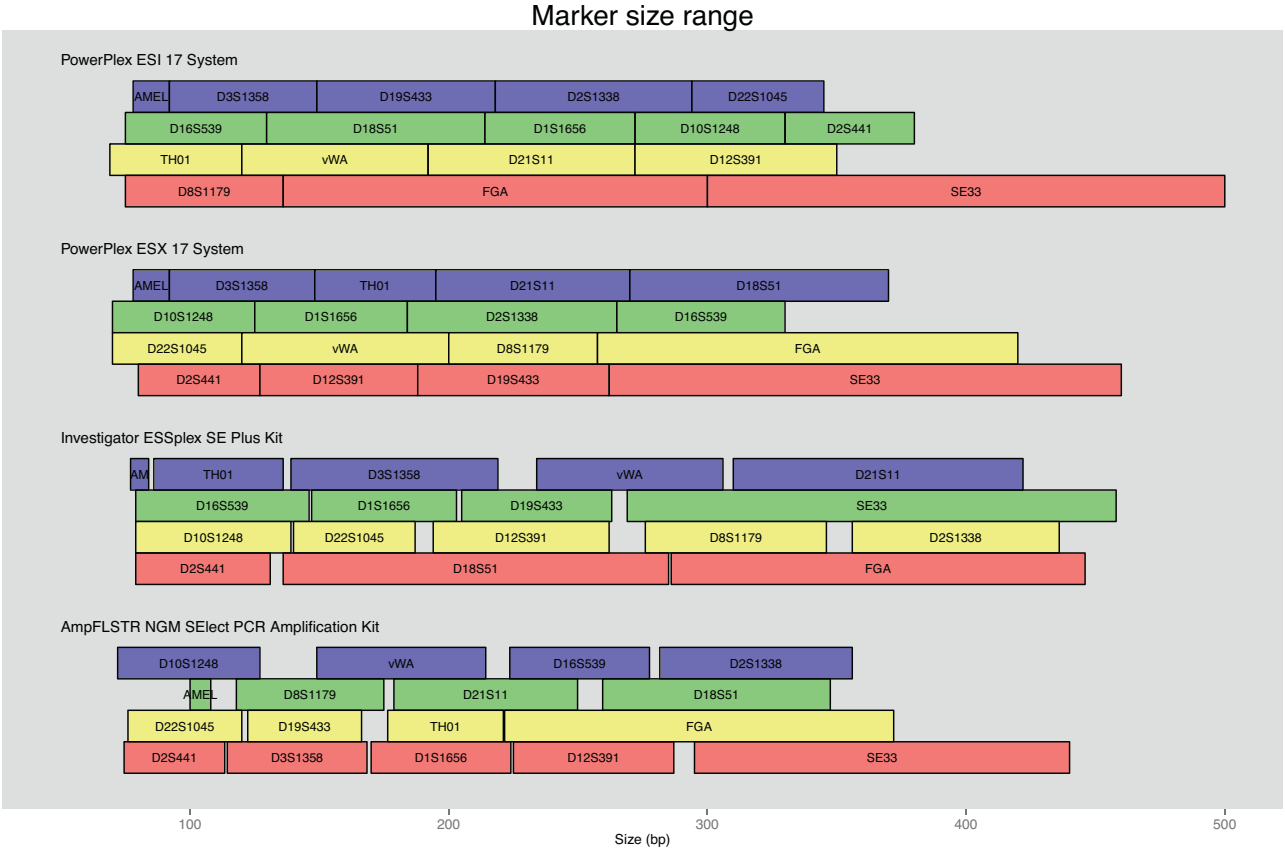

Fig. 1. Commonly used multiplex kits showing ESS loci relative to molecular weights (bp).

the North American CODIS loci has been agreed [27–29]. See the NIST website for lists of markers http://www.cstl.nist.gov/strbase/.

6. Population databases

Population databases are distinct from intelligence databases and are often referred to as 'frequency databases'. They are used to estimate the relative rarity of a profile in a population in order to give an indication to a court regarding the strength of the DNA evidence. Because allele frequencies differ between racial groups, it is the usual practice to collect databases for the major racial groups that comprise the commonest population groups of a country. For example, there are three different databases that are used in forensic casework in the UK: White Caucasian, Afro-Caribbean and Asian (Indian sub-continent). The greatest differences are found between broad racial groupings. Relatively minor differences are found between sub-groups within the same ethnic group but are from (for example) distinctly different geographical different locations. A key question is whether the frequency database that is utilised is representative, given that most databases used for forensic purposes are based on broad random collections of racial groups that do not usually take account of sub-population structure. The Asian database comprises people whose ancestors originated from a very wide geographical and cultural background. Can we be sure that a single database is representative for all sub-groups within the entire sub-continent?

The NRC report [30] took the view that the actual 'subgroup' to which the suspect belongs is irrelevant, since the consideration is the probability of the evidence if the suspect was not the source of the DNA. Accordingly, Foreman et al. [31] pointed out that it is the ethnicity of the offender that is relevant and not the ethnicity of the defendant. However, if the court wishes to evaluate the scenario where it is claimed that the sub-population of the offender is the same as that of the suspect (e.g. if all potential suspects are from a particular locality or a particular group of people) then the question does arise whether the database is representative?

To answer this question, fairly extensive studies have been carried out to measure genetic differences between different groups of people [32–35]. These studies support the notion that differences between subpopulations are low and discernible differences are unlikely within cosmopolitan populations. However, theoretical variation between sub-populations can be accommodated by the use of a correction factor (*Fst*) [36,37]. Measured differences between sub-populations appear minor and $Fst < 1\ \%$ (unless the population is highly inbred). This means that inferences derived about frequencies of alleles in a specific sub-population for which a database is not available, can be accommodated by using a general database so long as *Fst* is included in the calculation. Gill et al. [38] showed from a comparison of 24 different populations, that a single pan-European database could suffice for white Caucasians. See Welch et al. [26] for a more recent evaluation and review of European gene-frequencies using new generation ESS multiplexes.

7. Challenges of low-template DNA analysis

Forensic scientists have been keen to increase the sensitivity of their methods. Historically, the easiest way to do this was simply to raise the number of PCR amplification cycles. Findlay et al. [39] demonstrated that single cells (buccal) could be analysed by amplifying with 34 cycles using the AmpF*l*STR® SGM Plus® multiplex system. The interpretation was not straight-forward – additional alleles (known as drop-in products) were occasionally observed. The size of stutter artefacts was enhanced; missing alleles, known as allele drop-out, were common. However, such profiles can now be interpreted using statistical models that take these phenomena into account (section 9). Increasing the sensitivity of PCR by raising the number of cycles was used to increase the range of evidence types available to analysis. For example, some of the work in this area is as follows: Wiegand and Kleiber [40] analysed epithelial cells transferred from an assailant after strangulation using 30–31 cycles of PCR. Van Hoofstat et al. [41] analysed fingerprints from grips of tools with 28–40 cycles. Analysis of STRs from telogen hair roots and hair shafts in the absence of the root was reported [42–44].

Increased PCR cycles are routinely used by anthropologists and forensic scientists to identify ancient DNA from bones. Gill et al. [45] originally used 38–43 cycles to analyse STRs from 70 year old bone from the Romanov family. Schmerer et al. [46,47] and Burger et al. [48] analysed STRs from bone thousands of years old (60 and 50 PCR cycles respectively). Some authors used modified PCR methods, for example, a nested primer PCR strategy was used by Strom and Rechitsky [49]. This utilised a first round amplification with 40 cycles, with subsequent analysis of a portion with a further 20–30 cycles. This method was used to analyse DNA from charred human remains and minute amounts of blood.

For a comprehensive review of the literature relating to trace DNA evidence the reader is referred to Van Oorschot et al. [50].

All methods used to analyse low quantities of DNA suffered from the same basic disadvantages of stochastic (random) effects described in Section 16.2. If present in low copy-number, a DNA molecule will be delivered in variable quantities as a result of sampling variation. This leads to the preferential amplification of alleles. There are therefore several consequences that cannot be avoided:

- Locus drop-out, i.e. a whole locus fails to amplify.
- Allele drop out may occur because one of a pair of alleles at a heterozygote locus fails to be amplified to a detectable level.
- Stutters may increase in size relative to the progenitor allele.
- Allele drop-in results in additional alleles 'contaminating' the sample.

This means that different DNA profiles observed may not be fully representative. Tarbelet et al. [51] originally suggested a method of replicated analyses that comprised a rule that an allele could only be scored if observed at least twice in replicate samples. This theory was expanded by Gill et al. [13] who adopted Tarbelet's duplication rule. However, introduction of new software solutions that incorporate the drop-in/drop-out probabilities into calculations [5] superceded the need to derive consensus sequences and this approach is much to be preferred as it does not waste information (Section 9).

8. Low template vs. conventional DNA profiling (significance of new technology)

There has traditionally been some difficulty in defining the meaning of 'low-copy-number' or low-template DNA. The UK technical working group [52] observed (in Section 2.9.1):

> "We have demonstrated experimentally that some laboratories achieve results from c.50pg of DNA using standard 28 PCR cycles. Since these consequences are common to all methods of DNA analysis and are not restricted to 34 cycles, we do not consider the LCN label for 34 cycle work to be useful, or particularly helpful, and propose to abandon it as a scientific concept because a clear definition cannot be formulated"

In addition, rather than use arbitrary delineators to classify conventional vs. low-template DNA, it is preferable to use an interpretation strategy that can be used for all DNA profiles [53] so that the concept of delineation itself becomes redundant.

Table 1
Summary of available interpretation software.

Software	Approach	LTDNA	WoE	Deconvolution	License	Ref
DNAmixtures	Continuous	√	√	√	Commercial	[66]
Lab Retriever	Semi-continuous	√	√	✗	Open-source	[67]
LikeLTD	Semi-continuous	√	√	✗	Open-source	[65]
LRmix	Semi-continuous	√	√	✗	Open-source	[61]
LOCIM	Empirical	✗	✗	√	Non-commercial	[68]
STRmix	Continuous	√	√	√	Commercial	[63]
TrueAllele	Continuous	√	√	√	Commercial	[62]

Notes: WoE = weight of evidence. Note that DNAmixtures is a free of charge open-source R package, however it requires the HUGIN commercial software to run.

This means that there is no delineator that can be used (indeed there has never been a consensus definition). Instead, the term low-level (or low-template) is used as a broad generic term to describe partial profiles, independent of the method used. It is the quality of the result, rather than the method by which it was derived, that is important.

However, the new multiplex systems display sensitivities, in terms of detecting picograms of DNA, that are equivalent to or better than those that were originally developed more than ten years ago. For example, Pajnič et al. [54] used new multiplex systems to analyse degraded DNA from World War II skeletons, without using any enhancements or changes to the manufacturers protocol, therefore it can be stated with some confidence, that regardless of any definition, the forensic community has already moved to a position where all laboratories are currently analysing low-template DNA. Although the definition is not important, the method of interpretation is important – the lower the amount of DNA present in a sample, the greater the chance that it may not be associated with a crime-event. Therefore this change in technology is not without associated risks but can be addressed by suitable education and training.

9. Analysis of complex DNA profiles – mixtures and low template DNA

Low-level complex DNA mixtures are often encountered in casework. The statistical analysis of such mixtures is challenging: in single-source stains, only one genotype is possible at each locus, but several genotypic combinations are possible in DNA mixtures. Therefore, it is not straightforward to determine which genotypes contributed to the mixture.This is further complicated when samples are low-level, which makes them prone to PCR-stochastic effects, such as drop-out, drop-in and imbalanced heterozgotes.

Here we summarise the main categories of models and software that are available for mixture interpretation. Readers interested in discussions on the various software available are referred to comprehensive reviews and references to papers [55,37].

9.1. Approaches for mixtures interpretation

The different models used for mixture interpretation are typically classified into three groups:

I Binary models,
II Semi-continuous models,
III Continuous models.

This classification reflects the way peak heights are used. Binary models ignore peak height information completely, and are therefore not suited to interpret low-template mixtures. For this reason such models may be considered obsolete and are not discussed further. In semi-continuous models peak heights may be used to inform the model parameters, while continuous models incorporate peak heights fully. The classical models labeled as I are described by Buckleton et al. [56] while Steele and Balding [37] review software with emphasis on models II and III.

In general, there are two main approaches for statistical evaluation of forensic DNA samples: the LR approach, described in the next section, or calculation of the probability of exclusion (PE), or its converse, the probability of inclusion (PI), also termed random man not excluded (RMNE).

The preferred approach, according to the ISFG DNA commission [2], is to calculate the likelihood ratio (LR). However, whereas the computation of summary statistics, such as the RMNE, is straightforward, the complexity of likelihood ratios requires the use of specialised software to analyze complex DNA profiles. Although the theory to support the use of LRs has been available for several years [57–59], the introduction has been slow.

A number of new software, dedicated to the interpretation of low template DNA mixtures, have recently become available [60–65]. These software are anchored in a likelihood-ratio framework, but they all use different probabilistic models, and rely on different distributional assumptions (see Steele and Balding [37] for a review). Table 1 gives an overview of the available software (either open-source or commercial), and Table 2 further describes the different approaches for mixture interpretation.

9.2. Reporting DNA evidence using likelihood ratios

Several years ago, the International Society of Forensic Genetics DNA Commission reviewed the interpretation of complex DNA profiles [2]. They recommended the 'likelihood ratio' (*LR*)

Table 2
Advantages and disadvantages of the main interpretation approaches.

Model	Advantages	Disadvantages
Binary	Easy to use and to implement	Cannot be used for LT DNA
Semi-continuous	Can be used for LTDNA	Model-parameters have to be estimated
	Makes fewer assumptions than continuous models	Does not make use of peak heights
		Implementation requires specialized software
Continuous	Make use of peak heights	Numerous parameters need to be estimated
	Can be used for LTDNA	Implementation requires specialized software
		Require calibration for different STR kits and different conditions (e.g. PCR cycle no.)
Empirical	Simple to implement in casework	Require calibration for different STR kits and different conditions (e.g. PCR cycle no.)
		Can only be used to extract major profiles
		Cannot be used for weight of the evidence

approach. A typical analysis of crime sample evidence (E) requires the scientist or other evaluator to consider at least two alternative hypotheses–the prosecution hypothesis (Hp) and the defence hypothesis (Hd). For a profile with more than one contributor, the prosecution may hypothesise that the suspect (S) and one unknown (U) person were the contributors, whereas the defence may hypothesise that there were two unknown contributors U_1 and U_2 . The likelihood ratio (LR) compares the probabilities of the evidence under these alternative hypotheses:

$LR = \frac{Pr(E|H_p)}{Pr(E|H_d)}$ where $Pr(E \mid H)$ is calculated, based on a model, and in the example we use the notation:

$$H_p = S + U; H_d = U_1 + U_2$$

If the LR is greater than one, then the evidence favours H_p but if it is less than one then the evidence favours H_d. The evidence is then expressed in terms of the two alternative hypotheses. One of the attractive features of this approach is that it enables the scientist to simultaneously consider and to compare the alternative defence and prosecution scenarios. It enables a framework that allows for different hypotheses to be considered if necessary – for example the defence may contend that there are three unknown contributors instead of one. Much has been written on the subject of the likelihood ratio. However, it is important to point out that there are alternative methods, including versions of RMNE probabilities, to evaluate evidence, that are used in many jurisdictions. The reader is referred to Buckleton [56] (pp. 27–64) for a review.

9.3. More on continuous models

Generally, the likelihood ratio for *all* of the above models can be expressed as

$$LR_C = \frac{\sum_j w_j P(S_j|H_p)}{\sum_j w_j P(S_j)|H_d)} \quad (1)$$

for appropriate interpretation of the terms entering into the equation as explained below. We consider only one marker and one replication. The extension to several markers requires no fundamentally new ideas: the likelihoods for different markers are conditionally independent and can be used to express the combined likelihood as an integral. Here S_j is a set containing genotypes of the contributors and $w_j = P(G_C|S_j)$ where G_C is the crime stain. For continuous models, G_C records information on allele designation and corresponding peak information. The notation inevitably differs between between papers. Above we essentially follow Taylor et al. [63]. With no restrictions on drop-in, any evidence can be explained under any hypothesis (also without any contributors) and therefore the sum extends over all possible combinations of genotypes indexed by j. Restrictions on the mechanisms for drop-in are typically modelled and implemented as explained in Section 4.2 of [37] and therefore a large number of terms in Eq. (1) vanish.

For model I, $w_j = 1$ if the allele designations in G_C are consistent with the alleles specified by S_j and 0 otherwise. For models II and III, $0 \leq w_j \leq 1$. The peak heights only influence w_j via the drop-out and drop-in probabilities for model II. For III, a model for peak heights is specified. Large parts of papers on continuous models are typically devoted to estimating or maximising w_j, see [69,63,70,62] where different models and computational methods, including MCMC and Bayesian networks, have been implemented. The formulation of the likelihood ratio in Eq. (1) is useful as it expresses the likelihood for each hypotheses as a weighted average of the probabilities of the possible genotypes of the contributors. Furthermore, as we have seen Eq. (1) serves well to contrast models in categories I, II and III and also different versions within these categories.

We will not discuss these technical points further but rather focus on some fundamental issues. In general, it is desirable to use as much of the relevant information as possible and therefore it is reasonable to explore continuous models. There is an argument that ignoring information is wasteful and may itself lead to wrong conclusions. Conversely the incorporation of more data into the model leads to added complexity and more assumptions. For instance, a parametric model for peak height distribution must be specified. Both log–normal [63] and gamma [70] distributions have been suggested. It is important to verify the modelling assumptions since this directly influences the results. In many applications, inferred models can be checked against predictions (forecasting weather is one example). In crime cases, there is typically no undisputed truth to check against, hence the challenge is to use relevant experimental data that is representative of *all* DNA casework profiles.

Stochastic effects that lead to profile imbalance and dropout are not solely a function of the DNA quantity. Gill et al. [71] show that there are several 'causal' contributing factors especially: the amount of extract that is aliquoted by pipette; whether the stain contains haploid or diploid cells; different platforms for analysis; different PCR cycle number (Section 16.1). It is expected that prior distributions e.g. heterozygote balance will be highly dependent upon the 'causal' parameters, but they are difficult to control in experimental regimes. For example, analysis of simple serial dilutions of stock DNA do not strictly reflect the condition of typical casework stains.

By necessity, continuous models employ fixed generalised assumptions, but the *sensitivity* of the assumptions relative to the various individual causes of stochastic effects (listed above) is an area where more research is needed, especially when low-template DNA is considered. The interested reader is directed to a much more extensive discussion of this topic by Steele and Balding [37] in section 3.8 of their paper where they conclude:

> "Although discrete-model [semi-continuous] LRs may have similar drawbacks, the simpler data and modeling assumptions on which they are based diminish such concerns, possibly allowing these models to enjoy an advantage of robustness to laboratory-specific details in return for a loss of statistical efficiency."

Potential concerns of the continuous approach can therefore be listed as follows:

(a) Complexity of modelling parameters and requirement to use assumptions that are not easily verified.
(b) The continuous models are typically Bayesian and therefore formulated in a Bayesian framework with prior distributions specified, see for instance Section 3.2 of [70].
(c) The necessity to use different sets of assumptions for different PCR procedures such as increased cycle number, different kits etc that increases the amount of intra-laboratory validation, and may restrict the range of protocols, that may be utilised.

In contrast, the principle advantage of the semi-continuous model is that there are fewer assumptions, hence the same model can be used across different multiplexes and different cycling conditions.

9.4. Exploratory vs 'black box' approaches

A key goal of the continuous approach is to describe a DNA profile by modelling all sources of variation and to encapsulate evidence into a single likelihood ratio. Indeed, it may be tempting

to use new statistical methods as a convenient way to generate answers simply by feeding a program with numbers, running the program and reporting the result, but this does not circumvent a requirement for careful consideration of *all* of the DNA and non-DNA evidence in a case [72].

No statistical method can capture all of the uncertainty that is inherent to casework analysis of complex DNA profiles. This is why the ISFG DNA commission [5] stated: "we do not advocate a black box approach." Software is used as *part* of the overall evaluation of the evidence. Although the exploratory approach has been advocated in relation to semi-continuous models, there is no implicit restriction of the philosophy to any particular software. Apart from unpredicted stochastic effects, the following are important considerations that affect the reported likelihood ratio:

Biochemical phenomena such as primer-template mutations [73]; somatic mutations [74], may all affect the likelihood ratio in unpredictable ways.

Numbers of contributors [58] and associated defence and prosecution hypotheses may not be obvious and are subject to debate. There is no reason for numbers of contributors to be the same under alternative hypotheses. In the 'exploratory approach' advocated by Haned et al. [61,75], the biological basis of profiles are evaluated prior to any strength of evidence test.

Does the questioned profile fall within the scope of the software validation? Additional questions are relevant: does the profile fall within the limits defined by validation? For example, has the software been tested relative to 3/4/5 low template contributors if such mixtures are hypothesised?

The exploratory approach comprises a suite of software tools that can be used by the expert to evaluate evidence; the purpose is not restricted to solely reporting the strength of evidence as a likelihood ratio, it is also used to 'explore' the evidence itself. The statistical analysis may indicate 'aberrant loci' that require new tests employing different biochemistry. This is preferable to a 'blind' statistical analysis of samples. In addition, there will typically be several stains in a case that may be considered for evidential purposes - this will provide additional opportunity to cross-check inferences.

9.5. Concluding remarks – diversity of methodology

While there can be incorrectly calculated LRs, finding a perfect LR solution for all situations is not possible. Balding [65] states:

> "Likelihoods depend on modeling assumptions, and there can be no "true" statistical model for a phenomenon as complex as an LTDNA profile"

Consequently, there is no agreement within the forensic community on the best approach, and it is unrealistic to suppose that any single method will be universally adopted. This means that in practice a diversity of methods will be used for the foreseeable future. In principle, there is nothing wrong with this. It will encourage research. An inevitable outcome, to be encouraged, is that court-reports will be routinely prepared and challenged by different software that use different modeling assumptions. Typically, commercial software will not be available to defence experts and they will default to open source or non-commercial software. However, if similar answers are obtained, then confidence in results should increase. Here, we follow, Steele and Balding [37], and suggest that a difference in the order of one ban (one unit in log_{10} scale) is negligible.

Likelihood ratios are crucially dependent upon the assumptions or the propositions used in the models (e.g. the number of contributors). For this reason, there needs to be significant court involvement with the process of formulating alternative propositions – indeed several sets of propositions may be simultaneously proposed, tested and debated.

To summarise:

- There is no underlying 'true' likelihood ratio that can be formulated
- A diversity of models, that rely upon different modelling assumptions currently exist.
- Diversity of models is encouraged for court-reporting purposes – so that their results can be compared, if applicable with respect to similarities in assumptions, input-data and parameters.
- The 'black box' approach is strongly discouraged. There is no true LR; the veracity of the propositions is also uncertain and this will often indicate that the court needs to be proactive to ensure that the propositions addressed are appropriate to the case in question.
- A computer program does not replace the need to think carefully about the case.

10. Mixtures in kinship analysis

10.1. Mixtures with relatives

Kinship analysis is usually treated as a separate area within forensic genetics, and a number of different software are available. A comprehensive list of kinship analysis programs can be found on http://www.cstl.nist.gov/strbase/kinship.htm. A classic example of kinship analysis is paternity testing, while other uses include the identification of missing persons and disaster victim identification. However, there are cases where kinship and crime cases intervene, namely in mixtures where the contributors may be related. Although a wide range of software exists for the interpretation of mixtures, the usual assumption is that the involved individuals are unrelated, except for subpopulation effects which can be accounted for with F_{ST}. The terminology and notation differ (frequently θ is used). See [37] for a precise definition. However, if there are specific family relationships between contributors that need to be taken into account, this calls for alternative statistical methods. If a close relative of the perpetrator is disregarded as an alternative contributor, the evidence against the suspect may be overestimated [76]. We initially consider *Binary models*, see Section 9.3. The type of scenarios we envisage include mixtures where the contributors are believed to be related, e.g. hypotheses of the type – H_p: *The evidence is a mixture of the victim and her untyped grandfather vs.* H_d: *The evidence is a mixture of the victim and an unknown, unrelated individual.*

Another type of scenario involves mixtures where someone related to the questioned contributor is considered as an alternative contributor, e.g. H_p: *The evidence is a mixture of the victim and a suspect vs.* H_d: *The evidence is a mixture of the victim and an untyped brother of the suspect.*

Common to both scenarios is that one of the relatives is untyped. Fung and Hu [77] have derived kinship coefficients for cases with pairwise relationships like the two examples above. However, when there are more than two related contributors involved, another approach must be taken. In Egeland et al. [78] (which includes information on open software implementation), the problem of mixtures with related contributors was treated in generality with the use of pedigrees to describe relationships. This approach also allows for different family relationships to be specified under the opposing hypotheses, such as: H_p: *The evidence is a mixture of two typed individuals, who are siblings, and their untyped brother vs.* H_d: *The evidence is a mixture of two typed individuals, who are half siblings, and their untyped father.*

There is limited work to account for related contributors in mixture models accommodating peak heights. However, it is not so difficult to model relationships between pairs of individuals and

Table 3
The table shows expected number of adventitious matches with a large reference database size *N* when compared against a total of *n* single source stains. The results were simulated using the Norwegian population for the old Interpol standard (INTER) with 7 loci and the new ESS markers with 12 loci.

N=	1e+06	1e+07	1e+08	1e+09
INTER (*n* = 1e+05)	1.2e+02	1.2e+03	1.2e+04	1.2e+05
INTER (*n* = 5e+05)	5.9e+02	5.9e+03	5.9e+04	5.9e+05
INTER (*n* = 1e+06)	1.2e+03	1.2e+04	1.2e+05	1.2e+06
ESS (*n* = 1e+05)	7.4e−05	7.4e−04	7.4e−03	7.4e−02
ESS (*n* = 5e+05)	3.7e−04	3.7e−03	3.7e−02	3.7e−01
ESS (*n* = 1e+06)	7.2e−04	7.2e−03	7.2e−02	7.2e−01

this is also implemented in the programs likeLTD [79], TrueAllele [62] and STRmix [63], see Table 3 in Steele and Balding [37].

10.2. Drop-out in kinship cases

For mixtures in crime cases there exists a number of statistical methods and software that account for stochastic phenomena such as allelic drop-in and drop-out. Within kinship analysis however, literature that discusses the analysis of profiles that may contain drop-in and drop-out is scarce. Yet partial profiles may commonly appear in kinship problems such as missing person identification and disaster victim identification. Ignoring drop-out may lead to incorrect interpretation of profiles, loss of valuable information and biased results. Partial profiles were briefly treated in Brenner and Weir [80] in connection with identification work after the World Trade Center 9/11 disaster. A likelihood ratio model that accounts for drop-out in kinship cases involving pairwise relationships was formulated in Buckleton and Triggs [81], while Dørum et al. [82] describe a general likelihood ratio drop-out model for kinship described by pedigrees. Examples of software exist that can handle drop-out in kinship cases, including Bonaparte [83] (http://www.bonaparte-dvi.com/), DNA·VIEW (http://dna-view.com/dnaview.htm) and Familias 3 [84], where the latter is freely available (http://familias.no).

11. National DNA databases

With approximately 6 million records (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/252885/NDNAD_Annual_Report_2012-13.pdf), the UK National DNA database (NDNAD) [7] is no longer the largest database in the world–that honour now belongs to China with more than 32 million records as of November 2014, more than twice the size of the US database (greater than 13 million samples: http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics). However, it is interesting to note that as a proportion of the population, UK still ranks the highest with approximately 10% of the population on the national register, compared with approximately 4% of the US population. Reference samples are obtained from buccal (mouth) scrapes or hair roots taken from any individual arrested for any criminal offence. These are known as criminal justice (CJ) samples. Results are stored on computer in the form of an alphanumeric code that is based on the nomenclature of each STR allele. During criminal casework, operational laboratories carry out analysis of biological material such as semen or bloodstains. The STR profiles derived from these samples are compared against the CJ samples in the reference database. If a match is found then the investigating authorities are informed of the identity of the individual, to enable further investigations to be carried out. The NDNAD is primarily an intelligence database – it is used to discover potential perpetrators of crime – this is a separate exercise to the evaluation of the evidence.

Initially, DNA profiling evidence was confined to serious crimes, but now this has been widely extended (dependent upon jurisdiction) to include minor offences. As an example see the Home Office breakdown of database match statistics [85]. Many matches originate from volume crimes such as burglary. Databases can also be used to compare profiles from different crime-scenes to identify serial offenders. It is relatively common to find links between minor offences and more serious offences.

11.1. Familial searches

Novel applications are possible. In particular, the use of the database for familial searching [86]) has been implemented in the UK and elsewhere [87]. This extension of the utility of national DNA databases has proved ethically controversial – if a perpetrator is not recorded on the NDNADB, then no match will result. However close relatives e.g. brother or father will have many alleles in common. This can be used to good effect – rather than to search for a complete match, a search that relies on >50% alleles matching will yield a list of potential suspects that may be quite large. [88–90]. However, there is often information in the circumstances of the crime to suggest that the perpetrator may be local to a particular geographic area. This reduces the pool of suspects, which in turn reduces the number of database matches; therefore the investigation is much easier to manage. Additional confirmatory tests using Y-chromosome or mitochondrial DNA can be used to narrow the field further [91]. Once potential suspects are identified, then a sample may subsequently confirm a match with the crime stain. An indication of ethnicity is also possible, either from the genetic STR genotype of the perpetrator, or from the Y-chromosome – see ISFG DNA commission recommendations [92,1] and the reviews [93,94] for more details. Whereas these markers can give a useful indication of ethnicity, they are never 100%. Their use is primarily to prioritise a list of potential suspects for investigative purposes.

11.2. Conventional search strategies

If no suspect has been identified as a contributor to crime-stain evidence, then a search of the national DNA database is carried out.

A cursory analysis of discriminating potentials of multiplexed STRs systems is generally conditioned on the 'full' DNA profile. This assumption is never realistic [95–97]. Many case-stains are partial. This means that alleles or markers are usually missing from the profile.

The search of a national DNA database, where a crime stain profile is compared against reference samples, is a two-stage process:

- A match is declared if all the alleles contained in the crime stain profile match those in the reference sample – this condition may be relaxed to take account of possible allele mistyping and partial profiles (Section 11.2.1).
- The evaluation of the strength of the evidence is carried out with the putative 'match' and is a separate exercise to its discovery (outlined above).

11.2.1. Matching allele count (MAC) method

The MAC method is the standard method: a simple count of the number of matching alleles between an evidence sample and a corresponding reference sample on the NDNAD [98]. If every individual on the national DNA reference database is compared against the evidence, this would return a ranked list of matched candidates which satisfies the condition MAC$\geq T$ for a threshold, where T is a simple count of the number of matching alleles. If

I ='number of markers used in the analysis', then MAC= 2I is a complete match at every allele between the crime-stain and the reference sample minimises the probability of observing false positive matches. A 'reduced stringency' matching process (wild-card searches), described in section 11.3, allows for partial profiles with missing alleles and errors in allele designation so that T is less than the number of alleles in a full profile. Provided that the profile is a full single-contributor profile then the *MAC* is converted into a match probability *Pm* which can be used to measure the strength of evidence

For a selected threshold T, the false positive probability measure $p(T) = Pr(MAC >= T \mid E)$ evaluates the risk of false positive results based on the evidence E. The formula is provided in [99]. Similar problems were explored by Tvedebrink et al. [100].

If T is reduced, this will naturally increase the false positive rate. Consequently, the random match probability is increased (because of the loss of information), which means that the chance that it will match a sample that originates from a different individual is also increased. This is known as an 'adventitious match' that leads to an error of false inclusion. Adventitious matches can also occur with full profiles of course, but are less probable. The risks are directly estimated by the match probability. Conversely, if a profile is wrongly designated because of operator error (or because of the contamination event known as drop-in) then it will not match a reference or crime-stain profile. This could lead to an error of false exclusion. In order to minimize the level of false exclusions, NDNADs carry out low stringency searches. This means that a complete match is not needed for a putative match to be inferred – for example perhaps 23 out of 24 alleles match in the case stain match a reference sample, but there is one mismatch.

Low stringency tests intended to reduce false exclusion rates will paradoxically increase the false inclusion rate. Some false inclusions will be detected at the second stage of testing (where profiles are manually compared or re-worked with different multiplexes). Nevertheless, it follows that if the false inclusion rate is too high, then this compromises the efficiency of a NDNAD, simply because the increased amount of work to investigate multiple matches is prohibitive.

11.3. Reducing the test stringency threshold T for searches on national DNA databases

In order to reduce the stringency of a database search, wild cards are incorporated as part of a profile designation in order to accommodate the following phenomena:

- The locus may appear to be homozygous, but the allele peak is below the stochastic threshold (Section 16.2). Under laboratory rules, the possibility of drop-out is accommodated by an additional allele that cannot be visualised. (eg 17,F means the locus could be type 17,17 homozygote or 17,Q heterozyote where Q is an allele other than allele 17. The F or Q designation acts as a wild card that matches any other allele.
- If there is a locus with a wild-card e.g. 12,F; then it will match any other locus with a single 12 designation, including 12,14; 12,12; 8,12 etc.
- If comparisons are made between loci, the primer-sets made by two competing manufacturers will be different. If there is a primer binding site mutation that prevents amplification of the target molecule then the locus will appear to be homozygote with one set of primers, and heterozygous with the alternative set e.g. [73]
- There may be some ambiguity associated with the allele – it may be 'off-ladder' i.e. outside the usual range of alleles, so its identity cannot be reliably designated relative to an allelic ladder, or it may be rare so that it does not match a common allele (this may be given an 'R' designation – this is also treated as a 'wild-card').
- Transcription or operator errors may result in mis-designations of loci.

Low stringency tests operate by reducing the discriminating power of the multiplex systems. If this becomes so low that thousands of adventitious matches are expected to occur, then the utility of a database is compromised. The increased time-consuming aspects of manual inspection or investigation of adventitious matches makes the process unviable.

11.4. Summary of matching rules – the Interpol Gateway

The minimum match stringency [101,102]is a complete match between at least six of the seven ESS loci (the original set in Section 5). Wild cards are introduced in order to allow for errors of mis-designation outlined above so that a 'search' may introduce a 'near-match' report – bearing in mind that the search is for investigative purposes.

The performance of a multiplex is dependent upon a) the size of the reference database b) the inbuilt redundancy of the multiplex (i.e. the tolerance to wild-card designations). Mixtures are not allowed as the search strategies are based on simple MACs and this system has inherent limitations, examined below.

Database searches are challenging if the evidence is complex, with allele drop-out or/and multiple contributors. The increased discriminating power of the new DNA typing kits offered by ESS loci facilitates searches. Conversely, the inherent ambiguity in the complex partial profile and the necessity to introduce 'wild cards' into the match-strategy reduces the effectiveness.

12. Assessing the strength of the evidence of a match derived from the intelligence database

In current practice, a match of a crime stain with a reference sample during a database search is identified by the MAC method. The second step is to calculate a strength of evidence. This is always presented as a conditioned match probability or as a likelihood ratio, calculated by using a relevant population (frequency) database (Section 6). The question of whether a search of a large intelligence database for a match subsequently affected the strength of the evidence was addressed by the NRC report [30]; pp. 133–135. They originally recommended that an adjustment was applied by multiplying the match probability (*Pm*) by the number of people on the database [103]. Using an example of an intelligence database of $n = 1000$ and a multiplex with *Pm* of 10^{-6} this would result in a substantially reduced strength of evidence $n \times Pm = 0.001$. This suggestion became known as the 'np-rule' and led to major debate.

12.1. The DNA database search controversy: quantification of evidence after a database search

Balding and Donnelly [104] criticised the recommendation as follows. Consider the likelihood ratio between the hypothesis H_p: 'Suspect S is a contributor to the evidence' and the complement H_d: 'Suspect S is not a contributor to the evidence'. If the evidence is single-source and no other information is provided, the weight-of-evidence $LR = 1/p$ where p is the population match probability. Now let N be the known population size and n is the number of individuals in a reference database, so that that n and S are both drawn from the population N (we assume unrelatedness). Furthermore, if S was the only individual in the database that matched the evidence sample, the adjusted weight-of-evidence becomes $LR^* = ((N-1)/(N-n))(1/p)$ (equal prior probabilities $1/N$ are applied to all individuals). Stockmarr [103] presented the alternative hypotheses H'_p: 'The contributor is in the database' and

the complement H'_d: 'The contributor is not in the database'. This resulted in the weight-of-evidence $LR' = 1/np$ where the population size N is not relevant. This was the basis of the "np-rule" which was supported by NRC [30].

Fig. 2 shows that the two approaches differed greatly, since increased n will increase LR^* slightly, while LR' decreases rapidly. This difference led to a major debate. Meester and Sjerps [105] concluded that the two approaches described above were both valid in terms of posterior probability, provided that prior odds of guilt, based on the strength of non-DNA evidence were incorporated into the calculations above. The database search was essentially an exercise in intelligence gathering to identify potential suspects. At this stage, the probability of guilt per individual in the population is the same ($1/N$). If a suspect is identified then the investigation of the crime enters a second phase to search for additional evidence that may implicate or exonerate a suspect, hence the priors are continually updated to take account of new information.

The debate was recently reopened by Schneider et al. [106] who advocated use of the "np-rule" and drew responses from Biedermann et al. [107] and Gittelson et al. [108]. The argument followed similar lines to those briefly summarised above. There is an excellent summary provided by Nordgaard et al. [109] who asks the pertinent question:

> "why does this debate keep re-emerging?"

They also provide the answer:

> "..the risk behind the fear is that the court would not use prior odds for the individual to be the source of the recovered DNA. If that is the case there would be no differences between a database hit case and a probable cause case with respect to the decision about guilt, if the DNA match is the only evidence presented. In other words a conviction would be solely built on the DNA match."

This is the essence of the problem – there is an expectation that the court follows the rules – to take account of all the information provided and to incorporate probabilities and relevant priors in line with current theory. Unfortunately, in the UK (as an example), this is not always followed, Prosecutions can occur on the sole basis of database matches without corroboration [72]. The jury is not informed of the profile discovery via database search and the strength of the DNA evidence is not necessarily considered in correct context of the priors [110].

Nordgaard et al. [109] continue:

> "The seemingly rational thing to do is to enhance the communication between the forensic laboratory and the commissioners (police, prosecutors and court)."

Biedermann et al. [107] appears to concur:

> "there is need for more argumentation, perhaps using another language than that of mathematics".

The papers by Storvik [111] and Chung et al. [112] also provide a comprehensive summary of the database search debate for the interested reader; the latter extended the discussion to mixtures.

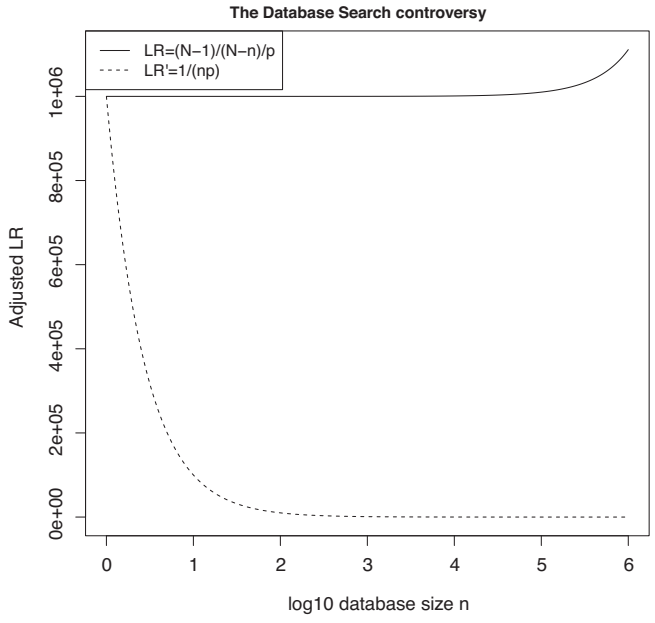

Fig. 2. By assuming population size $N = 10$ million, random match probability $p = 1e{-}6$, the plots show the different LR adjustment after database search for LR and LR' defined above.

13. Limitations of databases relative to their size and the discriminating power of the multiplex

Suppose that there are 1 million samples in an intelligence database. If a multiplex system is used that has an average match probability (Pm) of 2×10^{-8} (as for the original AmpF*l*STR® SGM™ system) then the chance of an adventitious match is conveniently defined by the 'np rule', noting that in this specific context the use is non-controversial: $n \times Pm = 0.02$. Taking the reciprocal demonstrates that approximately one in fifty samples that are compared to a database this size will match by chance.

As databases grow, it is implicit that the probability of a match needs to be reduced in order to keep the potential number of adventitious matches to a minimum. To fulfil this requirement, STR systems have been upgraded to include more loci (Section 3). Consequently, a much lower average random match probability was achieved.

To assess the impact of an adventitious match, the only relevant comparisons are between criminal justice (reference or known) samples and crime (or unknown) samples, rather than pairwise comparisons within the database itself. Approximately 6 million DNA profiles are currently retained on the UK database and to date they have been compared against approximately 415,000 samples taken from crime scenes [113] – this is $6.0\,\text{m} \times 415{,}000 = 2.5 \times 10^{12}$ pairwise comparisons in total. Applying a full profile match probability of 10^{-13} (the SGM Plus average), this gives a 92% chance that one or more comparisons have led to adventitious matches at all loci (ignoring relatedness) since inception. We can conclude that it is certainly possible that adventitious matches occur between crime samples and the database, but, (a) they will be very rare; (b) provided that the match is treated as an investigative lead, in the first instance, and not used as direct proof of guilt in lieu of other evidence – an adventitious match can be accommodated.

13.1. Adventitious matches affected by database size and multiple searching

In Table 3 an example is provided to show the number of expected adventitous matches when a large reference database (size N) is compared to crime stain profiles (size n) using the European standard set (ESS) (12 loci) which superceded the old Interpol standard set (INTER) (7 loci) described in Section 5. For the old Interpol standard it was expected that 120 adventitious matches will result if $n = 100$ thousand crime stain profiles are searched against a database of $N = 1$ million reference samples.

However, for the ESS loci, the discriminatory power is much greater so that only 0.7 false positive matches are expected when there are 1 million crime stains that are searched against a database of $N = 1$ billion reference samples (for example)

To counter the challenges of massive databases with many millions of individuals, the new generation multiplexes will significantly reduce risks of chance adventitious matches, but of the course the difficulty remains that there are 6 m samples in the UK, as an example, that require upgrading to the new system – recent EU legislation requires that DNA samples should not be stored indefinitely, hence upgrading retrospectively will be problematic – the expense is prohibitive and there appears to be no easy solution to this problem. However, if an existing match is believed to be adventitious then testing with further STR loci or using other typing systems should easily demonstrate this.

Databases will contain pairs of relatives (especially brothers) with increased probability of chance matching [114]. Between a pair of siblings the probability is approximately 10^{-7} for the 16 loci Life Technologies™ AmpfISTR® NGM™ system.

Partial DNA profiles will continue to have much higher match probabilities and therefore a much higher chance of adventitious match.

13.2. Reducing number of false positives

In order to reduce the number of false positive matches in a database search, an alternative method is to reduce the number of individuals in the database. During the course of an investigation, it is often possible to define the characteristics of the offender in terms of gender, age, geographical location etc. Gill [72], pages 105–108, describes the idea of defining a "Target population" which is a "slice" of the population which satisfies some of these criteria. If followed, it would be an effective way to reduce the number of false positive matches relative to the reference database (since N is reduced).

14. Searching databases with complex DNA profiling evidence samples

14.1. Alternative to the $\mathrm{MAC} > \mathrm{T}_{\mathrm{alleles}}$ approach using a likelihood ratio ($\mathrm{LR} > \mathrm{T}_{\mathrm{LR}}$) method

A probabilistic model $Pr(E \mid H)$ to the evidence stain E for a given hypothesis H, where known and unknown contributors may be specified, forms the basis of what is known as the likelihood ratio (LR) method (see Section 9.2). Such a model requires that the number of contributing individuals C is specified through H. Consider individual j to be a genotyped individual in the reference database, whom we want to investigate. Instead of counting matching alleles (MAC), a countinuous score given by the likelihood ratio (LR) between the two (typical) following competing hypotheses is considered:

H_j: Individual j and C-1 unknown individuals contribute to the evidence E
H_d: C unknown individuals contribute to the evidence E

LR is the most efficient way to evaluate evidence and this principle can be extended to database searches instead of using the MAC method which is demonstrably inefficient for these kinds of samples (see Bleka et al. [98] for low template DNA searching and Balding et al. [79] for familial searching).

Curran et al. [115] proposed a model to specify probabilities of allele drop-out and drop-in. If individuals in the database deviate from the prior allele frequencies (i.e. originate from a different population), this bias can be adjusted using the Fst correction [65].

Instead of selecting a threshold T based on $MAC > T_{alleles}$, an alternative is to extract all candidates which satisfy the condition $LR \geq T_{LR}$ (subscripts are introduced to make clear the difference between T defined as the number of alleles vs T defined as an LR) hence the false positive probability measure becomes $p(T_{LR}) = Pr(LR \geq T_{LR} \mid E,\ H_d)$. Whereas the MAC measure is a discrete match measure from 0 to $2I$, in contrast, the LR is continuous.

Another alternative method extracts a list of K candidates from a database with probability P that it contains the true donor subject to the conditional constraint that he is really in the database in the first place. This theory was introduced by Slooten et al. [87] and was evaluated by Bleka et al. [98].

14.2. Performance of database searching

The demonstration of improved performance using the $LR > T$ method was carried out by Bleka et al. [98]. A total of 4000 simulated partial two-person mixtures were searched against a 5 million person (simulated) database, for both the SGM Plus (10 loci) and the ESX 17 (16 loci). It was shown that the 95% rank quantile to extract the true donor with 6 allele drop-outs was 38864 and 18 respectively for SGM Plus vs. ESX 17 using $MAC > T_{alleles}$ model, versus 9832 and 2 when a $LR > T_{LR}$ model was used instead. This simultaneously demonstrated that increased discrimination power of complex DNA profiles was achieved by a) increasing the number of loci b) utilising the LR method to measure strength of evidence. Bright et al. [116] demonstrate an expansion to take account of allelic peak height and stutter if this information is available.

14.2.1. Software to carry out database searches

The R-package 'mastermix' (available from http://r-forge.r-project.org/) includes different user-friendly tools to carry out mixture intepretation. The GUI "Mixture Tool" incorporates the LRmix module, in addition to the MAC method, to provide efficient database searching when considering complex evidences. Besides allowing database searches, it also implements an extended version of the deconvolution model for STR DNA mixtures provided by Tvedebrink et al. [117].

15. Non-contributor tests of robustness for complex DNA profiles

A probabilistic LR model relies on a number of assumptions, including the propositions to be evaluated and parameters such as the drop-in and dropout probabilities. Likelihood ratios are based on comparing at least two propositions (Section 9.2). The choice of propositions is provided initially by the mandating authorities (reference: ENFSI guideline for evaluative reporting in forensic science.[1]) However, the proposition sets are not always obvious. A consideration of case pre-assessment and the expert's opinion may lead to the proposal of several secondary sets of propositions to test. Gill and Haned [75] advocate an exploratory approach and provide guidelines to organise relevant propositions to interpret complex mixtures. The proposal to evaluate different sets of proposition pairs is also supported by Gill et al. [58] and Buckleton et al. [118]. The robustness of the likelihood ratio model with regard to the modeling assumptions has been discussed in several papers [61,75,119,120].

[1] At the time of writing this document is still in preparation but will be available on the ENFSI website http://www.enfsi.eu/.

15.1. Non-contributor tests

Gill and Haned [75] proposed using non-contributor tests to investigate the robustness of a case specific LR. These tests are based on the 'Tippett test', introduced for DNA profiling on a specific *per case* basis by Gill et al. [119]. The test is used to indicate the probability of misleading evidence in favour of the prosecution. The idea is to replace the profile of the questioned contributor (POI), e.g. the suspect, by a number of random profiles, and for each random profile the likelihood ratio is recomputed. Consider a simple *LR* of one contributor, where *S* is the questioned contributor (e.g. suspect) in the numerator and *U* is an unknown unrelated person.

$$LR = \frac{Pr(Evidence|S)}{Pr(Evidence|U)} \quad (2)$$

In non-contributor testing, *S* is substituted by *n* random man profiles $R_{1..n}$, where *n* is usually a large number $\geq$1000 and this gives a distribution of random man *LR*s if the defense hypothesis is true.

$$LR_{i=1,\ldots,n} = \frac{Pr(Evidence|R_i)}{Pr(Evidence|U)} \quad (3)$$

Once a distribution of non-contributors has been propagated, some useful statistics can be provided:

- Quantile measurements e.g. median and 99 percentile,
- *p*-values.

If the *LR* of the suspect is large enough to be easily distinguished, and does not overlap the distribution of random man *LR*s, it gives support to the proposition that the suspect is a contributor. However, if replacement of the suspect's profile with random profiles results in *LR*s the same order of magnitude as the observed *LR*, then the suspect's profile data behaves no differently from a random man and gives support to the defence hypothesis of exclusion. The emphasis is that non-contributor tests are diagnostic tests to assist the interpretation and to understand any limitations of the observed *LR* that may prevail.

This is important since Haned et al. [121] showed that a large *LR* does not necessarily translate into probative evidence against a suspect when complex propositions are considered.

An example follows based on Table 3, Gill and Haned [75]. Two suspects are simultaneously accused of a crime against a victim. A mixture of three contributors is recovered and the victim can be conditioned under both *Hp* and *Hd*. The primary propositions provided by the mandating authorities for the scientist to evaluate are formulated:

$$LR_1 = \frac{Pr(Evidence|S_1, S_2, V)}{Pr(Evidence|U_1, U_2, V)} \quad (4)$$

For this example there is only one *LR*, but there are *two* separate non-contributor tests, or *p*-values, that must be evaluated.

This is because the *LR* is a *holistic* statistic that cannot distinguish *between* contributors within the construct. Their relative contributions are disproportionate. Non-contributor tests are used to 'dissect' the propositions to reveal this hidden property.

Consequently, the propositions can be simplified as shown in Eqs. (5) and (6):

$$LR_2 = \frac{Pr(Evidence|S_1, U_2, V)}{Pr(Evidence|U_1, U_2, V)} \quad (5)$$

$$LR_3 = \frac{Pr(Evidence|S_2, U_1, V)}{Pr(Evidence|U_1, U_2, V)} \quad (6)$$

Now there is only one *LR* and one non-contributor test per construct. Note that we evaluate the evidence in 'exploratory mode' since the proposition sets now deviate from the primary request of the mandating authorities illustrated in Eq. (4):

In the following, to continue the example provided in Table 3, by Gill and Haned [75] results are expressed as log_{10} values and the figures in parentheses are 99 percentiles from non-contributor distributions. It was demonstrated that the S_1 substitution in the non-contributor test(Eq. (3)) gave $LR = 5.5(-7)$ whereas S_2 substitution gave $LR = 5.5(8.2)$ i.e. the 99-percentile was greater than the 'combined' *LR* and therefore the S_2 result was exclusionary. When propositions were simplified (Eqs. (5) and (6), respectively) confirmation of the result was shown, $LR = 7.2(0.14)$ and -3 (0.14), respectively.

To summarise, when the propositions to be evaluated include several contributors, there is a chance of false inclusion of a questioned individual. This may be the case if a single large *LR* is used as simultaneous evidence against two suspects. If a non-contributor test is applied to a given suspect and the random man distribution overlaps the observed *LR* then this is exclusionary. The methods described above will capture such occurrences, which may otherwise be missed if total reliance is placed solely upon the value of the *LR*.

15.2. The role of the p-value

Due to the large number of possible random man profiles, the simulation based non-contributor test can only give an estimate of how much an observed *LR* differs from random man *LR*s, based on quantile measurement. An extension of the idea of the non-contributor test is to compute a *p*-value corresponding to the *LR*; an algorithm was presented by Dørum et al. [122]. The *p*-value asks a specific question, alternatively termed an 'exceedence probability' by Kruijver [123]:

> "What is the chance that a random (unrelated) man from a population will provide a likelihood ratio that is equivalent to or greater than that observed?"

In this context, the *p*-value algorithm considers all possible random man profiles and therefore gives the exact probability of providing misleading evidence in favour of the prosecution. Improvements in computation were suggested by Kruijver [123].

Another use of the non-contributor test or *p*-value algorithm is in database searches. In Bleka et al. [98] one approach to database search was to extract candidates with a LR above a selected threshold *T*. The probability of extracting false candidates in such searches, i.e. $Pr(LR \geq T)$, can be estimated with a non-contributor test or calculated exactly with the *p*-value algorithm.

Both Taylor et al. [124] and Kruijver et al. [125] recently discussed non-contributor (*p*-value) testing. However, [125] did not illustrate the issues with interpreting *LR*s for the multiple POI problem outlined above (Section 15.1).

Gill and Haned [75] did not advocate wholesale replacement of the *LR* by *p*-values. Rather, it was demonstrated that the interpretation of *LR*s derived from complex models has to proceed with much caution if the reciprocal *p*-value is less than the *LR*, as previously illustrated by the non-contributor tests in Section 15.1.

As already described, there is a *p*-value *per contributor*, whereas there is only one *LR* that can be calculated per pair of propositions. In addition, Gill and Haned [75] did not propose using a *p*-value to replace $LR < 1$ (i.e. clearly exclusionary), as suggested by Kruijver et al. [125].

On the contrary, once the propositions that make up the *LR* have been agreed, it is useful to carry out further analysis to ensure that the *LR* model proposed, is itself sound, to avoid the 'black-box'

approach and the associated potential dangers of 'garbage in, garbage out'.

15.3. Communicating ideas to the court

In the UK a number of court rulings disallow use of Bayesian statistics. The recent UK 'R. v. T.' judgment [126] amply illustrates the issues of poor communication between scientists and lawyers. Similar issues have arisen in relation to explaining the significance of database matches (section (12.1). Much more needs to be done to break down barriers and scientists have a responsibility to ensure that concepts are understood – this may require their simplification. There is little research that has been carried out to understand how lay-people conceive probabilistic thinking. A key contribution by Lindsey et al. [127] showed that the issue of understanding is related to cognitive effects. Lay-people have much trouble to understand probabilities, preferring to think in terms of 'natural frequencies'.[2] Indeed this method was proposed by the Doheny and Adams [128] court of appeal, in preference to use of Bayes theorem. New ways of thinking are needed to explain statistics in court. Pluralism, where multiple methods are utilised to explain and qualify evidence rather than a single dogmatic approach is needed. The application of non-contributor tests may well be useful as an adjunct to explain more complicated statistical concepts in the specific context of the courtroom – this is an area where much more research is now required.

To summarise

- If multiple POIs are present in the numerator of the likelihood ratio (eg two suspects) then a single likelihood ratio cannot be applied to determine individual strengths of evidence per contributor.
- The *LR* requires testing to ensure that it is robust. Non-contributor tests are the preferred method to do this. One test per POI in the numerator of the *LR* will identify potential exclusions that are otherwise hidden.
- Non-contributor statistics may play a role to explain evidence in the court-room.
- More research is needed to discover new methods to explain statistics to lay-persons. Non-contributor statistics will play a part in this endeavour.

16. Characterisation of STR profiles

The forensic community now has detailed understanding of the behaviour of STR multiplex systems. In order to interepret results, it is necessary to characterise loci by their key features, namely: heterozygote peak balance, inter-locus balance, stutter ratio, and the stochastic threshold. See validation recommendations of the Scientific Working Group on DNA Analysis Methods (SWGDAM) [129] and European Network of Forensic Science Institutes (ENFSI) DNA working group guidelines [130].

16.1. Heterozygote peak balance

Heterozygote balance (*Hb*) is the ratio of peak heights between the two alleles of a heterozygote. *In vivo*, there is perfect balance between the numbers of DNA molecules for a heterozygote locus. However, during forensic analysis, this balance is disrupted. Imbalance between two alleles of a single locus results from random (stochastic) sampling of DNA molecules [71]. During DNA extraction the link between a pair of (diploid) alleles is broken as the cell nucleus is disrupted. Pipetted aliquots of extracted material results in random sampling of alleles. As a consequence the variability of heterozygote balance increases as the template concentration decreases. [131–136].

Random sampling of alleles (i.e. the pipetting) and, to a much lesser extent, the PCR amplification[3] [71,139], account for the majority of variation in heterozygote peak balance. This has been confirmed by simulations [71,134,135]. There are several tools that can be used to simulate the extraction process and PCR: functions are available within the open source packages 'Forensim' [140] and 'PCRsim' [141]. A VBA Excel subroutine is provided in the supplement of [135]. These and other algebraic-based models [131,132,142] have been used to predict the heterozygote balance based on the quantity of input DNA.

Shorter DNA fragments are preferentially amplified. The effect is decreased peak height as the length of the amplified DNA fragment increases. Sample inhibition and degradation is common in casework samples so that high molecular weight alleles may drop-out completely. Tvedebrink et al. [132] showed a significant relationship between the difference in repeat units and the heterozygote balance in roughly half of the loci tested using two different kits. Kelly et al. [131] found that for each unit increase in repeat difference the natural logarithm of the heterozygote peak balance ($log_e(Hb)$) decreased on average 3%. Leclair et al. [143] observed reduced median and higher variability in casework samples compared to reference database samples where the latter was higher quality.

There are two common definitions of heterozygote balance: the high molecular weight allele peak height divided by the low molecular weight allele peak height (Eq. (7)) [131,132,136] (the converse may also be used [144,145]). The alternative method (Eq. (8)) calculates the smaller peak height divided by the larger peak height (irrespective of molecular weight of the allele) [135,146–149].

$$H'_b = \frac{\sigma_{HMW}}{\sigma_{LMW}} \tag{7}$$

$$H_b = \frac{\sigma_{smaller}}{\sigma_{larger}} \tag{8}$$

Where σ is the peak height; *HMW* and *LMW* refer to the high and low molecular weight allele, respectively; $\sigma_{smaller}$ and σ_{larger} represent the smaller and larger peak heights respectively.

Eq. (8) has been criticised as wasting information about the ordering of the alleles [131,150] but is often used for pragmatic simplicity. Both formulae ignore the repeat number difference between alleles.

Because heterozygote balance is essentially a product of the relative numbers of molecules that are randomly pipetted into the PCR mix, it is not surprising to find that no locus dependencies have been observed [131]. There is no effect of using different genetic analyzers of the same, or different models [139,133,135], or using different multiplex systems [136,132,135]. Further, it has been shown that the distributions of heterozygote balance for mixed and non-mixed stains for casework and artificial samples, created using pristine (extracted) DNA, are very similar [134].

A pragmatic guideline of $Hb \geq 0.60$ (Eq. (8)) can be used to define genotype combinations of good quality profiles. However it should be noted that primer binding site and somatic mutations can produce outliers. In general, *Hb* decreases as the average peak height decreases [151,152] (Fig. 3). However this relationship is

[2] e.g. How many people in a population could have contributed to the DNA profile?

[3] The PCR process itself is not 100% efficient. Some published values are 82% [71], 85% [137], and 82–97 [138].

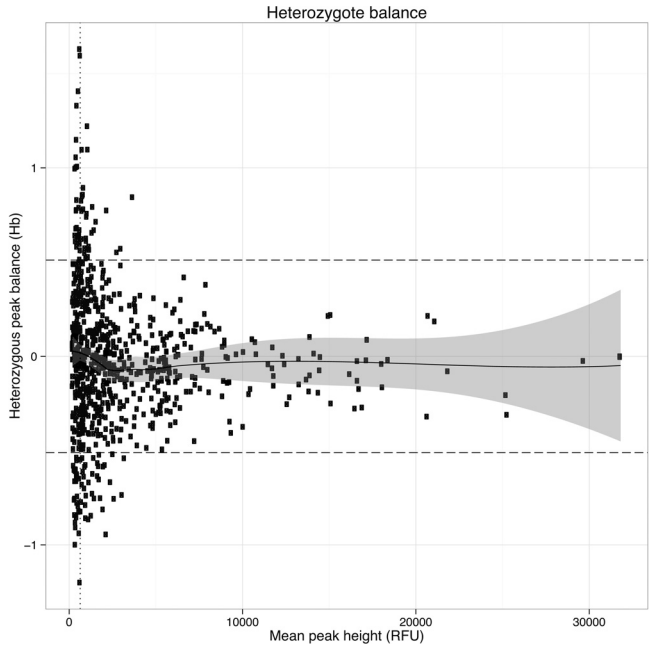

Fig. 3. Data generated from AB 3500xL Genetic Analyzer. Plot of average peak height vs. natural logarithm of *Hb* (764 data-points), showing increased variance as the stochastic threshold (*T* = 634 RFU) is approached (vertical dotted line). Data are the same used for the drop-out plot in Fig. 5 (but data below *LDT* = 200 RFU have been removed). The two horizontal dashed lines are the 60% *Hb* guideline that is used to interpret conventional DNA profiles. The *y*-axis is the natural logarithm of *Hb*. Analysis of data carried out using the balance module of R-package *strvalidator* (version 1.3), plot created using *ggplot2* (version 1.0) with the default loess regression smoothing.

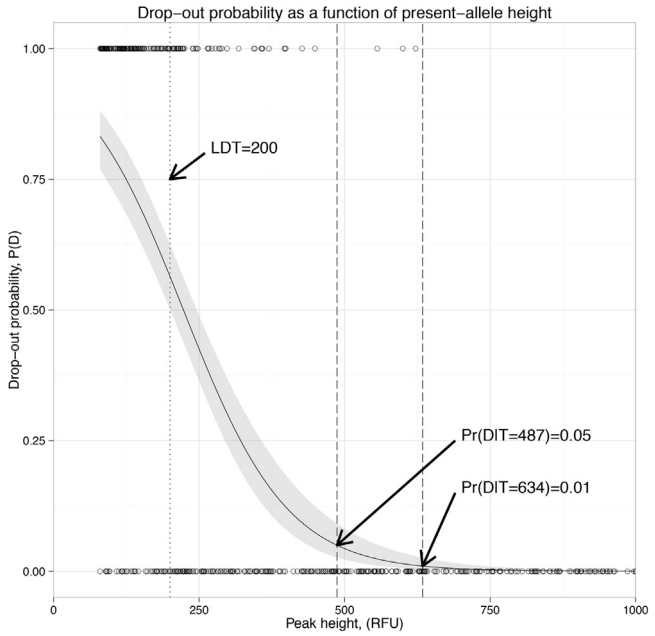

Fig. 5. Estimation of the stochastic threshold by logistic regression of known heterozygotes from the AB 3500xL Genetic Analyzer. The limit of detection (*LDT*) is 200 RFU with this instrument. The stochastic thresholds corresponding to *PrD* = 0.01 and 0.05 are at 634 and 487 RFUs respectively. These data are the same used to generate Fig. 3. Analysis of data was carried out using the drop-out module of the R-package *strvalidator* (version 1.3), plot customised using *ggplot2* (version 1.0).

not observed with Eq. (7), since increased variance is observed as DNA quantity decreases. This eventually leads to allele drop-out [131,139] and this is why the mean *Hb* decreases if data are analysed with equation 8.

16.2. Stochastic thresholds and logistic regression

Drop-out is an extreme form of heterozygote imbalance that is characteristic of low-template or partial DNA profile (Section 16.1) [2,5] and is specifically defined as an allelic signal that falls below the limit of detection threshold (*LDT*) [5] – this is the level where signals and background noise cannot be differentiated (Fig. 4). A stochastic or homozygote threshold (*T*) serves as an approximate delineation between a low-template (LT) and a conventional profile – however, a precise definition of LT is not possible (Section 8).

The stochastic threshold can be defined by estimating the probability of drop-out (e.g. *PrD* = 0.05) relative to peak height, determined by logistic regression of a series of samples of varying quantity, as recommended by the DNA Commission [5] (Fig. 5). A risk analysis associated with choice of PrD to designate the threshold is provided by Gill et al. [153] and demonstrated by Kirkham et al. [147]. Alternative ways to estimate the stochastic threshold have been published: empirical cumulative distribution of peak heights from single heterozygote peaks [154], peak height of the largest observed single heterozygote allele [136,155], and variance of heterozygote peak balance [156]. Butler [157], page 95, compares stochastic thresholds between the ABI 3130 and ABI 3500 instruments (the latter is much more sensitive).

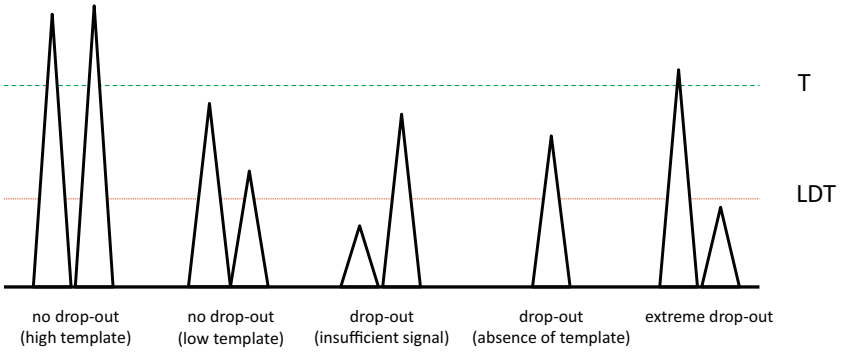

Fig. 4. High-template profiles usually have well balanced heterozygote peaks and no drop-out. Low-template profiles are characterised by the risk of drop-out. They may well show complete genotypes with no drop-out, but usually with increased imbalance. Upon analysis this can lead to three types of drop-out; insufficient number of molecules to generate a signal above the *LDT*, complete drop-out which is the absence of a molecule in the PCR reaction, and extreme drop-out with one allele above the stochastic threshold (*T*) and one below the *LDT*.

The probability of drop-out has been shown to be locus-dependent [158,159]. It is also affected by the limit of detection threshold [160,161]. Although drop-out and allele length have been shown to correlate [161], it was also suggested by Lohmueller et al. [161] that logistic regressions based on average drop-out estimation resulted in robust LRs. This relationship also held true with data from different STR multiplex systems [161]. However, the differences can be quite large between capillary electrophoresis instruments [147,155]; utilisation of different numbers of PCR cycles has a significant effect [132,159].

Drop-out probability has been modelled by Gill et al. [153] who used the peak height of the surviving heterozygote peak as a predictor. Tvedebrink et al. [162] used the average peak height of the profile instead, arguing that this has lower variability than using a single peak observation [158]. Both methods rely upon empirical data. There are also simulation based approaches that estimate the probability of drop-out using the crime scene profile and stated hypotheses. Balding et al. [65] used a simulated annealing algorithm to maximize the probability of evidence, while Haned et al. [61] used Monte-Carlo simulations to generate a distribution of drop-out probabilities that would result in the observed number of alleles.

16.3. Stutter peaks

Stutter peaks complicate DNA mixture interpretation and are therefore important to characterise [163]. Stutters are artefacts caused by mispairing of the DNA strands during the PCR [164]. The phenomenon is often referred to as the 'strand slippage model' or 'slipped strand mispairing/displacement model' and is a natural mechanism for DNA sequence evolution [165].

Stutter peak size is characterised by the stutter ratio (S_R) or less commonly the stutter proportion (S_x):

$$S_R = \frac{\sigma_S}{\sigma_A} \tag{9}$$

$$S_x = \left(\frac{\sigma_S}{\sigma_A + \sigma_S}\right) \tag{10}$$

where σ_S is the height of the stutter peak and σ_A is the height of the allelic peak.

The most common and pronounced stutter is one repeat unit shorter than the parental allelic peak (back stutter). Back stutters commonly have a 95th percentile $S_R \leq 0.15$ (Eq. (9)) [149,166]. Stochastic effects, especially associated with low-template samples, can produce outliers. In addition, somatic mutations are often found in stutter positions, and this may contribute to abnormally high peaks [74,167]. These effects can be investigated by replicate studies, in order to determine if such an event has occurred.

Forward stutters (one repeat longer), double back stutters (two repeats shorter), and intermediate stutters (e.g. two basepair shorter in D1S1656 and SE33) are also observed but to a much lesser extent – these complex stutters are observed with high signals where samples may be over-amplified [136,152,168–170]. All stutters complicate mixture interpretation. In addition, increased signal range of new, highly sensitive, capillary electrophoresis instruments (e.g. ABI 3500) lead to more frequent detection of stutters.

The general trend is that increased stutter is observed with increasing number of repeats for the parent allele [132,143,149,154,169]. However, there are microvariant alleles (e.g. allele 9.3 in TH01) and sequence variants (e.g. in SE33) interrupting the number of consecutive repeats of the same type. The effect is less stutter, and the longest uninterrupted repeat stretch is a better predictor than the total number of repeats [132,142,164,171]. In a clever and highly controlled study using synthetic STR fragments, Brookes et al. confirmed these findings [172]. Furthermore, they showed that high AT content in the synthetic fragments increased the stutter ratio. This is explained by the lower bond strength: there are two hydrogen bonds in an AT base pair compared to three in a GC base pair. However, the finding was contradicted by analysis of reference data. Nevertheless, it is well established that the repeat sequence is important and that the degree of stutter formation differ between loci [143,149,154].

The stutter ratio is also affected by the size of the repeat unit, hence the tri-nucleotide repeat locus D22S1045 is expected to show higher stutter ratios than tetra-nucleotide repeat locus [154]. The lower intensity of forward stutters compared to back stutters may be caused by structural limitations within the *Taq* enzyme, or the higher energy requirement for a forward shift to occur [137]. Leclair et al. [143] found similar stutter ratios in casework and database samples. Optimising PCR conditions by lowering the annealing and extension temperatures has been shown to decrease the heights of stutter peaks [173]. At low template levels, the stutter ratio gradually increased as the peak height of the parent allele decreased towards the *LDT*. This effect can be explained by the additive effect of stutter with background noise peaks [132,143].

Considerable efforts have been made to model stutter ratios [132,142,174]; PCR simulation of stutter formation was demonstrated by Gill et al. [71] and Weusten et al. [175].

16.4. The use of open source software to analyse characteristics

Validation of a new STR multiplexes, analysis instruments or extraction methods is required before routine use [129,130]. The process of validation is resource intensive and time consuming, which often delays the introduction of new technology. There are freely available open-source tools to perform the necessary calculations required, that can speed up this process.

Several Microsoft Excel/VBA based programs have been developed by David Duewer at the National Institute of Standards and Technology (NIST). The programs are freely available on the NIST STRBase website.[4] There are functions to calculate stutter percentage, characterise peak height ratios, calculate STR allele frequencies and population genetic metrics. Oskar Hansson at the National Institute of Public Health (NIPH) has developed an R program called *STR-validator* (R package *strvalidator*) with an easy-to-use graphical user interface [176]. The program performs all necessary calculations required for a validation study, including concordance and mixture studies. A complete manual can be found on-line at https://sites.google.com/site/forensicapps/strvalidator.

17. The evolving interpretation strategy

The drive to report complex DNA profiles using innovative software is supported by the ISFG DNA commission, which published a number of recommendations for users to interpret complex DNA profiles where drop-out and drop-in are considerations [5]. These recommendations are a result of consensus international agreement on ways to interpret difficult DNA profiles, especially those that are low level, and subject to the phenomenon of allele drop-out/ drop-in or where stutters may confuse interpretation.

This work continues. The aim is to consult widely between the major scientific societies, including ENFSI, EDNAP in order to produce additional authoritative documents. Training initiatives are supported by the ISFG and the EU funded Euroforgen (Network

[4] http://www.cstl.nist.gov/strbase/software.htm.

of Excellence) http://www.euroforgen.eu/ and collaborative exercises are in progress and have been completed [177].

Acknowlegements

We are grateful to an anonymous referee who greatly improved the manuscript. PG, TE, OH, OB, GD have received funding support from the European Union Seventh Framework Programme (FP7/2007-2013), Euroforgen-NOE, under Grant agreement No. 285487.

References

[1] L. Gusmao, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider, DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, Forensic Sci. Int. 157 (2–3) (2006) 187–197.
[2] 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 (2–3) (2006) 90–101.
[5] P. Gill, L. Gusmao, 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. 6 (6) (2012) 679–688.
[4] C.P. Kimpton, P. Gill, A. Walton, A. Urquhart, E.S. Millican, M. Adams, Automated DNA profiling employing multiplex amplification of short tandem repeat loci., Genome Res. 3 (1) (1993) 13–22.
[5] K.M. Sullivan, A. Mannucci, C.P. Kimpton, P. Gill, A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin, Biotechniques 15 (4) (1993) 636–638, 640–641.
[6] K.A. Mills, D. Even, J.C. Murray, Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA), Hum. Mol. Genet. 1 (9) (1992) 779.
[7] D. Werrett, The national DNA database, Forensic Sci. Int. 88 (1997) 33–42.
[8] E. Cotton, R. Allsop, J. Guest, R. Frazier, P. Koumi, I. Callow, A. Seager, R. Sparkes, Validation of the AMPFlSTR®SGM Plus™ system for use in forensic casework, Forensic Sci. Int. 112 (2) (2000) 151–161.
[9] C. Kimpton, P. Gill, E. D'Aloja, J.F. Andersen, W. Bar, S. Holgersson, S. Jacobsen, V. Johnsson, A.D. Kloosterman, M.V. Lareu, et al., Report on the second EDNAP collaborative STR exercise. European DNA Profiling Group, Forensic Sci. Int. 71 (2) (1995) 137–152.
[10] P. Gill, E. d'Aloja, J. Andersen, B. Dupuy, M. Jangblad, V. Johnsson, A.D. Kloosterman, A. Kratzer, M.V. Lareu, M. Meldegaard, C. Phillips, H. Pfitzinger, S. Rand, M. Sabatier, R. Scheithauer, H. Schmitter, P. Schneider, M.C. Vide, Report of the European DNA profiling group (EDNAP): an investigation of the complex STR loci D21S11 and HUMFIBRA (FGA), Forensic Sci. Int. 86 (1–2) (1997) 25–33.
[11] W. Bar, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, W. Mayr, B. Olaisen, DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics, Int. J. Leg. Med. 110 (4) (1997) 175–176.
[12] P.M. Schneider, P.D. Martin, Criminal DNA databases: the European situation, Forensic Sci. Int. 119 (2) (2001) 232–238.
[13] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Sci. Int. 112 (1) (2000) 17–40.
[14] P.D. Martin, National DNA databases - practice and practicability. a forum for discussion., Prog. Forensic Genet. 10 (2004) 1–8.
[15] C.J. Fregeau, National casework and the national DNA database: the Royal Canadian Mounted Police perspective, Prog. Forensic Genet. 7 (1998) 541–543.
[16] J. Walsh, Canada's proposed forensic DNA evidence bank, Can. Soc. Forensic Sci. J. 31 (1998) 113–125.
[17] R. Hoyle, Forensics. The FBI's national DNA database, Nat. Biotechnol. 16 (11) (1998) 987.
[18] J. Butler, U.S. Initiatives to Strengthen Forensic Science and International Standards in Forensic DNA, Forensic Sci. Int. Genet. (2015) (in press).
[19] A. Leriche, Final report of the Interpol Working Party on DNA profiling., in: Proceedings from the 2nd European Symposium on Human Identification, Promega Corporation, 1998, pp. 48–54.
[20] Prum Decision http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/police-cooperation/prum-decision/index_en.htm.
[21] M.D. Coble, J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA, J. Forensic Sci. 50 (1) (2005) 43–53.
[22] L.A. Dixon, A.E. Dobbins, H.K. Pulker, J.M. Butler, P.M. Vallone, M.D. Coble, W. Parson, B. Berger, P. Grubwieser, H.S. Mogensen, N. Morling, K. Nielsen, J.J. Sanchez, E. Petkovski, A. Carracedo, P. Sanchez-Diz, E. Ramos-Luis, M. Brion, J.A. Irwin, R.S. Just, O. Loreille, T.J. Parsons, D. Syndercombe-Court, H. Schmitter, B. Stradmann-Bellinghausen, K. Bender, P. Gill, Analysis of artificially degraded DNA using STRs and SNPs-results of a collaborative European (EDNAP) exercise, Forensic Sci. Int. 164 (1) (2006) 33–44.
[23] P. Gill, L. Fereday, N. Morling, P.M. Schneider, The evolution of DNA databases–recommendations for new European STR loci, Forensic Sci. Int. 156 (2) (2006) 242–244.
[24] P. Gill, L. Fereday, N. Morling, P.M. Schneider, New multiplexes for Europe-amendments and clarification of strategic development, Forensic Sci. Int. 163 (1–2) (2006) 155–157.
[25] Council Resolution of 30 November 2009 on the exchange of DNA analysis results. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2009:296:0001:0003:EN:PDF (2009).
[26] L. Welch, P. Gill, C. Phillips, R. Ansell, N. Morling, W. Parson, J. Palo, I. Bastisch, European Network of Forensic Science Institutes (ENFSI): evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers, Forensic Sci. Int. Genet. 6 (6) (2012) 819–826.
[27] J. Ge, A. Eisenberg, B. Budowle, Developing criteria and data to determine best options for expanding the core CODIS loci, Investig. Genet. 3 (2012) 1.
[28] D.R. Hares, Addendum to expanding the CODIS core loci in the United States, Forensic Sci. Int. Genet. 6 (5) (2012) e135.
[29] D.R. Hares, Expanding the CODIS core loci in the United States, Forensic Sci. Int. Genet. 6 (1) (2012) e52–e54.
[30] National Research Council, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996.
[31] L.A. Foreman, J.A. Lambert, I.W. Evett, Regional genetic variation in Caucasians, Forensic Sci. Int. 95 (1) (1998) 27–37.
[32] D.J. Balding, M. Greenhalgh, R.A. Nichols, Population genetics of STR loci in caucasians, Int. J. Leg. Med. 108 (6) (1996) 300–305.
[33] B. Budowle, T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians, J. Forensic Sci. 44 (6) (1999) 1277–1286.
[34] L.A. Foreman, I.W. Evett, Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system, Int. J. Leg. Med. 114 (3) (2001) 147–155.
[35] P. Gill, I. Evett, Population genetics of short tandem repeat (STR) loci, Genetica 96 (1–2) (1995) 69–87.
[36] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (2–3) (1994) 125–140.
[37] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Ann. Rev. Stat. Appl. 1 (2014) 361–384.
[38] P. Gill, L. Foreman, J.S. Buckleton, C.M. Triggs, H. Allen, A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations, Forensic Sci. Int. 131 (2) (2003) 184–196.
[39] I. Findlay, A. Taylor, P. Quirke, R. Frazier, A. Urquhart, DNA fingerprinting from single cells, Nature 389 (6651) (1997) 555–556.
[40] P. Wiegand, M. Kleiber, DNA typing of epithelial cells after strangulation, Int. J. Leg. Med. 110 (4) (1997) 181–183.
[41] D. Van Hoofstat, D. Deforce, V. Brochez, I. De Pauw, K. Janssens, M. Mestdagh, R. Millecamps, E. Van Geldre, E. Van Den Eeckhout, DNA typing of fingerprints and skin debris: sensitivity of capillary electrophoresis in forensic applications using multiplex PCR, in: Proceedings of the Second European Symposium on Human Identification, 1998, pp. 131–137.
[42] A. Barbaro, G. Falcone, A. Barbaro, DNA typing from hair shaft, Prog. Forensic Genet. 8 (2000) 523–525.
[43] A. Hellmann, U. Rohleder, H. Schmitter, M. Wittig, STR typing of human telogen hairs-a new approach, Int. J. Leg. Med. 114 (4–5) (2001) 269–273.
[44] R. Szibor, I. Plate, H. Schmitter, H. Wittig, D. Krause, Forensic mass screening using mtDNA, Int. J. Leg. Med. 120 (6) (2006) 372–376.
[45] P. Gill, P.L. Ivanov, C. Kimpton, R. Piercy, N. Benson, G. Tully, I. Evett, E. Hagelberg, K. Sullivan, Identification of the remains of the Romanov family by DNA analysis, Nat. Genet. 6 (2) (1994) 130–135.
[46] W.M. Schmerer, S. Hummel, B. Herrmann, Optimized DNA extraction to improve reproducibility of short tandem repeat genotyping with highly degraded DNA as target, Electrophoresis 20 (8) (1999) 1712–1716.
[47] W.M. Schmerer, S. Hummel, B. Herrmann, STR-genotyping of archaeological human bone: experimental design to improve reproducibility by optimisation of DNA extraction, Anthropol. Anz. 58 (1) (2000) 29–35.
[48] J. Burger, S. Hummel, B. Hermann, W. Henke, DNA preservation: a microsatellite-DNA study on ancient skeletal remains, Electrophoresis 20 (8) (1999) 1722–1728.
[49] C.M. Strom, S. Rechitsky, Use of nested PCR to identify charred human remains and minute amounts of blood, J. Forensic Sci. 43 (3) (1998) 696–700.
[50] R. Van Oorschot, K.N. Ballantyne, R.J. Mitchell, Forensic trace DNA: a review, Investig. Genet. 1 (1) (2010) 14.
[51] 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.
[52] P. Gill, R. Brown, M. Fairley, L. Lee, M. Smyth, M. Simpson, B. Irwin, J. Dunlop, M. Greenhalgh, K. Way, E. Westacott, S. Ferguson, L. Ford, T. Clayton, J. Guiness, National recommendations of the technical UK DNA working group on mixture interpretation for the NDNADB and for court going purposes, Forensic Sci. Int. Genet. 2 (2008) 76–82.
[53] P. Gill, J. Buckleton, A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number, Forensic Sci. Int. Genet. 4 (4) (2010) 221–227.
[54] I. Zupanič Pajnič, B. Gornjak Pogorelc, J. Balažic, T. Zupanc, B. Štefanič, Highly efficient nuclear DNA typing of the World War II skeletal remains using three

new autosomal short tandem repeat amplification kits with the extended European Standard Set of loci, Croat. Med. J. 53 (1) (2012) 17–23.
[55] H. Kelly, J.-A. Bright, J.S. Buckleton, J.M. Curran, A comparison of statistical models for the analysis of complex forensic dna profiles, Sci. Justice 54 (1) (2014) 66–70.
[56] J. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, London, 2005.
[57] M. Bill, P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy, J. Buckleton, PENDULUM-a guideline-based approach to the interpretation of STR mixtures, Forensic Sci. Int. 148 (2–3) (2005) 181–189.
[58] P. Gill, A. Kirkham, J. Curran, LoComatioN: A software tool for the analysis of low copy number DNA profiles, Forensic Sci. Int. 166 (2007) 128–138.
[59] M.W. Perlin, B. Szabady, Linear mixture analysis: a mathematical approach to resolving mixed DNA samples, J. Forensic Sci. 46 (6) (2001) 1372–1378.
[60] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, T. Caragine, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (6) (2012) 749–761.
[61] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (6) (2012) 762–774.
[62] M. Perlin, M. Legler, C. Spencer, J. Smith, W. Allan, J. Belrose, B. Duceman, Validating Trueallele® DNA mixture interpretation, J. Forensic Sci. 56 (6) (2011) 1430–1447.
[63] D. Taylor, J.A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[64] R. Puch-Solis, T. Clayton, Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRa software, Forensic Sci. Int. Genet. 11 (2014) 220–228.
[65] D.J. Balding, Evaluation of mixed-source, low-template DNA profiles in forensic science, Proc. Natl. Acad. Sci. U. S. A. 110 (30) (2013) 12241–12246.
[66] T. Graversen, S. Lauritzen, Computational aspects of DNA mixture analysis, Stat. Comput., 2014, pp. 1–15.
[67] Lab retriever. http://scieg.org/lab_retriever.html.
[68] C. Benschop, T. Sijen, LoCIM-tool: an expert's assistant for inferring the major contributor's alleles in mixed consensus DNA profiles, Forensic Sci. Int. Genet. 11 (2014) 154–165.
[69] R. Puch-Solis, L. Rodgers, A. Mazumder, S. Pope, I.W. Evett, J. Curran, D. Balding, Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int. Genet. 7 (5) (2013) 555–563.
[70] R.G. Cowell, T. Graversen, S.L. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts, Appl. Stat. 64 (1) (2015) 1–32.
[71] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2) (2005) 632–643.
[72] P. Gill, Misleading DNA Evidence: Reasons for Miscarriages of Justice, Elsevier, 2014.
[73] T. Clayton, S. Hill, L. Denton, S. Watson, A. Urquhart, Primer binding site mutations affecting the typing of STR loci contained within the AMPFlSTR®SGM Plus kit, Forensic Sci. Int. 139 (2) (2004) 255–259.
[74] T.M. Clayton, J.L. Guest, A.J. Urquhart, P.D. Gill, A genetic basis for anomalous band patterns encountered during DNA STR profiling, J. Forensic Sci. 49 (6) (2004) 1207–1214.
[75] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2) (2013) 251–263.
[76] I. Evett, Evaluating DNA profiles in a case where the defence is "it was my brother", J. Forensic Sci. Soc. 32 (1) (1992) 5–14.
[77] W.K. Fung, Y.Q. Hu, Statistical DNA Forensics: Theory, Methods and Computation, Wiley, England, 2008.
[78] T. Egeland, G. Dorum, M.D. Vigeland, N.A. Sheehan, Mixtures with relatives: a pedigree perspective, Forensic Sci. Int. Genet. 10 (2014) 49–54.
[79] D. Balding, M. Krawczak, J. Buckleton, J. Curran, Decision-making in familial database searching: KI alone or not alone? Forensic Sci. Int. Genet. 7 (2013) 52–54.
[80] C. Brenner, B. Weir, Issues and strategies in the DNA identification of World Trade Center victims, Theor. Popul. Biol. 63 (3) (2003) 173–178.
[81] J. Buckleton, C. Triggs, Dealing with allelic dropout when reporting the evidential value in DNA relatedness analysis, Forensic Sci. Int. Genet. 160 (2–3) (2006) 134–139.
[82] G. Dørum, D. Kling, C. Baeza-Richer, M. García-Magariños, S. Sæbø, S. Desmyter, T. Egeland, Models and implementation for relationship problems with dropout, Int. J. Leg. Med. (2014) 1–13.
[83] C.B. van Dongen, K. Slooten, W. Burgers, W. Wiegerinck, Bayesian networks for victim identification on the basis of DNA profiles, Forensic Sci. Int. Genet. Suppl. Ser. 2 (1) (2009) 466–468.
[84] D. Kling, A.O. Tillmar, T. Egeland, Familias 3: extensions and new functionality, Forensic Sci. Int. Genet. 13 (2014) 121–127.
[85] Home Office National DNA database strategy board annual report 2013–2014 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/387581/NationalDNAdatabase201314.pdf.
[86] F.R. Bieber, C.H. Brenner, D. Lazer, Human genetics. finding criminals through DNA of their relatives, Science 312 (5778) (2006) 1315–1316.
[87] K. Slooten, R. Meester, Probabilistic strategies for familial DNA searching, J. R. Stat. Soc. Ser. C: Appl. Stat. 63 (3) (2014) 361–384.
[88] S. Cowen, J. Thomson, A likelihood ratio approach to familial searching of large DNA databases, Forensic Sci. Int. Genet. Suppl. Ser. 1 (2008) 643–645.
[89] Y. Chung, W. Fung, Y. Hu, Familial database search on two-person mixture, Comput. Stat. Data Anal. 54 (2010) 2046–2051.
[90] J. Ge, R. Chakraborty, A. Eisenberg, B. Budowle, Comparisons of familial DNA database searching strategies, J. Forensic Sci. 56 (6) (2011) 1448–1456.
[91] S.P. Myers, M.D. Timken, M.L. Piucci, G.A. Sims, M.A. Greenwald, J.J. Weigand, K.C. Konzak, M.R. Buoncristiani, Searching for first-degree familial relationships in california's offender dna database: validation of a likelihood ratio-based approach, Forensic Sci. Int. Genet. 5 (5) (2011) 493–500.
[92] P. Gill, C. Brenner, B. Brinkmann, B. Budowle, A. Carracedo, M.A. Jobling, P. de Knijff, M. Kayser, M. Krawczak, W.R. Mayr, N. Morling, B. Olaisen, V. Pascali, M. Prinz, L. Roewer, P.M. Schneider, A. Sajantila, C. Tyler-Smith, DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs, Forensic Sci. Int. 124 (1) (2001) 5–10.
[93] M.A. Jobling, P. Gill, Encoded evidence: DNA in forensic analysis, Nat. Rev. Genet. 5 (10) (2004) 739–751.
[94] C. Phillips, Bio-geographical ancestry, Forensic Sci. Int. Genet. (2015) (in press).
[95] P. Gill, A. Kirkham, Development of a simulation model to assess the impact of contamination in casework using STRs., J. Forensic Sci. 49 (3) (2004) 485–491.
[96] T. Hicks, F. Taroni, J. Curran, J. Buckleton, O. Ribaux, V. Castella, Use of DNA profiles for investigation using a simulated national DNA database: Part I. Partial SGM Plus profiles., Forensic Sci. Int. Genet. 4 (4) (2010) 232–238.
[97] T. Hicks, F. Taroni, J. Curran, J. Buckleton, V. Castella, O. Ribaux, Use of DNA profiles for investigation using a simulated national DNA database: Part II. Statistical and ethical considerations on familial searching, Forensic Sci. Int. Genet. 4 (5) (2010) 316–322.
[98] O. Bleka, G. Dorum, P. Gill, H. Haned, Database extraction strategies for low-template evidence, Forensic Sci. Int. Genet. 9 (2014) 134–141.
[99] F. Van Nieuwerburgh, E. Goetghebeur, M. Vandewoestyne, D. Deforce, Impact of allelic dropout on evidential value of forensic DNA profiles using RMNE, Bioinformatics 25 (2009) 225–229.
[100] T. Tvedebrink, P.S. Eriksen, J.M. Curran, H.S. Mogensen, N. Morling, Analysis of matches and partial-matches in a Danish STR data set, Forensic Sci. Int. Genet. 6 (3) (2012) 387–392.
[101] Interpol Handbook on DNA Data Exchange: Recommendations from the Interpol DNA monitoring expert group. http://www.interpol.int/content/download/10460/74503/version/7/file/handbookpublic2009[2].pdf (2009).
[102] The Council of The European Union: COUNCIL DECISION 2008/616/JHA of 23 June 2008 on the implementation of Decision 2008/615/JHA on the stepping up of cross-border cooperation, particularly in combating terrorism and cross-border crime. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008D0616 (2008).
[103] A. Stockmarr, Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search, Biometrics 55 (1999) 671–677.
[104] D.J. Balding, P. Donnelly, Evaluating DNA profile evidence when the suspect is identified through a database search, J. Forensic Sci. 41 (4) (1996) 603–607.
[105] R. Meester, M. Sjerps, The evidential value in the DNA database search controversy and the two-stain problem., Biometrics 59 (3) (2003) 727–732.
[106] P.M. Schneider, H. Schneider, R. Fimmers, W. Keil, G. Molsberger, W. Pflug, T. Rothämel, M. Eckert, H. Pfeiffer, B. Brinkmann, et al., Allgemeine Empfehlungen der Spurenkommission zur statistischen Bewertung von DNA-Datenbank-Treffern, Rechtsmedizin 20 (2) (2010) 111–115.
[107] A. Biedermann, S. Gittelson, F. Taroni, Recent misconceptions about the 'database search problem': a probabilistic analysis using Bayesian networks, Forensic Sci. Int. 212 (1) (2011) 51–60.
[108] S. Gittelson, A. Biedermann, S. Bozza, F. Taroni, The database search problem: A question of rational decision making, Forensic Sci. Int. 222 (1) (2012) 186–199.
[109] A. Nordgaard, K. Hedberg, C. Widén, R. Ansell, Comments on 'the database search problem' with respect to a recent publication in Forensic Sci. Int., Forensic Sci. Int. 217 (1) (2012) e32–e33.
[110] P. Gill, Ø. Bleka, T. Egeland, Does an english appeal court ruling increase the risks of miscarriages of justice when complex dna profiles are searched against the national dna database? Forensic Sci. Int. Genet. 13 (2014) 167–175.
[111] T.E Geir Storvik, The DNA database search controversy revisited: bridging the Bayesian-frequentist gap, Biometrics 63 (2007) 922–925.
[112] Y. Chung, Y. Hu, W. Fung, Evaluation of DNA mixtures from database search., Biometrics 66 (2009) 233–238.
[113] UK National DNA Database website. https://www.gov.uk/government/statistics/national-dna-database-statistics (2013).
[114] N. Zaken, U. Motro, R. Berdugo, L.E. Sapir, A. Zamir, Can brothers share the same STR profile? Forensic Sci. Int. Genet. 7 (5) (2013) 494–498.
[115] J. Curran, P. Gill, M. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53.
[116] J.A. Bright, D. Taylor, J. Curran, J. Buckleton, Searching mixed DNA profiles directly against profile databases, Forensic Sci. Int. Genet. 9 (2014) 102–110.
[117] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Identifying contributors of DNA mixtures by means of quantitative information of STR typing, J. Comput. Biol. 19 (7) (2012) 887–902.
[118] J. Buckleton, J.-A. Bright, D. Taylor, I. Evett, T. Hicks, G. Jackson, J.M. Curran, Helping formulate propositions in forensic DNA analysis, Sci. Justice 54 (2014) 258–261.
[119] P. Gill, J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, J. Lambert, Interpretation of complex DNA profiles using empirical models and a method to measure their robustness, Forensic Sci. Int. Genet. 2 (2008) 91–103.
[120] D.J. Balding, The likeLTD software: an illustrative analysis, explanation of the model, results of performance tests and version history: https://sites.google.com/site/baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code.

[121] H. Haned, G. Dørum, T. Egeland, P. Gill, On the meaning of the likelihood ratio: Is a large number always an indication of strength of evidence? Forensic Sci. Int. Genet. Suppl. Ser. 4 (1) (2013) e176–e177.

[122] G. Dorum, O. Bleka, P. Gill, H. Haned, L. Snipen, S. Sabo, T. Egeland, Exact computation of the distribution of likelihood ratios with forensic applications, Forensic Sci. Int. Genet. 9 (2014) 93–101.

[123] M. Kruijver, Efficient computations with the likelihood ratio distribution, Forensic Sci. Int. Genet. 14 (2015) 116–124.

[124] D. Taylor, J. Buckleton, I. Evett, Testing likelihood ratios produced from complex {DNA} profiles, Forensic Sci. Int. Genet. 16 (0) (2015) 165–171. , http://dx.doi.org/10.1016/j.fsigen.2015.01.008, http://www.sciencedirect.com/science/article/pii/S1872497315000265.

[125] M. Kruijver, R. Meester, K. Slooten, p-Values should not be used for evaluating the strength of {DNA} evidence, Forensic Sci. Int. Genet. 16 (2015) 226–231. , http://dx.doi.org/10.1016/j.fsigen.2015.01.005, http://www.sciencedirect.com/science/article/pii/S1872497315000150.

[126] C.E. Berger, J. Buckleton, C. Champod, I.W. Evett, G. Jackson, Evidence evaluation: a response to the court of appeal judgment in r v t, Sci. Justice 51 (2) (2011) 43–49.

[127] S. Lindsey, R. Hertwig, G. Gigerenzer, Communicating statistical dna evidence, Jurimetrics (2003) 147–163.

[128] R v Doheny and Adams [1997] 1 Cr App. R. 369.

[129] SWGDAM, Validation guidelines for forensic DNA analysis methods (2012) http://swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf.

[130] ENFSI recommended minimum criteria for the validation of various aspects of the DNA profiling process http://www.enfsi.eu/documents/minimum-validation-guidelines-dna-profiling-v2010.

[131] H. Kelly, J.-A. Bright, J.M. Curran, J. Buckleton, Modelling heterozygote balance in forensic DNA profiles, Forensic Sci. Int. Genet. 6 (6) (2012) 729–734.

[132] T. Tvedebrink, H.S. Mogensen, M.C. Stene, N. Morling, Performance of two 17 locus forensic identification STR kits–Applied Biosystems's AmpFℓSTR® NGMSElect™ and Promega's PowerPlex®ESI17 kits, Forensic Sci. Int. Genet. 6 (5) (2012) 523–531.

[133] J.-A. Bright, S. Neville, J.M. Curran, J.S. Buckleton, Variability of mixed DNA profiles separated on a 3130 and 3500 capillary electrophoresis instrument, Aust. J. Forensic Sci. 46 (3) (2013) 304–312.

[134] J.-A. Bright, K. McManus, S. Harbison, P. Gill, J. Buckleton, A comparison of stochastic variation in mixed and unmixed casework and synthetic samples, Forensic Sci. Int. Genet. 6 (2) (2012) 180–184.

[135] M.D. Timken, S.B. Klein, M.R. Buoncristiani, Stochastic sampling effects in STR typing: implications for analysis and interpretation, Forensic Sci. Int. Genet. 11 (2014) 195–204.

[136] J.-A. Bright, E. Huizing, L. Melia, J. Buckleton, Determination of the variables affecting mixed MiniFiler™ DNA profiles, Forensic Sci. Int. Genet. 5 (5) (2011) 381–385.

[137] D. Shinde, Y. Lai, F. Sun, N. Arnheim, Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites, Nucleic Acids Res. 31 (3) (2003) 974–980.

[138] W.R. Hudlow, M.D. Chong, K.L. Swango, M.D. Timken, M.R. Buoncristiani, A quadruplex real-time qPCR assay for the simultaneous assessment of total human DNA, human male DNA, DNA degradation and the presence of PCR inhibitors in forensic samples: a diagnostic tool for STR typing, Forensic Sci. Int. Genet. 2 (2) (2008) 108–125.

[139] A. Debernardi, E. Suzanne, A. Formant, L. Pene, A.B. Dufour, J.R. Lobry, One year variability of peak heights, heterozygous balance and inter-locus balance for the DNA positive control of AmpFℓSTR© Identifiler©STR kit, Forensic Sci. Int. Genet. 5 (1) (2011) 43–49.

[140] H. Haned, Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (4) (2011) 265–268.

[141] O. Hansson, P. Gill, Free open source software for internal validation of forensic STR typing kits, Forensic Sci. Int. Genet. Suppl. Ser. 4 (1) (2013) e300–e301.

[142] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2) (2013) 296–304.

[143] B. Leclair, C.J. Frégeau, K.L. Bowen, R.M. Fourney, Systematic analysis of stutter percentages and allele peak height and peak area ratios at heterozygous STR loci for forensic casework and database samples, J. Forensic Sci. 49 (5) (2004) 968–980.

[144] J. Whitaker, E. Cotton, P. Gill, A comparison of the characteristics of profiles produced with the AMPFlSTR®SGM Plus™ multiplex system for both standard and low copy number (LCN) STR DNA analysis, Forensic Sci. Int. 123 (2) (2001) 215–223.

[145] P. Gill, R. Sparkes, L. Fereday, D.J. Werrett, Report of the European Network of Forensic Science Institutes (ENSFI): formulation and testing of principles to evaluate STR multiplexes, Forensic Sci. Int. 108 (1) (2000) 1–29.

[146] T. Clayton, J. Whitaker, R. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int. 91 (1) (1998) 55–70.

[147] A. Kirkham, J. Haley, Y. Haile, A. Grout, C. Kimpton, A. Al-Marzouqi, P. Gill, High-throughput analysis using AmpFlSTR® identifiler® with the applied biosystems 3500 Genetic Analyser, Forensic Sci. Int. Genet. 1 (2013) 92–97.

[148] S. Petricevic, J. Whitaker, J. Buckleton, S. Vintiner, J. Patel, P. Simon, H. Ferraby, W. Hermiz, A. Russell, Validation and development of interpretation guidelines for low copy number (LCN) DNA profiling in New Zealand using the AmpF lSTR®SGM Plus™ multiplex, Forensic Sci. Int. Genet. 4 (5) (2010) 305–310.

[149] C.R. Hill, D.L. Duewer, M.C. Kline, C.J. Sprecher, R.S. McLaren, D.R. Rabbach, B.E. Krenke, M.G. Ensenberger, P.M. Fulmer, D.R. Storts, et al., Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex® ESX 17 and ESI 17 systems, Forensic Sci. Int. Genet. 5 (4) (2011) 269–275.

[150] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler™ multiplex, Forensic Sci. Int. Genet. 4 (2) (2010) 111–114.

[151] J.R. Gilder, K. Inman, W. Shields, D.E. Krane, Magnitude-dependent variation in peak height balance at heterozygous STR loci, Int. J. Leg. Med. 125 (1) (2011) 87–94.

[152] V.C. Tucker, A.J. Kirkham, A.J. Hopwood, Forensic validation of the Powerplex® ESI 16 STR multiplex and comparison of performance with AmpFlSTR® SGM plus®, Int. J. Leg. Med. 126 (3) (2012) 345–356.

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

[154] A.A. Westen, L.J. Grol, J. Harteveld, A.S. Matai, P. de Knijff, T. Sijen, Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM, Forensic Sci. Int. Genet. 6 (6) (2012) 708–715.

[155] C. Luce, S. Montpetit, D. Gangitano, P. O'Donnell, Validation of the AMPFℓSTR® MiniFiler™ PCR Amplification Kit for use in forensic casework, J. Forensic Sci. 54 (5) (2009) 1046–1054.

[156] L. Albinsson, J. Hedman, R. Ansell, Verification of alleles by using peak height thresholds and quality control of STR profiling kits, Forensic Sci. Int. Genet. Suppl. Ser. 3 (1) (2011) e251–e252.

[157] J.M. Butler, Advanced Topics in Forensic DNA Typing, Interpretation, Academic Press, 2014.

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

[159] T. Tvedebrink, P.S. Eriksen, M. Asplund, H.S. Mogensen, N. Morling, Allelic drop-out probabilities estimated by logistic regression–further considerations and practical implementation, Forensic Sci. Int. Genet. 6 (2) (2012) 263–267.

[160] C.A. Rakay, J. Bregu, C.M. Grgicak, Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and locus drop-out, Forensic Sci. Int. Genet. 6 (6) (2012) 723–728.

[161] K.E. Lohmueller, N. Rudin, K. Inman, Analysis of allelic drop-out using the Identifiler ® and PowerPlex® 16 forensic STR typing systems, Forensic Sci. Int. Genet. 12 (2014) 1–11.

[162] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures, J. R. Stat. Soc. Ser. C Appl. Stat. 59 (5) (2010) 855–874.

[163] P. Gill, B. Sparkes, J. Buckleton, Interpretation of simple mixtures of when artefacts such as stutters are present: with special reference to multiplex STRs used by the Forensic Science Service, Forensic Sci. Int. 95 (3) (1998) 213–224.

[164] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA., Nucleic Acids Res. 24 (14) (1996) 2807–2812.

[165] G. Levinson, G.A. Gutman, Slipped-strand mispairing: a major mechanism for DNA sequence evolution., Mol. Biol. Evol. 4 (3) (1987) 203–221.

[166] P. Gill, R. Sparkes, C. Kimpton, Development of guidelines to designate alleles using an STR multiplex system, Forensic Sci. Int. 89 (3) (1997) 185–197.

[167] G. Shutler, T. Roy, Genetic anomalies consistent with gonadal mosaicism encountered in a sexual assault-homicide, Forensic Sci. Int. Genet. 6 (6) (2012) e159–e160.

[168] A.J. Gibb, A.-L. Huell, M.C. Simmons, R.M. Brown, Characterisation of forward stutter in the AmpFlSTR®SGM Plus® PCR, Sci. Justice 49 (1) (2009) 24–31.

[169] K. Oostdik, K. Lenz, J. Nye, K. Schelling, D. Yet, S. Bruski, J. Strong, C. Buchanan, J. Sutton, J. Linner, et al., Developmental validation of the PowerPlex® fusion system for analysis of casework and reference samples: A 24-locus multiplex for new database standards, Forensic Sci. Int. Genet. 12 (2014) 69–76.

[170] R.L. Green, R.E. Lagacé, N.J. Oldroyd, L.K. Hennessy, J.J. Mulero, Developmental validation of the AmpFℓSTR®NGM SElect™ PCR amplification kit: a next-generation str multiplex with the SE33 locus, Forensic Sci. Int. Genet. 7 (1) (2013) 41–51.

[171] M. Klintschar, P. Wiegand, Polymerase slippage in relation to the uniformity of tetrameric repeat stretches, Forensic Sci. Int. 135 (2) (2003) 163–166.

[172] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (1) (2012) 58–63.

[173] S.B. Seo, J. Ge, J.L. King, B. Budowle, Reduction of stutter ratios in short tandem repeat loci typing of low copy number DNA samples, Forensic Sci. Int. Genet. 8 (1) (2014) 213–218.

[174] J.-A. Bright, J.M. Curran, J.S. Buckleton, Investigation into the performance of different models for predicting stutter, Forensic Sci. Int. Genet. 7 (4) (2013) 422–427.

[175] J. Weusten, J. Herbergs, A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications, Forensic Sci. Int. Genet. 6 (1) (2012) 17–25.

[176] O. Hansson, P. Gill, T. Egeland, STR-validator. An open source platform for validation and process control, Forensic Sci. Int. Genet. 13 (2014) 154–166.

[177] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, F. Alvarez, C. Baeza-Richer, A. Dominguez, C. Doutremepuich, et al., Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.