UNIVERSITY OF COPENHAGEN

Analysis of matches and partial-matches in a Danish STR data set

Tvedebrink, Torben; Eriksen, Poul Svante; Curan, James Michael; Mogensen, Helle Smidt; Morling, Niels

Published in:
Forensic Science International: Genetics

DOI:
doi:10.1016/j.fsigen.2011.08.001

Publication date:
2012

Citation for published version (APA):
Tvedebrink, T., Eriksen, P. S., Curan, J. M., Mogensen, H. S., & Morling, N. (2012). Analysis of matches and partial-matches in a Danish STR data set. *Forensic Science International: Genetics*, *6*(3), 387-392. https://doi.org/doi:10.1016/j.fsigen.2011.08.001

Download date: 20. Apr. 2021

Forensic Science International: Genetics 6 (2012) 387–392

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Analysis of matches and partial-matches in a Danish STR data set

Torben Tvedebrink [a,*], Poul Svante Eriksen [a], James Michael Curran [b], Helle Smidt Mogensen [c], Niels Morling [c]

[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] Department of Statistics, University of Auckland, New Zealand
[c] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 6 December 2010
Received in revised form 19 July 2011
Accepted 4 August 2011

Keywords:
DNA database
θ-Correction
Subpopulation
Close relatives
Covariance matrix

Abstract: Over the recent years, the national databases of STR profiles have grown in size due to the success of forensic DNA analysis in solving crimes. The accumulation of DNA profiles implies that the probability of a random match or near match of two randomly selected DNA profiles in the database increases.

We analysed 53,295 STR profiles from individuals investigated in relation to crime case investigations at the Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark. Incomplete STR profiles (437 circa 0.8% of the total), 48 redundant STR profiles from monozygotic twins (0.09%), 6 redundant STR profiles of unknown cause and 1283 STR profiles from repeated testing of individuals were removed leaving 51,517 complete 10 locus STR profiles for analysis. The number corresponds to approximately 1% of the Danish population. We compared all STR profiles to each other, i.e. 1.3×10^9 comparisons.

With these large number of comparisons, it is likely to observe DNA profiles that coincide on many loci, which has concerned some commentators and raised questions about "overstating" the power of DNA evidence. We used the method of Weir [11,12] and Curran et al. [3] to compare the observed and expected number of matches and near matches in the data set. We extended the methods by computing the covariance matrix of the summary statistic and used it for the estimation of the identical-by-descent parameter, θ. The analysis demonstrated a number of close relatives in the Danish data set and substructure. The main contribution to the substructure comes from close relatives. An overall θ-value of 1% compensated for the observed substructure, when close familial relationships were accounted for.
© 2011 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The STR data accumulated by forensic laboratories are an important source of information for verifying the population genetic models used when reporting the evidential weight related to DNA evidence. These models may be used to compute the probability that a pair of STR profiles share a particular genotype or only a given number of alleles. Hence, irrespective of the number of coinciding alleles, the pairwise comparison of STR profiles contain information about the validity of the population genetic models.

Over the recent years, the national databases of STR profiles have grown in size due to the success of forensic DNA analysis in solving crimes. With these vast numbers of profiles available, it is possible to test the validity and applicability of population models to forensic genetics [5,11,12]. Furthermore, the accumulation of DNA profiles implies that the probability of a random match or near match of two randomly selected DNA profiles in the database increases. If all pairs of profiles are compared to each other in the database this corresponds to $\binom{n}{2} = n(n-1)/2$ pairwise comparisons in a database with n DNA profiles.

With these large number of comparisons, it is likely to observe DNA profiles that coincide on many loci which has concerned some commentators and raised questions about "overstating" the power of DNA evidence. Hence, it is important to demonstrate that the observed and expected number of matches are sufficiently close to each other in order to retain the confidence in DNA typing in general and the population genetic models used for evidential calculations in particular.

The discriminatory power of a set of genetic markers is related to the number of alleles that two STR profiles share. Hence, the exercise of making all pairwise comparisons of STR profiles in the data set gives a summary statistic, which can be compared to what is expected under the population genetic model.

Weir [11,12] presented models for the expected number of pairs of DNA profiles matching and partially-matching on a given number of loci. Curran et al. [3] extended the work and discussed

* Corresponding author at: Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark. Tel.: +45 99408860; fax: +45 98158129.
E-mail address: tvede@math.aau.dk (T. Tvedebrink).

1872-4973/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2011.08.001

how the deviation between the observed and expected counts can be used to estimate model parameters.

In the present paper, we derive computationally efficient formulae for the expected value and covariance matrix of the summary statistic. Furthermore, these two quantities are used in order to improve the parameter estimation procedure. We demonstrate our methods using a Danish STR data set, which constitute approximately 1% of the Danish population.

2. Materials and methods

2.1. Data

The Danish STR data set contains 53,295 STR profiles accumulated from 1999 to 2009 typed at the 10 autosomal loci and amelogenin included in the AmpFℓSTR® SGM Plus kit (Applied Biosystems, CA, USA). The results of amelogenin were discarded after the preliminary analysis (cf. below) to avoid complications from its differences to the autosomal STR markers.

In the preliminary analysis, 48 STR profiles (0.09%) were removed since they originated from one of the twins in pairs of monozygotic twins. A total of 1283 duplicated DNA profiles of individuals typed twice or more were removed such that only a single copy of their profile remained in the data set. A total of 437 STR profiles (0.8%) with incomplete genotypes (missing allelic information at some STR loci) were discarded. Furthermore, 6 (0.01%) full STR profiles were excluded because it could not be determined if six pairs of identical STR profiles came from the same individual, identical twins, close relatives, etc. Hence, the STR data set analysed in this paper included 51,517 DNA profiles, which corresponds to about 1% of the Danish population.

The data set was subjected to a matching exercise where every STR profile was compared to any other STR profile in the data set. For each pairwise comparison, the number of matching (agreement on both alleles), m, and partially-matching loci (sharing one allele of two), p, were recorded. With n DNA profiles in the data set, this corresponds to $N = \binom{n}{2} = n(n-1)/2$ pairwise comparisons. For the Danish STR data set ($n = 51{,}517$), this results in $N = 1{,}326{,}974{,}886$ STR profile comparisons. We denote the resulting matrix $M = \{M_{m/p}\}_{m,p}$, where entry $M_{m/p}$ has the number of pairs with m matching and p partially matching loci, respectively.

The result of analysing the Danish data set is summarised in Table 1, where we find that, e.g. the number of profile pairs with 5 matching loci and 4 partially-matching loci out of ten autosomal loci is $M_{5/4} = 17{,}060$. Fig. 1 shows the summary statistic for the Danish data set in an informative way, where we have plotted the observed counts on a $\log_{10}$-scale. Superimposed are expected counts when substructure ($\theta = 0.0118$) and close familial relationships are accounted for (see Table 3).

Two of the authors (Tvedebrink and Curran) implemented computationally efficient functions for constructing the M-table in the statistical software R [7]. The `dbCompare`-function from the `DNAtools`-package [10] used less than 5 min to perform all 1,326,974,886 pairwise comparisons on a 2.50 GHz laptop computer. Most of the methodology in this paper has been implemented in the `DNAtools`-package together with specialised plotting functions. The R-package is described in more detail in the package's vignette, which is obtained by running `vignette(''DNAtools'')` after installation in R.

2.2. Population genetic model

The model proposed by Weir [12] defines for each of L typed loci three probabilities ($P_{0/0}$, $P_{0/1}$, $P_{1/0}$), which are the probabilities for two randomly selected profiles sharing none, one or both alleles at a given locus (Weir denoted the probabilities P_0, P_1, P_2. The change of subscript will hopefully be clear from Section 2.2.1). The probabilities $P_{m/p}$ depend on the coancestry coefficient, θ, through the match probability equations [6] that are derived using the recursion formula: $P(A_i|\boldsymbol{x}^n) = [x_i^n\theta + (1-\theta)\,p_i]/[1+(n-1)\theta]$, where $\boldsymbol{x}^n$ is the sufficient statistic [8], and $P(A_i \mid \boldsymbol{x}^n)$ denotes the probability of observing an i allele after having seen x_i^n alleles of type i among n sampled alleles.

Furthermore, Weir [12] showed that for a specified family relationship of a pair of profiles, $P_{m/p}$ is updated using the probabilities, k_I, that the two individuals share I alleles identical-by-descent (IBD):

$$\tilde{P}_{0/0} = k_0 P_{0/0} \quad \tilde{P}_{0/1} = k_1(1-\theta)(1-S_2) + k_0 P_{0/1} \quad \text{and}$$
$$\tilde{P}_{1/0} = k_2 + k_1[\theta + (1-\theta)S_2] + k_0 P_{1/0},$$

where $S_2 = \sum_{i=1}^{K} p_i^2$ is the sum of squared allele probabilities at a given locus with K different alleles, and $\tilde{P}_{m/p}$ denotes the probability that two individuals with the specified family relationship will match as m/p at a given locus. From $\tilde{P}_{m/p}$, it is clear that close relatives have an increased probability of sharing alleles due to alleles being IBD.

In Table 2, we have listed the five types of relatedness considered in this paper. The avuncular class covers half-sibling, grandparent–grandchild and uncle–nephew (independent of gender) since these have identical $\boldsymbol{k}$-vectors and are indistinguishable using unlinked genetic markers.

2.2.1. Expected value and covariance matrix

Since the summary statistic is formed by adding identically distributed elements, the expected count $\mathbb{E}[M] = N\mathbb{E}[M(G_{i_1}, G_{i_2})]$,

Table 1
Summary matrix, M, for the Danish STR data set with 51,517 DNA profiles. $M_{m/p}$ is the number pairs of profiles with m matching (where m is the row number) and p partially-matching (where p is the column number) loci.

$M_{m/p}$	0	1	2	3	4	5	6	7	8	9	10
0	906,881	8,707,969	37,632,872	96,157,037	160,570,778	182,820,115	143,627,613	76,852,119	26,786,782	5,486,572	501,671
1	1,100,493	9,484,061	36,229,766	80,292,877	113,733,413	106,635,954	66,164,365	26,183,818	5,992,415	604,900	
2	595,135	4,531,792	14,996,133	28,165,271	32,810,688	24,271,278	11,132,519	2,887,555	325,493		
3	188,146	1,237,733	3,467,281	5,353,738	4,913,791	2,683,854	805,798	103,305			
4	38,094	212,192	487,484	592,929	401,832	143,202	21,490				
5	5114	23,490	42,459	37,933	17,060	3100					
6	470	1685	2272	1414	378						
7	26	96	91	64							
8	3	6	21								
9	0	0									
10	0										

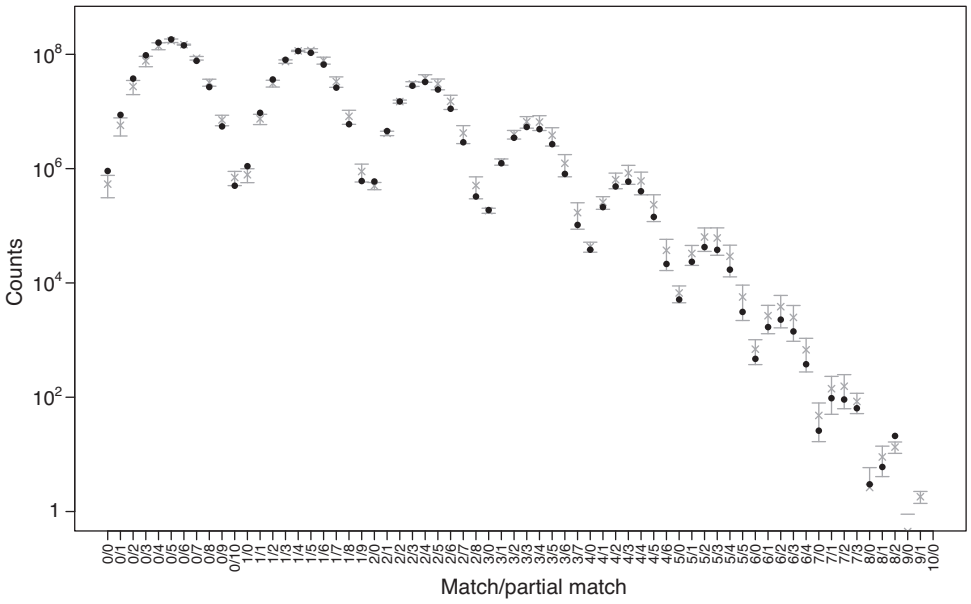

Fig. 1. Plot of observed counts (marked by •) versus the number of matching and partially-matching loci (counts on log $_{10}$-scale) for the Danish STR data set. The superimposed points (×) represents the expected counts when substructure ($\theta = 0.01$) and close familial relationships are accounted for. The vertical bars indicate an approximated 95%-confidence interval.

Table 2
Probability of sharing I alleles IBD for the specified relationship [12, Table 4].

Relationship	Full-siblings (FS)	First-cousins (FC)	Parent–child (PC)	Avuncular (AV)	Unrelated (UN)
$\boldsymbol{k} = (k_2, k_1, k_0)$	(0.25, 0.5, 0.25)	(0, 0.25, 0.75)	(0, 1, 0)	(0, 0.5, 0.5)	(0, 0, 1)

where $N = \binom{n}{2} = n(n-1)/2$ is number of pairwise DNA profile comparisons, and G_{i_1} and G_{i_2} are any two DNA profiles from different individuals in the data set. Let $\pi = \mathbb{E}[M(G_{i_1}, G_{i_2})]$ be a matrix of match/partially-match probabilities. That is, $\pi = \{\pi_{m/p}\}_{m,p}$ is the matrix of probabilities for the match/partially-match events (m, p), where $m = 0, \ldots, L$ and $p = 0, \ldots, L - m$.

The elements of π may be computed using recursion over loci. Let π^{ℓ} denote the probability based on ℓ loci, i.e. using only a subset of size ℓ of the L loci such that $m = 0, \ldots, \ell$, and $p = 0, \ldots, \ell - m$. Furthermore, let $P^{\ell}_{m/p}$ refer to the $P_{m/p}$ probabilities for the ℓ th added locus, then the following equation denote how to compute $\pi^{\ell+1}_{m/p}$ by recursion for $\ell = 1, \ldots, L - 1$:

$$\pi^{\ell+1}_{m/p} = \begin{cases} P^{\ell+1}_{0/0}\pi^{\ell}_{m/p} + P^{\ell+1}_{0/1}\pi^{\ell}_{m/p-1} + P^{\ell+1}_{1/0}\pi^{\ell}_{m-1/p}, & m > 0 \text{ and } p > 0, \\ P^{\ell+1}_{0/0}\pi^{\ell}_{0/p} + P^{\ell+1}_{0/1}\pi^{\ell}_{0/p-1}, & m = 0 \text{ and } p > 0, \\ P^{\ell+1}_{0/0}\pi^{\ell}_{m/0} + P^{\ell+1}_{1/0}\pi^{\ell}_{m-1/0}, & m > 0 \text{ and } p = 0, \\ P^{\ell+1}_{0/0}\pi^{\ell}_{0/0}, & m = 0 \text{ and } p = 0, \end{cases} \quad (1)$$

where the "sum" of the subscripts for each term on the right hand side equals the subscript on the left hand side, e.g. the subscripts of the last term in the first equation gives $1/0 + m - 1/p = m/p$. The initial step of the recursion has $\pi^1_{1/0} = P^1_{1/0}$, $\pi^1_{0/1} = P^1_{0/1}$ and $\pi^1_{0/0} = P^1_{0/0}$. Eq. (1) is readily implemented in computer software and efficiently computes the expected numbers for various θ-values.

For a pair of R-relatives (close relatives of type R), the expected numbers of matching/partially-matching loci, $\tilde{\pi}_R$, are calculated by replacing $P_{m/p}$ with $\tilde{P}_{m/p}$ in (1). The effect of relatedness on the expected number of matching/partially-matching loci is plotted in Fig. 2. Note that parent–child (marked by + in Fig. 2) must share at least one allele per locus implying that $\tilde{\pi}_{\mathrm{PC}} = 0$ when $m + p \neq L$.

Weir [12] focused in his survey paper primarily on comparison between the observed counts and the expected number, $N\pi(\theta)$, for different values of θ. However, as Curran et al. [3] discussed one needs to consider normalisation of these differences for a proper comparison between the observed and expected counts. One way of normalising the difference between the observed and expected counts is by the covariance matrix of the summary statistic. As for the expected value, the covariance matrix can be computed using recursion over loci. In the on-line supplementary material we give the full details on computing the covariance matrix, $\Sigma(\theta)$.

The expected value and covariance matrix can be used to construct confidence intervals for the cell counts. Inserting the estimated parameter values in $\vec{\tilde{\boldsymbol{\pi}}}$ and $\Sigma(\theta)$ gives the fitted expected values and covariance matrix. Given these quantities, we compute marginal 95%-confidence intervals by $N\vec{\tilde{\boldsymbol{\pi}}} \pm 2\sqrt{\mathrm{diag}\{\Sigma(\theta)\}}$ (superimposed in Fig. 1), where $\vec{\tilde{\boldsymbol{\pi}}}$ is defined in (2). The construction of the confidence intervals rely on an approximation to normality for the cell counts. The performance of this approximation increases with the cell counts, i.e. the smaller the counts, the less accurate is the approximation.

2.3. Estimating model parameters

From Fig. 2, it is evident that close relatives in the data set may induce more near-matching pairs of profiles than having a data set of unrelated individuals. Hence, not considering close relatives may erroneously increase the estimate of θ to accommodate the increased similarity of the DNA profiles.

The inclusion of related pairs of profiles was investigated by Curran et al. [3] using Australian data with Caucasian and Aborigine origin. Curran et al. [3] defined α_R as the proportion of pairwise comparisons between R-relatives. Hence, the expected

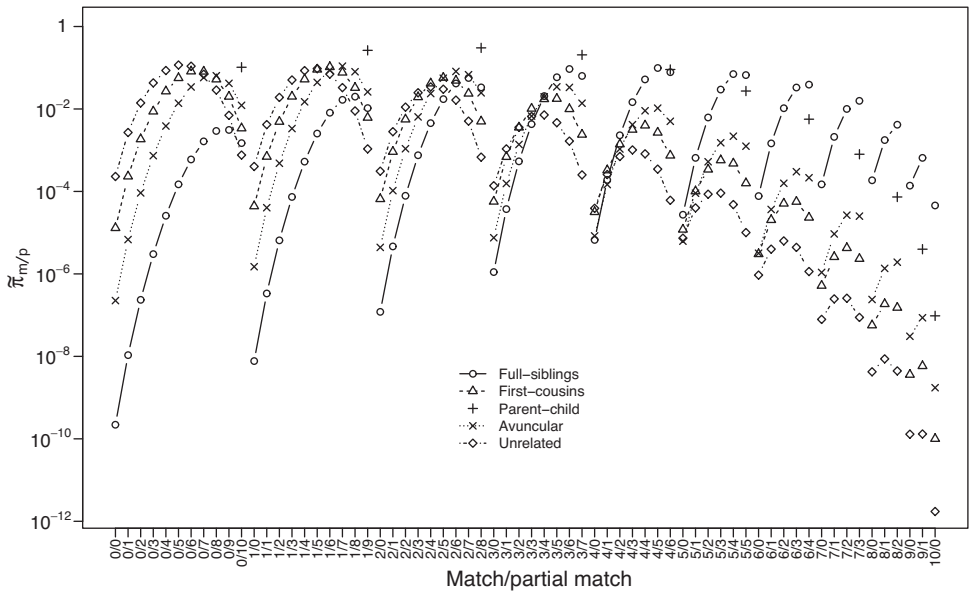

Fig. 2. Effect on $\tilde{\pi}$ for the five types of relatedness (see Table 2) with $\theta = 0.03$. The legend explains the plot characters.

number of matching/partially-matching loci can be written as:

$$N\vec{\tilde{\pi}} = N(\alpha_{FS}\tilde{\pi}_{FS} + \alpha_{FC}\tilde{\pi}_{FC} + \alpha_{PC}\tilde{\pi}_{PC} + \alpha_{AV}\tilde{\pi}_{AV} + \alpha_{UN}\pi), \quad \text{with} \quad \alpha_{UN} = 1 - \sum_R \alpha_R, \tag{2}$$

where the abbreviations in the subscripts are defined in Table 2. Note that Curran et al. [3] did not include the avuncular relationship (AV) in their original formulation.

In order to fit the model to the data, Curran et al. [3] investigated the object functions in (3) as a mean to compare the expected and observed counts:

$$\begin{aligned} C_1(\theta) &= \sqrt{\sum_{m,p} \left(M_{m/p} - N\tilde{\pi}_{m/p}(\theta)\right)^2}; \\ C_2(\theta) &= \sum_{m,p} \frac{\left(M_{m/p} - N\tilde{\pi}_{m/p}(\theta)\right)^2}{N\tilde{\pi}_{m/p}(\theta)}; \\ C_3(\theta) &= \sum_{m,p} \frac{|M_{m/p} - N\tilde{\pi}_{m/p}(\theta)|}{N\tilde{\pi}_{m/p}(\theta)}, \end{aligned} \tag{3}$$

where summations are over the matrix entries indexed by $m = 0, \ldots, L$ and $p = 0, \ldots, L - m$. Curran et al. [3] argued that numerical work indicated that $C_3(\theta)$ yielded good results since special emphasis is placed on the upper tail of the distribution (large number of matching loci).

An alternative approach for estimating the parameters in the model would be to use the covariance matrix, $\Sigma(\theta)$, in order to compensate for the variance of the counts and the correlation between cell counts:

$$\begin{aligned} T_1(\theta) &= \sum_{m,p} \frac{\left(M_{m/p} - N\tilde{\pi}_{m/p}(\theta)\right)^2}{\Sigma_{m/p,m/p}(\theta)}; \\ T_2(\theta) &= \{\vec{M} - N\vec{\tilde{\pi}}(\theta)\}^{\top} \Sigma(\theta)^{-} \{\vec{M} - N\vec{\tilde{\pi}}(\theta)\}, \end{aligned} \tag{4}$$

where $\vec{M}$ and $\vec{\tilde{\pi}}$ are vector versions of the matrices M and $\tilde{\pi}$, respectively. $T_2(\theta)$ is a so-called Mahalanobis distance, which is an often used measure of divergence between observed and expected quantities. Note that we use the generalised inverse of $\Sigma(\theta)$ due to the linear constraint that $\sum_{m,p} M_{m/p} = N$. Simulations of data sets with and without relatives (see [9, Chapter 3]) indicate that $T_2(\theta)$ is the most efficient estimator of θ among the object functions considered here.

3. Results

The Danish STR data set was analysed using the described methods and gave the summary statistic presented in Table 1 and Fig. 1. We used the object functions in (3) and (4) for comparing the observed and expected cell counts for estimating θ and α_R. For $T_2(\theta)$, the minimum was obtained with $\hat{\theta} = 0.0118$ and $\hat{\boldsymbol{\alpha}}_R$ as reported in Table 3. It is noteworthy that $\hat{\theta} = 0$ for all of the $C_i(\theta)$-methods, $i = 1, 2, 3$. It seems rather unlikely that there is no effect of subpopulation stratification after allowing for close relatives. Simulation studies (see [9]) suggest that the θ estimates are stable, whereas the estimates of α are subject to variability.

Note that the estimated α_{FS} for $T_2(\theta)$ is about a factor 10 larger than 2×10^{-7} which is the approximate value obtained if one assumes that every individual of the Danish adult population has exactly one full-sibling. However, it is likely that the frequency of full-siblings is larger in the Danish STR data set than in the population due to various factors, e.g. sampling criteria and social factors.

3.1. Accounting for close relatives when evaluating the weight of evidence

The argument for using the θ-correction when assessing the evidential weight of a given DNA profile, is to adjust for possible subpopulation effects in the population from which the suspect

Table 3
Estimated values for the Danish STR data set using various object functions of (3) and (4).

Method	θ	α_{FS}	α_{FC}	α_{PC}	α_{AV}
$C_1(\theta)$	0.0000	5.0E−07	2.0E−15	1.6E−09	7.8E−09
$C_2(\theta)$	0.0000	2.6E−09	1.0E−09	2.1E−08	1.4E−14
$C_3(\theta)$	0.0000	5.0E−06	7.9E−06	1.4E−19	5.0E−07
$T_1(\theta)$	0.0137	1.3E−06	5.9E−09	2.5E−07	1.8E−17
$T_2(\theta)$	0.0118	2.6E−06	1.2E−08	5.1E−07	3.5E−17

and profiles for estimating allele probabilities are drawn. A structured population causes the probability of observing a specific DNA profile to be heterogeneous, since the prevalence of its constituting alleles may be higher in some subpopulation relative to the entire population.

Close relatives is another violation of the most simple population genetic models. However, the probability that a specified pair of R-relatives shares a DNA profile can be computed for heterozygous and homozygous loci, respectively, by [2, Table 4.5]:

$$P(E_l = A_{l,i}A_{l,j}|H_d, R) = k_2 + \frac{k_1}{2}\frac{2\theta + (1-\theta)(p_{l,i} + p_j)}{1+\theta} + 2k_0\frac{\theta^2 + \theta(1-\theta)(p_{l,i} + p_{l,j}) + (1-\theta)^2 p_{l,i}p_{l,j}}{(1+2\theta)(1+\theta)}$$

$$P(E_l = A_{l,i}A_{l,i}|H_d, R) = k_2 + k_1\frac{2\theta + (1-\theta)p_{l,i}}{1+\theta} + k_0\frac{6\theta^2 + 5\theta(1-\theta)p_{l,i} + (1-\theta)^2 p_{l,i}^2}{(1+2\theta)(1+\theta)},$$

where R determines $\boldsymbol{k} = (k_2, k_1, k_0)$ and E_l is evidence from locus l with $p_{l,i}$ being the probability of allele i at that locus.

Since all individuals in a population may have close relatives, it is important to consider the possibility that a close relative of the suspect is the true perpetrator. Hence, when forming the likelihood ratio, LR, the hypothesis in the denominator could be H_d: "A man possibly related to the suspect is the true donor of the biological stain" [1,4]. The probability α_R represents the probability that two randomly selected individuals are R-relatives. The probability that the "random man" of the defence hypothesis is a R-relative of the suspects can therefore be expressed by α_R. The probability of the evidence taking these close relationships into account can be evaluated by summing over R yielding $P(E \mid H_{d,R}) = \sum_R P(E \mid H_d, R)\alpha_R$, such that $H_{d,R}$ explicitly takes close relatives into account.

In Fig. 3, $P(E \mid H_{d,R})$ is plotted against $P(E \mid H_d)$ for all 51,517 DNA profiles of the Danish STR data set. The relationship is close to log–log linear: $\log_{10}P(E \mid H_{d,R}) = \beta_0 + \beta_1 \log_{10}P(E \mid H_d)$. In Fig. 3, we have superimposed the expected relationship (solid line) with the uncertainty represented by the predictive interval (dashed lines). The estimated mean and standard deviation of $\log_{10}P(E \mid H_{d,R})/P(E \mid H_d)$ are 3.22 and 0.95, respectively. Hence, an approximative confidence interval for the ratio is given as $10^{3.22 \pm 2 \times 0.95} \approx [10; 10^5]$ with a mean of 1660, i.e. taking close relatives into account increases the probability of the evidence with up to five orders of magnitude. The dominating contribution to the sum of $P(E \mid H_{d,R})$ is that of full-siblings, $P(E|H_d, R = \mathrm{FS})\hat{\alpha}_{\mathrm{FS}}$, which accounts for approximately 99.5% of $P(E \mid H_{d,R})$. In Fig. 2, this was also the category with the largest $\tilde{\pi}_{10/0}$. Hence, for practical purposes the only relevant type of close relatedness to include in the calculations when reporting likelihood ratios on routine basis is full-siblings, since the decrease in $P(E \mid H_d, R)$ for the remaining types of relatives is minimal relative to $\hat{\alpha}_R$ making their contribution to $P(E \mid H_{d,R})$ negligible.

4. Discussion

It is evident from the analysis of the Danish STR data set that a θ-correction close to 1% is sufficient to capture the effects from substructure among the typed STR profiles. Furthermore, the analysis demonstrated the presence of close relatives in the data set. A fact that was suspected beforehand, but the number of close relatives was unknown.

It is doubtful that the present STR data set is completely representative of the distribution of STR profiles among people living in Denmark. The STR data set investigated is also different from the one in the Danish crime DNA database. It was impossible to conclude the cause of identity of 6 pairs of STR profiles (12 STR profiles, 0.02%). The STR data set included a significant number of known, identical twins with identical STR profiles. Approximately every 1 in 250–300 births give rise to identical twins, so it was not surprising that 96 twins (0.18%) = 48 twin pairs (0.09%) were observed. These facts may be of value for investigators who search large crime DNA databases.

The investigations showed that there is substructuring in the Danish STR data set. Close relatives seem to contribute to a significant degree, although mechanisms like immigration may contribute. The degree of substructure is, however limited, and an

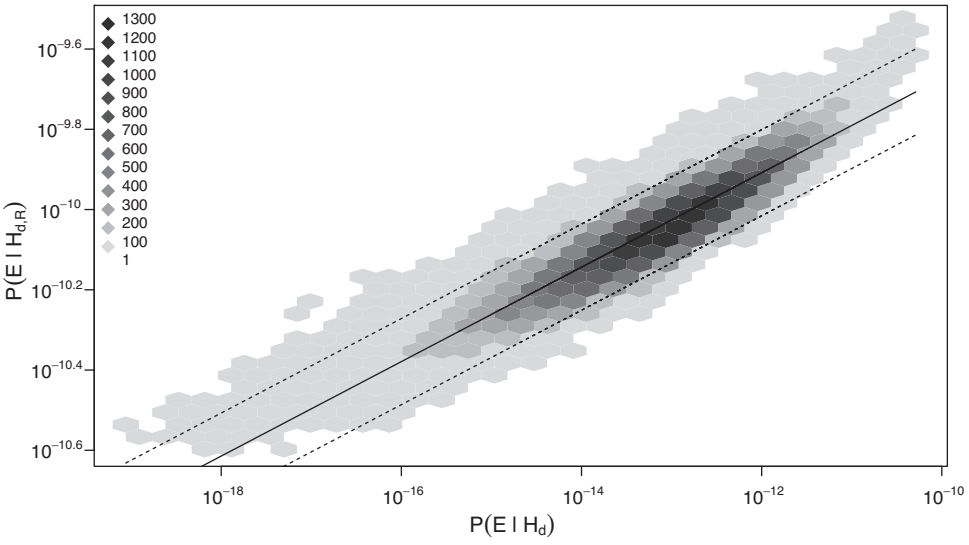

Fig. 3. Relationship between $P(E \mid H_d)$ and $P(E \mid H_{d,R})$ with a predictive interval superimposed (solid line: mean, dashed lines: predictive limits). The shading colour of the hexagons indicate bin counts, i.e. the number of cases with an $(P(E \mid H_d), P(E \mid H_{d,R}))$-value in this region.

overall θ correction of 0.01 will compensate for the substructure observed. The θ-correction should be taken into consideration when calculating the weight of the evidence, since it offers a better estimate of the weight of the evidence than the often used *LR* which compares the DNA evidence to a random, unrelated person. However, it will still be necessary to address the problem with identical twins and other close relatives in database searches and evaluation of the evidence in case of matches, family matches, etc.

5. Conclusion

The main objective with the work presented in this paper was to analyse a Danish STR data set of 51,517 different individuals. This was to accommodate the fact that at some point two apparently unrelated individuals will share DNA profiles for all ten autosomal SGM Plus loci in the Danish population. If a specified relationship is assumed it is straight forward to calculate the probability of identical DNA profiles. However, one still needs to account for remote coancestry for both related and unrelated pairs of STR profiles.

Only modelling the expected value or calculating the mean is never satisfactory in statistics. A measure of precision or variability is needed in order to discuss the extremity of an observation relative to the expectation under a given model. Hence, deriving and computing the covariance matrix of M was essential. Simulation studies showed that the object function $T_2(\theta)$, which uses the covariance matrix, was the most efficient estimator of θ among the object functions considered here.

Acknowledgements

The authors would like to thank Ms. Kirstine Kristensen and Ms. Line Maria Irlund Pedersen (both from The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen) for assistance in verifying the familial relationships of the twins in the data set and validating some near matches due to typing errors.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2011.08.001.

References

[1] 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 (1994) 125–140.
[2] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, FL, 2005, pp. 217–274.
[3] J.M. Curran, S.J. Walsh, J. Buckleton, Empirical testing of estimated DNA frequencies, Forensic Sci. Int.: Genet. 1 (2007) 267–272.
[4] I.W. Evett, Evaluating dna profiles in a case where the defence is "it was my brother", J. Forensic Sci. Soc. 32 (1) (1992) 5–14.
[5] L.D. Mueller, Can simple populations genetic models reconcile partial match frequencies observed in large forensic databases? J. Genet. 87 (2) (2008) 101–107.
[6] R.A. Nichols, D.J. Balding, Effects of population structure on DNA fingerprint analysis in forensic science, Heredity 66 (1991) 297–302.
[7] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2009 , , ISBN: 3-900051-07-0.
[8] T. Tvedebrink, Overdispersion in allelic counts and θ-correction in forensic genetics, Theor. Popul. Biol. 78 (3) (2010) 200–210.
[9] T. Tvedebrink, Statistical Aspects of Forensic Genetics – Model for Qualitative and Quantitative STR Data. Ph. D. thesis, Department of Mathematical Sciences, Aalborg University, 2010b. URL: http://vbn.aau.dk/files/48415084/thesisEmbedded.pdf.
[10] T. Tvedebrink, J. Curran, DNAtools: statistical functions for analysing forensic DNA databases. R package version 0.1-2, 2011. Available at CRAN: http://www.cran.r-project.org/web/packages/DNAtools.
[11] B.S. Weir, Matching and partially-matching DNA profiles, J. Forensic Sci. 49 (5) (2004) 1–6.
[12] B.S. Weir, The rarity of DNA profiles, Ann. Appl. Stat. 1 (2) (2007) 358–370.