Forensic Science International: Genetics 13 (2014) 128–133

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Choosing supplementary markers in forensic casework

Andreas O. Tillmar [a,b,*], Petter Mostad [c,d]

[a] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[b] Department of Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, Linköping, Sweden
[c] Mathematical Sciences, Chalmers University of Technology and Mathematical Sciences, University of Gothenburg, Sweden
[d] Department for Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway

ARTICLE INFO

Article history:
Received 18 March 2014
Received in revised form 18 June 2014
Accepted 30 June 2014

Keywords:
Decision making
Error rate
Paternity testing
Simulation

ABSTRACT

The vast majority of human familial identifications based on DNA end up with a well founded conclusion, normally using a standard set of genetic short tandem repeat (STR) loci. There are, however, a proportion of cases that show ambiguous results. For such occasions a number of different supplementary markers could be typed in order to gain further information. There are numerous markers available for such supplementary DNA typing, including STRs, deletion and insertion polymorphisms (DIPs), and single nucleotide polymorphisms (SNPs). The purpose of this work was to describe a precise method for decision making, aiming to aid the comparison of different sets of markers for different case scenarios in order to find the most efficient set for routine casework. Comparisons are based on a particular function relating the expected additional value of information from new data to the amount of information already obtained from initial data. The function can be computed approximately by approximating likelihood-based error rates using simulation. In this paper we focused on paternity investigations, more specifically the use of supplementary markers in cases where a smaller number of genetic inconsistencies make the matter inconclusive. We applied the method to a comparison of three different kits: Investigator HDplex (STRs), Investigator DIPplex (DIPs), and the SNPforID-plex (SNPs) to study their efficiencies in gaining information in different case scenarios involving various alternative relationships between the tested man and the tested child. We show that the Investigator HDplex was the most efficient set of supplementary markers for the standard paternity case. However, for paternity cases with a close relative being the alternative father, the Investigator HDplex and the SNPforID-plex showed similar patterns in their ability to deliver a well-founded conclusion. The Investigator DIPplex was the least efficient set.

1. Introduction

For a long period of time the CODIS short tandem repeat (STR) loci [1] have provided the basis for DNA analysis for human identification. Until a few years ago, there were only a smaller number of kits commercially available that included other DNA markers. During the past couple of years there has been an expansion on the forensic STR kit market, which now offers more diverse kits with a larger number of new genetic loci. One reason for this increase is the expansion of the core European standard set (ESS) [2]. Apart from the standard STR markers, there are also kits available for forensic use that include biallelic markers such as DIPs (Deletions and Insertion Polymorphisms) and SNPs (Single Nucleotide Polymorphisms) [3,4].

Although generally a good thing that a larger number of sets of markers are available to choose from, there is a challenge in choosing the best kit for a forensic laboratory's routine casework. Apart from technical issues with each new kit, the usefulness (in terms of efficiency to solve specific cases) of the additional markers depends on the type of marker, population uniqueness, number of alleles, mutation rates, and which types of genetic relationships are to be tested, among other things. Even though there are various general measurements of the genetic diversity of a given set of markers (e.g. overall match probability, power of discrimination, power of exclusion), it can, for numerous applications, be difficult to compare different sets of markers only on the basis of such general efficiency parameters.

The purpose of this paper is to describe a method for decision making when having multiple sets of markers to choose from. Furthermore, we have applied our method in a comparison of three

* Corresponding author at: National Board of Forensic Medicine, Department of Forensic Genetics and Forensic Toxicology, Linköping, Sweden.
Tel.: +46 13 25 21 43; fax: +46 13 13 60 05.
E-mail address: andreas.tillmar@rmv.se (A.O. Tillmar).

http://dx.doi.org/10.1016/j.fsigen.2014.06.019
1872-4973/

different sets of markers; one set that includes 11 STRs (Investigator HDplex, Qiagen), a second that includes 30 biallelic markers of the type Insertion/Deletions (InDels) (Investigator DIPplex, Qiagen), and a third set that includes 52 biallelic markers of the type Single Nucleotide Polymorphisms (SNPs) (SNPforID). The goal was to find the most efficient set to be used as supplementary markers in paternity investigations.

The chosen sets of markers represent two different types, namely biallelic markers (SNPs [3] or InDels [4]) and additional STRs. These types have different characteristics. The main advantage of using SNPs/Indels in paternity testing is their low mutation rate, thus confirming a genetic inconsistency as a true genetic exclusion. The disadvantage is that generally only two alleles exist per locus making it harder to find true genetic exclusions. For example, in paternity duo cases, the alleged father (AF) and the child need to be opposite homozygous in order to yield a genetic exclusion. STR loci on the other hand have larger number of alleles but also a higher mutation rate, increasing the uncertainty of whether a genetic inconsistency really is a true exclusion or in fact a mutation. There are studies promoting the use of biallelic markers as supplementary markers in relationships testing [5,6] while other, more recent studies, have concluded that smaller sets of biallelic markers should be used with caution, especially when relatives might be involved [7,8].

When it comes to paternity investigations, the vast majority of such cases end up with a well-founded conclusion: If H_1 and H_2 are the two competing hypotheses, with H_1 representing paternity, the likelihood ratio $L(D) = Pr(D|H_1)/Pr(D|H_2)$ in favor of H_1 is computed, with D representing the test data. If $L(D) > L_H$ for some cutoff value, for example $L_H = 10{,}000$, paternity is declared, whereas if there are numerous genetic inconsistencies, non-paternity is declared [9]. Different laboratories may use different rules to declare non-paternity. In this paper, we will assume that computational models that allow mutations and silent alleles are used, and that the rule for declaring non-paternity is in fact formulated as $L(D) < L_L$ for some L_L. For inconclusive cases where $L_L < L(D) < L_H$, a number of different supplementary markers can be typed in order to reach a definite conclusion.

To investigate the usefulness of acquiring data D, or additional data D_1 or D_2 if the initial test is inconclusive, a central issue is the error rates for the data. Assuming that one of the considered hypotheses is true, what is the probability that the likelihood ratio will lead to an erroneous conclusion, when compared to the cutoff values. Specifically, we need to consider the functions

$$E_1(\ell) = Pr(L(D) < \ell | H_1)$$

and

$$E_2(\ell) = Pr(L(D) > \ell | H_2)$$

for various types of data. We will see how these functions can be approximated using simulation for several specific data types and hypotheses H_1 and H_2.

Decision theory is a general way to aid decisions in situations of uncertainty, by specifying costs of various possible outcomes from decisions, and then selecting the decision with the lowest expected cost [10]. Application of decision theory is often hampered by the difficulty in assigning costs to outcomes. In our case, it is not obvious how to assign a monetary cost to falsely concluding with paternity when there is no paternity, or to the opposite type of error. However, when laboratories decide on cutoff values L_H and L_D, they are implicitly making decisions about the relative costs of various errors. We show how we can use these cutoff values together with prior odds for paternity and likelihood values computed from initial data analyses to obtain indirect cost estimates whose numerical values can be used as part of guides for decisions.

The data used in the case examples were based on simulations. Simulation of families and their DNA profiles gives the opportunity to rather simply investigate different issues and also test the impact of model change. Both mutations and silent alleles were modeled and accounted for in our simulations.

2. Material and methods

2.1. Decision theory

Assume two competing hypotheses H_1 and H_2 have probabilities $Pr(H_1)$ and $Pr(H_2) = 1 - Pr(H_1)$, and that a choice should be made between H_1, H_2, or possibly making no decision. To facilitate a choice one may assign costs to various outcomes. Without loss of generality, we assume that making no decision has unit cost, so that all other costs are measured relative to this. If H_1 is true and we decide on H_2 we assume a cost $1 + c_1$ is incurred ($c_1 > -1$), while we assume a cost of $1 + c_2$ in the opposite case ($c_2 > -1$). The expected costs of deciding on H_1, H_2, and making no choice are $Pr(H_2)(1 + c_2)$, $Pr(H_1)(1 + c_1)$, and 1, respectively. If $c_1 c_2 \le 1$, minimizing expected costs leads to choosing H_1 if $Pr(H_1) > (1 + c_2)/(2 + c_1 + c_2)$ and otherwise H_2. A more interesting case for us is when $c_1 c_2 > 1$, which also implies that $c_1 > 0$ and $c_2 > 0$. In this case one should choose H_1 if $Pr(H_2)(1 + c_2) < 1$, H_2 if $Pr(H_1)(1 + c_1) < 1$, and otherwise one should make no decision. In terms of the odds ratio $o = Pr(H_1)/(1 - Pr(H_1))$, where $Pr(H_1) = o/(o + 1)$ and $Pr(H_2) = 1/(o + 1)$, the decision rules are as follows: If $c_1 c_2 \le 1$, decide on H_1 if $o > (1 + c_2)/(1 + c_1)$, otherwise on H_2. If $c_1 c_2 > 1$, decide on H_1 if $o < c_2$, H_2 if $o < 1/c_1$, and otherwise make no decision. Generally o will be the posterior odds based on data D. According to Bayes formula on odds form we have $o = L(D)o_0$, where $L(D)$ is the likelihood ratio and o_0 the prior odds.

In DNA testing laboratories, decisions are generally not made by first estimating costs c_1 and c_2. Instead, one often uses fixed lower and upper bounds to which $L(D)$ is compared: If $L(D)$ is very high, H_1 is declared true, if $L(D)$ is very low, H_2 is declared true, and in between one delays a decision. This paper focuses on helping laboratories choose additional data sets when $L(D)$ gives an inconclusive result. That inconclusive results are possible means as we saw above that $c_1 c_2 > 1$ and that both c_1 and c_2 are positive.

When making a decision based on D, the theoretically most sound approach is to compare $o = L(D)o_0$ to cutoff values L_H and L_L, and declare H_1 true if $o > L_H$, declare H_2 true if $o < L_L$, and otherwise declare the result inconclusive. In practice, however, laboratories for convenience often ignore the prior odds o_0, comparing $L(D)$ directly with L_H and L_L to make the decision. Let us define $L_H^* = L_H$ for the first type of decision making, and $L_H{}^* = o_0 L_H$ for the second type, and similar for L_L^*. Then in all cases, H_1 is chosen if $o > L_H^*$, H_2 is chosen if $o < L_L^*$, and in between no decision is made. Comparing with the above, we see that $c_1 = 1/L_L^*$ and $c_2 = L_H^*$, in other words, the decision routines of the laboratories implicitly correspond to estimating costs c_1 and c_2 at these values.

We now consider the situation where results using data D are inconclusive, i.e., $L_L^* < o < L_H^*$, and we would like to optimally choose between producing additional data sets $D_1, D_2, \ldots, D_N$, or possibly produce none of these datasets. We start by defining for $i = 1, \ldots, N$ and any $\ell > 0$ functions

$$E_{i1}(\ell) = Pr(L(D_i) < \ell | H_1)$$

$$E_{i2}(\ell) = Pr(L(D_i) > \ell | H_2)$$

As these cumulative probability distribution functions measure the probability to make "errors" in the sense of making a wrong conclusion, we will refer to them as error rates. As we assume the different data sets are independent given the hypotheses, the decision after D_i has been acquired should be based on comparing