Legal Medicine 12 (2010) 137–143

Contents lists available at ScienceDirect

Legal Medicine

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

Interpreting Y chromosome STR haplotype mixture

Jianye Ge [a,b,*], Bruce Budowle [a,b], Ranajit Chakraborty [a,b,c]

[a] Institute of Investigative Genetics, University of North Texas Health Science Center, Fort Worth, TX 76107, USA
[b] Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Ft Worth, TX 76107, USA
[c] Center for Genome Information, Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, OH, 45267, USA

ARTICLE INFO

Article history:
Received 4 September 2009
Received in revised form 9 November 2009
Accepted 3 February 2010
Available online 25 March 2010

Keywords:
Y chromosome
Short tandem repeats
DNA mixture
Likelihood ratio
Probability of exclusion
Evidence interpretation

ABSTRACT

Mixture interpretation is a challenging problem in forensic DNA analyses. The interpretation of Y short tandem repeat (STR) haplotype mixtures, due to a lack of recombination, differs somewhat from that of the autosomal DNA markers and is more complex. We describe approaches for calculating the probability of exclusion (PE) and likelihood ratio (LR) methods to interpret Y-STR mixture evidence with population substructure incorporated. For a mixture sample, first, all possible contributor haplotypes in a reference database are listed as a candidate list. The PE is the complement of the summation of the frequencies of haplotypes in the candidate list. The LR method compares the probabilities of the evidence given alternative hypotheses. The hypotheses are possible explanations for the mixture. Population substructure may be further incorporated in likelihood calculation. The maximum number of contributors is based on the candidate list and the computing complexity is polynomial. Additionally, mixtures were simulated by combining two or three 16 Y-STR marker haplotypes derived from the US forensic Y-STR database. The average PE was related to the size of database. With a database comprised of 500 haplotypes an average PE value of at least 0.995 can be obtained for two-person mixtures. The PE decreases with an increasing number of contributors to the mixture. Using the total sample population, the average number of candidate haplotypes of two-person mixtures is 3.73 and 95% mixtures have less than or equal to 10 candidate haplotypes. More than 98.7% of two-person mixtures can only be explained by the haplotype combinations that mixtures are composed. These values are generally higher for three-person mixtures. A small proportion of three-person mixture can also be explained by only two haplotypes.

1. Introduction

The human Y chromosome, specific to males, occurs in single copy and is inherited through the paternal lineage. Typing short tandem repeat (STR) markers within the non-recombining region of the Y chromosome is a powerful tool for characterizing certain types of forensic biological evidence and for identifying possible male lineages [1–12]. Due to a lack of recombination, a Y-STR profile is evaluated as a haplotype. The power of discrimination (PD) of Y-STR haplotypes is approximately 99.9% with 10 forensically-selected Y-STRs, and 99.99% with 16 Y-STR markers [10,11 and manuscript in review]. While less discriminatory than autosomal STR-based DNA profiles collectively, Y-STR haplotypes, nonetheless, are useful in the investigation of sexual offenses, since the female DNA contribution in a mixed sample does not compromise analysis of the male contributors DNA in the sample [13,14].

Interpretation of Y-STR markers in forensic evidence is necessarily different from that of the autosomal STR markers that are currently being used in DNA forensics [15]. The estimation of the rarity of a single source Y-STR haplotype is based on the count of the specific target haplotype in a reference population database(s) [11–13]. Haplotypic organization of Y-STR markers also must be considered when quantifying the significance of mixture evidence [16–18].

The general principles for mixture interpretation have been discussed for autosomal short tandem repeat (STR) [15,19–22]. Interpretation of Y-STR mixtures also has been presented. Fukshansky and Bär [23] analyzed Y-STR mixture by recursive searches of all possible haplotype combinations given a number of contributors. However, they did not take into account occurrences of multiple alleles at some of the Y-STR markers. One of the commonly used Y-STR markers, DYS385, often presents two alleles. In US forensic Y-STR database (manuscript in review), 114 out of a total of 17,447 haplotypes had more than one allele at a single locus (excluding the DYS385 locus unless it presented more than two alleles). Therefore, Y-STR profiles with at least one locus having multiple alleles are not rare and should be considered in mixture evidence interpretation. Additionally, Fukshansky and Bär [23] did not describe how the number of contributors was determined. Wolf et al. [24] and Krawczak [25] described a similar analytical framework, in which alleles in the evidence (i.e., *V*) were separated

* Corresponding author. Tel.: +1 817 735 2773; fax: +1 817 735 5016.
E-mail address: jge@hsc.unt.edu (J. Ge).

1344-6223/$ - see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.legalmed.2010.02.003

into alleles observed in known contributors (i.e., W) and alleles not observed in known contributors. However, known contributors and unknown contributors of a mixture could share common alleles. Their framework also did not consider multiple allele situations. Both methods start from the mixture profile and recursively seek all possible subset haplotypes of the mixture profile. The computational complexity of this approach becomes very high, particularly when the number of alleles at each locus and the number of loci are large and the possibility of multiple alleles at single locus is considered. More importantly, most haplotype profiles will not be seen in a database because of the strong linkage among Y-STR loci and the limited size of current databases. Hence, this approach is cumbersome and not practically useful. Shrestha et al. [26] suggested an approach for mixture evaluation based on that used for autosomal markers and extended the approach to Y-STR haplotypes; but again the possibility of multiple alleles at single locus was ignored. Buckleton et al. [27] discussed the likelihood ratio (LR) method for interpreting Y-STR mixtures. Unfortunately, their equations incorporating the effect of substructure were incorrect.

Budowle et al. [12] proposed the logic of starting from Y-STR database, instead of the mixture profile itself, to search for the possible haplotype combinations. Both probability of exclusion (PE) and LR approaches were suggested for assessing the significance of mixture samples. In this study, both methods are further formalized, in which all possible contributor haplotypes of mixtures from database are listed first, and then the possible explanations for the mixture are exhaustively searched based on the list. Population substructure is also incorporated in both methods. Examples are used to explain the details of calculation. Furthermore, in order to investigate the pattern of mixtures, a simulation study is conducted by combining two and three haplotypes as a mixture based on the Y haplotype data excerpted from US forensic Y-STR database (manuscript in review). The average PE is calculated for each population; possible haplotype combination to explain the two and three persons mixtures are searched in a reference data set.

2. Methods

The theory and logic of both the PE and LR have been well-described and commonly used in interpreting autosomal STR mixture profile [15,19]. The PE provides an estimate of the portion of the population that has a genotype composed of at least one allele not observed in the mixed profile. The LR provides the ratio of the probability of observing the DNA evidence for two competing hypotheses, given the mixture evidence. The details of both methods for interpreting a Y-STR haplotype mixture are described as follows.

2.1. Probability of exclusion (PE)

Begin with the simple logic that the alleles in a Y-STR haplotype are a subset of the alleles (at all loci) seen in the mixture. Assume that the haplotype cannot be excluded as a part contributor of the Y-STR mixture. This leads to the following procedure for computing the PE.

First, the entire Y-STR haplotype database is searched and all haplotypes (i.e., candidate list; L_u) are listed that can not be excluded as contributors of the mixture M. For example, the two-locus haplotypes [{10}, {13}], [{10, 11}, {13}], [{null}, {13}] and [{10}, {−}] are possible contributors of the mixture [{10, 11}, {12, 13}], and haplotypes [{9}, {13}] and [{10, 11, 12}, {13}] are not contributors of the mixture. Here, "{10, 11}" presents two alleles 10 and 11 at single locus, "{−}" is an untyped locus and "{null}" is null allele. Note that "null" alleles and untyped loci are different. Null alleles are usually due to the variants at PCR primer binding regions of the STR markers. In contrast, low quantity and/or quality of samples may cause some loci to be untypable.

Second, sum the frequencies of all haplotypes in candidate list L_u; then the PE of the mixture is the complement of the haplotype frequencies summation for a specific reference database or population (Eq. (1), in which $\Pr(H)$ is the frequency of haplotype H).

$$\mathrm{PE} = 1 - \sum_{H \in L_u} \Pr(H) \tag{1}$$

The PE depends on the population chosen, database size and the number of loci in the mixture. Generally, the greater the number of loci that there is the higher the PE will be. A larger database size provides a more accurate estimation.

2.2. Likelihood ratio (LR)

As in the assessment of mixture profiles of autosomal STRs, the PE evaluation does not consider the Y-STR profiles of the known persons tested. Therefore, the full strength of the observations (such as, a subset of the tested persons explains all alleles observed in the mixture profile) is not fully captured in the PE evaluation. The LR maximally uses the information from the database and all known samples.

First, all possible contributors of the mixture are searched in a Y-STR database as unknown contributors, as is done in the first step in the PE calculation. Together with the haplotype(s) of the known contributor(s), such as victim(s) and suspects(s), a candidate list (denoted as L; $L = L_u + known\ contributors$) of possible known and unknown contributor haplotypes of the mixture is generated.

Second, an assessment is made about the possible number of contributors in the mixture. Since the number in candidate list L is limited, we suggest setting the maximum number of contributors at the size of L or the number of candidate haplotypes. In reality, the actual number of contributors may be much smaller. From these considerations, the approach to evaluate the LR will include computations for a reasonable range of a varying number of contributors in the mixture.

Third, list all non-empty subsets of L which can explain all alleles seen in the mixture profile as CL, namely, all possible haplotype combinations that can explain the mixture from the candidate list L. For example, there are 4 haplotypes, {8}, {9}, {10} and {8, 10} at a single locus (the null allele is ignored in this example since all null alleles at a multiple-locus haplotype are very unlikely). For a mixed sample {8, 9, 10}, all possible haplotype contributor combinations that can explain the mixture are

a. {9} + {8, 10}.
b. {8} + {9} + {10}.
c. {8} + {9} + {8, 10}.
d. {9} + {10} + {8, 10}.
e. {8} + {9} + {10} + {8, 10}.

The likelihood of each combination only depends on the haplotypes in L_u, if population substructure is not considered. Suppose there are l unknown contributor haplotypes in L_u, the likelihood of the combination can be calculated by

$$L(H_1, H_2, \ldots, H_l) = P(l, l) \prod_{i=1}^{l} \Pr(H_i), H_i \in L_u, \tag{2}$$

where H_i is the haplotype, $\Pr(H_i)$ is the frequency of haplotype H_i, and $P(l, l)$ is the permutation of l. For example, the likelihood of the combination including three unknown contributors {8}, {9}, and {10} is $6 * \Pr(\{8\}) * \Pr(\{9\}) * \Pr(\{10\})$.