Simulated approach to estimate the number and combination of known/unknown contributors in mixed DNA samples using 15 short tandem repeat loci

Sho Manabe, Chihiro Kawai, Keiji Tamaki*

Department of Forensic Medicine, Kyoto University, Graduate School of Medicine, Kyoto, Japan

Abstract. The calculation of likelihood ratios (LRs) for DNA mixture analysis is necessary to establish an appropriate hypothesis based on the estimated number of contributors and known contributor genotypes. In this paper, we recommend a relevant analytical method from the 15 short tandem repeat typing system (the Identifiler multiplex), which is used as a standard in Japanese forensic practice and incorporates a flowchart that facilitates hypothesis formulation. We postulate that: (1) all detected alleles need to be above the analytical threshold (e.g., 150 relative fluorescence unit (RFU)); (2) alleles of all known contributors should be detected in the mixture profile; (3) there should be no contribution from close relatives. Furthermore, we deduce that mixtures of four or more persons should not be interpreted by Identifiler as the LR values of 100,000 simulated cases have a lower expectation of exceeding our temporal LR threshold (10,000) which strongly supports the prosecution hypothesis. We validated the method using various computer-based simulations. The estimated number of contributors is most likely equal to the actual number if all alleles detected in the mixture can be assigned to those from the known contributors. By contrast, if an unknown contributor(s) needs to be designated, LRs should be calculated from both two-person and three-person contributions. We also consider some cases in which the unknown contributor(s) is genetically related to the known contributor(s).

Keywords: Short tandem repeat (STR); Likelihood ratio (LR); Mixture analysis; Number of contributors; Combination of contributors

1. Introduction

Several countries have developed guidelines for mixture interpretation, and a recent recommendation by the DNA Commission of the International Society for Forensic Genetics (ISFG) stipulated the analytical method for low-template DNA samples [1]. However, in Japan, mixed stains are rarely analyzed because of the complexity of the interpretational process. In particular, we hesitate to analyze low-template DNA samples because they are prone to be misinterpreted owing to stochastic effects such as allelic imbalance, drop-out, drop-in, and laboratory-based contamination. Even if the mixture contains high-template DNA, the formulation of alternative hypotheses for likelihood ratio (LR) calculation remains challenging. This is because we need to estimate the number and combination of contributors, using not only the genotypes of a mixture, but also those of known contributor(s) such as suspect(s) and victim(s). Hence, we should establish a relevant analytical method derived from the 15 short tandem repeat (STR) typing system (AmpFℓSTR® Identifiler® PCR Amplification Kit (Life Technologies, Carlsbad, CA)), which is used as a standard in Japanese forensic practice.

In this study, we recommended a process for estimating the number and combination of contributors in a mixture by considering the known contributor genotypes, and validated the process using various computer-based mixtures.

2. Recommended process for estimating the number and combination of contributors in a mixture

Determining the number and combination of contributors proceeds according to the flowchart in Fig. 1. This process assumes that: (1) all detected alleles are above the analytical threshold (e.g., 150 relative fluorescence unit (RFU)); (2) alleles of all known contributors will be detected in a mixture profile; and (3) there is no contribution from close-relatives. Furthermore, we deduce that mixtures of four or more persons should not be interpreted by Identifiler as the LR values of 100,000 simulated cases have a lower expectation of exceeding our temporal LR threshold (10,000), which strongly supports the prosecution hypothesis [2]. In a previous study, the percentage of samples for which the number of contributors was correctly estimated decreased dramatically for mixtures with four or more contributors [3].

Let *K* and *U* denote a known contributor and an unknown contributor, respectively. First, we select a value of *K* (up to and including three). The second step is to investigate the maximum number of extra alleles per locus. For example, when there are four alleles in one locus (named *A, B, C, D*) and one known contributor (K_1) of genotype (*A, B*), we have two extra alleles in this locus (*C, D*; these are from *U*(*s*)). If the maximum number of extra alleles in all loci is one or two, the minimum number of contributors (*MNC*) is two ($K_1 + U_1$). However, there actually may be three ($K_1 + U_1 + U_2$) or more contributors.

Thus, we need to estimate the number and combination of contributors probabilistically. The third step is to compare the two likelihoods of the observed DNA evidence (*E*) under the hypotheses that the number of contributors is *MNC* and *MNC* + 1. The case of *MNC* + 2 or more is not considered, because this probability was found to be very low in a past study [4]. For example, when *MNC* is two and the number of known contributors is one (K_1), we calculate the ratio $\Pr(E \mid K_1 + U_1) / \Pr(E \mid K_1 + U_1 + U_2)$. If the ratio is greater than one, the combination of contributors tends to be determined as $K_1 + U_1$. The determined combination of contributors is then used to calculate the LR for the suspect's contribution.

[*] Corresponding author. Tel: +81 75 753 4472. Fax: +81 75 761 9591. *E-mail address*: ktamaki@fp.med.kyoto-u.ac.jp.
Department of Forensic Medicine, Kyoto University, Graduate School of Medicine, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan.

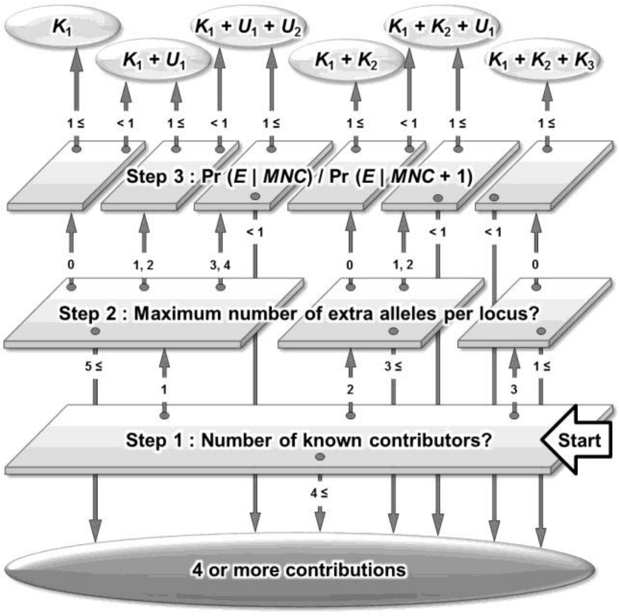

Fig. 1. Recommended process for estimating the number and combination of contributors in a mixture.

3. Method of validation for our recommended process

First, based on previously reported Japanese allele frequencies investigated by Identifiler, we computationally generated the genotypes of 600,000 unrelated individuals. Using these individuals, we synthesized 100,000 mixtures of two to six persons. Next, we estimated the combination of contributors in each mixture through the process illustrated in Fig. 1. This process was repeated for the number of Ks = 1, 2, and 3 (other than for two-person mixtures). The $K(s)$ was selected randomly from all contributors in a mixture. We then calculated the proportion of correctly identified mixtures in terms of the combination of contributors, and evaluated our process.

All programs used for the simulations were written using the statistical software R (version 3.0.1).

4. Results and Discussion

Using the recommended process, the probability of a correct estimation was always >90% (Table 1). In particular, if all alleles detected in the mixture could be assigned to the known contributors, the estimated number of contributors was most likely to be correct. By contrast, if any unknown contributors are designated, some misinterpretation occurred: e.g., estimated combination of contributors was K_1 + U_1 + U_2, but actual combination was K_1 + U_1 (0.187%). Even if the estimated combination was correct, Pr(E | MNC) might not differ significantly from Pr(E | MNC + 1). Thus, if the estimated combination contains $U(s)$, the LRs should be calculated assuming plural possibilities, such as two- and three-person contributions.

We also considered some cases in which the $U(s)$ is genetically related to the $K(s)$. We synthesized computer-based mixtures containing one sibling pair and determined the combination of contributors according to the procedure in Fig. 1. The results suggested a tendency to underestimate the number of contributors. For example, when the actual combination was K_1 + U_1 + U_2 (U_1 is a sibling of K_1), the probability of estimating K_1 + U_1 (14.4%) was much greater than that when U_1 was unrelated to K_1 (0.586%). Therefore, we should determine the combination of contributors cautiously in consideration of the possibility of relatives' contributions.

Table 1: Counting the estimated combination of contributors through our recommended process for each of the 100,000 mixtures.

Number of known contributors	Estimated combination of contributors	Total number of contributors				
		2	3	4	5	6
	K_1	0	0	0	0	0
1	$K_1 + U_1$	99,813	586	0	0	0
	$K_1 + U_1 + U_2$	187	92,424	7,712	32	0
	4 or more contributions	0	6,990	92,288	99,968	100,000
	$K_1 + K_2$	100,000	0	0	0	0
2	$K_1 + K_2 + U_1$	0	97,102	4,554	6	0
	4 or more contributions	0	2,898	95,446	99,994	100,000
3	$K_1 + K_2 + K_3$	-	100,000	3	0	0
	4 or more contributions	-	0	99,997	100,000	100,000

5. Role of funding: none

6. Conflict of interest: none

7. References

[1] P. Gill, L. Gusmao, H. Haned, et al., 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 (2012) 679-688.

[2] S. Manabe, Y. Mori, C. Kawai, et al., Mixture interpretation: Experimental and simulated reevaluation of qualitative analysis, Leg. Med. (Tokyo) 15 (2013) 66-71.

[3] H. Haned, L. Pene, J.R. Lobry, et al., Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count?, J. Forensic Sci. 56 (2011) 23-28.

[4] D.R. Paoletti, D.E. Krane, T.E. Doom, et al., Inferring the Number of Contributors to Mixed DNA Profiles, IEEE/ACM Trans. Comput. Biol. Bioinform. (2011).