

상염색체 STR 유전자형 시뮬레이션을 이용한 한국인에서의 혈연지수 분포 분석

양인석¹ · 이환영¹ · 박수정²
양우익¹ · 신경진¹

¹연세대학교 의과대학 법의학과
²대검찰청 DNA수사담당관실
DNA분석실

접 수 : 2013년 4월 29일
게재승인 : 2013년 5월 21일

이 논문은 2012년도 대검찰청의 DNA감식 기술의 국산화 및 선진화 연구사업의 지원을 받아 수행된 연구임(1333-304).

책임저자 : 신경진
(120-752) 서울시 서대문구 연세로 50
연세대학교 의과대학 법의학과
전화 : +82-2-2228-2481
FAX : +82-2-362-0860
E-mail : kjshin@yuhs.ac

Analysis of Kinship Index Distributions in Koreans Using Simulated Autosomal STR Profiles

In Seok Yang¹, Hwan Young Lee¹, Su Jeong Park², Woo Ick Yang¹,
Kyoung-Jin Shin¹

¹Department of Forensic Medicine, Yonsei University College of Medicine, Seoul, Korea
²DNA Analysis Laboratory, DNA Forensic Division, Supreme Prosecutors' Office, Seoul, Korea

Kinship testing in forensic casework is largely based on a likelihood ratio (LR) approach with short tandem repeat (STR) markers; however, in order to efficiently identify potential relatives in a specific population, the threshold values for kinship prediction should be determined by analyzing the kinship index distributions of the population in question. In this study, 250,000 DNA profiles were simulated using allele frequencies at 20 autosomal STR loci in Koreans, then the LRs were calculated for true close relatives and unrelated pairs. The LR distributions in related and unrelated pairs under a given relationship were compared in 2 sets of 13 Combined DNA Index System (CODIS) and 20 STR profiles. Using 13 CODIS STRs, true relatives in parent/child and full-sibling relationships were sufficiently discriminated from unrelated pairs with LR thresholds of 1,000 and 100, respectively. However, the CODIS STRs lacked the discriminatory power to differentiate between related and unrelated pairs in uncle/nephew and first cousin relationships due to high false-positive and false-negative rates with a LR threshold of 10. Increasing the number of STR loci to 20 increased discrimination of close relatives, but high false results remained in uncle/nephew and first cousin relationships. The kinship index data from this study will help make decisions on various kinship testing and familial searching in Koreans.

Key Words : STR, Kinship index, Simulation, Familial searching, Koreans

서 론

짧은연쇄반복 (short tandem repeat; STR)은 2~7 base pair (bp)로 구성된 반복단위 (repeat unit)가 연속적으로 나타나는 표식자로, 이들의 길이가 사람마다 다르고 독립적으로 유전되기 때문에 법과학 영역에서 개인식별과 혈연관계의 확인에 널리 사용되고 있다.¹⁾ 우리나라를 비롯한 많은 나라에서는 미국 연방수사국에서 사용하고 있는 13개의 Combined DNA

Index System (CODIS) STR (D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, CSF1PO, FGA, TH01, TPOX, vWA)을 기본적으로 분석하여 범죄수사에 활용하고 있으며,²⁾ 최근에 유럽과 미국에서는 다양한 혈연관계의 확인이나 분해된 시료로부터 더욱 정확한 유전자분석 결과를 얻기 위하여 새로운 상염색체 STR을 추가하여 사용하는 추세에 있다.

개인식별이나 혈연관계의 확인을 위한 상염색체 STR 분석 결과는 통계적 확률값을 산출하여 평가하게 되는데, 이때 정확

률(match probability)²⁾과 우도비(likelihood ratio; LR)^{3, 4)}를 대표적으로 사용한다. 혈연관계의 평가에서는 혈연지수(kinship index)라고도 불리는 우도비를 주로 활용하고 있으며, 특히 부모/자식 관계에서의 우도비는 부권지수(paternal index), 형제 관계에서의 우도비는 형제지수(sibling index)라고 한다. 두 사람 사이의 혈연관계 여부를 판단하기 위해 i) 서로 혈연관계인 경우와 ii) 서로 아무런 관계가 없는 경우의 두 가지 상황을 가정하여 두 사람의 유전자형이 특정 인구집단에서 나타날 확률을 계산하고 이들의 비율인 우도비를 산출한다. 우도비가 1 보다 큰 경우에는 서로 혈연관계가 사실이라는 가정에 힘을 실어주는 것이며, 1 보다 작은 경우에는 서로 아무런 관계가 없다는 가정에 힘이 실리게 된다.

혈연관계의 확인을 위한 상염색체 STR 분석에서 부모/자식 간에는 모든 유전자에서 최소한 1개의 대립유전자를 공유하기 때문에 우도비를 평가하여 그 관계를 판정하는 데 대체로 무리가 없지만, 형제나 3촌 이상의 관계에서는 두 사람 사이의 촌수가 멀어질수록 서로 대립유전자를 공유할 확률이 감소하여 우도비에 기초하여 혈연관계를 확정하는데 어려움이 있다.³⁾ 실제 가족 시료를 대상으로 다양한 혈연관계에서 나타나는 우도비를 조사하여 이러한 문제를 해결하는데 참고할 수 있겠으나, 현실적으로 특정 인구집단에 나타나는 혈연지수 분포를 보여주기에 충분한 가족 시료를 확보하기가 쉽지 않아 주로 유전자형 시뮬레이션을 통해 혈연지수 분포를 분석하여 왔다.⁵⁻⁸⁾

국내에서도 재외 한국인의 국적 취득과 6.25 전사자 유가족 찾기, DNA 데이터베이스 기반의 가족 검색(familial searching) 등과 같은 다양한 혈연관계의 확인에 필요한 혈연지수 분포 분석이 요구되고 있다. 본 연구에서는 13개 CODIS STR 및 20개 상염색체 STR 유전자형 생성 시뮬레이션을 통하여 한국인 1촌~4촌 관계에서 나타나는 혈연지수 분포를 분석하여, 다양한 혈연관계를 평가하는데 가이드라인이 되는 자료를 제시해 보고자 하였다.

연구 방법

1. 상염색체 STR 유전자의 대립유전자 빈도

시뮬레이션으로 임의의 상염색체 STR 유전자형을 생성하기 위해 한국인에서 관찰된 총 20개의 상염색체 STR을 대립유전자 빈도를 이용하였다. 이들 중에서 13개의 CODIS STR 및 D2S1338과 D19S433의 대립유전자 빈도는 대검찰청에서 제공한 자료를 사용하였으며[미발표 자료], 나머지 5개 STR(D1S1656, D2S441, D10S1248, D12S391, D22S1045)에 대해서는 박 등이 발표한 자료를 참조하였다.⁹⁾

2. 상염색체 STR 유전자형 생성

Microsoft® Excel® 프로그램을 기반으로 VBA (Visual Basic for Application) 언어로 작성한 Macro 기능을 이용하여 상염색체 STR 유전자형을 생성하는 시뮬레이션을 구현하였다. 우선 한국인의 상염색체 STR 대립유전자 빈도를 바탕으로 무작위 함수를 통해서 20개의 STR로 이루어진 250,000개의 임의의 유전자형을 생성시켰다. 다음으로 하나의 가족 유전자형을 구성하기 위해서 생성된 250,000개의 유전자형 중에 임의로 4개의 유전자형을 선택하여 이들을 각각 할아버지, 할머니, 어머니, 숙모의 유전자형으로 할당하였으며, 멘델의 유전법칙에 따라 할아버지와 할머니의 유전자형으로부터 아버지와 삼촌의 유전자형을, 아버지와 어머니의 유전자형으로부터 아들의 유전자형을, 그리고 삼촌과 숙모의 유전자형으로부터 사촌의 유전자형을 생성시켰다(Fig. 1). 이러한 과정을 반복하여 50,000쌍의 가상 가계도를 구축했다. 우도비 계산을 위한 유전자형 짝(genotype pair)은 서로 혈연관계가 사실인 경우에는 특정 관계(부모/자식, 형제, 삼촌/조카, 사촌)에 해당하는 유전자형을 가상 가계도로부터 얻었으며, 아무런 혈연관계가 없는 경우에는 임의의 유전자형을 추가로 생성시킨 후 이들을 무작위로 짝지어 얻었다.

3. 우도비 계산

혈연관계에서의 우도비는 i) 서로 특정 혈연관계에 있는 경우(related, H_p)와 ii) 서로 혈연관계가 없는 경우(unrelated, H_d)의 두 가지 상반되는 가정 하에 특정 인구집단에서 주어진 유전자형 짝이 나타날 확률의 비로 계산된다(Eq. 1). 이들 확률은 다시 특정 관계에서의 Identity-By-Descent (IBD) 확률 [$Pr(Z_2), Pr(Z_1), Pr(Z_0)$]¹⁰⁾과 하나의 유전자형이 G_1 으로 주어진 상황에서 다른 하나의 유전자형이 G_2 가 되는 조건부 확률

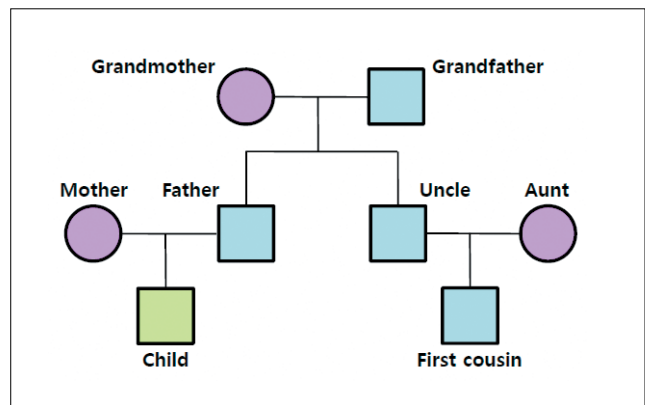


Fig. 1. Virtual pedigree constructed by simulated autosomal STR profiles in this study.

$[\Pr(G_2|G_1, Z_2), \Pr(G_2|G_1, Z_1), \Pr(G_2|G_1, Z_0)]$ 로 표현된다 (Eq. 2). 여기서 유전자형에 따른 조건부 확률은 ITO 행렬¹⁰⁾을 이용하여 계산하였다.

$$LR = \frac{\Pr(G_2|G_1, H_p)}{\Pr(G_2|G_1, H_d)} \quad - \text{Eq. 1}$$

$$\begin{aligned} \Pr(G_2|G_1) &= \Pr(G_2|G_1, Z_2) \times \Pr(Z_2) \\ &+ \Pr(G_2|G_1, Z_1) \times \Pr(Z_1) \\ &+ \Pr(G_2|G_1, Z_0) \times \Pr(Z_0) \end{aligned} \quad - \text{Eq. 2}$$

이러한 수식에 따라 가상 가계도와 무작위로 조합된 50,000쌍의 유전자형 짝으로부터 부모/자식, 형제, 삼촌/조카, 사촌 관계에서의 우도비를 각각 5회 반복하여 산출하였다.

4. 돌연변이를 고려한 우도비 계산

임의의 상염색체 STR 유전자형을 생성하는 시뮬레이션 과정에서 돌연변이는 고려하지 않았다. 하지만 임의의 유전자형 짝으로 이루어진 부모/자식 관계에서는 대부분의 유전자에서 공유하는 대립유전자가 없기 때문에 우도비 산출을 위하여 두 개의 유전자형 사이에 +1 증가(gain) 또는 -1 감소(loss)가 있는 경우를 돌연변이로 가정하였다. 이에 대한 우도비 계산은 American Association of Blood Bank에서 권장하는 방법¹¹⁾에 따라 계산하였으며, 한국인에 대한 상염색체 STR의 돌연변이율에 대한 자료가 보고된 것이 없어 비교적 거리가 가까운 민족인 중국의 한족(Chinese Han population)에서 조사된 돌연변이율¹²⁾을 사용하였다.

5. 통계 분석

부모/자식, 형제, 삼촌/조카, 사촌 관계에서의 우도비 분포를 비교하기 위해 5회 반복하여 산출된 우도비의 로그 중앙값 [$\text{Log}_{10}(\text{median LR})$]에 대한 평균(mean)과 표준편차(standard deviation)를 Microsoft® Excel® 로 계산하였고, 각각의 관계에서 1~100,000 범위의 우도비 값을 기준으로 거짓 양성률(false-positive rate)과 거짓 음성률(false-negative rate)을 조사하였다.

결 과

한국인에서 관찰된 13개의 CODIS 및 20개 상염색체 STR의 대립유전자 빈도를 이용하여 시뮬레이션으로 부모/자식, 형제, 삼촌/조카, 사촌의 관계인 경우와 그렇지 않은 경우에 해당하는 각각 50,000쌍의 유전자형 짝을 얻었고, 이를 통해 각각의 특정 관계에 대한 우도비를 산출하여 그 분포를 얻을 수 있었다(Figs. 2-5). 시뮬레이션에 사용된 상염색체 STR의 수가 동일한 경우에 1촌에서 4촌으로 촌수가 멀어짐에 따라서 서로 특정 혈연관계인 경우와 그렇지 않은 경우의 우도비 분포가 점점 더 겹쳐지는 것을 확인할 수 있었으며, 상염색체 STR을 13개에서 20개로 증가시켰을 때 혈연관계인 경우에는 우도비 값이 증가하는 방향(오른쪽)으로 우도비 분포 그래프가 이동하였으며, 그렇지 않은 경우에는 우도비 값이 감소하는 방향(왼쪽)으로 이동하였다.

1. 부모/자식 관계에서의 우도비 분포

부모/자식 관계에서 13개의 CODIS STR을 사용했을 때 서로 관계가 사실인 경우에는 모든 우도비 값이 1보다 컸으며

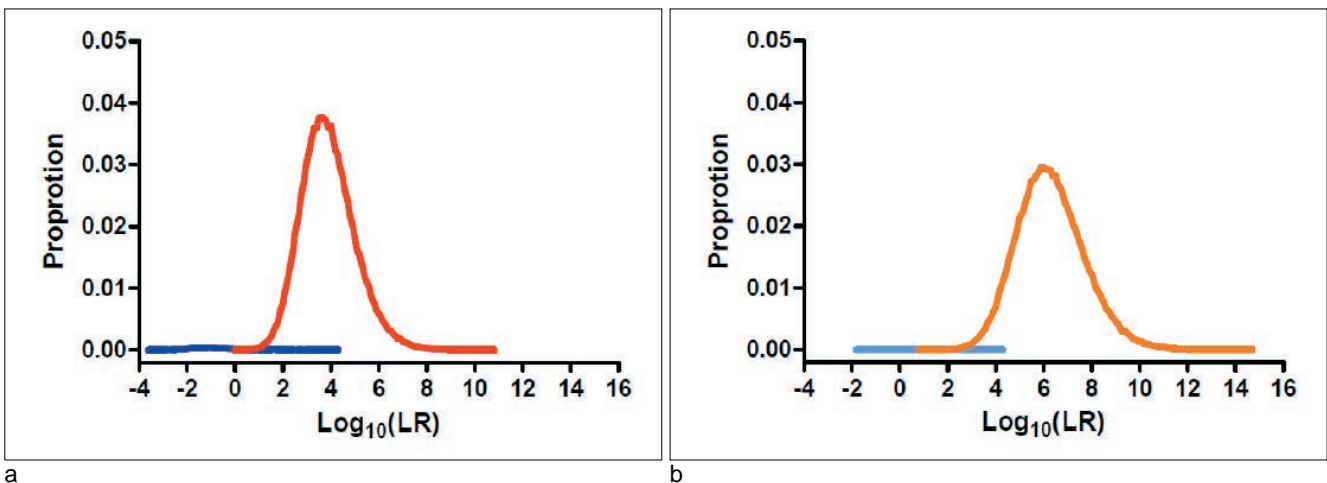


Fig. 2. Likelihood ratio distributions for parent/child relationship with 13 (a) and 20 autosomal STR loci (b). Red and orange lines represent related distributions. Blue and skyblue lines indicate unrelated distributions.

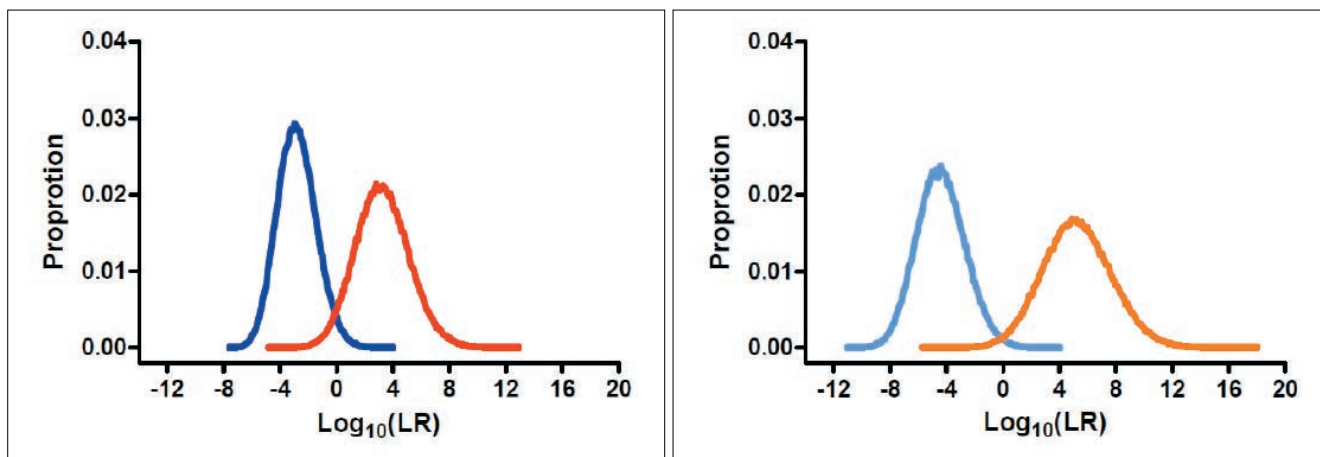


Fig. 3. Likelihood ratio distributions for full-siblings relationship with 13 (a) and 20 autosomal STR loci (b). Red and orange lines represent related distributions. Blue and skyblue lines indicate unrelated distributions.

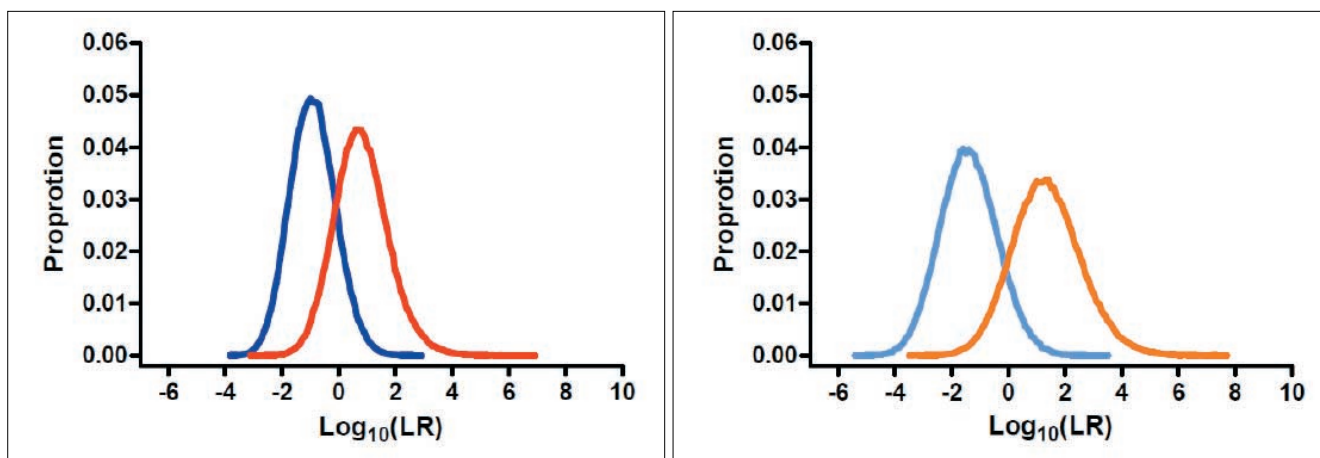


Fig. 4. Likelihood ratio distributions for uncle/nephew relationship with 13 (a) and 20 autosomal STR loci (b). Red and orange lines represent related distributions. Blue and skyblue lines indicate unrelated distributions.

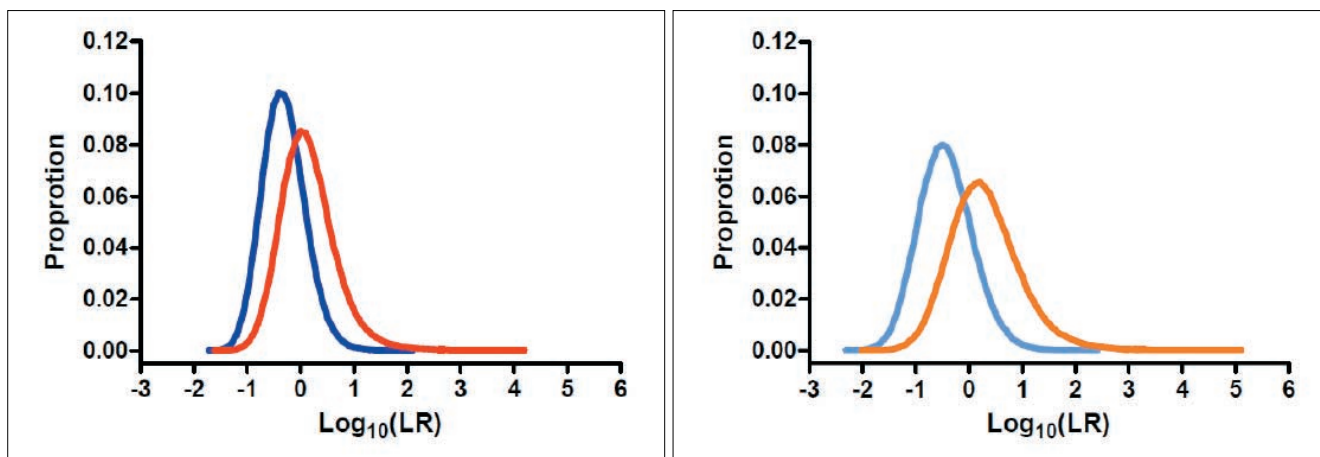


Fig. 5. Likelihood ratio distributions for first cousins relationship with 13 (a) and 20 autosomal STR loci (b). Red and orange lines represent related distributions. Blue and skyblue lines indicate unrelated distributions.

(Fig. 2a 붉은색 선), 이때의 우도비 분포는 로그 중앙값 3.90을 중심으로 형성되었다(Table 1). 그리고 아무런 관계가 없지만 부모/자식의 관계로 설정한 경우에는 대부분의 STR 유전좌에서 공유하는 대립유전자의 부재로 우도비 값이 0.00이 되기 때문에, 이에 대한 우도비 분포는 비율이 아주 낮았으며(Fig. 2a 파란색 선) 우도비의 로그 중앙값은 -0.83으로 나왔다(Table 1). 그럼에도 일부에서 우도비 값이 10,000 이상으로 나오는 경우도 확인할 수 있었다. 상염색체 STR을 13개에서 20개로 증가함에 따라 서로 관계가 사실인 경우의 우도비 분포가 오른쪽으로 이동하여(Fig. 2b 주황색 선) 우도비의 로그 중앙값이 3.90에서 6.31로 증가하였다(Table 1). 그리고 아무런 관계가 없는 경우에는 극히 일부에서만 우도비 값을 계산할 수 있었으며, 우도비의 로그 중앙값이 0.36으로 조금 증가했다(Fig. 2b 하늘색 선).

2. 형제 관계에서의 우도비 분포

형제 관계에서는 13개의 CODIS STR을 사용했을 때 서로 관계가 사실인 경우의 우도비 분포는 부모/자식 관계의 경우에서 더욱 넓은 구간에서 형성되고 있으며(Fig. 3a 붉은색 선), 우도비의 로그 중앙값은 3.30으로 비슷한 수준이었다(Table 1). 부모/자식 관계와 다르게 형제 관계에서는 모든 상염색체 STR 유전좌에서 대립유전자를 반드시 공유하지 않아도 되기 때문에, 아무런 관계가 없지만 형제 관계로 설정한 경우의 우도비도 정규 분포를 형성하였으며(Fig. 3a 파란색 선) 우도비의 로그 중앙값은 -2.76이었다(Table 1). 상염색체 STR을 20개로 증가시키기에 따라 서로 관계가 사실인 경우의 우도비 분포는 오른쪽으로 이동하였으며, 그렇지 않은 경우의 우도비 분포는 왼쪽으로 이동하였다(Fig. 3b). 이로 인하여 서로 관계가 사실인 경우의 우도비의 로그 중앙값은 3.30에서 5.30으로 증가했고, 아무런 관계가 없는 경우에는 -2.76에서 -4.34로 감소했다(Table 1).

3. 삼촌/조카 관계에서의 우도비 분포

삼촌/조카 관계에서 13개의 CODIS STR을 사용했을 때 서로 관계가 사실인 경우와 그렇지 않은 경우의 우도비 분포가

형제 관계의 경우보다 많은 부분이 겹치는 것을 확인할 수 있다(Fig. 4a). 게다가 우도비 분포가 나타나는 구간도 더 좁은 범위에서 나타났다. 우도비의 로그 중앙값은 서로 관계가 사실인 경우에는 0.86이었으며, 아무런 관계가 없는 경우에는 -0.83으로 나왔다(Table 1). 상염색체 STR을 20개로 증가시켰을 때 삼촌/조카 관계에서도 형제 관계에서 볼 수 있었던 것처럼 우도비 분포의 이동이 있었으며(Fig. 4B), 이에 따라 우도비의 로그 중앙값도 서로 관계가 사실일 때에는 0.86에서 1.41로 커졌으며, 아무런 관계가 없는 경우에는 -0.83에서 -1.32로 작아졌다(Table 1). 확률은 낮았지만 서로 관계가 사실인 경우에서 충분한 수의 공유하는 대립유전자가 존재함으로 우도비 값이 10,000 이상으로 나오는 경우도 볼 수 있었다.

4. 사촌 관계에서의 우도비 분포

사촌 관계에서 13개의 CODIS STR을 사용했을 때 서로 관계가 사실인 경우와 그렇지 않은 경우의 우도비 분포가 서로 상당 부분이 겹치는 것을 볼 수 있었다(Fig. 5a). 우도비의 로그 중앙값이 서로 관계가 사실인 경우에는 0.20이고, 아무런 관계가 없는 경우에는 -0.23으로 이들 값 간의 차이가 매우 작다는 것으로 두 개의 우도비 분포가 얼마나 겹치고 있는지를 말해 준다(Table 1). 상염색체 STR을 20개로 증가시켰음에도 여전히 두 개의 우도비 분포가 상당히 겹치고 있었다(Fig. 5b). 우도비의 로그 중앙값도 서로 관계가 사실인 경우에는 0.20에서 0.35로 약간 증가했으며, 아무런 관계가 없는 경우에는 -0.23에서 -0.36으로 조금 감소했다(Table 1). 그럼에도 서로 관계가 사실인 경우에 우도비 값이 1,000 이상으로 나오는 경우도 볼 수 있었다.

5. 거짓 양성률 및 거짓 음성률을 이용한 우도비 분포의 평가

시뮬레이션을 통해 얻은 우도비 분포를 이용하여 혈연관계 확인을 위한 적절한 우도비의 범위를 결정하기 위해서 1~100,000까지의 우도비 값을 기준으로 거짓 양성률과 거짓 음성률을 조사하였다(Table 2). 상염색체 STR을 13개에서 20개로 증가시켰을 때 부모/자식 및 형제 관계에서 우도비 분포의 변화가 가장 뚜렷하게 나타났으며, 거짓 양성률 및 거짓 음

Table 1. Log₁₀ (Median Likelihood Ratio) Values for Each Relationship in Koreans

	No. of STRs	Relationship			
		Parent/Child	Full-siblings	Uncle/Nephew	First cousins
Related	13	3.90 ± 0.01	3.30 ± 0.00	0.86 ± 0.00	0.20 ± 0.00
	20	6.31 ± 0.01	5.30 ± 0.01	1.41 ± 0.01	0.35 ± 0.00
Unrelated	13	-0.83 ± 0.06	-2.76 ± 0.01	-0.83 ± 0.00	-0.23 ± 0.00
	20	0.36 ± 0.48	-4.34 ± 0.01	-1.32 ± 0.00	-0.36 ± 0.00

Table 2. False-positive and False-negative Rates on Kinship Index (KI) Distributions for Each Relationship

No. of STRs	Threshold KI	False-positive rate (%)			
		Parent/Child	Full-siblings	Uncle/Nephew	First cousins
13	1	0.213 ± 0.005	2.951 ± 0.040	15.656 ± 0.137	28.455 ± 0.189
	10	0.126 ± 0.016	0.562 ± 0.029	1.457 ± 0.039	0.392 ± 0.027
	100	0.071 ± 0.013	0.090 ± 0.008	0.056 ± 0.014	0.002 ± 0.002
	1,000	0.019 ± 0.008	0.010 ± 0.006	0.000 ± 0.000	0.000 ± 0.000
	10,000	0.002 ± 0.003	0.000 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
	100,000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
20	1	0.015 ± 0.004	0.856 ± 0.047	10.176 ± 0.114	24.248 ± 0.157
	10	0.006 ± 0.004	0.166 ± 0.017	1.348 ± 0.077	0.757 ± 0.047
	100	0.002 ± 0.001	0.025 ± 0.009	0.084 ± 0.013	0.006 ± 0.004
	1,000	0.001 ± 0.001	0.003 ± 0.002	0.002 ± 0.002	0.000 ± 0.000
	10,000	0.000 ± 0.001	0.001 ± 0.001	0.000 ± 0.000	0.000 ± 0.000
	100,000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
No. of STRs	Threshold KI	False-negative rate (%)			
		Parent/Child	Full-siblings	Uncle/Nephew	First cousins
13	1	0.000 ± 0.000	3.374 ± 0.045	16.865 ± 0.067	33.170 ± 0.211
	10	0.059 ± 0.010	10.496 ± 0.069	56.012 ± 0.106	92.333 ± 0.103
	100	2.193 ± 0.037	24.353 ± 0.273	87.720 ± 0.187	99.374 ± 0.012
	1,000	18.428 ± 0.095	43.853 ± 0.139	97.750 ± 0.091	99.950 ± 0.010
	10,000	53.511 ± 0.264	64.384 ± 0.189	99.684 ± 0.028	99.998 ± 0.002
	100,000	82.622 ± 0.166	81.030 ± 0.134	99.973 ± 0.006	100.000 ± 0.000
20	1	0.000 ± 0.000	1.026 ± 0.025	10.933 ± 0.117	27.921 ± 0.303
	10	0.000 ± 0.001	3.116 ± 0.115	36.296 ± 0.204	83.170 ± 0.096
	100	0.008 ± 0.004	7.870 ± 0.114	68.610 ± 0.175	98.028 ± 0.044
	1,000	0.265 ± 0.028	16.444 ± 0.123	89.549 ± 0.104	99.787 ± 0.022
	10,000	3.084 ± 0.082	29.145 ± 0.121	97.460 ± 0.058	99.984 ± 0.002
	100,000	15.639 ± 0.133	44.950 ± 0.206	99.484 ± 0.012	99.999 ± 0.001

성률 간의 차이도 확연히 드러났다.

부모/자식 관계에서 13개의 CODIS STR을 사용했을 때 우도비 값이 1,000에서 거짓 양성률과 거짓 음성률이 각각 0.019%와 18.428%였으며, 상염색체 STR을 20개로 증가시켰을 때에는 비슷한 수준의 거짓 양성률(0.015%)과 거짓 음성률(15.639%)이 각각 우도비 값 1과 100,000에서 나왔다. 특히 우도비 값 1,000에서 거짓 양성률이 0.001%였으며, 거짓 음성률이 0.265%로 상당히 낮아졌다. 따라서 부모/자식의 관계에서는 13개 및 20개의 상염색체 STR을 사용하여 계산된 우도비 값이 1,000 이상이라면 부모/자식의 관계를 확인하기에 적합할 것이다.

형제 관계에서는 13개의 CODIS STR을 사용했을 때 낮은 거짓 양성률(0.090% 및 0.010%)이 우도비 값 10과 1,000에서 각각 보였지만, 낮은 거짓 음성률(10.496% 및 24.353%)은 우도비 값 10과 100에서 나타났다. 상염색체 STR을 20개로 증가시켰을 때 우도비 값 10과 1,000에서 거짓 양성률이 0.025%와 0.003%였으며, 거짓 음성률이 7.870%와 16.444%로 낮아졌다. 그리고 우도비 값 10,000에서는 거짓 양성률은 0.001%로 값으로 우도비 값 1,000에서와

비슷한 수준이었지만, 거짓 음성률이 29.145%로 두 배 가까이 컸다. 이러한 점에서 형제 관계를 구분하기 위한 우도비 값은 13개 및 20개의 상염색체 STR로 얻어진 우도비 값이 100 이상이라면 충분할 것으로 보인다.

삼촌/조카 관계의 경우 13개의 CODIS STR을 사용했을 때 거짓 양성률이 우도비 값 10에서는 1.457%로 높은 편이었으며, 100에서는 0.056%로 낮은 편이었다. 거짓 음성률은 우도비 값 10과 100에서 각각 56.012%와 87.720%로 상당히 높았다. 상염색체 STR을 20개로 증가시켰을 때 거짓 양성률은 우도비 값 10에서 1.348%로 조금 낮아졌지만, 100에서는 0.084%로 조금 높아졌다. 거짓 음성률은 우도비 값 10과 100에서 각각 36.296%와 68.610%로 낮아졌지만, 여전히 높은 편이었다. 따라서 삼촌/조카 관계에서는 높은 거짓 음성률로 인해 이들의 관계 확인이 힘들 수 있겠지만, 20개의 상염색체 STR을 이용하여 우도비 값 10 이상을 기준으로 한 경우에서 가능성을 볼 수 있었다.

사촌 관계의 경우는 13개의 CODIS STR을 사용했을 때 우도비 값 10과 100에서 거짓 양성률이 각각 0.392%와 0.002%로 비교적 낮은 수준이었지만, 거짓 음성률의 경우는 각각

Table 3. Log₁₀ (Median Likelihood Ratio) Values for Each Relationship in US Caucasian⁶⁾

	No. of STRs	Relationship		
		Parent/Child*	Full-siblings	Half-siblings [†]
Related	13	4.00	3.38	0.84
	20	6.95	5.83	1.51
Unrelated	13	-16.57	-2.82	-0.85
	20	-	-4.62	-1.44

* Possible mutation in unrelated pairs identified as parent/child was differently treated on kinship index calculation compared with the case of this study (see <http://dna-view.com/mudisc.htm>).

[†]Half-siblings relationship is genetically identical with uncle/nephew relationship.

92.333%와 99.374%로 상당히 높았다. 또한, 상염색체 STR을 20개로 증가시켰을 때 거짓 양성률은 우도비 값 10과 100에서 0.757%과 0.006%로 이들 관계에서 모두 조금 커졌으며, 거짓 음성률은 각각 83.170%와 98.028%로 값이 약간 작아졌다. 결과적으로 13개의 STR로는 사촌 관계를 구분하기 위한 우도비 값을 찾을 수 없었으며, 심지어 20개의 상염색체 STR을 사용한다고 하더라도 확연한 차이는 없었다.

고 찰

법과학 영역에서 개인식별과 혈연관계의 확인에 기본적으로 사용되는 13개 CODIS STR이 나타내는 한국인에서의 우도비의 분포를 살펴보면, 부모/자식 및 형제 관계에서는 서로 관계가 사실인 경우와 그렇지 않은 경우의 우도비 분포는 비교적 잘 분리되었으며(Table 1), 각각의 관계에서 우도비 값 1,000과 100을 기준으로 낮은 수준의 거짓 양성률과 거짓 음성률을 가짐으로써 이들의 관계를 어렵지 않게 구분할 수 있겠다(Table 2). 하지만 삼촌/조카 및 사촌 관계에서는 서로 관계가 사실인 경우와 그렇지 않은 경우의 우도비 분포 그래프가 형제 관계와 비교해서 많은 부분에서 겹쳤으며(Figs. 4a, 5a) 거짓 음성률도 함께 커졌다(Table 2). 따라서 13개의 CODIS STR 유전좌로 3촌 이상의 혈연관계를 확인하는 것에는 한계가 있음을 확인할 수 있었다. 이러한 점을 극복해보고자 7개의 상염색체 STR 유전좌를 추가하여 총 20개의 STR을 사용함으로써 우도비 분포에 대한 영향을 조사하였다. 전체적으로 사용된 상염색체 STR 수가 증가함에 따라 우도비 분포의 이동(Figs. 2-5)과 함께 우도비의 로그 중앙값 변화(Table 1)도 확인되었다. 부모/자식 및 형제 관계에서는 20개 STR 유전좌를 사용함으로써 13개 STR 유전좌를 사용했을 때보다 우도비 값 1,000과 100에서 거짓 양성률과 거짓 음성률이 더 떨어졌기 때문에 STR 유전좌의 수를 증가시킴으로써 이들의 관계를 구분하는 능력이 더 향상됐다는 것을 보여 주었다. 하지만 20개의 STR 유전좌를 사용하더라도 삼촌/조카 및 사촌 관계는 우도비 값 10을 기준에서도 각각 36.296%와 83.170%로 여전히 높은

거짓 음성률을 보여주었다(Table 2). 결국, 20개의 STR 유전좌를 사용했음에도 여전히 삼촌/조카 및 사촌 관계의 확인이 어렵다는 것을 말해준다.

혈연관계가 멀어질수록 이들 간에 유전정보를 공유할 확률(IRD 확률)이 점점 감소하기 때문에, 형제, 삼촌/조카, 사촌 관계에서는 부모/자식의 관계의 경우와 비교하여 우도비 값이 작아질 것이라고 예상을 할 수 있다. 그렇지만 형제 관계에서의 우도비 값은 상당수가 부모/자식 관계와 유사하거나 오히려 크게 나오는 것을 볼 수 있었다(Figs. 2, 3). 13개 CODIS STR에 대한 우도비 분포에서는 약 56.1%가 1,000 이상의 우도비 값을 가졌으며, 20개 STR 유전좌를 사용한 경우에는 약 55.1%가 100,000 이상으로 크게 나왔다(Table 2). 부모/자식 및 형제 관계에서 두 개의 유전자형 사이에서 공유하는 대립유전자(Identity-By-State; IBS)의 수에 대한 분포를 조사했을 때, 부모/자식 및 형제 관계에서 IBS 중앙값이 13개 STR을 기준으로 각각 16개와 17개였으며, 20개 STR에서는 각각 24개와 26개로 나타났다(자료 제시 없음). 분명히 형제 관계에서는 공유하는 대립유전자가 없는 상염색체 STR 유전좌가 나올 가능성이 있음에도 부모/자식의 관계보다 평균적으로 공유하는 대립유전자의 수가 더 많았다. 이러한 점이 형제 관계에서의 우도비 값이 부모/자식 관계의 경우보다 크게 나올 수 있다는 것을 설명해준다.

본 연구에서는 20개의 상염색체 STR을 사용했음에도 3촌 및 4촌의 혈연관계를 갖는 사람과 아무런 관계가 없는 사람과의 구분이 충분치 않다는 것을 확인할 수 있었다. 이러한 점을 극복하기 위한 방법으로 다음과 같은 몇 가지를 더 고려해볼 수 있다. 첫 번째로 O' Connor 등⁶⁾ 및 Phillips 등¹³⁾의 연구에서 처럼 더욱 많은 수의 STR 유전좌를 사용하는 방법이다. 다만 STR 유전좌를 추가로 사용하는 경우에는 STR의 돌연변이율이 10⁻³~10⁻⁴ 정도로 높기 때문에, 유전자검사에 이용되는 STR 유전좌의 수가 증가할수록 돌연변이로 인해 공유하는 대립유전자를 찾지 못하는 경우도 많아질 수 있다는 점을 고려해야 할 것이다. 두 번째로 STR 유전좌보다 돌연변이율이 낮은 single nucleotide polymorphism (SNP) 마커²⁾를 이용하는 것

이다. STR 유전좌와 함께 SNP 마커를 추가로 사용함으로써 식별력(discrimination power)이 증가할 수 있을 것이다.¹³⁾ 다만 다수의 대립유전자를 갖는 STR에 비해서 SNP는 대부분 2개의 대립유전자만을 갖기 때문에, STR과 동일한 식별력을 보여주기 위해서는 상대적으로 많은 수의 SNP 마커가 필요하게 된다.¹⁴⁾ 마지막으로 Y 염색체 STR이나 미토콘드리아 DNA 염기서열 정보와 같은 lineage 마커를 사용하는 것이다. 이들 마커는 돌연변이의 가능성을 배제하면 한 세대에서 다음 세대로 동일한 유전정보가 전달된다.^{15, 16)} 때문에 이들 마커는 두 사람 사이의 혈연관계가 멀다고 할지라도 동일한 부계나 모계 관계를 확인하는데 유용하게 적용될 수 있다.

STR 유전자에서의 대립유전자 분포 및 이들의 빈도는 민족 또는 집단마다 다르게 나타난다.¹⁷⁾ 이러한 민족 간의 특징이 우도비 분포에는 어떠한 영향을 미치는지를 알아보기 위해 본 연구에서 조사된 한국인에서의 우도비 분포와 미국에 거주하는 백인에 대해 O'Connor 등이 조사한 우도비 분포 결과⁸⁾를 비교해 보았다(Tables 1, 3). 서로의 관계가 사실인 경우에는 미국 백인 집단에서 우도비의 로그 중앙값이 조금 높게 나왔으며, 아무런 관계가 없는 경우에는 부모/자식의 관계를 제외하고 미국 백인 집단에서 조금 낮게 나왔다. 또한, 혈연관계가 멀어질수록 한국인 집단과 미국 백인 집단 모두에서 우도비의 로그 중앙값 간의 차이는 더 줄어들었다. 이것은 비록 두 집단 간의 STR 대립유전자 빈도가 다른 양상을 나타낸다고 해도 이러한 차이가 우도비 분포에 미치는 영향은 그렇게 크지 않다는 것을 알 수 있었다.

염색체 상에서 두 개의 유전좌가 서로 가깝게 위치할 경우에는 재조합에 의한 독립적인 유전이 일어나기 어렵기 때문에 연관 불균형(linkage disequilibrium)이 나타날 수 있다.¹⁸⁾ 여러 개의 상염색체 STR 유전좌에 대한 우도비 계산에서 곱의 법칙(product rule)을 이용하는 것은 사용된 STR 유전좌가 독립적으로 유전된다는 점을 전제로 한다. 이상적으로 각 유전좌가 재조합에 의한 독립적인 유전이 가능하기 위해서는 서로 간의 거리가 50 megabase (Mb) 이상은 떨어져 있어야 한다.⁶⁾ 최근 유럽에서 유전자검사에 사용되는 European Standard Set의 STR을 7개에서 12개로 확장하였으며, 미국에서도 CODIS STR을 13개에서 20개로 확장하여 사용할 것을 권고하고 있다.¹⁾ 이들 STR 유전좌 중에서 vWA와 D12S391 간의 거리는 6.3 Mb 밖에 되지 않는다. 본 연구에서 20개의 상염색체 STR을 이용한 우도비 분포 분석에서는 vWA와 D12S391이 함께 사용되었으나, 이들 간에 존재하는 연관 불균형을 고려하지 않고 우도비를 계산하였다. 이전의 연구에서 모든 혈연관계에 대해서 연관을 무시하고 우도비를 계산하는 경우에는 평균적으로 과소평가되는 현상을 보이지만, 일부 특이한 경우에서 우도비가 3~7배 정도 과대평가되기도 하며, 심한 경우에는 25~30배까지 나타나기도 하였다.¹⁸⁾ 따라서 혈연관계의 평가에서

더욱 정확한 우도비의 산출 뿐만 아니라 추후에 다른 STR 유전좌가 추가되는 경우를 대비하여 유전좌 간의 연관 불균형에 대한 고려가 있어야 할 것으로 본다.

현재 미국(캘리포니아, 콜로라도, 플로리다, 뉴욕, 텍사스 및 버지니아), 영국, 네덜란드, 뉴질랜드에서는 DNA 데이터베이스를 대상으로 한 가족 검색이 시행되고 있다. 우리나라에서도 2010년부터 범죄자 DNA 데이터베이스가 구축되고 있으며 범죄수사에서 DNA 데이터베이스의 효율성을 증대하기 위하여 가족 검색에 대한 논의가 진행되는 단계에 있다.¹⁹⁾ 가족 검색은 범죄현장에서 수집된 증거로부터 얻은 DNA profile을 DNA 데이터베이스에서 검색하여 일치하는 결과를 얻지 못한 경우에 DNA 데이터베이스에서 유전적으로 유사한 DNA profile을 가진 사람들을 찾아내고 그들과 혈연관계에 있는 특징인이 연관될 가능성을 조사하여 범죄수사에 활용하는 방법이다.²⁰⁾ DNA 데이터베이스 기반의 가족 검색 결과 평가에서는 우도비와 IBS에 기초한 두 가지 통계적 방법이 사용될 수 있다.^{8, 21)} IBS를 이용하는 방법은 두 개의 유전자형 사이에 공유하는 대립유전자를 조사하고 이들의 수 또는 이들을 포함하는 유전좌의 수를 계산하는 것이다. 최근 Ge 등⁸⁾은 13 CODIS STR에 대해서 이러한 두 가지 방법을 이용하여 얻은 결과들을 비교하여 이들을 결합하여 사용하는 것이 가족 검색에서 정확성을 높일 수 있다고 보고하였다. 이에 따라 한국인에서 혈연관계 확인을 위한 평가에서도 우도비에 IBS를 접목하여 사용하는 것도 고려해 볼 필요가 있을 것이다.

결론적으로 본 연구에서 조사한 1촌~4촌 관계에 있는 한국인 두 사람 사이에서 나타날 수 있는 우도비의 분포는 친자 확인을 비롯하여 대형재난이나 재해로 인한 희생자와 실종자에 대한 신원 파악, 국적취득을 위한 심사, 그리고 DNA 데이터베이스를 대상으로 한 가족 검색 등과 같은 다양한 혈연관계의 확인을 위한 유전자검사 결과의 평가에 가이드라인이 되는 자료로 활용될 수 있을 것이다.

Acknowledgments

본 연구를 위해서 대검찰청의 상염색체 STR의 대립유전자 빈도 자료를 제공해 주시고, 연구 과정에서 조언을 아끼지 않으신 대검찰청 DNA수사담당관실 DNA분석실의 이승환 박사님께 감사의 마음을 전합니다.

참 고 문 헌

1. Butler JM, Hill CR. Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. *Forensic Sci Rev* 2012;24:15-26.
2. Butler JM. *Fundamentals of forensic DNA typing*. 2nd ed.

- San Diego: Elsevier Academic Press; 2009. p. 489-94.
3. Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* 2006;7:771-80.
 4. Shin KJ, Lee HY, Yang WI, et al. Statistical Interpretation in Making DNA-based Identifications of Mass Victims. *Korean J Leg Med* 2008;32:55-60.
 5. Tamaki K, Kaszynski RH, Yuan QH, et al. Likelihood evaluation using 15 common short tandem repeat loci: a practical and simulated approach to establishing personal identification via sibling/parental assessments. *Transfusion* 2009;49:578-84.
 6. O' Connor KL, Butts E, Hill CR, et al. Evaluating the effect of additional forensic loci on likelihood ratio values for complex kinship analysis. *Proceedings of the Twenty-first International Symposium on Human Identification*. 2010.
 7. Ge J, Budowle B, Chakraborty R. Choosing relatives for DNA identification of missing persons. *J Forensic Sci* 2011;56 Suppl 1:S23-8.
 8. Ge J, Chakraborty R, Eisenberg A, et al. Comparisons of familial DNA database searching strategies. *J Forensic Sci* 2011;56:1448-56.
 9. Park JH, Hong SB, Kim JY, et al. Genetic variation of 23 autosomal STR loci in Korean population. *Forensic Sci Int Genet*. 2013;7:e76-7.
 10. Li CC, Sacks L. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* 1954;10:347-60.
 11. Gjerston DW. The effect of an isolated single-locus inconsistency in the statistical evaluation of paternity. In: Walker RH, editor. *Paternity testing accreditation requirements manual of the American Association of Blood Banks*. 3rd ed. Bethesda: American Association of Blood Banks, 1998:98-9.
 12. Yan J, Liu Y, Tang H, et al. Mutations at 17 STR loci in Chinese population. *Forensic Sci Int* 2006;162:53-4.
 13. Phillips C, García-Magariños M, Salas A, et al. SNPs as supplements in simple kinship analysis or as core markers in distant pairwise relationship tests: When do SNPs add value or replace well-established and powerful STR tests? *Transfus Med Hemother* 2012;39:202-10.
 14. Schneider PM. Beyond STRs: The role of diallelic markers in forensic genetics. *Transfus Med Hemother* 2012;39:176-80.
 15. Jobling MA, Tyler-Smith C. Fathers and sons: the Y chromosome and human evolution. *Trends Genet* 1995;11:449-56.
 16. Butler JM, Levin BC. Forensic applications of mitochondrial DNA. *Trends Biotechnol* 1998;16:158-62.
 17. Budowle B, Shea B, Niezgoda S, et al. CODIS STR loci data from 41 sample populations. *J Forensic Sci* 2001;46:453-89.
 18. O' Connor KL, Tillmar AO. Effect of linkage between vWA and D12S391 in kinship analysis. *Forensic Sci Int Genet* 2012;6:840-4.
 19. Korean Speaking Working Group of ISFG, Seo SB, Lee SH, et al. DNA database searching using genetic relationship. *Korean J Leg Med* 2011;35:92-7.
 20. Bieber FR, Brenner CH, Lazer D. Human genetics. Finding criminals through DNA of their relatives. *Science* 2006;312:1315-6.
 21. Ge J, Budowle B. Kinship index variations among populations and thresholds for familial searching. *PLoS One* 2012;7:e37474.