JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, November 2011, Vol. 56, No. 6
doi: 10.1111/j.1556-4029.2011.01867.x
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Jianye Ge,[1,2] *Ph.D.; Ranajit Chakraborty,*[1,2] *Ph.D.; Arthur Eisenberg,*[1,2] *Ph.D.; and Bruce Budowle,*[1,2] *Ph.D.*

Comparisons of Familial DNA Database Searching Strategies

ABSTRACT: The current familial searching strategies are generally based on either Identity-By-State (IBS) (i.e., number of shared alleles) or likelihood ratio (i.e., kinship index [KI]) assessments. In this study, the expected IBS match probabilities given relationships and the logic of the likelihood ratio method were addressed. Further, the false-positive and false-negative rates of the strategies were compared analytically or by simulations using Caucasian population data of the 13 CODIS Short Tandem Repeat (STR). IBS ≥ 15, IBS ≥ 16, KI ≥ 1000, or KI ≥ 10,000 were found to be good thresholds for balancing false-positive and false-negative rates. IBS ≥ 17 and/or KI ≥ 1,000,000 can exclude the majority of candidate profiles in the database, either related or not, and may be an initial screening option if a small candidate list is desired. Polices combining both IBS and KI can provide higher accuracy. Typing additional STRs can provide better searching performance, and lineage markers can be extremely useful for reducing false rates.

KEYWORDS: forensic science, DNA database, kinship analysis, familial searching, Identity-By-State, likelihood ratio, moderate stringency match

Forensic DNA database searches have become an essential part of DNA-based forensic investigations. Generally, a forensic DNA sample is collected from a crime scene, typed, and then searched in a database (of, e.g., convicted felon profiles) to seek a direct match (i.e., all alleles at all interpretable evidentiary loci are the same as those in the candidate sample in the database). In some cases, a direct match is not obtained because the database does not contain all people in the population. To extend the investigative lead value of current databases, an alternative approach to determine the source of the forensic sample (e.g., a suspect) is to search the database for possible relatives of the true source of the sample. The concept of familial searching has been successfully used in a number of forensic investigations (1–4). The utility of familial searching of large DNA databases was highlighted recently because of its potential ability to develop investigative leads (5–12).

There are two general methods for searching for a relative (typically a sib or parent-offspring relationship) in a felon database, Identity-By-State (IBS) based or likelihood ratio based. Introduced in 1993, the IBS-based method simply compares the number of shared alleles between the forensic profile and the candidate profile(s) from a database (13). Potential candidate relatives of the forensic profile are suggested if the number of shared alleles reaches a predefined threshold. For example, the California Department of Justice used to require at least 15 Short Tandem Repeat (STR) alleles to be considered a potential familial hit for further investigation (14). Some other states (e.g., Nebraska, Oregon, and Washington) require at least one allele shared at all available loci (15). The SWGDAM Ad Hoc Committee on Partial Matches recommendations (10) also considers moderate stringency-matched profiles, along with a likelihood ratio-based method. A moderate stringency match is defined as follows: (i) if the forensic profile exhibits only a single allele at a locus, say A, the candidate profile has to have at least one copy of the allele A at the same locus; or (ii) if the forensic profile is heterozygous, say A, B, the candidate profile may only contain a single representation of either allele A or B. High stringency match (i.e., all alleles matched at each locus) is included (at some loci) in the moderate stringency match.

A likelihood ratio-based method compares the joint probabilities of the forensic, and candidate profiles given that the donors are related (e.g., parent–child or full-sib) versus unrelated. Essentially, this is the likelihood ratio or kinship index (KI) for a given relationship (e.g., paternity index for father–child relationship) (16,17). The KI can be directly used to evaluate candidates in familial searching or can be further modified. The SWGDAM Ad Hoc Committee on Partial Matches recommended a likelihood ratio-based measure, that is, Expected Kinship Ratio (EKR), for familial searching, which is the KI divided by the sample size of the searched database SWGDAM (10). Another measure, Expected Match Ratio (EMR), recommended by (10), depends only on the forensic profile itself, and is not directly applicable for familial detection for a specific candidate profile. The likelihood ratio-based method includes the allele frequency data and gives a relatively higher accuracy than that of the IBS-based method (5,7–9,12), especially when profiles share rare allele(s). However, the likelihood ratio-based method depends on specific population data and use of nonrelevant population data increase or reduce the likelihood ratio leading to false inclusions or false exclusions. Thus, the SWGDAM Ad Hoc Committee on Partial Matches (10)

[1]Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107.
[2]Institute of Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107.

Received 14 June 2010; and in revised form 13 Sept. 2010; accepted 10 Oct. 2010.

recommended that the maximum and minimum EKRs among Caucasians, African Americans, Southwestern Hispanics, and Southeastern Hispanics should be greater than 1 and 0.1, respectively.

In most felon database searches, familial searching typically seeks parent–child or full-sib relationships. Distant pairwise relationships (e.g., half-sib, uncle–nephew, etc.) have a relatively low number of shared alleles and a low probability of kinship determination based on a likelihood ratio with commonly used forensic STRs (18).

In this study, the analytical details are provided of expected IBS match probabilities between a pair of profiles given Identity-By-Descent (IBD) of the relationships and the expected IBS distributions of unrelated, parent–child and full-sib pair relationships with 13 CODIS STRs using allele frequency data from the Caucasian population as an example. Further, the logic of the likelihood ratio method is also addressed. Millions of unrelated pairs, parent–child and full-sib genotype data are simulated, and the KI of each pair is calculated by MPKin (19). The joint distributions of IBS and KI are summarized to compare the false-positive and false-negative rates of common familial searching strategies. The KI variations among populations are also evaluated to validate the SWGDAM recommendations. Finally, some recommendations or guidelines are made to facilitate familial searching.

Material and Methods

IBS-Based Method

Weir (20,21) gave the equations for the expected probabilities that two unrelated individuals share 0, 1, or 2 alleles at a locus in terms of allele frequencies and the population substructure parameter (θ). Herein, the probabilities are extended to any relationships given IBD distributions according to Balding and Nichols (22). Table 1 shows the ordered joint genotype probabilities with population substructure correction of two individuals given a relationship described by IBD distribution. Table 2 lists the IBD probabilities for pairwise kinships commonly used in familial searching. Using these two tables together, along with IBS values for all possible distinct genotype pairs listed in Table 1, the expected probability that two individuals share i alleles (i = 0, 1, or 2), given IBD = d, $P_i(\Phi_d)$, can be computed as the sum of all possible genotypic combinations for these two individuals sharing i alleles given IBD = d, as shown in equations set (1), where $P(A_iA_i,\ A_jA_k|\Phi_d)$ represents the joint probabilities of the ordered genotypes A_iA_i, A_jA_k for a given IBD coefficient φ_d. The details of expected probabilities of IBS given IBD distribution for a single locus are listed in Table 3.

$$
\begin{aligned}
P_0(\Phi_d) = &\sum_i \sum_{j(i\neq j)} P(A_iA_i, A_jA_j|\Phi_d) \\
&+ 1/2 \sum_i \sum_{j(i\neq j)} \sum_{k(i\neq j\neq k)} P(A_iA_i, A_jA_k|\Phi_d) \\
&+ 1/2 \sum_i \sum_{j(i\neq j)} \sum_{k(i\neq j\neq k)} P(A_jA_k, A_iA_i|\Phi_d) \\
&+ 1/4 \sum_i \sum_{j(i\neq j)} \sum_{k(i\neq j\neq k)} \sum_{l(i\neq j\neq k\neq l)} P(A_iA_j, A_kA_l|\Phi_d) \\
P_1(\Phi_d) = &\ 1/2 \sum_i \sum_{j(i\neq j)} P(A_iA_i, A_iA_j|\Phi_d) \\
&+ 1/2 \sum_i \sum_{j(i\neq j)} P(A_iA_j, A_iA_i|\Phi_d) \\
&+ \sum_i \sum_{j(i\neq j)} \sum_{k(i\neq j\neq k)} P(A_iA_j, A_iA_k|\Phi_d) \\
P_2(\Phi_d) = &\sum_i P(A_iA_i, A_iA_i|\Phi_d) \\
&+ 1/2 \sum_i \sum_{j(i\neq j)} P(A_iA_j, A_iA_j|\Phi_d)
\end{aligned}
\qquad (1)
$$

The expected IBS distribution of multiple loci profiles can be computed by a dynamic programming approach (23,24). Suppose, there are L loci in profile comparisons. Let P_{il} be the expected

TABLE 2—*IBD distributions for relevant pairwise kinships for familial searching.*

	IBD		
Pairwise Kinship	Φ_2	Φ_1	Φ_0
Unrelated	0	0	1
Parent–child	0	1	0
Full-sib	1/4	1/2	1/4

IBD, Identity-By-Descent.

TABLE 1—*Joint genotypic probabilities of two-ordered individuals (X,Y) given the IBD distribution, Pr(X,Y\ Φ_i), where Φ_0: IBD = 0, Φ_1: IBD = 1, Φ_2: IBD = 2, and $\Phi_0 + \Phi_1 + \Phi_2 = 1$. A_i, A_j, A_k, A_l are alleles at the locus. ϑ is the population substructure parameter. IBS is Identity-By-State, namely, the number of shared alleles.*

Ordered Genotypes (X, Y)	IBS	Joint Probabilities		
		Φ_2	Φ_1	Φ_0
A_iA_i, A_iA_i	2	$p_i^2 + \theta p_i(1-p_i)$	$\frac{p_i[\theta+p_i(1-\theta)][2\theta+p_i(1-\theta)]}{1+\theta}$	$\frac{p_i[\theta+p_i(1-\theta)][2\theta+p_i(1-\theta)][3\theta+p_i(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_iA_i, A_jA_j	0	0	0	$\frac{(1-\theta)p_ip_j[\theta+p_i(1-\theta)][\theta+p_j(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_iA_i, A_iA_j	1	0	$\frac{p_ip_j(1-\theta)[\theta+p_i(1-\theta)]}{1+\theta}$	$\frac{2(1-\theta)p_ip_j[\theta+p_i(1-\theta)][2\theta+p_i(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_iA_j, A_iA_i	1	0	$\frac{p_ip_j(1-\theta)[\theta+p_i(1-\theta)]}{1+\theta}$	$\frac{2(1-\theta)p_ip_j[\theta+p_i(1-\theta)][2\theta+p_i(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_iA_i, A_jA_k	0	0	0	$\frac{2(1-\theta)^2p_ip_jp_k[\theta+p_i(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_jA_k, A_iA_i	0	0	0	$\frac{2(1-\theta)^2p_ip_jp_k[\theta+p_i(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_iA_j, A_iA_j	2	$2p_ip_j(1-\theta)$	$\frac{p_ip_j(1-\theta)[(p_i+p_j)(1-\theta)+2\theta]}{1+\theta}$	$\frac{4(1-\theta)p_ip_j[\theta+p_i(1-\theta)][\theta+p_j(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_iA_j, A_iA_k	1	0	$\frac{p_ip_jp_k(1-\theta)^2}{1+\theta}$	$\frac{4(1-\theta)^2p_ip_jp_k[\theta+p_i(1-\theta)]}{(1+\theta)(1+2\theta)}$
A_iA_j, A_kA_l	0	0	0	$\frac{4(1-\theta)^3p_ip_jp_kp_l}{(1+\theta)(1+2\theta)}$

IBD, Identity-By-Descent.

TABLE 3—*Expected probability of two individuals sharing i number of alleles (i.e., IBS) given IBD = j,* $P_i(\Phi_j)$*, where* i *is IBS (*i *= 0, 1, 2) and j is IBD (*j *= 0, 1, 2).* α_r *is the sum of the r-th power of allele frequencies at the locus, namely,* $a_r = \sum_{i=1}^{k} p_i^r$*, where* p_i *is the allele frequency, k is the number of alleles at this locus.* ϑ *is the population substructure parameter.*

	IBD		
IBS	Φ_2	Φ_1	Φ_0
0	0	0	$\frac{\theta^2(1-\theta)(1-a_2)+2\theta(1-\theta)^2(1-2a_2+a_3)+(1-\theta)^3(1-4a_2+4a_3+2a_2^2-3a_4)}{(1+\theta)(1+2\theta)}$
1	0	$(1-\theta)(1-a_2)$	$\frac{8\theta^2(1-\theta)(1-a_2)+4\theta(1-\theta)^2(1-a_3)+4(1-\theta)^3(a_2-a_3-a_2^2+a_4)}{(1+\theta)(1+2\theta)}$
2	1	$\theta+(1-\theta)a_2$	$\frac{6\theta^3+\theta^2(1-\theta)(2+9a_2)+2\theta(1-\theta)^2(2a_2+a_3)+(1-\theta)^3(2a_2^2-a_4)}{(1+\theta)(1+2\theta)}$

IBD, Identity-By-Descent; IBS, Identity-By-State.

probability to have i shared alleles ($i = 0, 1, 2$) at the l-th locus ($l = 0, 1, \ldots, L$), which is actually the weighted sum of each row for a given relationship in Table 3 as Eq. 2.

$$P_{il} = \sum_{j=0,1,2} \Phi_j P_i(\Phi_j) \tag{2}$$

Based on Table 3, the expected probability of having x shared alleles on first l loci, *IBSP(x,l),* can be computed by the dynamic programming approach as follows (Eq. [3]).

$$IBSP(x,l) = IBSP(x,l-1) \times P_{0l} + IBSP(x-1,l-1) \times P_{1l} + IBSP(x-2,l-1) \times P_{2l} \tag{3}$$

For the strategy that requires at least one allele shared at each locus, instead of counting number of shared alleles, the number of loci with at least one shared allele are counted, and the expected probability of having x loci with at least one shared allele on first l loci, *OSP(x,l)*, is similar to Eq. 3 but merging P_{01} and P_{11} (Eq. [4]).

$$OSP(x,l) = OSP(x,l-1) \times P_{0l} + OSP(x-1,l-1) \times (P_{1l} + P_{2l}) \tag{4}$$

For the strategy of only considering a moderate stringency match, the expected probabilities of a moderate stringency match given IBD distribution, $MP_t(\varphi_d)$, can be calculated in the same way as in Table 1 but only (A_iA_i, A_iA_i), (A_iA_i, A_iA_j), and (A_iA_j, A_iA_j) match pairs are included (Table 4). The expected probabilities of having x moderate stringency-matched loci on first l loci, *MSP(x, l)*, are as Eq. 5

$$MSP(x,l) = MSP(x,l-1) \times MP_{0l} + MSP(x-1,l-1) \times MP_{1l} \tag{5}$$

where $MP_{il} = \sum_{j=0,1,2} \Phi_j MP_i(\Phi_j)$, calculated for the l-th locus.

Likelihood Ratio-Based Method

The likelihood ratio-based method basically calculates the pairwise kinship ratio or KI for the forensic profile (X) and candidate profile (Y) (Eq. [6]). For a father–child relationship, the KI is the Paternity Index.

$$KI = \frac{\sum_{i=0,1,2} \Pr(X,Y|\Phi_i)\Pr(\Phi_i|\text{Relationship})}{\sum_{i=0,1,2} \Pr(X,Y|\Phi_i)\Pr(\Phi_i|\text{Unrelated})} \tag{6}$$

The SWGDAM Ad Hoc Committee on Partial Matches (10) recommended a measure, EKR, which is the KI divided by the sample size (N) of the searched database (i.e., EKR = KI/N). The recommendation also requires that the maximum and minimum EKRs among four U.S. major populations should be greater than 1 and 0.1, respectively.

In this study, the software MPKin (19) is used to calculate the KI. Population substructure and mutation are incorporated as options. Allowing mutations reduces the false exclusion of parent–child relationship when mutational events may occur. A realistic mutation model for STRs, the two-phase model, was implemented in MPKin (19). The mutation rates from AABB are used in the calculations (25). Incorporating population substructure could change the kinship ratio by several orders of magnitudes in some cases (19) and generally provides more conservative results.

Simulation

Millions of unrelated, parent–child, and full-sib DNA profiles were simulated using Caucasian population data on the 13 CODIS STR loci (26). To generate simulated data, the genotypes of founders (i.e., individuals without parents in the pedigree) were randomly assigned according to the conditional genotype frequencies given observed alleles of the founders of each locus and each locus was treated independently. The conditional genotype frequencies can be calculated by the theory described in Balding and Nichols (22). Founders transmitted with equal probability a single allele at each locus to his/her offspring. Mutations were allowed according to the two-phase Model during the transmissions. The paternal and maternal mutation rates were different (25). This study mainly used allele frequency data from Caucasian population to illustrate different distributions. Allele frequencies for other populations (i.e., Caucasian, African American, Southwestern Hispanic, Southeastern Hispanic, Navajo, and Asian) were also used when comparing the EKRs among populations.

Results

IBS-Based Measures

The expected distributions of the number of shared alleles, loci with at least one allele shared, and loci with a moderate

TABLE 4—*Expected probability of moderate stringency match given IBD distribution,* $MP_t(\varphi_i)$*. t = 1 means the compared genotypes are moderate stringency matched, otherwise t = 0.*

Moderate Stringency Match (t)	IBD		
	Φ_2	Φ_1	Φ_0
1	1	$\frac{\theta(3-\theta)+3(1-\theta)^2\alpha_2-2(1-\theta)^2\alpha_3}{(1+\theta)}$	$\frac{2\theta^2(5-2\theta)+\theta(1-\theta)(16-15\theta)\alpha_2+2(1-\theta)^2(2-7\theta)\alpha_3+2(1-\theta)^3\alpha_2^2-5(1-\theta)^3\alpha_4}{(1+\theta)(1+2\theta)}$
0	0	1- $MP_1(\Phi_1)$	1- $MP_1(\Phi_0)$

IBD, Identity-By-Descent.

stringency match for unrelated, parent–child and full-sib pairs were calculated according to (Eqs [1–5]) based on 13 CODIS STR loci Caucasian data (26), assuming no population substructure and mutations.

For the strategy considering number of shared alleles (Fig. 1), the threshold with the least false rate (i.e., both false negative and false positive) was 13 shared alleles. The California strategy (i.e., at least 15 shared alleles) accepted only 0.45% unrelated pairs as potential-related pairs for further investigations. However, this strategy falsely excluded 17.8% true full-sib and 18.4% true parent–child relationships. The shared allele approach favored false exclusion over false inclusion. A lower threshold increased the false-positive rate. For example, 1.5% unrelated profiles were falsely identified as potential relatives of the candidate profile for the threshold of ≥ 14 shared alleles, which may not be acceptable (or practical) for large database searches. A higher threshold could be an option. Only 0.11% unrelated profiles were falsely included for a threshold of ≥ 16 shared alleles, but 31.1% true full-sib and 43.3% true parent–child relationships were excluded. The threshold chosen likely will favor generation of a manageable number of candidates thus favoring fewer false positives and greater false negatives. No false-positive hits are expected to be observed for a database with 1 million samples, if at least 20 shared alleles are required for a hit with 13 CODIS loci; although the consequence is that the majority of true relatives would be excluded.

If at least one allele shared at each locus was required as the searching strategy, only 0.077% unrelated profiles were falsely included (Fig. 2), which was similar to the strategy with at least 15 shared alleles, but 76.0% true full-sib pairs were falsely excluded. Almost all true parent–child relationships were included, except a small proportion (i.e., about 1%) with mutations. This strategy obviously is better suited to search for parent–child relationships, compared with the strategy with at least 15 shared alleles.

If familial searching only considers the profiles with all loci matched with moderate stringency, the chance to include unrelated profiles as relatives was extremely low (i.e., 2.69×10^{-9}). Unfortunately, more than 99.9% true relatives were also excluded. Even if multiple nonmoderate-stringency-matched loci were allowed in the search, the false-positive rates were not comparable with the strategy of simply counting the number of shared alleles. Roughly 98.3% of parent–child and 96.0% of full-sib pairs were excluded if up to two loci were allowed to not meet the moderate (and/or high) stringency criterion (Fig. 3). The findings herein support the SWGDAM recommendation (10) that the concept of a moderate stringency match has "little useful probative value" in familial searching.

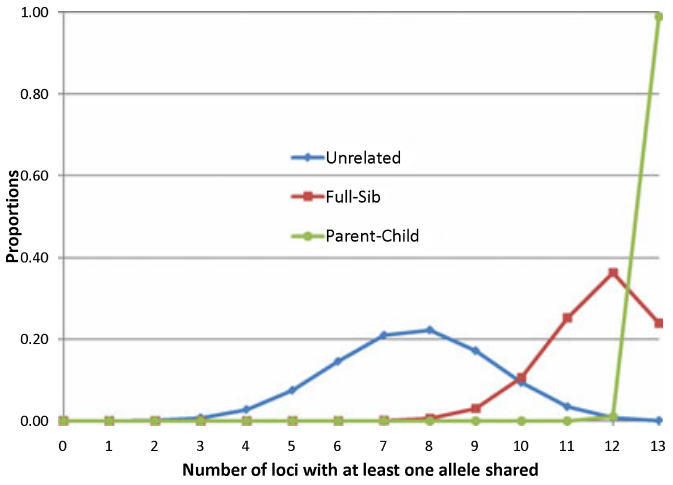

FIG. 2—*Distribution of loci with at least one allele shared for unrelated and full-sib pairs (13 CODIS Short Tandem Repeat; Caucasian population; $\theta = 0$; no mutation), and parent–child pairs (simulated using Caucasian population data with mutation rates from AABB [24]).*

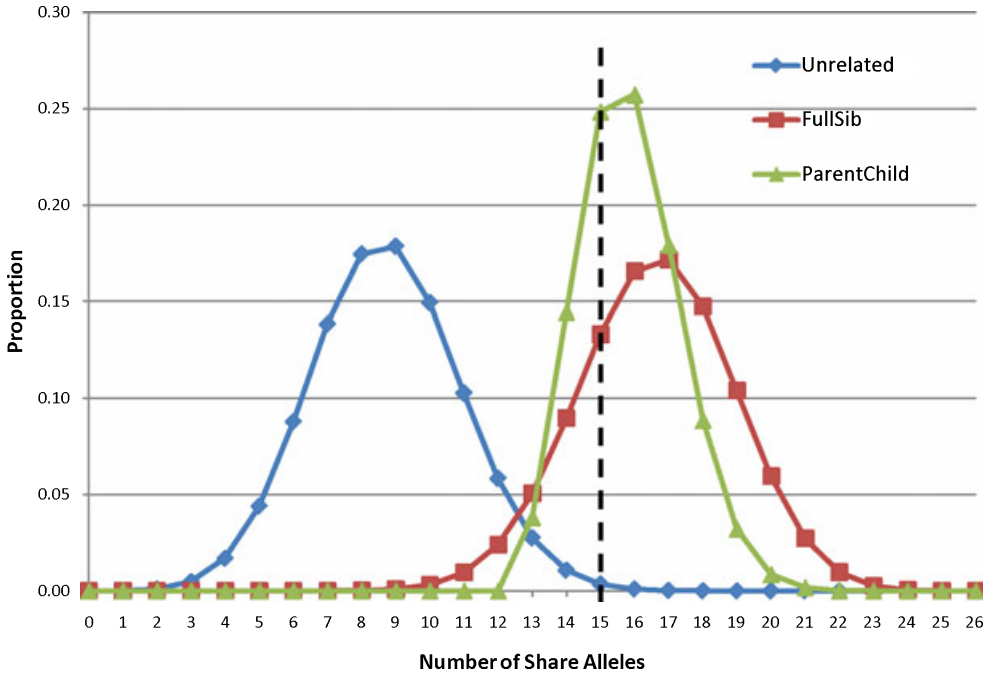

FIG. 1—*Distributions of number of shared alleles for unrelated, parent–child and full-sib pairs (13 CODIS Short Tandem Repeat; Caucasian population; $\theta = 0$; no mutation).*

Likelihood Ratio-Based Measures

Unrelated (2 million), parent–child (1 million), and full-sib (1 million) pairs were simulated using 13 CODIS STRs Caucasian population data (26). The KIs were calculated for the true parent–child and full-sib pairs with mutation rates from AABB (25), and the unrelated pairs were also calculated as parent–child or full-sib relationships (1 million each) (Fig. 4). Only one true parent–child pair of a million had KI <1 (Table 5). Full-sib pairs generally are expected to have lower KIs than those of parent–child, but there is still a substantial proportion of high value KIs (e.g., 57.5% KIs were >1000). For the unrelated pairs which were identified as parent–child, only 0.013% pairs had KIs >1000 (i.e., $Log_{10}(1000) = 3$) (Table 5*c*). The maximum KI observed was 1.57×10^5. For the unrelated pairs identified as full-sib, about 0.009% pairs had KIs >1000 (Table 5*d*), and the maximum KI was 1.51×10^5.

The SWGDAM suggested EKR, which adjusts for database size searched, was a very stringent measure for large databases. For a

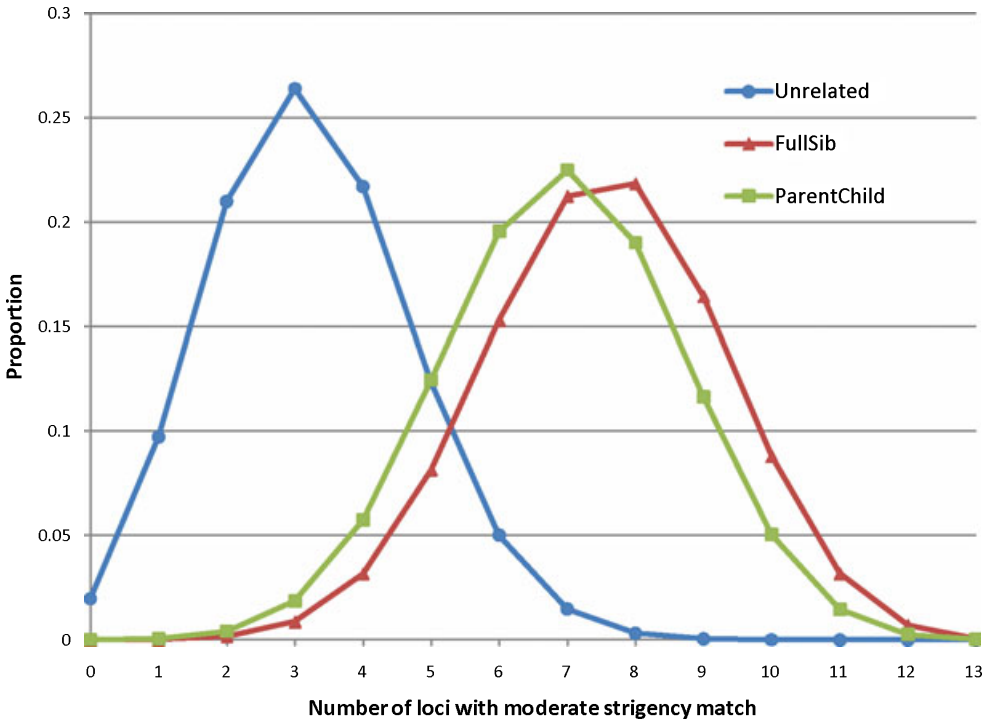

FIG. 3—*Distribution of loci with moderate stringency matches for unrelated, parent–child and full-sib pairs (13 CODIS Short Tandem Repeat; Caucasian population; θ = 0; no mutation).*

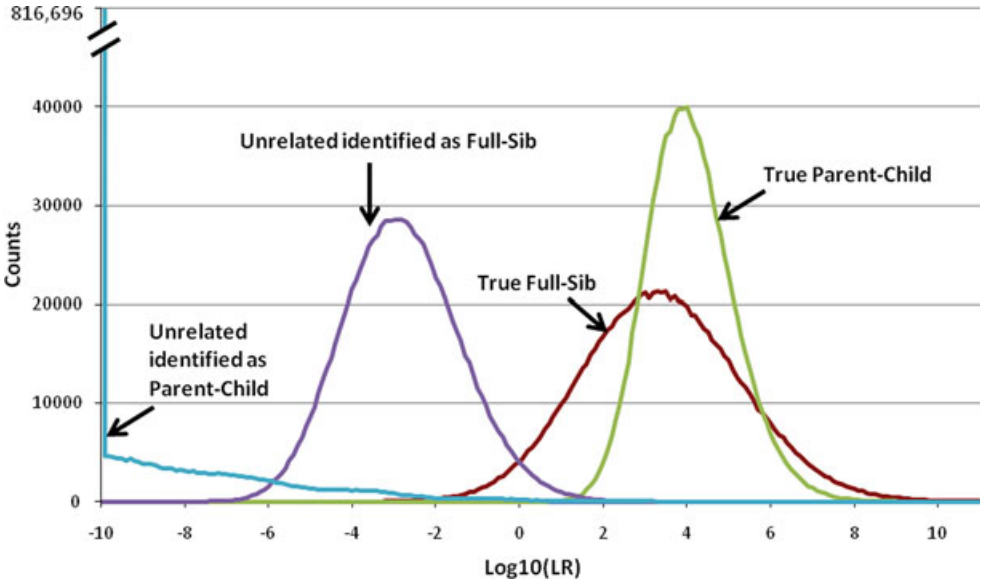

FIG. 4—*Distributions of $Log_{10}(KI)$ for simulated unrelated, parent–child and full-sib pairs (13 CODIS Short Tandem Repeat; Caucasian population; θ = 0; no mutation in simulation). Unrelated pairs were identified as full-sib or parent–child relationships.*

TABLE 5—*Joint distributions of IBS and KI for (a) true parent–child identified as parent–child, (b) true full-sib identified as full-sib, (c) unrelated identified as parent–child, and (d) unrelated identified as full-sib. One million pairs were simulated with Caucasian population data for each relationship. Mutation or population substructure was not included in simulations, and only mutation was considered in KI calculations to avoid zeros when identifying unrelated as parent–child.*

	IBS												
Log_{10}(KI)	13	14	15	16	17	18	19	20	21	22	23	24	Sum
(a)													
<=0	1	0	0	0	0	0	0	0	0	0	0	0	1
(0,1)	84	53	19	3	0	0	0	0	0	0	0	0	159
(1,2)	2427	4688	3379	1211	190	16	0	0	0	0	0	0	11,911
(2,3)	11,084	34,535	42,201	28,115	10,851	2489	325	26	2	0	0	0	129,628
(3,4)	14,299	56,957	98,298	95,792	57,893	22,620	5745	940	87	1	0	0	352,632
(4,5)	7454	35,034	71,884	87,138	68,138	36,750	13,916	3481	592	67	1	0	324,455
(5,6)	2321	11,706	25,774	35,095	31,504	19,711	8835	2948	677	103	8	1	138,683
(6,7)	473	2533	5973	8541	8238	5530	2683	996	286	51	3	1	35,308
(7,8)	91	433	940	1469	1482	1037	550	209	62	16	2	1	6292
(8,9)	15	40	136	187	182	147	75	36	6	3	1	1	829
(9,10)	1	8	12	26	12	24	3	3	2	1	0	0	92
(10,11)	0	0	2	2	3	2	1	0	0	0	0	0	10
Sum	38,250	145,987	248,618	257,579	178,493	88,326	32,133	8639	1714	242	15	4	1,000,000

	IBS															
Log_{10}(KI)	≤11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	Sum
(b)																	
<=0	11,558	11,115	7490	1526	57	0	0	0	0	0	0	0	0	0	0	0	31,746
(0,1)	1885	9399	23,709	24,279	7995	656	3	0	0	0	0	0	0	0	0	0	67,926
(1,2)	361	2865	14,599	39,688	49,134	23,247	3350	100	0	0	0	0	0	0	0	0	133,344
(2,3)	47	545	4058	18,702	49,809	69,105	40,378	8427	487	4	0	0	0	0	0	0	191,562
(3,4)	8	66	776	4710	20,001	50,378	70,594	47,232	12,571	1070	20	0	0	0	0	0	207,426
(4,5)	0	7	102	866	4708	17,665	40,506	54,814	38,660	11,813	1305	42	1	0	0	0	170,489
(5,6)	0	2	4	112	878	4028	13,064	26,673	33,698	22,532	7465	964	28	0	0	0	109,448
(6,7)	0	0	1	10	112	714	2881	7931	14,583	15,876	10,018	3190	390	15	0	0	55,721
(7,8)	0	0	0	2	19	93	507	1696	3954	5910	5791	3235	948	117	1	0	22,273
(8,9)	0	0	0	0	2	6	60	243	828	1549	2126	1663	751	174	11	0	7413
(9,10)	0	0	0	0	0	2	3	41	144	301	541	545	372	107	15	1	2072
(10,11)	0	0	0	0	0	0	0	8	13	46	101	126	117	44	10	1	466
(11,12)	0	0	0	0	0	0	0	0	4	7	11	30	25	11	6	0	94
(12,13)	0	0	0	0	0	0	0	0	0	0	0	6	5	6	1	0	18
(13,14)	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2
Sum	13,859	23,999	50,739	89,895	132,715	165,894	171,346	147,165	104,942	59,108	27,379	9801	2637	474	45	2	1,000,000

	IBS									
Log_{10}(KI)	≤11	12	13	14	15	16	17	18	19	Sum
(c)										
<=0	898,638	58,480	27,185	10,041	2833	633	100	16	0	997,926
(0,1)	0	93	242	387	289	116	39	9	0	1175
(1,2)	0	11	91	154	113	45	12	5	1	432
(2,3)	0	0	42	88	117	59	23	6	0	335
(3,4)	0	0	4	26	32	33	21	2	0	118
(4,5)	0	0	0	4	3	2	1	2	1	13
(5,6)	0	0	0	0	0	0	0	0	1	1
Sum	898,638	58,584	27,564	10,700	3387	888	196	40	3	1,000,000

	IBS									
Log_{10}(KI)	≤11	12	13	14	15	16	17	18	19	Sum
(d)										
<=0	897,655	54,299	18,383	2512	65	0	0	0	0	972,914
(0,1)	923	4027	8601	6652	1593	101	0	0	0	21,897
(1,2)	18	119	620	1502	1580	518	62	0	0	4419
(2,3)	1	1	26	83	206	228	120	11	1	677
(3,4)	0	0	0	4	14	29	19	19	1	86
(4,5)	0	0	0	0	0	1	1	1	1	4
(5,6)	0	0	0	0	0	0	1	1	1	3
Sum	898,597	58,446	27,630	10,753	3458	877	203	32	4	1,000,000

IBS, Identity-By-State.

database containing a million samples, almost no unrelated relatives were included, but about 91.2% true full-sib and 95.7% true parent–child were excluded (Table 5). An absolute KI = 1000 or 10,000 might be a good measure for the 13 CODIS loci to balance the false-positive and false-negative rates. With KI ≥ 1000, on average, less than about 0.013% unrelated pairs were included, and more than half of relatives were not excluded.

Joint Measures

Usually, either the IBS or the likelihood ratio measure is used as the familial searching strategy. However, these measures can be jointly considered to improve the accuracy of searching. Both the number of shared alleles (i.e., IBS) and KI were recorded in the aforementioned simulations (Table 5). If at least 15 shared alleles was the searching threshold, around 4600 unrelated pairs were found as possible hits in 1 million unrelated pairs. With a KI as a further screening measure, say KI ≥ 1000, 97.8% or 98.1% of those IBS searched possible hits were excluded as parent–child or full-sib relationships, respectively. At the same time, only 30.8% true full-sib and 10.9% true parent–child relationships in the possible candidate list were further excluded by the KI ≥ 1000 threshold. Thus, a strategy with both IBS and KI combined can exclude a substantial proportion of the unrelated profiles with only a relatively small proportion of relatives, eventually improving the efficiency of familial searching.

Table 6 summarizes false-positive and false-negative rates of some reasonable familial searching thresholds based on Table 5. An IBS ≥ 14 was the least stringent threshold, which included most true relatives but also included a great number of unrelated profiles for a large database search. Fortunately, adding a KI threshold excluded most unrelated profiles from a possible candidate list. A KI threshold alone also excluded the majority of unrelated profiles. For example, a KI ≥ 10 or KI ≥ 100 excluded 99.79% or 99.91% unrelated profiles. An IBS ≥ 15 or IBS ≥ 16 with a KI ≥ 1000 or KI ≥ 10,000 combined are practical searching strategies with good balance between false-positive and false-negative rates. An IBS ≥ 17 and/or KI ≥ 1,000,000 can exclude the majority of profiles in the database, either related or not, and initially may be good start options to produce a small, but manageable possible candidate list.

TABLE 6—*False-positive and false-negative rates for some common strategies. For unrelated identified as related, the rates may not be accurate because only a small number of simulated pairs (based on 1 million simulations) qualified with the high thresholds of the strategies. Higher accuracy can be achieved with a greater number of simulations, but 1 million simulations were sufficient to compare the strategies.*

	False Positive	False Negative	
Strategy	Unrelated (%)	Parent–Child (%)	Full-sib (%)
(a) IBS-based strategies			
IBS ≥ 14	1.5	3.8	8.9
IBS ≥ 15	0.45	18.4	17.8
IBS ≥ 16	0.11	43.3	31.1
IBS ≥ 17	0.024	69.0	47.7

	Unrelated Identified As		True	True
	Parent–Child (%)	Full-sib (%)	Parent–Child (%)	Full-sib (%)
(b) Likelihood-based strategies				
KI ≥ 1000	0.0132	0.0093	14.1	42.5
KI ≥ 10,000	0.0014	0.0007	49.4	63.2
KI ≥ 100,000	0.0001	0.0003	81.9	80.2
KI ≥ 1,000,000	<0.0001	<0.0001	95.7	91.2
(c) Polices with both IBS and likelihood combined				
IBS ≥ 14; KI ≥ 100	0.0421	0.0742	4.8	23.8
IBS ≥ 14; KI ≥ 1000	0.0128	0.0093	16.6	42.6
IBS ≥ 14; KI ≥ 10,000	0.0014	0.0007	50.5	63.2
IBS ≥ 15; KI ≥ 1000	0.0098	0.0089	27.3	43.1
IBS ≥ 15; KI ≥ 10,000	0.0010	0.0007	21.8	63.3
IBS ≥ 16; KI ≥ 1000	0.0063	0.0075	55.4	45.7
IBS ≥ 16; KI ≥ 10,000	0.0007	0.0007	65.9	63.9
IBS ≥ 16; KI ≥ 100,000	0.0001	0.0003	86.9	80.4

IBS, Identity-By-State.

EKR Variation Among Populations

SWGDAM (10) recommended the maximum and minimum EKRs among Caucasians, African Americans, Southwestern Hispanics, and Southeastern Hispanics should be >1 and 0.1, respectively. To investigate the variations of the EKR among populations, 1 million unrelated, parent–child, and full-sib pairs each using Caucasian population data were simulated, and Log_{10}(Min. EKR/Max. EKR) for the four populations or six populations (i.e., the above four populations, Navajo, and Asian) (Fig. 5) were calculated. Interestingly, the distributions of the Log_{10}(Min. EKR/Max. EKR) for true parent–child and true full-sib with given populations were very close to each other. More than 96.2% or 58.5% true relationship pairs had Min.EKR/Max.EKR <0.1 for the four or six populations, respectively. About 0.2% true relationships had the ratios even <0.001 for the four populations. The distributions were similar for unrelated pairs if they were identified as parent–child or full-sib relationships, although full-sibs had relatively lower variations than those of parent–child. The long tails of the distributions were because of the allele frequency differences between populations, especially of rare alleles which have not been observed in all population(s) (e.g., Caucasian in these simulations). Rare allele matches between profiles are usually flags for high likelihood of close relationships. The SWGDAM cutoff threshold may exclude a large proportion of possible relatives with rare alleles.

Discussion

According to the Bureau of Justice Statistics "Correctional populations in United States, 1996" report (27), at least 42.8% of jail inmates had close relatives (i.e., father, mother, brother, sister, child; spouse was not counted in the 42.8% because couples are usually not related) who were incarcerated. As DNA profiles of most inmates are entered in the CODIS system, familial searching has a great potential to assist law enforcement by identifying individuals in CODIS who may be close relatives of the true source of forensic samples. In this study, different familial searching strategies that are currently adopted by the authorities were investigated. The false-negative and false-positive rates of IBS and KI measures were compared and it was concluded that combining both IBS and KI may be a better approach than IBS or KI alone. The strategy that requires at least one shared allele at each locus has a similar false-positive rate as the IBS ≥ 15 strategy, and the majority of parent–child will not be excluded. However, according to the Bureau of Justice Statistics (27), about 36.5% inmates had full-sibs who had been incarcerated and less inmates (i.e., 22.8%) had parents or children incarcerated, so that a 76.0% exclusion rate of the strategy for true full-sibs may not make effective use of a database search functionality.

There was a case in the U.K. involving a rare allele to support a familial search association to identify a suspect (28). A girl was murdered in 1988 and 15 years later a search of the U.K. National DNA database found a single rare allele match between the profiles

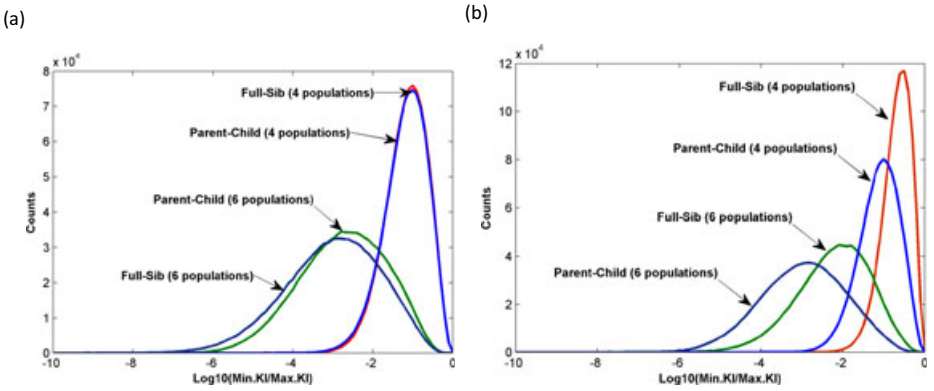

FIG. 5—*Distributions of Minimum Expected Kinship Ratio (EKR)/Maximum EKR among four or six populations for (a) true relationships (i.e., full-sib or parent–child) and (b) false relationships (i.e., unrelated). The four populations are Caucasian, African American, Southwestern Hispanic, and Southeastern Hispanic. The six populations include the four populations in (a), Navajo, and Asians. The pedigree data were simulated based on Caucasian population data.*

of the evidence and a young boy who was born after 1988. This rare allele match led to the investigation of the relatives of the boy, and the uncle of the boy was linked to the murder of this cold case. While it may be difficult to define how rare a "rare" allele should be, generally, the weight of rare alleles has been taken into account in the KI calculation. Because of the limitation of the genotyping technologies, some rare alleles may simply be assigned as greater or less than the limits of the allelic ladder (e.g., >20 or <5). Hence, some power of rare allele may not be fully exploited for familial searching.

A likelihood calculation with population substructure and mutations usually render more conservative KIs (18). However, incorporating mutation only slightly changes the KI of the full-sib and parent–child without mutation (i.e., <5% using the mutation rates from AABB [25]). It is desirable, though, to avoid excluding true parent–child relationships with mutations. Population substructure has relatively higher effects than mutation, but the differences are within 10-fold in 99% of the parent–child and full-sib cases (Fig. 6).

The EMR (Eq. [7]) recommended by SWGDAM (10) was originally proposed to validate the target profile (e.g., too many [apparently] homozygous loci in a profile could be attributed to multiple allele drop out). It is not directly applicable to familial searching, as the EMR is a measure for an expected familial match (given a target or evidentiary profile), which depends only on the forensic profile itself, as shown in Eq. 7, where X is the forensic profile and Y is all moderate stringency-matched profiles. For example, if X is {12,12}, then Y could be {12,12}, {12,13}, {12,15}, …, etc.

$$\mathrm{EMR} = \frac{\sum_{Y}\sum_{i=0,1,2} \Pr(Y|X,\Phi_i)\Pr(\Phi_i|R)}{\sum_{Y}\sum_{i=0,1,2} \Pr(Y|X,\Phi_i)\Pr(\Phi_i|\mathrm{Unrelated}) * N} \qquad (7)$$

Many new database profiles may contain more than 13 STRs. Therefore, the same simulations as aforementioned were performed with 15 STRs (i.e., 13 CODIS STRs, along with D2S1338, and D19S433). As expected, distributions of the unrelated and the related relationships became more resolved compared to those with 13 STRs (Table 7). With more loci, the false-negative and false-positive rates were reduced, providing a higher accuracy for familial searching.

When profiles in the databases contain lineage-based markers, such as Y chromosome STRs and mitochondrial DNA (mtDNA) sequences, more adventitious profile hits can be eliminated. Autosomal STRs biologically assort independently of mitochondrial sequences and Y chromosomal haplotypes. Budowle et al. (29) and Walsh et al. (30) did not detect statistically significant departures from independence between the Y-STR haplotypes, mtDNA haplotypes, and autosomal STR loci used in DNA forensics. As most convicted felons are males, Y-STRs may be a more practical tool than mtDNA for resolving adventitious hits.

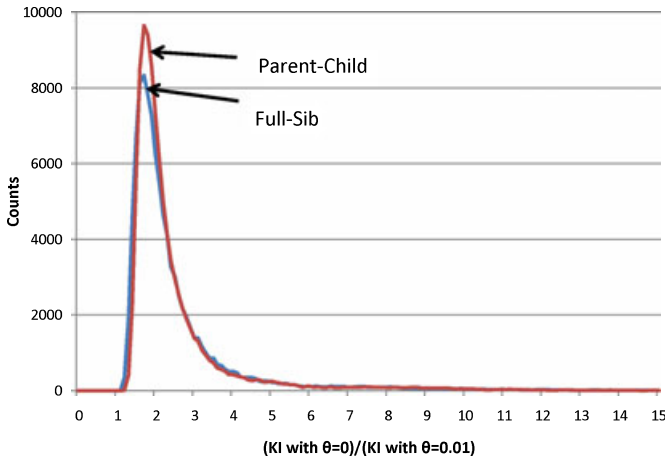

FIG. 6—*Distributions of ratios of KI with or without population substructure correction (i.e., θ = 0 or 0.01). The ratios were calculated from 100,000 simulated parent–child and full-sib pairs with 13 Caucasian CODIS Short Tandem Repeat population data.*

TABLE 7—*Comparison of the mean values of the IBS and Log$_{10}$(KI) distributions for simulated unrelated (UN), parent–child (PC) and full-sib (FS) pairs (15 STRS; Caucasian population data; ϑ = 0; no mutation in simulation). Unrelated pairs were identified as full-sib or parent–child relationships. The 15 STRs include 13 CODIS STRs, D2S1338, and D19S433. The allele frequencies of D2S1338 and D19S433 were kindly provided by New York State Police.*

	Means with 13 STRs		Means with 15 STRs	
Identifications	IBS	Log$_{10}$(KI)	IBS	Log$_{10}$(KI)
FS → FS	16.5996	3.4012	19.064	3.9999
PC → PC	15.8409	4.0833	18.2018	4.8119
UN → FS	8.7088	−2.8043	9.8483	−3.3029
UN → PC	8.7122	−15.5829	9.8431	−19.0476

IBS, Identity-By-State; STR, Short Tandem Repeat.

There are two approaches to consider when employing Y-STRs for familial searching. The first is to identify candidates by autosomal STR profile searches and then resolve adventitious hits using Y-STRs. Currently, this approach is the only tenable one because most profiles in the CODIS database do not contain Y-STR data, and Y-STR typing would have to be performed on samples from the candidate list. The second approach is initially to screen the database by Y-STR profiles, so that true male lineage relatives will not be excluded because of relatively low IBS or KI values. This approach would require that the database samples be typed for Y-STRs for entry into the CODIS database.

The position of a true relative in a familial searching candidate list should be considered as the resources for further investigations could be limited. The position of a true relative (if in the database) mainly depends on the database size and the specific alleles in the profiles. In a considerable proportion of cases, a true relative may be at the bottom of the list, or even not on the list. However, the expected position can be estimated by the expected measures of a relationship (Figs 1 and 4, and Table 5). For example, a full-sib pair is expected to have about 16.6 shared alleles and a KI of roughly 2500 for 13 CODIS loci. For an evidence profile searched against a database containing 1 million unrelated samples, around 70 unrelated samples are expected to have higher positions in the candidate list than the true full-sib.

In summary, this study compared the familial searching strategies used or proposed in the U.S. and summarized the false-positive and false-negative rates of the thresholds of the strategies based on 13 CODIS STRs. IBS $\geq$ 15, IBS $\geq$ 16, KI $\geq$ 1000, or KI $\geq$ 10,000 may be good thresholds for balancing false-positive and false-negative rates. IBS $\geq$ 17 and/or KI $\geq$ 1,000,000 can exclude the majority of candidate profiles in the database, either related or not, and may be an initial screening option if one criterion is to generate a small or manageable candidate hit list. Polices combining both IBS and KI can provide higher accuracy. The SWGDAM suggested EKR and the threshold of the EKR variations among the populations may be too stringent for large databases. Additional STRs, beyond the required CODIS 13, can provide better searching performance, and lineage markers can be extremely useful for reducing false rates.

A familial searching software was developed, which allows all strategies described earlier. Visit http://sites.google.com/site/gejianye/ for more details of the software.

Acknowledgment

All authors would like thank Dr. George Carmody for his valuable comments. This paper is dedicated to the memory of Dr. George Carmody (1939–2011).

References

1. Evett IW. Evaluating DNA profiles in a case where the defence is ''It was my brother.'' J Forensic Sci Soc 1992;32:5–14.
2. Sjerps M, Kloosterman AD. On the consequences of DNA mismatches for close relatives of an excluded suspect. Int J Legal Med 1999; 112:176–80.
3. Belin TR, Gjertson DW, Hu M. Summarizing DNA evidence when relatives are possible suspects. J Am Stat Assoc 1997;92(438):706–16.
4. Brookfield JF. The effect of relatives on the likelihood ratio associated with DNA profile evidence in criminal cases. J Forensic Sci Soc 1994; 34(3):193–7.
5. Bieber FR, Brenner CH, Lazer D. Finding criminals through DNA of their relatives. Science 2006;312:1315–6.
6. Paoletti DR, Doom TE, Raymer ML, Krane DE. Assessing the implications for close relatives in the event of similar but nonmatching DNA profiles. Jurimetrics J 2006;42:161–75.
7. Reid TM, Baird ML, Reid JP, Lee SC, Lee RF. Use of sibling pairs to determine the familial searching efficiency of forensic databases. Forensic Sci Int Genet 2008;2:340–2.
8. Curran JM, Buckleton J. Effectiveness of familial searches. Sci Justice 2008;48:164–7.
9. Cowen S, Thomson J. A likelihood ratio approach to familial searching of large DNA databases. Forensic Sci Int Genet 2008;1:643–5.
10. Scientific Working Group on DNA Analysis Methods Ad Hoc Committee on Partial Matches. SWGDAM Recommendations to the FBI Director on the ''Interim Plan for the Release of Information in the Event of a 'Partial Match' at NDIS.'' Forensic Sci Comm 2009;11(4).
11. Hicks T, Taroni F, Curran JM, Buckleton J, Ribaux O, Castella V. Use of DNA profiles for investigation using a simulated Swiss national DNA database: part I, partial SGM Plus1 profiles. Forensic Sci Int Genet 2010;4(4):232–8.
12. Hicks T, Taroni F, Curran JM, Buckleton J, Castella V, Ribaux O. Use of DNA profiles for investigation using a simulated Swiss national DNA database: part II, statistical and ethical considerations on familial searching. Forensic Sci Int Genet 2010;4(5):316–22.
13. Chakraborty R, Jin L. Determination of relatedness between individuals by DNA fingerprinting. Hum Biol 1993;65:875–95.
14. California Department of Justice, Division of Law Enforcement. DNA Partial Match (Crime Scene DNA Profile to Offender) Policy, 2008, http://ag.ca.gov/cms_attachments/press/pdfs/n1548_08-bfs-01.pdf (accessed on August 23, 2010).
15. http://www.scienceprogress.org/2009/11/map-state-dna-policies/ (accessed on August 23, 2010).
16. Li CC, Sacks L. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics 1954;10: 347–60.
17. Thompson EA. The estimation of pairwise relationships. Ann Hum Genet 1975;39:173–88.
18. Ge J, Budowle B, Chakraborty R. Choosing relatives for DNA identification of missing persons. J Forensic Sci 2011;56(s1):S23–8.
19. Ge J, Budowle B, Chakraborty R. DNA identification by pedigree likelihood ratio accommodating population substructure and mutations. Investig Genet 2010;1:8.
20. Weir BS. Matching and partially-matching DNA profiles. Ann Appl Stat 2007;1(2):358–70.
21. Weir BS. The rarity of DNA profiles. J Forensic Sci 2004;49:1009–14.
22. Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 1994;64:125–40.
23. Chakraborty R, Schull WJ. A note on the distribution of the number of exclusions to be expected in paternity testing. Am J Hum Genet 1976;28:615–8.
24. Chakraborty R. The distribution of the number of heterozygous loci in an individual in natural populations. Genetics 1981;98:461–6.
25. AABB annual report 2008, http://www.aabb.org/sa/facilities/Documents/rtannrpt08.pdf (accessed on August 23, 2010).
26. Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001;46:453–89.
27. Bureau of Justice Statistics. Correctional populations in United States, 1996, http:// bjs.ojp.usdoj.gov/content/pub/pdf/cpius964.pdf (accessed on August 23, 2010).
28. BBC News. ''How police found Gafoor,'' 4 July 2003, http://news.bbc.co.uk/1/hi/wales/3038138.stm (accessed on August 23, 2010).
29. Budowle B, Ge J, Aranda X, Planz J, Eisenberg A, Chakraborty R. Texas population substructure and its impact on estimating the rarity of Y STR haplotypes from DNA evidence. J Forensic Sci 2009;54(5): 1016–21.
30. Walsh B, Redd A, Hammer M. Joint match probabilities for Y chromosomal and autosomal markers. Forensic Sci Int 2008;174(2):234–8.

Additional information and reprint requests:
Jianye Ge, Ph.D.
Institute of Investigative Genetics
University of North Texas, Health Science Center
3500 Camp Bowie Blvd.
Fort Worth, TX 76107
E-mail: Jianye.Ge@unthsc.edu