

Comparing DNA based familial searching policies

Jianye Ge^{1,2,*}, Bruce Budowle^{1,2}, Arthur Eisenberg^{1,2}, Ranajit Chakraborty^{1,2}

¹Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Ft Worth, Texas 76107, USA

²Institute of Investigative Genetics, University of North Texas Health Science Center, Ft Worth, Texas 76107, USA

Familial searching of large DNA databases is another approach that has a huge potential capability to assist in solving cases in which there is biological evidence. Familial searching is used to develop investigative leads regarding the source of a forensic sample by searching in a forensic database(s) for possible relatives of the sample source. This concept has been proposed and is being used in forensic investigations in several States of the USA and in the United Kingdom.

Generally, there are two methods for familial searching: Identity-by-State (IBS) based or likelihood ratio (LR) based. The IBS based method simply compares the number of shared alleles between the forensic profile and the profiles from a database(s). There are mainly three policies in the USA based on allele sharing: 1) California Department of Justice requires sharing of at least 15 Short Tandem Repeat (STR) alleles to be considered a potential familial hit warranting further investigation; 2) some other states (e.g., Florida, Nebraska, Oregon, Washington) require at least one allele shared at all typed loci (similar to a moderate stringency approach); 3) SWGDAM Ad Hoc Committee on Partial Matches recommendations also considers a moderate stringency approach with additional conditions.

The LR based method compares the probabilities of the forensic and candidate profiles given that they are related (e.g., parent-child or full-sib) or unrelated, which is the LR or kinship index (KI) for the given relationship. The KI can be directly used in familial searching or further adjusted by database size. The SWGDAM Ad Hoc Committee on Partial Matches recommended the Expected Kinship Ratio (EKR) for familial searching, which is the KI divided by the sample size of the searched database.

In this study, we introduce the analytical details of expected IBS match probabilities between a pair of profiles given Identity-By-Descent (IBD) of the relationships, and present the expected IBS distributions of unrelated, parent-child and full-sib pair relationships. The logic of the LR ratio method is also addressed. Further, we investigated different familial searching policies which are currently adopted, compared the false negative and false positive rates of IBS and KI measures, and suggested that combining both IBS and KI may be a better approach than IBS or KI alone. Using the 13 core CODIS STR data of the Caucasian population, the analytical and simulation studies show that

1. The California policy (i.e., at least 15 shared alleles) is a good threshold balancing between false negative and false positive rates, which only accepted 0.046% unrelated pairs as potential related pairs for further investigations; however it may falsely exclude 17.8% true full-sib and 18.2% true parent-child relationships. A higher threshold may be an option. Only 0.12% unrelated profiles were falsely included for a threshold with ≥ 16 shared alleles; but 31.1% true full-sib and 43.2% true parent-child relationships were excluded.

2. The policy which requires at least one shared allele at each locus has a similar false positive rate as the $IBS \geq 15$ policy, and the majority of parent-child will not be excluded. However, according to Bureau of Justice Statistics 1996 report, about 36.5% of inmates had full-sibs incarcerated compared to 22.8% inmates with parents or children incarcerated. Therefore, a 76.0% exclusion rate for true full-sibs may not be practically acceptable.
3. If familial searching only considers the profiles with all loci that are moderate stringency matched, the chance to include unrelated profile as relatives was extremely low (i.e., 2.69×10^{-9}) with the consequence that more than 99.9% of true relatives were also excluded. In accordance with the SWGDAM recommendations, the use of moderate stringency match has "little useful probative value" in familial searching.
4. The SWGDAM suggested EKR is a very stringent measure for large databases. For a database containing a million profiles, almost no unrelated individuals were included, but about 90.1% true full-sib and 81.3% true parent-child pairs were excluded. For the unrelated pairs that were identified as parent-child or full-sib, only 0.038% or 0.008% pairs, respectively, had KIs greater than 1,000. A $KI \geq 1000$ might be a good measure for 13 CODIS loci to balance the false positive and false negative rates.
5. SWGDAM recommended that variations of EKR among populations (i.e., Caucasian, African American, Southwest Hispanic, and Southeast Hispanic) should not be less than 10-fold. However, this 10-fold threshold may be too stringent. More than 88% of true full-sib or parent-child relationship pairs had Min.EKR/Max.EKR less than 0.1 for the 4 reference populations suggested in the SWGDAM recommendations.
6. The SWGDAM recommended measure, Expected Match Ratio (EMR), only depends on the forensic profile itself for given population, and is not directly applicable for familial searching. It may be helpful to determine how common the forensic profile is.
7. The LR calculation accommodating population substructure and mutation usually yields more conservative KIs. Although incorporating mutation only slightly changes the KI of true relationships, it is crucial if one chooses to avoid excluding true parent-child pairs with mutations. Population substructure has relatively higher effects than mutation, but the differences are within 10-fold in 99% cases.
8. IBS and LR measures can be jointly considered to improve the accuracy of searching. Table 1 summarizes false positive and false negative rates of some reasonable familial searching thresholds. $IBS \geq 15$ or $IBS \geq 16$ with $KI \geq 1,000$ or $KI \geq 10,000$ combined are practical searching policies with good balance between false positive and false negative rates. $IBS \geq 17$ 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 possible hit list.
9. False negative and false positive rates will be reduced and higher accuracy of familial searching will be provided with more loci.
10. 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. A full-sib pair is expected to have about 16 shared alleles and a KI of roughly 1,000 for 13 CODIS loci. For an evidence profile searched against a database containing 1 million unrelated samples, more than 200 unrelated samples are expected to have higher positions in the candidate list than the true full-sib.

Y chromosome STR and mitochondrial DNA can be used as additional markers for familial searching to exclude unrelated candidate profiles. The discrimination power of current multiplex Y-STR systems can reach 0.9999 with 16 Y-STRs, which means that, even if there are 10,000 potential hits after autosomal based familial searching, approximately one is expected to remain after Y-STR filtering. However, it may be better to first screen the database by Y-STRs, so that true male lineage relatives, particularly sibs, will not be excluded because of a relatively low IBS or KI. The mtDNA has a comparatively lower discrimination power than that of Y-STR haplotypes because of higher cumulative mutation rates and heteroplasmy.

Table 1. False positive and false negative rates for some common policies base on one million simulation pairs.

(a) IBS based

Policy	False positive		False negative	
	Unrelated	Parent-Child	Parent-Child	Full-Sib
IBS \geq 14	1.5%	3.8%	3.8%	8.8%
IBS \geq 15	0.46%	18.2%	18.2%	17.8%
IBS \geq 16	0.12%	43.2%	43.2%	31.1%
IBS \geq 17	0.024%	68.9%	68.9%	47.7%

(b) Likelihood ratio based

Policy	Unrelated identified as		True Parent-Child	True Full-Sib
	Parent-Child	Full-Sib		
KI \geq 1,000	0.0379%	0.0079%	4.4%	42.2%
KI \geq 10,000	0.0127%	0.0005%	21.8%	63.0%
KI \geq 100,000	0.0017%	<0.0001%	53.4%	80.1%
KI \geq 1,000,000	0.0001%	<0.0001%	81.3%	90.1%

(c) IBS and likelihood ratio combined

Policy	Unrelated identified as		True Parent-Child	True Full-Sib
	Parent-Child	Full-Sib		
IBS \geq 14; KI \geq 100	0.0720%	0.0773%	3.8%	23.1%
IBS \geq 14; KI \geq 1,000	0.0348%	0.0078%	4.4%	42.2%
IBS \geq 14; KI \geq 10,000	0.0123%	0.0005%	21.8%	63.0%
IBS \geq 15; KI \geq 1,000	0.0490%	0.0076%	18.3%	23.1%
IBS \geq 15; KI \geq 10,000	0.0102%	0.0005%	21.8%	63.0%
IBS \geq 16; KI \geq 1,000	0.0157%	0.0064%	43.2%	42.2%
IBS \geq 16; KI \geq 10,000	0.0064%	0.0004%	43.2%	63.0%
IBS \geq 16; KI \geq 100,000	0.0012%	<0.0001%	53.4%	80.1%