Forensic Science International: Genetics 5 (2011) 493–500

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

# Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach

Steven P. Myers [a,*], Mark D. Timken [a], Matthew L. Piucci [a], Gary A. Sims [a], Michael A. Greenwald [b], James J. Weigand [a], Kenneth C. Konzak [a], Martin R. Buoncristiani [a]

[a] *California Department of Justice Jan Bashinski DNA Laboratory, 1001 West Cutting Blvd., Suite 110, Richmond, CA 94804, United States*
[b] *2615 Spencer Lane, Hayward, CA 94542, United States*

ARTICLE INFO

*Article history:*
Received 14 August 2010
Received in revised form 6 October 2010
Accepted 7 October 2010

*Keywords:*
Familial search
Forensic science
DNA database
Combined DNA Index System (CODIS)
Short tandem repeat (STR)
Y-chromosomal STR

ABSTRACT

A validation study was performed to measure the effectiveness of using a likelihood ratio-based approach to search for possible first-degree familial relationships (full-sibling and parent–child) by comparing an evidence autosomal short tandem repeat (STR) profile to California's ~1,000,000-profile State DNA Index System (SDIS) database. Test searches used autosomal STR and Y-STR profiles generated for 100 artificial test families. When the test sample and the first-degree relative in the database were characterized at the 15 Identifiler® (Applied Biosystems®, Foster City, CA) STR loci, the search procedure included 96% of the fathers and 72% of the full-siblings. When the relative profile was limited to the 13 Combined DNA Index System (CODIS) core loci, the search procedure included 93% of the fathers and 61% of the full-siblings. These results, combined with those of functional tests using three real families, support the effectiveness of this tool. Based upon these results, the validated approach was implemented as a key, pragmatic and demonstrably practical component of the California Department of Justice's Familial Search Program. An investigative lead created through this process recently led to an arrest in the Los Angeles Grim Sleeper serial murders.

Published by Elsevier Ireland Ltd.

## 1. Introduction

The use of autosomal short tandem repeat (STR) DNA genotyping information to aid in identifying familial relationships is widely accepted in forensic science and is commonly applied to civil and criminal paternity cases, to missing persons cases, and in mass disaster and mass burial situations [1]. There has been interest in applying similar familial searching methods in a systematic way to forensic cases for which an evidence-associated STR profile has been generated but where no matching profile has been found in a relevant database of offender DNA profiles, e.g., in the CODIS (Combined DNA Index System) database [2]. For such circumstances, the use of familial searching techniques could provide investigative leads to potential relatives of the evidence source who may be in the database.

Several recent studies have demonstrated the effectiveness of STR-based likelihood ratio (LR) calculations to search for potential familial relationships in a DNA database [2–5]. Those studies used either large simulated databases or real databases that were orders of magnitude smaller than the California convicted-offender SDIS (State DNA Index System) database. We report here results from a study performed to validate an STR-based familial searching procedure for the California SDIS database, the largest SDIS database in the United States and comprised of ~1,000,000 STR profiles at the time of the study.

Bieber et al. recommended the broad application of familial-searching techniques for all cases in which a direct hit to the database did not occur [2]. This approach is being pursued by the Colorado Bureau of Investigations with some success [6]. In contrast, the California Department of Justice designed a Familial Search Program that would be used only upon the request of law enforcement agencies investigating major violent crimes where there is a serious risk to public safety, and where all other investigative leads have been exhausted. Our goal was to ensure that we were employing an effective investigative tool, given current and readily available technology, as part of a pragmatic program that would strike a balance between privacy concerns and the need to provide information that may solve a violent crime or series of crimes [7]. In addition to the successful detection of authentic relationships, and in recognition of privacy concerns, the program had the important goal of avoiding the further investigation of individuals identified because of coincidental associations.

---

* Corresponding author. Tel.: +1 510 620 3300; fax: +1 510 231 8744.
*E-mail address:* Steven.Myers@doj.ca.gov (S.P. Myers).

1872-4973/$ – see front matter. Published by Elsevier Ireland Ltd.
doi:10.1016/j.fsigen.2010.10.010

## 2. Material studied, methods, techniques

### 2.1. The 100 test families

A set of 100 unique, artificial families was created using profiles from a publicly available population database [8,9] of genuine Identifiler® and Yfiler® (Applied Biosystems®, Foster City, CA) profiles. Pairs of profiles were randomly assigned to represent the parents, and artificial matings were performed using Mendelian principles to create two male offspring. No mutations were simulated. For each family, the autosomal STR profile of one offspring was selected as the "test" or "evidence" profile to be searched against the convicted offender SDIS profiles. The STR profiles for the father and the remaining offspring served as positive controls for the "authentic" parent–child and full-sibling relationships, respectively. Although the STR profiles for the father and brother positive-control samples were not actually placed into the SDIS, comparisons to the test sample were performed as if they were in the database (i.e., using the same LR formulae, *vide infra*).

### 2.2. STR profile comparisons—likelihood ratio calculations

LRs for the autosomal STR test-to-offender comparison were calculated for three racial/ethnic population groups (African American, Caucasian, SW Hispanic) using standard formulae for parent–child and full-sibling duos [10,11] and FBI databases [12,13]. Calculations were limited to first-degree relationships in recognition of the low expectation that more distant relatives would yield LRs sufficient to differentiate them from such a large pool of unrelated individuals [14]. The profiles of our authentic relatives were tested separately as 13-locus profiles (the CODIS core loci required for inclusion in the national database: D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, vWA, TPOX, D18S51, D5S818, and FGA) and as 15-locus profiles (the Identifiler® kit loci consisting of the 13 CODIS core loci plus D2S1338 and D19S433) in recognition of the locus-count bimodality in our offender database. Consistent with the approach used in other studies [2,4], no $\theta$-corrections for population substructure were included in the LR calculations.

In a departure from those studies, no attempts were made to correct for the possibility of meiotic mutations or deletions. While our approach to mutations has the intention of simplifying the calculation and avoiding false associations, it does, however, mean that a true parent–child relationship could be excluded. In practice, even when the only relative in the database is a father or son of the perpetrator, only a small proportion of searches should be negatively impacted by this decision. Average meiotic mutation rates for autosomal STR loci [15] are approximately $10^{-3}$ when testing paternity trios, a comparison that commonly allows for the determination of one specific obligate paternal allele. In parent–child familial searching comparisons, essentially paternity duos, either allele of the evidence profile could be considered obligate. This leaves the possibility that a shared allele identical-by-state between the father and son might mask a mutation in a shared allele identical-by-descent. Similarly, the mutated allele might now be identical-by-state to the maternal allele of the son. Both scenarios suggest that our rate of detecting mutations when performing familial searching should be lower than predicted by paternity trio mutation rates.

Rather than ignore the statistical strength of the Y-STR testing, a Y-STR LR was calculated based upon the known haplotype of the evidence assuming *a priori* that a match to the offender exists. The Y-STR LR is the inverse of the test sample's overall (i.e., "All" population groups combined) Y-haplotype frequency expressed as the upper 95% confidence interval [16] in the US Y-haplotype database [17]. This is combined with the autosomal LR [18,19].

Consistent with the recommendations of Bieber et al. [2], as well as the practices of agencies performing mass disaster identifications [20–22], the overall LR value was adjusted for the size of the database ($N$). In this case, "$1/N$" is analogous to the approximate prior odds of a database sample being the true relative under the assumption that a relative exists in the database, making the final value proportional to the approximate posterior odds of a Bayesian analysis. It should be noted that, in our approach, familial searches use only California's database of *convicted* offenders. Arrestees collected pursuant to Proposition 69 [23] are not included in a familial search and are not included in $N$.

$$\text{The final calculation : “Odds”} = \text{Autosomal STR LR} \times \text{Y} - \text{STR LR} \times 1/N$$

To measure the effect of including the Y-STR LR in the calculation, a comparison was made to "Odds" based solely upon the autosomal STR LR and the database size.

### 2.3. Autosomal STR allele frequencies for a structured database

For reasons unrelated to this study, the racial/ethnic structure of our database was not assessed. To address this unknown racial/ethnic structure, the "odds" were based upon the minimum LR calculated for the three population groups (African American, Caucasian, SW Hispanic). This approach presupposes that an individual's alleles tend to be more common in their own group than in other groups, and is consistent with other work showing that the use of the cognate ethnic allele frequencies will, on average, give minimum calculated LR values [24]. To measure the effectiveness of using the minimum LR in the calculation, a comparison was made to "Odds" based upon each of the three population groups.

### 2.4. Statistical thresholds

Statistical thresholds were established to be in line with those recommended by the SWGDAM Ad Hoc Committee on Partial Matches for the investigation of CODIS partial matches [18]. For an offender to be investigated further as a possible familial lead, "odds" for either the parent–child relationship or the full-sibling relationship must be greater than or equal to 1 for at least one population *and* no less than 0.1 for the remaining two populations. Treating 1 and 0.1 as approximate posterior odds, these thresholds mean that at least one population group had attained a posterior probability of relatedness greater than or equal to 50%, and neither of the other two had posterior probabilities lower than 9.1%. The magnitude of these thresholds is appropriate given the intended purpose is to develop a lead for further investigation, not to directly identify an individual.

We acknowledge that including the *same* Y-STR LR and the *same* $N$ for each calculation in a search simply re-scales the LR values. However, combining all of the statistical information in advance allowed us to eliminate associations that would not reach our thresholds even if the test and database samples were to share the same Y-STR profile. While not modeled in this study, in practice the initial Y-STR LR would be recalculated for any offender-to-evidence comparison later observed to have very similar but still discordant Y-STR profiles (e.g., a one-locus discordance within a Yfiler® profile that may be due to a mutation). In such cases, the pair's revised "odds" would be evaluated in relation to the SWGDAM thresholds.