Familial DNA Database Search System-Hardware/Software Integration Project
Award - 2012-DN-BX-K036

Bonnie Mountain, M.A.[1], Stephanie A. Santorico, Ph.D.[2,3] and Greggory LaBerge, Ph.D.[1, 3]
[1]Denver Police Crime Laboratory, Denver, CO
[2]University of Colorado Denver, Department of Mathematical & Statistical Sciences, Denver, CO
[3]University of Colorado Anschutz Medical Center, Human Medical Genetics Program, Aurora, CO

Abstract:

Familial DNA searching is an investigative tool used in Australia, France, the Netherlands, New Zealand, United Kingdom, and the United States and derives forensic DNA-based intelligence to assist in criminal investigations. Currently, ten states (Arizona, California, Colorado, Louisiana, New York, Texas, Utah, Virginia, Wisconsin, and Wyoming) have implemented familial search programs. Familial searching has the potential to significantly leverage forensic DNA databases to detect siblings and parent-child relatives with existing short tandem repeat (STR) based data, which can assist in the identification of perpetrators of crimes where probative DNA evidence has been recovered.

However, the current ability of the criminal justice field to conduct familial DNA searches is limited. At this time, a Familial Search Program exists which was collaboratively developed by the Denver Police Department and the Denver District Attorney's Office (DA) to compile, search, and report potential familial relationships from existing forensic DNA data stored in the CODIS databank. The existing system is a standalone system that requires installation of local hardware and software. Although the current system has generated significant interest (now being used by Arizona, Louisiana, New York, States, Virginia, Wisconsin, and Wyoming), it lacks the ability to reach large audiences of users and the computational power to handle large datasets and searches.

The proposed effort has two main goals to enhance current familial search approaches. To achieve these goals, the Denver Crime Laboratory partnered with a Post Doctorate from the University of Colorado Denver, Department of Mathematical & Statistical Sciences and DRC Computer Corporation to produce the following objectives:

1. Develop a web-based familial search system that operates with DNA data allowing secure data transmission via a graphical interface for determination of relatedness which can be used by law enforcement/crime laboratories nationally.
2. Statistical evaluations of familial searching to generate additional recommendations for interpretation of results and potential enhancements to likelihood ratio calculations.
    a. Determining the utility of using Expected Match Ratio (EMR) and Estimated Kinship Ratio (EKR) calculations to enhance interpretation of familial search results.
    b. Methods from graph theory will be assessed to determine if familial clusters can be identified.
    c. Evaluating the power, specificity, and sensitivity of the methods identified in Goals 2a and 2b will be studied.

Table of Contents

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Executive Summary:

I. Introduction:

Statement of the problem:

The success of the Combined DNA Index System (CODIS) system (the United States national forensic DNA database) since its inception in 1998 has been well established, as reflected by the fact that the national database (NDIS) has assisted in more than 77,000 cases through exact matches. The National CODIS system contains more than six million Short Tandem Repeat (STR) DNA profiles. All 50 states and the District of Columbia require offenders convicted of certain crimes to submit DNA samples, and in 2005 forensic crime laboratories contributed more than 800,000 profiles to the national system (Durose 2008). Most states require samples from all felons, and many states require collection of DNA samples from certain arrestees. The number of DNA profiles in state and local databases continues to increase as states enact legislation expanding eligibility into the system, doubling the total in CODIS since 2005.

Searches into DNA databases beyond exact matches - to identify offenders' close relatives - have been conducted in only a very small number of cases, primarily in the United Kingdom or for the identification of human remains (Leclair et al., 2004). However, in 2008 the California Department of Justice issued policy guidelines under which partial match searches could be performed in California's state DNA database (Anderson 2008). Their policy requires at least 15 shared STR alleles and a threshold this high could result in missing potential kinship matches. The first case under that new policy was not successful in detecting a familial relationship in the California convicted offender DNA database. Recently they have had success with familial DNA searching leading to the identification of a serial killer in Los Angeles (Myers et al. 2010).

Success in detecting close relatives in a database depends on a relative's DNA profile actually being included in the DNA database. Moreover, research has demonstrated the continuity of antisocial behavior across generations (Doumas et al., 1994), and a government sponsored study reported that 46 percent of incarcerated people stated they had at least one close relative who had been incarcerated (Doumas et al. 1994; Correctional Populations in the United States 1996**)** indicating that the chances of finding relatives in the DNA database are very good.

In October of 2009, the Scientific Working Group on DNA Analysis and Methods (SWGDAM) Ad Hoc Committee released a communication regarding the recommendations for the use of partial matches (Li et al. 2006) (familial searching). Their recommendation included performing an Expected Match Ratio (EMR) and an Expected Kinship Ratio (EKR) calculation. An assumption is made that most offender databases consist of African American, Caucasian, Southeastern Hispanic and Southwestern Hispanic and suggest using allele frequencies for these population groups. Partial (i.e. potential familial) match is considered useful if the EMR or EKR value of at least one of the four populations is greater than or equal to 1.0 and all of the others are greater than or equal to 0.1. To date, there have not been any published studies comparing the EMR/EKR calculation to traditional likelihood ratio calculations. EMR/EKR has not been utilized by the Denver Crime Laboratory and its rate of accuracy for detecting true relatives is untested. It is unknown if this calculation method provides a higher true positive rate than traditional likelihood ratio. This research project will help determine the best application of EMR/EKR and how it might be used in conjunction with the current likelihood method to further narrow results of potential matches.

The weakness of any developed system or method of calculation is that some true relatives

which exist in a database queried against a forensic unknown sample will not be detected due to sharing of only common alleles. In large many-to-many searches, hundreds of millions of results are possible, and taking the top candidates based on either locus/allele sharing or magnitude-of-likelihood ratio will result in missing some true relatives. Comparing EMR/EKR calculations with traditional likelihood ratio calculations could help generate additional recommendations to interpret familial search results and potentially maximize the identification of true relatives.

The current ability of the criminal justice field to conduct familial DNA searches is limited. At this time, a Familial Search Program exists which was collaboratively developed by the Denver Police Department and the Denver District Attorney's Office to compile, search, and report potential familial relationships from existing forensic DNA data stored in the Combined DNA Index System (CODIS) databank. However, the existing system is a standalone version and requires installation of local hardware and software. Although the system as-is has generated significant interest (the standalone version is now in use by Arizona, Louisiana, New York, Utah, Wisconsin, Wyoming, and Virginia), this version lacks the ability to reach large audiences of users and the computational power to handle large datasets and searches. At this time, the largest database size in the United States for a single state is in California, which has approximately 1,779,967 convicted offender profiles, so even a single comparison to this database would result in over a million pair-wise comparisons.

Literature citations and review:

The identification of close relatives through kinship analyses or paternity testing relying on a likelihood ratio approach has strong theoretical and statistical foundations. Research related to this project includes the identification of relatives in mass disasters (Durose 2008, Dudbridge 2007, Brenner and Weir 2003), the identification (or elimination) of a relative of a known suspect (Anderson and Weir 2006), and paternity analyses. The probabilistic foundation for the likelihood ratio approach has also been established (Leclair et al. 2004, Anderson 2008, Myers et al. 2010, Anderson and Weir 2006, Reid et al. 2008, Curran and Buckleton 2008). The specificity of STR-based DNA systems has been tested for detecting sibships and shown to minimize false-positive results when the correct likelihood ratios are calculated (Reid et al. 2008).

Several publications have outlined familial search strategies in an attempt to identify best practices. Suggestions range from scoring the number of alleles shared between suggested pairs (Ge et al. 2013) to in-depth probabilistic strategies (Balding et al. 2013; Slooten and Meester 2014). A validation study (Myers et al. 2010) examined using a Y-STR likelihood value to be incorporated into the resulting LR. Myers et al. also looked at the impact of 15 locus profiles compared to 13 loci on detecting true relatives and sorted results lists by the max LR with a focus on the top 168 (two 96 well plates) results for additional follow up. The SWGDAM Ad Hoc Committee provided recommendations for the use of partial matches in familial searching (Interim Plan for the Release of Information in the Event of a "Partial Match" in 2009). Their recommendation included performing an Expected Match Ratio and an Estimated Kinship Ratio calculation.

Classical statistical theory (by way of the Neyman-Pearson lemma) instructs that the likelihood ratio based on the full evidence achieves maximal power for a fixed false positive rate (Casella and Berger 2001). This has been explored in Balding et al. 2013 where simulations backed up statistical theory by demonstrating that the kinship ratio outperforms the approach that combined both kinship information and identity by state (IBS). Specifically, it was shown that for the same level of false-positive rate, the LR had a higher true-positive rate than the combined kinship and IBS statistic.

Statement of hypothesis or rationale for the research:

Familial searching has not been implemented within the CODIS system and is not recommended to be conducted at the NDIS level. The current ability of the criminal justice field to conduct familial DNA searches is limited. At this time, a Familial Search Program exists which was collaboratively developed by the Denver Police Department and the Denver District Attorney's Office to compile, search, and report potential familial relationships from existing forensic DNA data exported from the Combined DNA Index System (CODIS) databank. However, the existing system is a standalone version and requires installation of local hardware and software. Although the system has generated significant interest (the standalone version is now in use by Arizona, Louisiana, New York, Utah, Wisconsin, Wyoming and Virginia), this version lacks the ability to reach large audiences of users and the computational power to handle large datasets and searches. Development of a web-based system would extend familial searching to any agency and allow international searches and use within the intelligence community.

At this time, the largest State database size in the United States is California, which has approximately 1,779,967 convicted offender profiles, so even a single comparison to this database would result in over a million pair-wise comparisons. Evaluation of these pair-wise results needs further refinement to minimize false positives and increase the identification of true positives. Evaluation of the SWGDAM recommendation could provide additional methods to reduce false positives and increase true positives.

II. Methods

Web-based Familial Search System Method – DRC

For the web-based system, DRC had previously developed a proof of concept system that initially used a commodity 1U (1.75 inch high) server and included a DRC coprocessor. For the prototype, DRC modified the existing Denver software application to execute on the DRC coprocessor. Specifically, the computationally intense Match routine (refer to the blue box in Figure 1) was ported to the DRC coprocessor.

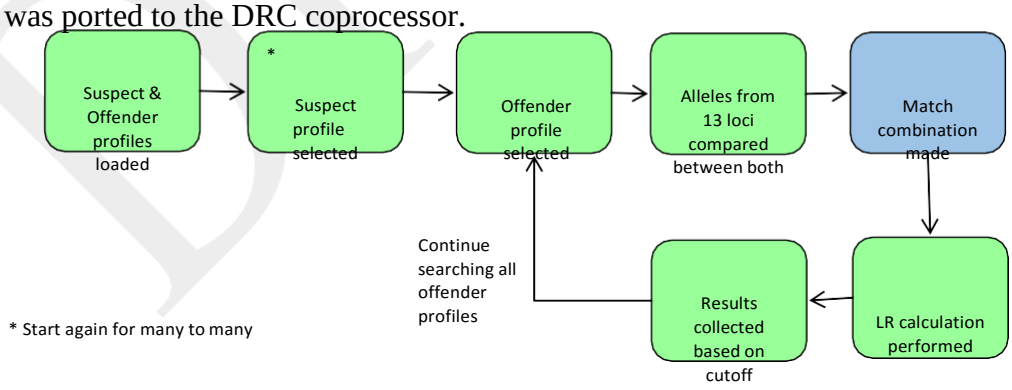

Figure 1. Enhanced Familial program data flow with DRC technology.

DRC leveraged the prototype technology and developed the familial search software on a web-based cloud system that will be made available to users through a secure gateway over the Internet (or intranet). The system uses existing strong encryption (Security First) technologies to secure upward and downward genetic data flow to allow rapid familial searches between forensic samples, offender databases of individuals or unknown samples, and other datasets using the

CODIS core Short Tandem Repeat (STR) genetic loci, used routinely for human identification.

This development phase required five key activities to advance the prototype into production:

1. Migration of the prototype from Windows to a Linux based environment for improved performance. The Linux operating system has a lower overhead structure than Windows and supports multiple users, processes more efficiently with better security, and will scale linearly with the number of DRC Accelium Accelerator's on a single machine.
2. Customize existing Security First Corporation data assurance technologies to provide a highly secure user access and communication portal using strong encryption technologies using SPxConnect™.
3. Explore increasing the number of match engines executing simultaneously in the DRC coprocessor to extend performance in a secure data transmission environment. Increasing to 6 match engines can be accomplished, which would allow multiple simultaneous suspect searches to be conducted and enable a split of offender DNA database information into multiple subsets (based on ethnicity, geography, or other factors).
4. Expand the number of DRC accelerators in the server and/or increasing the number of servers. This would linearly increase the number of simultaneous searches or reduce the time to complete one complex familial DNA search. These systems will result in a robust computational system for deriving important genetic intelligence from DNA data.
5. Addition of browser based front end GUI for end user interaction with customization of search run parameters and visualization of search results. Allow for the selection of result thresholds and allele frequencies applicable to the search region.

Statistical Evaluation Methods

Determine the utility of SWGDAM recommended EMR/EKR Method

Our review of the SWGDAM Recommendations for partial matches identified two differences compared to Denver's current familial search methods. The EMR/EKR method excludes low stringency matches (Table 1) and the resulting LRs are divided by the database size. These differences served as a focus for simulation and comparative analysis. The Denver familial search software system was modified to include EMR/EKR calculations as detailed in the SWGDAM recommendations, and the updated program was validated using the hypothetical partial match outlined in the SWGDAM publication (Interim Plan for the Release of Information in the Event of a "Partial Match" at NDIS 2009).

Three comparisons were completed. The original EMR/EKR method was compared to the DLR method. In the second comparison, a low stringency calculation was added to the EMR/EKR method. For the final comparison, division by the database size was removed. To determine the performance of the EMR/EKR method compared to the DLR method, true positive, false negative, and false positive rates were calculated.

SWGAM EMR/EKR Method

SWGDAM defines the Expected Match Ratio (EMR) as a statistic that is used to determine which is more likely to occur: a match between the forensic unknown and one relative in a DNA database or a match between the forensic unknown and one or more unrelated persons in a DNA database. The final EMR calculation is a likelihood ratio where the numerator is the probability of a moderate stringency match, denoted msMatch, in the CODIS database between the forensic unknown and one relative including an estimation of alleles being identical by descent. A moderate stringency (Table 1) match occurs with a homozygous forensic unknown matching at least one allele of a heterozygous candidate offender or when a heterozygous forensic unknown

has at least one allele matching a homozygous candidate offender profile. The denominator is the probability of a moderate stringency match between the forensic unknown and one unrelated person, reflecting alleles being identical by state. The final EMR value is the product of the individual locus ratios, where *L* is the number of loci included in the calculation, divided by the size (Z) of the DNA database being searched presented in equation (1).

$$\frac{1}{Z} * \prod_{x=1}^{L} \frac{\text{P(msMatch | Forensic Unknown \& 1 relative)}}{\text{P(msMatch | Forensic Unknown \& 1 non - relative)}} \quad (1)$$

Table 1: Examples of stringency matches where locus F is low stringency and is not included in the EMR/EKR calculations.

Locus	Forensic Unknown Genotype	Candidate Offender Genotype	CODIS Match Stringency
A	7	7	High
B	7	7, 8	Moderate
C	13,14	13,14	High
D	13,14	13	Moderate
E	13,14	14	Moderate
F	13,14	14,16	Low

In contrast, the Estimated Kinship Ratio (EKR) compares the probability of a specific pair of profiles given that the individuals are related versus unrelated. The EKR is based only on loci with moderate stringency matches between the forensic unknown and a candidate offender. EKR contrasts the probability of alleles being identical by descent versus identical by state. The final EKR value is the product of the individual locus ratios, where *L* is the number of loci included in the calculation, divided by the size (Z) of the DNA database being searched, presented in equation (2).

$$1 * \prod_{x=1}^{L} \frac{\text{P(msMatch | Forensic Unknown \& Candidate offender are related)}}{\text{P(msMatch | Forensic Unknown \& Candidate offender are unrelated)}} \quad (2)$$

The SWGDAM method uses five formulas to derive the probability estimates. Two formulas (1a and 1b) are used when the forensic profile is homozygous at a locus and three formulas (2a, 2b, and 2c) are used when the forensic profile is heterozygous. These formulas are outlined in Table 2 where k_0, k_1, and k_2 are kinship coefficients, p and q are allele frequencies, and θ is a subpopulation correction with a recommended value of 0.01 (National Research Council 1996) The frequency (Freq) used for q differs for EMR and EKR. Table 3 identifies the various match scenarios and corresponding EMR and EKR formulas used with p and q identified.

The ethnicity of the perpetrator is often not known at the time a DNA search is performed. SWGDAM has suggested that most offender databases consist of African American, Caucasian, Southeastern Hispanic, and Southwestern Hispanic ethnicities, and recommends using allele frequencies for these population groups (Budowle et al. 2001). As a result, four individual EMRs and EKRs are calculated. A partial match (i.e., potential familial lead) is considered useful if the EMR or EKR value of at least one of the four populations tested is greater than or equal to 1.0 and all of the others are greater than or equal to 0.1.

Table 2. SWGDAM recommended formulas for the calculation of EMR and EKR. See the supplemental table 6B for examples of when each formula is used. *Formula 2d was derived using the formulas from Popstats and is not an original recommendation from SWGDAM.

Formula	P(msMatch|Relative)	P(msMatch|Non-Relative)
1a	$k_2 + \frac{k_1[2\theta+(1-\theta)p]}{(1+\theta)} + \frac{k_0[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$
1b	$\frac{k_1(1-\theta)q}{(1+\theta)} + \frac{k_0 2[2\theta+(1-\theta)p](1-\theta)q}{(1+\theta)(1+2\theta)}$	$\frac{2[2\theta+(1-\theta)p](1-\theta)q}{(1+\theta)(1+2\theta)}$
2a	$k2 + \frac{k1[\theta+(1-\theta)p]}{2(1+\theta)} + \frac{k1[\theta+(1-\theta)q]}{2(1+\theta)} + \frac{k0\ 2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$
2b	$\frac{k1[\theta+(1-\theta)p]}{2(1+\theta)} + \frac{k0[\theta+(1-\theta)p][2\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	$\frac{[\theta+(1-\theta)p][2\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$
2c	$\frac{k_1[\theta+(1-\theta)q]}{2(1+\theta)} + \frac{k_0[\theta+(1-\theta)q][2\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	$\frac{[\theta+(1-\theta)q][2\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$
2d*	$\frac{k_1[\theta+(1-\theta)q]}{2(1+\theta)} + \frac{k_2[\theta+(1-\theta)p][\theta(1-\theta)q]}{(1+\theta)(1+2\theta)}$	$\frac{2[\theta+(1+\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$

Table 3. Example of match scenarios and EMR/EKR formulas used with frequency (Freq) values for alleles p or q.

Forensic Alleles	Offender Alleles	Matched Allele	EMR			EKR		
			Formula	p	q	Formula	p	q
7	7	7	1a+1b	Freq(7)	1- Freq(7)	1a	Freq(7)	N/A
7	7,8	7	1a+1b	Freq(7)	1- Freq(7)	1b	Freq(7)	Freq(8)
8	7,8	8	1a+1b	Freq(8)	1-Freq(8))	1b	Freq(8)	Freq(7)
7,8	7,8	7,8	2a+2b+2c	Freq(7)	Freq(8)	2a	Freq(7)	Freq(8)
13,14	13	13	2a+2b+2c	Freq(13)	Freq(14)	2b	Freq(13)	N/A
13,14	14	14	2a+2b+2c	Freq(14)	Freq(13)	2c	N/A	Freq(14)

DLR Method

Denver's familial search program (DLR Method) calculates parent-child and sibling likelihood ratios (Equation 3) according to formulae presented in Table 13 of Presciuttini et al. 2002 and makes use of all loci without a subpopulation correction. These formulas are presented in Tables 4 and 5 where pc is the parent-child index and sib represents the sibling index. The DLR method does not adjust for the size of the database being searched. The calculations use three populations commonly used in the United States for determining allele frequencies in forensic DNA profile estimation: U.S. Caucasians, U.S. African-Americans, and U.S. Hispanics as well as an average frequency denoted by FBI Average (Budowle at al. 1999). Four likelihood ratio values are calculated for potential parent-child pairs and four for sibling pairs. The program uses a default likelihood ratio threshold value of 100,000. This threshold is based on results from the first familial search conducted for Denver in 2008. Matches confirmed to be related with subsequent Y-STR analysis had likelihood ratio values greater than 100,000. Table 5 identifies the various familial match scenarios and corresponding DLR formulas.

$$\prod_{x=1}^{13} \frac{\text{P(msMatch} \mid \text{Forensic Unknown \& Candidate offender are related)}}{\text{P(msMatch} \mid \text{Forensic Unknown \& Candidate offender are unrelated)}} \quad (3)$$

Table 4. Denver recommended likelihood ratio formulas. (1) No allele is shared, (2) Forensic and offender locus are heterozygous sharing both alleles, (3) Forensic and offender locus are heterozygous sharing one allele a or allele b, (4) Forensic locus is heterozygous and offender locus is homozygous sharing one allele a, (5) Forensic locus is homozygous and offender locus is heterozygous sharing one allele b, (6) Forensic and offender loci are homozygous sharing single allele.

Formula	Parent-Child	Sibling
1	$\frac{1}{1000}$	$\frac{1}{4}$
2	$\frac{0.25}{q}+\frac{0.25}{p}$	$\frac{\frac{0.125}{p}}{q}+\frac{\frac{0.25}{q}+\frac{0.25}{p}}{2}+0.5-0.25$
3	$\frac{0.25}{p}$	$\frac{\frac{0.25}{p}}{2}+0.25$
4	$\frac{0.5}{p}$	$\frac{\left(\frac{0.5}{p}+0.5\right)}{2}$
5	$\frac{0.5}{p}$	$\frac{\left(\frac{0.5}{p}+0.5\right)}{2}$
6	$\frac{1}{p}$	$\frac{\left(\frac{1}{p}+1\right)}{2}*\frac{\left(\frac{1}{p}+1\right)}{2}$

Table 5. Examples of match scenarios and DLR formulas used.

Forensic Alleles	Offender Alleles	Matched Allele	Formula
7	7	7	6
7	7,8	7	5
8	7,8	8	5
7,8	7,8	7,8	2
7,8	7,9	7	3
7,8	6,7	7	3
7,8	8,9	8	3
7,8	6,8	8	3
13,14	13	13	4
13,14	14	14	4
13,14	11,12	N/A	1

Comparative Research Methods

Analysis A – Baseline performance of EMR/EKR for familial searching

Simulation studies of parent-child and sibling pairs in artificially created STR-DNA datasets were used to determine the baseline performance of the EMR/EKR method. To generate simulation profiles, a pool of 2 million alleles was created to match the frequency distribution of alleles observed in Denver's Local DNA Index System (LDIS) of known individuals without regard to ethnicity. This allele pool was used to randomly generate 100,000 profiles with the 13 core CODIS loci. Random pairs were selected to "mate" and one allele from each parent was randomly selected for inheritance. This method was used to generate offspring per mating pair for a total of 5,000 children. Two simulations randomly selected 500 of these children with a corresponding parent and two simulations selected 500 children with a known sibling. The child profiles were loaded as the "forensic unknown/perpetrator" profile to search against a known

offender database. The known offender database consisted of 1,000,000 simulated offender profiles as well as the 500 known parents. The sibling simulations were conducted with 500 "forensic unknown/perpetrator" siblings searched against a 1,000,000 offender database plus the 500 known siblings. A many-to-many comparison was performed for each simulation by comparing the 500 profiles with 1,000,500 profiles for a total of 500,250,000 pair wise comparisons. The familial search script was programmed to report results when any of the four EMR or EKR values were greater than or equal to 0.1, allowing more results to be studied. The same profiles were then used to perform simulations using the DLR method and served as a comparison standard.

Additional analyses were performed to determine if the EMR/EKR method assisted in narrowing potential investigative leads. Many laboratories have resources to follow up leads with additional testing, e.g., Y-STR typing on approximately 100 samples. The Denver Crime Laboratory currently uses a ranking system to determine leads for follow-up. This is accomplished by sorting the result list by the maximum LR value from the four population groups. The top 100 in this ranked list are evaluated for additional investigation, often involving the genotyping of Y-STR loci for comparison where male relatives are indicated (Bieber et al. 2006). The same approach was applied to the EMR/EKR results to determine the number of simulated pairs that were located in the top 100 of the ranked list. The maximum of the four EMR and four EKR values for each match pair was calculated and the list sorted by the maximum EMR/EKR value. Each simulated pair was flagged in the ranked list and a count of the number of simulated pairs found in the top 100 was documented for each simulation experiment. This analysis did not account for the threshold cut-off values and was only based on the ranked position in the results list.

The final analysis used to compare the baseline performance of the EMR/EKR method to the DLR method was to calculate the true positive, false negative, and false positive rates. These values were visualized with the use of a confusion matrix or contingency table (Powers 2007). The definitions for true positive, false positive, false negative, and true negative can be seen in Table 6.

Table 6. Definitions of true positive, false positive, true negative and false negative used in the confusion matrix analysis.

Condition	EMR/EKR	DLR
True Positive	Simulated familial match with EMR or EKR value of at least one of the four populations is greater than or equal to 1.0 and all of the others are greater than or equal to 0.1	Simulated familial match with one of the four populations LR value greater than or equal to 100,000.
False Positive	Match that is not a simulated familial pair but has EMR or EKR value of at least one of the four populations is greater than or equal to 1.0 and all of the others are greater than or equal to 0.1	Match that is not a simulated familial pair but has one of the four populations LR value greater than or equal to 100,000.
True Negative	Match that is not a simulated familial pair and does not have at least one of the four populations greater than or equal to 1.0 with all the others greater than or equal to 0.1.	Match that is not a simulated familial pair but has all four population LR values less than 100,000
False Negative	Simulated familial match that does not have at least one of the four populations greater than or equal to 1.0 with all of the others greater than or equal to 0.1.	Simulated familial match with all four population LR values less than 100,000

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Analysis B - Low Stringency Matches

The SWGDAM calculations focus on moderate and high stringency matches as defined in the CODIS software, with low stringency matches being excluded from the EMR/EKR analysis (http://www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet). High stringency matches are defined where the forensic unknown alleles searched are identical to the candidate offender, for example loci A and C shown in Table 1. A moderate stringency match occurs with a homozygous forensic unknown matching at least one allele of a heterozygous candidate offender or when a heterozygous forensic unknown has at least one allele matching a homozygous candidate offender profile. Examples of a moderate stringency match can be seen in loci B, D, and E, and low stringency matches, as seen in Locus F (heterozygous to heterozygous matching one allele), are not considered in the EMR/EKR calculations.

To determine the impact of excluding low stringency matches in the EMR/EKR calculations, the simulation program script was modified to count the number of loci with low stringency matches that exist in the 500 simulated familial pairs. This count was averaged for each group of 500 simulated familial pairs. Using real DNA profile data from the Denver LDIS database, 1,035 forensic unknowns were compared to 1,199 candidate offenders and a count of low stringency matches was averaged. Both analyses provided insight into how much potential genetic information is being excluded in the EMR/EKR familial search.

To further evaluate the exclusion of low stringency matches, the EMR/EKR method was updated to include a low stringency calculation and use all loci. As stated in the SWGDAM guidelines, when a forensic unknown profile has 2 alleles, EMR is calculated as a sum of formulas 2a, 2b, and 2c (Table 3). The formula used for EKR is determined based on the matching criteria with the offender candidate DNA profile. An additional formula (2d) was derived from general kinship calculations used by the statistical program Popstats (CODIS version 7.0 with service pack 4). Formula 2d used for low stringency matches is noted in Table 2, where θ is a subpopulation correction with a recommendation value of 0.01 and includes kinship coefficients k_1 and k_2.

EMR was calculated as a sum of 2a, 2b, 2c, and 2d. EKR was determined using the 2d formula for low stringency matches; otherwise, the formula used was based on the matching criteria with the offender candidate DNA profile. Table 7 demonstrates the impact of the additional formulas on final EMR and EKR calculations. The familial program script was updated to include the low stringency match calculation and all four simulations were repeated.

Table 7. Match scenarios and updated EMR/EKR formulas used when low stringency matches are included. The values used for p and q remain the same.

Forensic Alleles	Offender Alleles	Matched Allele	EMR			EKR		
			Formula	p	q	Formula	p	q
7	7	7	1a+1b	Freq(7)	1-Freq(7)	1a	Freq(7)	N/A
7	7,8	7	1a+1b	Freq(7)	1-Freq(7)	1b	Freq(7)	Freq(8)
8	7,8	8	1a+1b	Freq(8)	1-Freq(8)	1b	Freq(8)	Freq(7)
7,8	7,8	7,8	2a+2b+2c+2d	Freq(7)	Freq(8)	2a	Freq(7)	Freq(8)
7,8	7,9	7	2a+2b+2c+2d	Freq(7)	Freq(8)	2d	Freq(7)	Freq(9)
7,8	8,9	8	2a+2b+2c+2d	Freq(8)	Freq(7)	2d	Freq(8)	Freq(9)
7,8	6,7	7	2a+2b+2c+2d	Freq(7)	Freq(8)	2d	Freq(7)	Freq(6)
7,8	6,8	8	2a+2b+2c+2d	Freq(8)	Freq(7)	2d	Freq(8)	Freq(6)

13,14	13	13	2a+2b+2c+2d	Freq(13)	Freq(14)	2b	Freq(13)	N/A
13,14	14	14	2a+2b+2c+2d	Freq(14)	Freq(13)	2c	N/A	Freq(14)

The resulting data from the modified EMR/EKR method was sorted into a ranked list and each simulated familial pair flagged. The familial pairs located in the top 100 were counted and the true positive, false positive, and false negative rates were also calculated.

Analysis C - The SWGDAM calculations include normalization factor, accomplished by dividing the EMR/EKR value by the database size and threshold.

The combination of the threshold cut off of 1.0 (with remaining values greater than 0.1) and the suggestion to divide EMR/EKR by the database size has the potential to produce large numbers of false negatives. The resulting EMR/EKR value would need to be larger than the database size, 1,000,500 in our simulations for the match to be considered useful. To further analyze the impact of dividing by the database size, a parent-child simulation was completed without the normalization factor. Once the division by the database size is removed, the threshold recommendation becomes irrelevant. An alternative threshold was determined using false positive rates. The false positive rate for the same parent simulation using the DLR method was determined. An EMR/EKR cut off value was determined that would produce the same false positive rate (5.6E-3%); that threshold is 13,750.

Evaluating the Power, Specificity and Sensitivity methods

DLR Evaluation

Given that the ethnicity for the unknown profile and offenders in the crime database are unknown at the time of a familial search, there are several approaches with respect to allele frequencies and evaluation of likelihood ratio results that can be compared. Five different approaches were used. The first method used LRs calculated from three populations commonly used in the United States: U.S. Caucasians (EUR), U.S. African-Americans (AA), and U.S. Hispanics (HISP) (Hill et al. 2013) and focused on the maximum LR value from these three results (LRMAX1). The second method used allele frequencies derived from the local offender database (LAF) being searched which resulted in one LR value being calculated. The third method looked at the maximum LR (LRMAX2) when EUR, AA, HISP, and LAF are used. A weighted average based on the prior probabilities of ethnicity corresponding to local crime statistics (LRWAVG) was the fourth method. The fifth looked at an average (LRAVG) based on equal probabilities over EUR, AA, HISP. Simulation studies were used to evaluate these five approaches.

Additionally, for each of the five comparisons, the power for detection of a relative will be computed based on a false-positive rate of 8.5×10^{-5}. This false positive rate is based on the idea that when familial search results are obtained, most laboratories are setup to run 96 well sample plates to obtain Y-STR profiles to determine matches to the evidence profile. After control samples and allelic ladders, it leaves about 85 samples that could be run on a 96 plate. Assuming a large offender database of 1 million, we obtain the 8.5×10^{-5} false positive rate. Follow up thresholds will be calculated to correspond to this false-positive rate.

The simulated population was considered to be a mixture of 20% African-American, 48% Caucasian, and 32% Non-White Hispanic. Based on scaling, the estimated proportions of the Colorado state prison population provided by the Colorado Criminal Justice Reform Coalition

(2010 Colorado Quick Facts), individuals were randomly assigned to a racial group and the corresponding allele frequency distribution used. Unrelated and related pairs (1 to 5 million) were simulated using the Pedantics R package (Morrissey et. al. 2010). Related pairs were assumed to be from the same racial group. Unrelated pairs had a race assigned randomly for each individual. 300 million unrelated pairs were simulated from which the 99.7% confidence interval for the 8.5×10^{-5} percentile was computed. Unrelated pairs were simulated based on the allele frequency distributions of Hill et al.; however, the LRs are calculated using a minimum allele frequency of 5/(2N) for rare alleles observed in a racial group which are more rare than 5/(2N). Here N is the number of individuals sampled for estimating the allele frequencies (Butler 2009).

III. Results

Web- based System Improvement

DRC successfully converted the familial search program from Windows to Linux operating system (Figure 2). The number of match engines executing simultaneously in the DRC coprocessor were increased to 6 to extend performance in a secure data transmission environment. Increasing to 6 match engines allows multiple simultaneous suspect searches to be conducted and enable a split of offender DNA database information into multiple subsets (based on ethnicity, geography, or other factors). Additionally, expanding the number of DRC coprocessors to 3 in the server linearly increased the number of simultaneous searches and reduced the time to complete one complex familial DNA search. These enhancements to the system resulted in a robust computational system for deriving important genetic intelligence from DNA data.

The overall architecture of the system and key attributes are outlined in Figure 2. This streamlined architecture sets the basis for secure information exchange and future expansion. Data pushed to the servers for processing is transferred to, and resides on, the server in an encrypted format. Data is deleted once processed. No application software is required on the client side.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

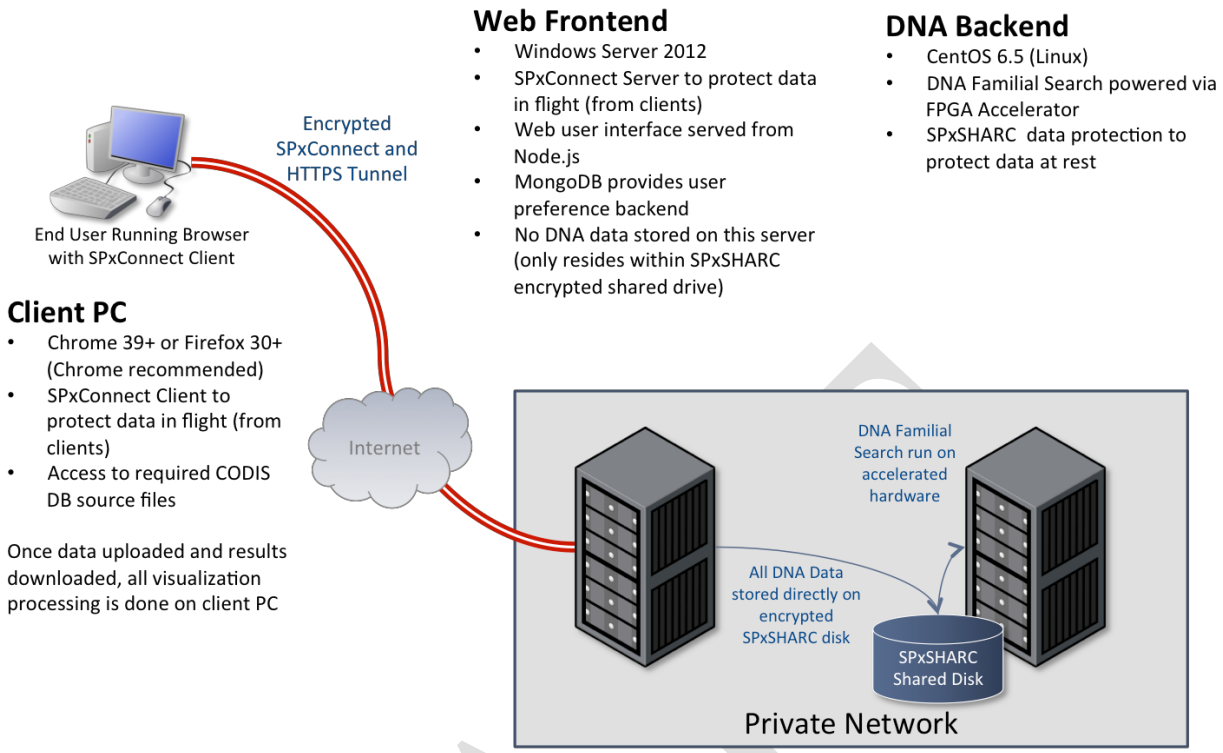

Figure 2. Overall Familial DNA search system architecture.

Strong encryption technologies from Security First were integrated and ensure secure upward and downward genetic data flow (Figure 3). DRC enhanced the Security First SecureParser™ (SFC) capabilities. There are two components to this; data connection capability as a front end and a secure data storage system as a back end. The data connection capability involves the use of a tightly encrypted communications link between the remote user site and the host system. This product is called SPxConnect™. The secure data storage system is implemented using the SFC SPxBitFiler™ product, now called SPxSHARC™. SpxConnect® encryption uses cryptographic splitting technology and is National Security Agency Suite B compliant and certified to Federal Information Processing Standard 140-2.

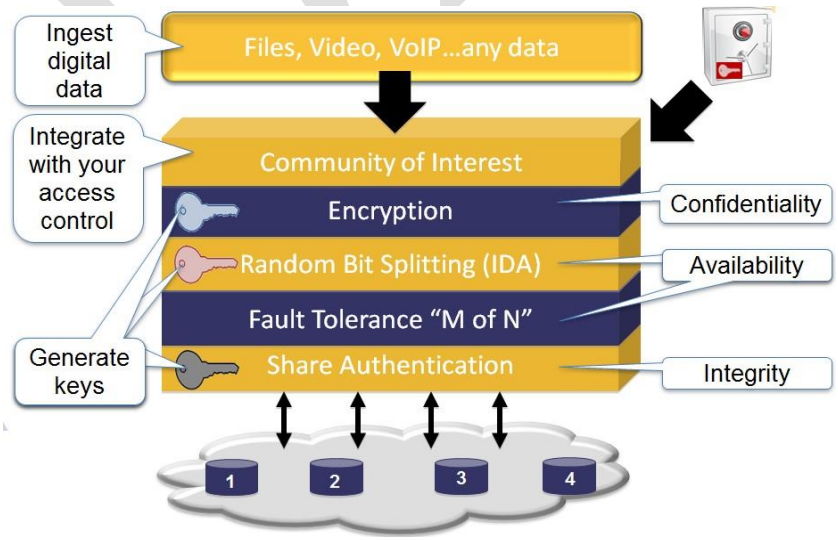

Figure 3. Security First encryption.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

DRC collaborated with Solidyn and developed the browser-based front-end (GUI) to the Denver Crime Lab/DRC DNA Familial Search Algorithm system. The GUI-based front end delivers user-based authentication, secure file transfer of CODIS profiles in XML format to the server and secure data management on the server side. Agencies can customize algorithm execution options to enhance search results. Results are returned for client-side data visualization and display with the option to download the results on their local workstation/laptop. All search activity is captured with detailed audit logging. Agencies are only required to load a browser security feature on their desktop to utilize the web based search system

From a visualization perspective, the analyst is provided with a user friendly layout with a wide array of visualization methods. This includes a basic user interface layout, as depicted in Figure 4. The analyst can pick from a menu of visualization tools and arrange them in the multi-frame layout. The user can drag the same visualization into multiple places, allowing different "views" of the data with the same plot type. Dragging visualization on top of an existing visualization will replace its contents with the new visualization (selections and filtering will be persisted).

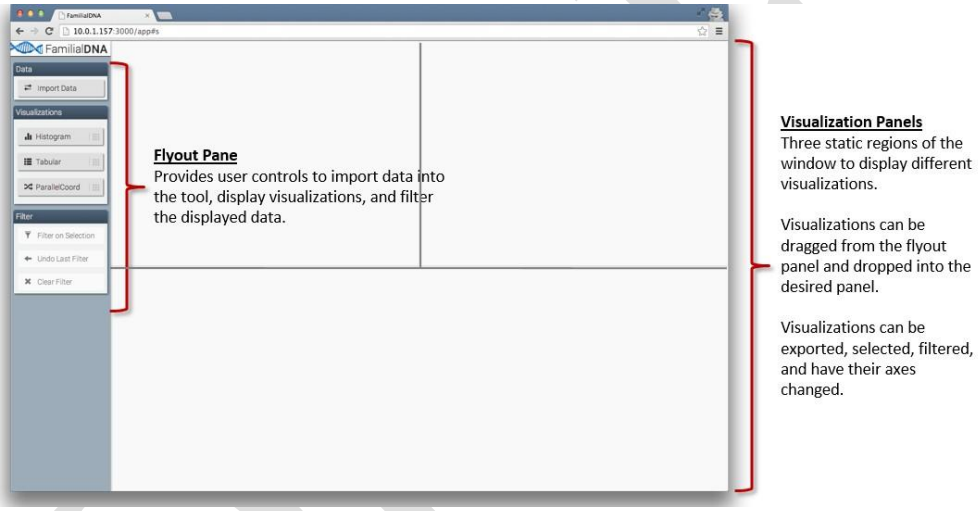

Figure 4. Familial DNA search GUI layout with visualization panels.

The analyst can chose from histograms, tabular, and parallel coordinate displays. Data sets can be filtered by data type and sorted by each value. Data can be filtered via cursor selection. This will hide and/or highlight the corresponding data in all the other visualization panels. Additional filters can be applied, and each excludes more data. Each filter can be cleared individually using the "Undo Last Filter" button. The filter counter will decrease after each filter is undone. All filters can be cleared using the "Clear Filter" button. The filtering and rendering is accomplished in real-time to maximize the analyst's efficiency. Sample displays are shown in Figures 5, 6, 7, and 8.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

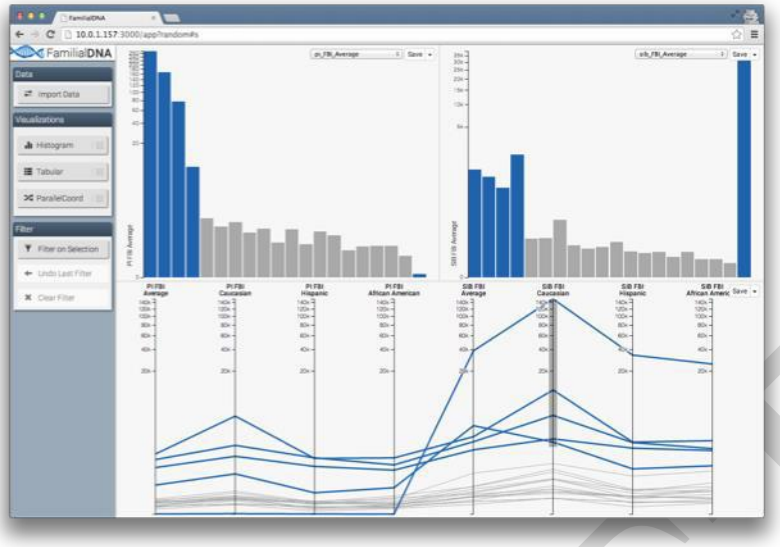

Figure 5. Sample of histogram and parallel coordinate visualization of search results.

- **Histogram**
  - Plots a basic histogram with a logarithmic axis.
  - Column of data to be plotted can be selected with the dropdown menu in the upper right-hand corner
  - Bars display a tooltip when hovered over
  - Updated in real time when other data is filtered
- **Selecting Data**
  - Individual bars can be selected by clicking on them
  - Contiguous bars can be selected by clicking and dragging along the bottom of the plot

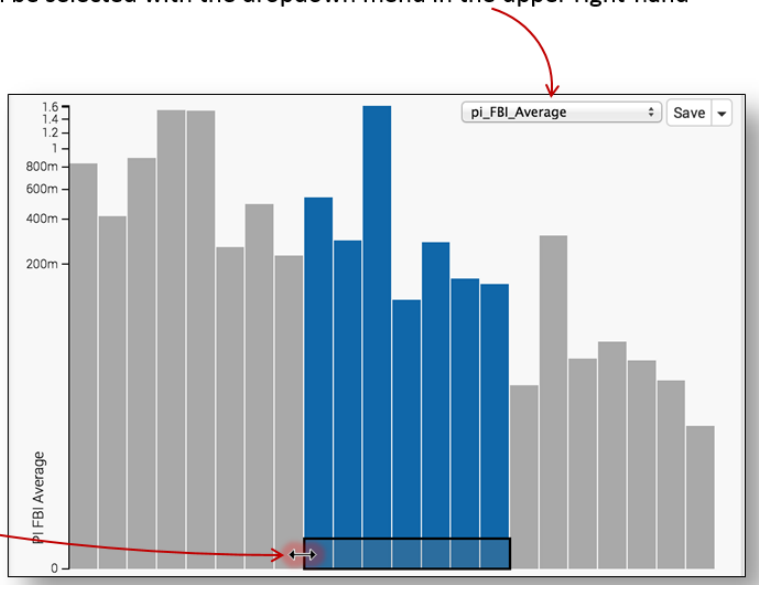

Figure 6. Options for histogram visualization of search results.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

- **Tabular Display**
  - Simple table of values contained in the results
  - Data can be exported as a CSV (Comma Separated Values) file by selecting the Save menu
  - Each column can be sorted by clicking the small up/down arrows in the headers
  - Reflects the full (unfiltered) data set – selections and filters do not modify this plot
  - Results are paginated and can be paged through with the controls at the bottom

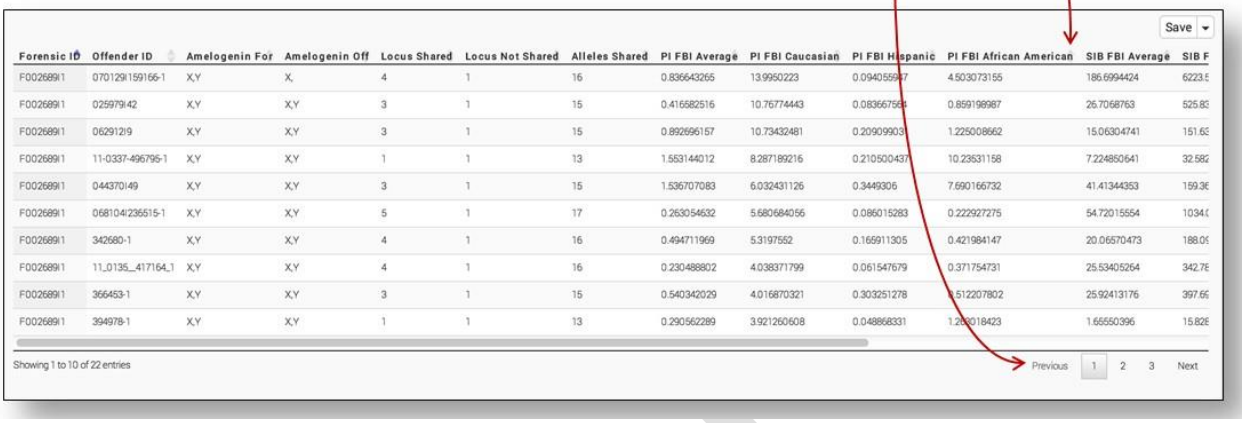

Figure 7. Options for tabular display of search results.

- **Parallel Coordinates**
  - Each line represents a single Offender + Forensic row of data
  - Shows all data plotted on multiple sets of logarithmic axes
- **Selecting Data**
  - Data lines can be selected by clicking on an axis and dragging a selection box. Selection refinement can be performed by clicking and dragging on additional axis

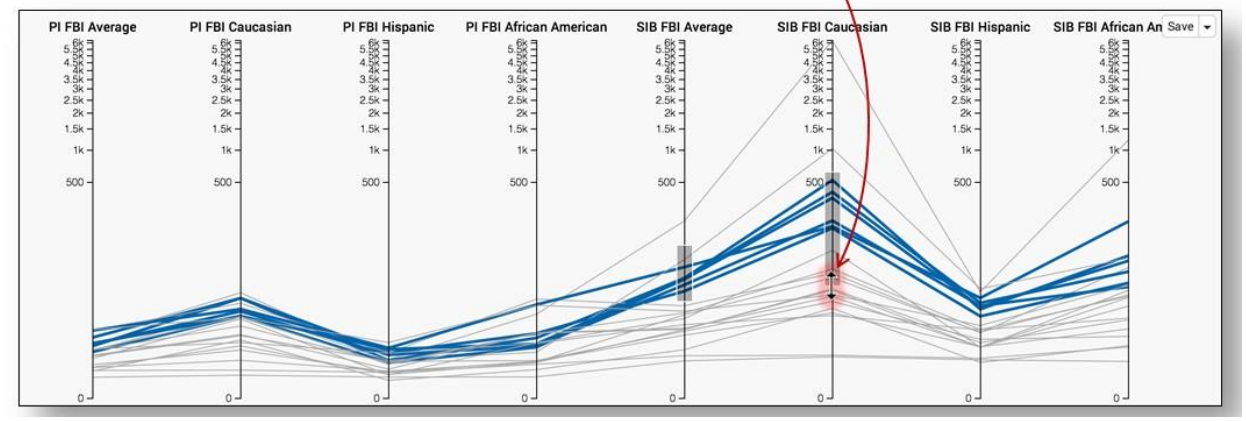

Figure 8. Options for parallel coordinates visualization of search results.

The GUI has undergone extensive testing of the security features from both local and remote sites with no issues found. The web-based search algorithm was validated. Several mock familial searches were completed on the local version and web based system. One-to-many searches with varying number of offender profiles, many-to-many searches, and setting various thresholds were used to ensure correct performance of the web system. Upload and search speeds

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

were also documented. LR values calculated by both systems were compared to ensure they matched to 6-7 significant digits with the percentage difference being less than $1x10^{-5}$ % .

Baseline performance of the EMR/EKR compared to the DLR method results

Results of the baseline performance of the EMR/EKR method are presented in Table 8. The true positive rate for DLR averaged 55.1% for parent-child familial pairs and 36% for sibling pairs, which was higher than the EMR/EKR simulations that averaged 0.3% for parent-child familial pairs and 20.4% for siblings. The EMR/EKR false negative rates were higher than the DLR with 99.7% for parent-child pairs and 79.6% for sibling pair simulations. The false positive rates for the EMR/EKR method averaged 5.9E-7% for parent-child pairs and 1.5E-4% for sibling pairs and were smaller than the DLR method (5.6E-3% for parent-child pairs and 5.0E-4% for siblings). These results indicate that the DLR method is substantially more effective in locating true relatives in the results versus the EMR/EKR method. Although the EMR/EKR method is more effective at controlling the false positive rate, it comes at the cost of a higher false negative rate.

Table 8. Count of true positives, false negatives, and false positives from each simulation. True Positive Rate = TP/(TP+FN), False Negative Rate = FN/(FN+TP) and False Positive Rate =FP/(FP+TN).

Simulation	True Positives (TP)	False Negatives(FN)	False Positives(FP)	True Positive Rate	False Negative Rate	False Positive Rate
Parent-Child EMR/EKR Original	3	997	6	0.3%	99.7%	5.9E-7%
Sibling Pair EMR/EKR Original	204	796	1487	20.4%	79.6%	1.5E-4%
Parent-Child EMR/EKR Low Stringency	157	843	5796	15.7%	84.3%	6.0E-4%
Sibling Pair EMR/EKR Low Stringency	302	698	728639	30.2%	69.8%	0.073%
Parent-Child DLR	551	449	56259	55.1%	44.9%	5.6E-3%
Sibling Pair DLR	360	640	5317	36%	64%	5.0E-4%

Table 9. Number of simulated pairs found in the top 100 of the ranked list.

Simulation	Number of pairs in Top 100 of ranked list
EMR/EKR Parent-Child Original	4
EMR/EKR Parent-Child with Low Stringency	21
DLR Parent-Child	44.5
EMR/EKR Sibling Original	34.5
EMR/EKR Sibling with Low Stringency	6
DLR Sibling	62

The EMR/EKR parent-child simulations averaged four simulated familial pairs in the top 100 (Table 9) of the ranked result list. The siblings averaged 34.5 simulated familial pairs in the top 100. With the DLR method there was a 79.7% increase of simulated sibling pairs located in the top 100 when compared to the EMR/EKR and a 1012.5% increase for the parent-child familial comparison.

Low Stringency Matches- Modified EMR/EKR

The EMR/EKR calculations exclude low stringency loci matches (10,**12** to **12**,14) and ignores valuable genetic information. The impact of this exclusion was explored. The average number of low stringency matches (Table 10) in the 500 parent-child pairs was ~6.5 (at least 6 loci) or 46% of the STR profile information was excluded from the EMR/EKR calculations. The sibling pairs had ~4.2 low stringency matches or at least 4 loci (30%) excluded from likelihood ratio calculations.

Table 10. Average Number of Low Stringency Matches.

Simulation	Average Low Stringency Matches
Parent-Child 1	6.486
Parent-Child 2	6.464
Sibling 1	4.286
Sibling 2	4.174
Real Denver	3.844

The result lists from the modified EMR/EKR simulations were also ranked and the number of simulated pairs found in the top 100 was tabulated (Table 9). An average of 21 parent-child pairs were found, showing a 425% increase in detection of true relative pairs in the top 100 pairs compared to the unmodified EMR/EKR results. The modified code had a negative impact on the sibling simulations. An average of 6 sibling pairs was found in the top 100 (an 82% decrease). A comparison of the results from the modified EMR/EKR to the DLR still indicated that the DLR method detects more pairs in the top 100 paired results list. The additional calculation did increase the EMR/EKR true positive rate by 15.4% for the parent-child and 9.8% for the sibling simulations. The false negative rate was reduced to 84.3% for the parent-child and 69.8% for sibling pairs (Table 8).

Dividing the EMR/EKR value by the database size and suggested threshold

The SWGDAM recommendation does not discuss the reasons for dividing EMR/EKR by the database size; we assume it is an attempt to reduce false positive findings. SWGDAM recommends that at least one of the EMR/EKR values must be greater than or equal to 1 with remaining values greater than 0.1, therefore indicating that likelihood ratios values must be greater than the database size. No supporting information was provided to justify this recommended threshold.

The combination of the threshold and the suggestion to divide EMR/EKR by the database size produced a higher false negative rate (99.7%) compared to the DLR method (44.9 %). 997 false negatives occurred in the parent-child simulations with only three pairs meeting the SWGDAM threshold. This can be explained by examining the EMR/EKR values prior to division by database size (1,000,500 in all simulations). The three true positives all had EMR/EKR values greater than 1,600,000 with all remaining being lower than 800,000. The maximum EMR/EKR values are presented in Table 11. The sibling simulations had 796 false negatives (Table 8).

Table 11. The maximum EMR/EKR values for each simulation calculated without division by database size.

Simulation	Maximum EMR	Maximum EKR
Parent-Child 1	162,039	782,999
Parent-Child 2	348,178	1,857,980
Siblings 1	10,143,359	1,340,762,944
Siblings 2	125,679,281.6	41,150,394,688

Comparison of EMR/EKR to DLR method without division by database size at a matching false positive rate

Matching the false positive rate of the DLR method (5.6E-3%) for the parent-child simulation, a threshold of 13,750 (without division by database size) was used for the EMR/EKR. Results of this simulation can be seen in Table 8. An additional 177 simulated parent-child pairs were found in the results with 323 false negatives. The results were also ranked by the maximum EMR/EKR and the number of simulated pairs detected in the top 100 was tallied and one additional pair was found in the top 100 for a total of five. The false negative rate decreased to 88.2 % in the parent-child simulation; however, this was higher than the false negative rate for DLR (44.9%) The true positive rate increased to 11.8% for EMR/EKR with the DLR at 55.1% (Figure 9 and Table 8).

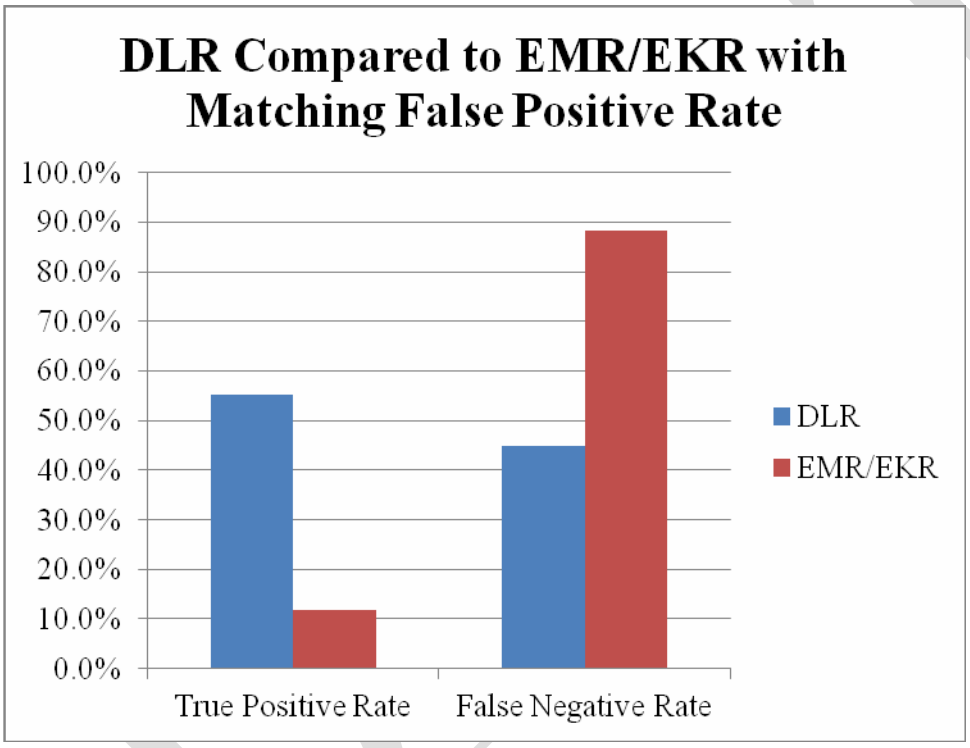

Figure 9. True Positive and False Negative rate of DLR and EMR/EKR methods for the Parent-Child simulation with matching False Positive rate.

Power, Specificity and Sensitivity

The power of the five statistics (LRMAX1, LRMAX2, LRLAF, LRWAVG and LRAVG) was computed for both parent-offspring pairs and sibling-sibling pairs. Within each group of comparisons, the results are highly correlated (Table 12) with the exception of the LRLAF being more distinct. Correlations were calculated for a random 1,000,000 simulated pairs.

Table 12. Power comparison matrix of the five statistics where PC= parent-child and SS = sibling.

		PC					SS				
		LRMAX1	LRLAF	LRMAX2	LRWAVG	LRAVG	LRMAX1	LRLAF	LRMAX2	LRWAVG	LRAVG
PC	LRMAX1		**0.64**	**1.00**	**0.97**	**0.99**	0.12	0.15	0.12	0.17	0.14

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

	LRLAF	**0.64**		**0.64**	**0.76**	**0.67**	0.15	0.31	0.15	0.22	0.18
	LRMAX2	**1.00**	**0.64**		**0.97**	**0.99**	0.12	0.15	0.12	0.17	0.14
	LRWAVG	**0.97**	**0.76**	**0.97**		**0.99**	0.16	0.23	0.16	0.23	0.19
	LRAVG	**0.99**	**0.67**	**0.99**	**0.99**		0.14	0.19	0.14	0.21	0.17
SS	LRMAX1	0.12	0.15	0.12	0.16	0.14		**0.72**	**1.00**	**0.98**	**0.99**
	LRLAF	0.15	0.31	0.15	0.23	0.19	**0.72**		**0.72**	**0.82**	**0.78**
	LRMAX2	0.12	0.15	0.12	0.16	0.14	**1.00**	**0.72**		**0.98**	**0.99**
	LRWAVG	0.17	0.22	0.17	0.23	0.21	**0.98**	**0.82**	**0.98**		**0.99**
	LRAVG	0.14	0.18	0.14	0.19	0.17	**0.99**	**0.78**	**0.99**	**0.99**	

Thresholds (Table 13) were calculated for the false positive rate with a 99.7% confidence interval for each of the five approaches. 5 million simulated related pairs were used to estimate the power.

Table 13. Follow up thresholds and power for each of the five test statistics. PC= Parent-child and SS= Siblings.

		Threshold for FPR= 8.5×10^{-5}	L99.7%	U99.7%	Power for FPR= 8.5×10^{-5}
PC	LRMAX1	11705	11385	12042	0.0001286
	LRLAF	1723	1685	1766	0.0001270
	LRMAX2	11717	11393	12057	0.0001288
	LRWAVG	4246	4145	4356	0.0001312
	LRAVG	5361	5221	5506	0.0001294
SS	LRMAX1	10716	10488	10958	0.0000142
	LRLAF	1155	1134	1179	0.0008902
	LRMAX2	10745	10517	10996	0.0000200
	LRWAVG	3391	3326	3460	0.0003836
	LRAVG	4417	4324	4507	0.0001356

The LRMAX2 statistic is closest to the current evaluation method used by the Denver Crime Laboratory. DLR looks at the maximum LR value calculated using U.S. Caucasians, U.S. African-Americans, and U.S. Hispanics as well as an average frequency denoted by FBI Average (Budowle at al. 1999). The identified thresholds for the FPR do not match with the current DLR threshold of 100,000. This is due to the difference in simulating samples from the population. In the power analysis, a mixed population is simulated using race specific allele frequencies within sub populations. Simulation studies previously performed by the Denver Crime Laboratory used the local allele frequency distribution representing a single randomly mating population and each pair of individuals is independent.

The distribution of the LR comparison differs based on the race of the two individuals being compared (p-value=0.000271 for 1,000,000 randomly drawn pairs). Pairwise comparisons (Figure 12, Tables 14, 15) shows the LRs for an HH comparison are significantly larger than the LRs for an EE, AE, and an EH comparison. Below the results of a Tukey's HSD procedure are given with a family-wise error rate of $\alpha = 0.05$.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

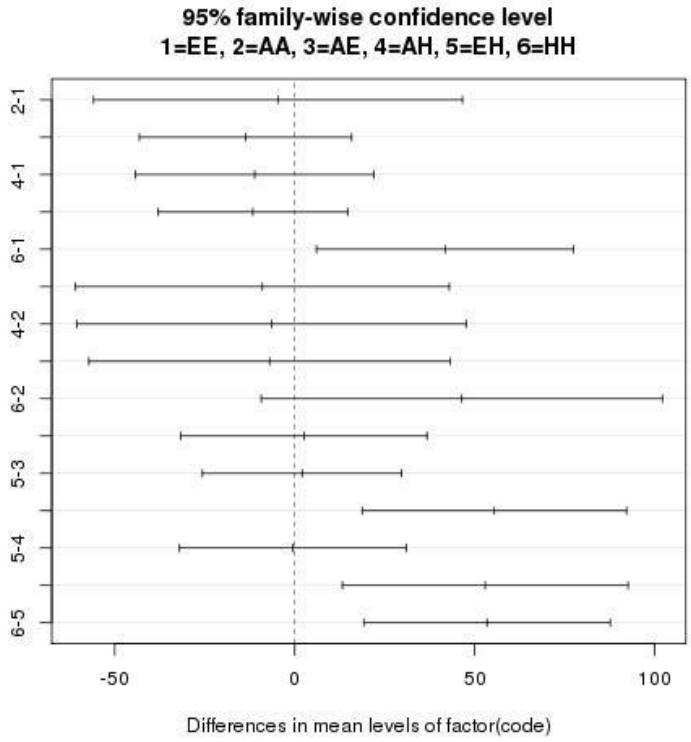

Figure 12. Differences in mean levels with pair wise comparisons. EE= European to European, AA= African American to African American, AE = African American to European, AH= African American to Hispanic, EH= European to Hispanic, and HH= Hispanic to Hispanic.

Table 14. Adjusted P values for the pair wise comparisons.

Groups	LR Difference	Lower	Upper	Adjusted P value
2-1	-4.6596	-55.8867	46.56753	0.999842
3-1	-13.6809	-43.1598	15.79807	0.772629
4-1	-11.1473	-44.2373	21.94271	0.930448
5-1	-11.6508	-37.9804	14.67879	0.806259
6-1	41.74404	6.03595	77.45213	0.011168
3-2	-9.02125	-60.957	42.91453	0.996364
4-2	-6.48771	-60.555	47.57954	0.999386
5-2	-6.99121	-57.2064	43.22399	0.998738
6-2	46.40364	-9.3044	102.1117	0.165455
4-3	2.533537	-31.6433	36.71036	0.999943
5-3	2.03004	-25.6531	29.71313	0.999946
6-3	55.42489	18.70741	92.14237	0.000245
5-4	-0.5035	-32.0042	30.99724	1
6-4	52.89135	13.21623	92.56648	0.002015
6-5	53.39485	19.15432	87.63539	0.000129

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 15. Summary of LRs per compared ethnic group. EE= European to European, AA= African American to African American, AE = African American to European, AH= African American to Hispanic, EH= European to Hispanic, and HH= Hispanic to Hispanic.

Group		Proportion of Population	Mean	Var	Quantile for $FPR = 5 \times 10^{-5}$	Quantile for $FPR = 8.5 \times 10^{-5}$
1	EE	0.23	14.57936	2053227	1222.994	36718.59
2	AA	0.04	9.919757	1438737	165.6693	8765.884
3	AE	0.19	0.89851	19878.83	5.664872	1482.488
4	AH	0.13	3.432047	877505.9	11.00877	2425.924
5	EH	0.31	2.92855	135410.2	124.5585	6068.769
6	**HH**	**0.10**	**56.3234**	**101372969**	543.3944	35957.87

IV. Conclusions

Results indicate at a comparable false positive rate, the current implementation of the DLR method detected more true positives than the SWGDAM recommended statistics for familial searching (Figure 10 and Table 8). The DLR method had larger true positive rates, though this came at a cost of higher false positive rates. While the use of the SWGDAM recommendation more stringently controls the false positive rate, this is accomplished with high false-negative rates (Table 8). The trade off of more false positives with a higher true positive rate can be prudent if a link can be found to the true perpetrator of the crime in question using other downstream verification steps such as Y-STR genotyping or other investigative methods.

The DLR method makes use of all the available genetic information in the forensic profile being searched while the SWGDAM recommendation excludes loci with low stringency matches (10,**12** to **12**,13). The use of low stringency matches in a familial search has been proven successful as a decade old rape/homicide was recently solved in France using this approach (Phan-Hoai et al. 2014). A CODIS low stringency search was performed. The 18 loci forensic profile was searched against 1.8 million convicted offender profiles, resulting in one individual matching at least one allele at each locus. The resulting profile was further tested with Y-STR analysis that matched to the crime scene profile, demonstrating a possible familial linkage. Subsequently, a single male with two sons was identified. The father and youngest son were eliminated due to their age at the time of the crime, which placed the focus on the older son. The older son had died a few months after the crime and in order to complete DNA analysis, the body was exhumed. The DNA results identified the older son as the homicide suspect. Our low stringency simulations demonstrate that the inclusion of low stringency matches improves the EMR/EKR true positive rate from 0.3% to 15.7% for parent-child searches, and for siblings from 20.4% to 30.2% (Table 3).

Large DNA database searches are likely to yield a higher percent of false positive relative pairs. The SWGDAM recommendations attempted to take into account that a large database search was conducted by recommending that the EMR and EKR values be divided by the size of the database. Additionally, the recommended threshold is only achieved if the EMR/EKR is greater than the database size. Of the 51 state CODIS databases, including Washington D.C. (http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics), at this time there is one state that has more than 1,000,000 offender profiles, an additional four that have more than 500,000, 23 that have more than 100,000, nine with 50,000, five with 25,000, eight with 10,000, and one with less than 10,000. With this large distribution of database sizes, a familial pair in one state database may meet the threshold to be considered, while if the same familial pair occurred in another state database it may not be discovered. Further, the EKR and EMR measure two different likelihood ratios and a statistically relevant threshold should be developed for each.

For a mixed population, the Hispanic comparison is going to lead to a higher threshold than is needed for other comparisons. This will result in differential power among the comparisons. Given this, a threshold should be derived for each of the four major ethnic population groups, such that they are treated to the same false positive rate.

The results presented here clearly demonstrate that evaluation of all genetic information increases the identification of true familial pairs when conducting familial searches. Our data also indicates that adjustment of likelihood ratios to account for database size can be counterproductive. Selecting a statistically relevant threshold also plays a critical role in the determination of familial matches and selecting one based on a specific false positive rate should be pursued more formally. The current implementation of familial searching ranking statistics used by the Denver Police Crime Laboratory is recommended for maximizing useful results.

Implications of policy and practice:

The use of familial searching is a powerful tool that can increase the number of investigative leads and potential suspects identified. Familial searching takes advantage of existing DNA technology and DNA profiles. Bieber et al. (2006) indicated that familial searches could increase cold hit rates by 40%. An automated web-based familial search system is now available to any agency wanting to conduct these types of searches as a result of the work conducted with this funding. Forensic and offender profiles (XML format) are exported from CODIS software and are uploaded to the search system. Agencies can determine which allele frequencies and follow up likelihood ratio statistical thresholds to use for the search. Profiles can be stored in the system for performance of multiple searches and reduce the requirement to upload the offender profiles for each search. The system can perform many-to-many and one-to many searches. An agency could compare forensic unknowns against themselves or perform a population to population search. With the availability of a web system, searches could be performed between agencies, states, or countries.

Our research indicates that additional investigation should be conducted to determine thresholds needed to optimize statistical power. These thresholds could be based on a specific false positive rate or be determined for each of the four major ethnic population groups. Most agencies focus on results in a ranked list for additional Y-STR follow up based on the capacity of their laboratory (top 100 or top 168).

Policy and practice for the use of a web-based search system have not been developed. Transportation of DNA profile information over the internet for familial searching has never been done and is being explored by our group and will be reported when complete. Policies regarding the security and protection of this data will need to be developed and understood by the forensic science community. Our system has been designed with a solution to this concern with the implementation of SPXConnect, a distributed encryption solution that is National Security Agency Suite B compliant and certified to Federal Information Processing Standard 140-2.

Implications for further research:

This research indicates that additional investigation is needed to determine the best follow up threshold values for implementation. Thresholds that improve true positive rates or meet a specific false positive rate should be considered. Our results also indicate that a threshold should be derived for each of the four major ethnic population group allele frequencies being used. An additional approach being considered is simulating all possible related profiles based on the evidence profile being searched and determining the range of LR.

The web-based system could be further enhanced by moving the offender databases into the internal DRAM memory on the Accelerator and spreading the data across multiple Accelerator's and servers, eliminating the I/O overhead of reading the databases for every search. An additional enhancement would be moving the filtering of the search results from the software module to inside the Accelerator (hardware). The Accelerator would only output requested data and leave the software module to deal with the actual data the user desires. The possibilities of various visualizations on the web interface are endless. One could consider a cluster approach to visualize entire families if DNA profiles exist for an entire village, town, region, or target population.

V. References:

Anderson AL, Weir, B.S.2006. It was one of my brothers. *International Journal of Legal Medicine*. **120**:95-104.

Anderson GB.2008. DNA partial match (crime scene DNA profile to offender) policy. California Department of Justice-Division of Law Enforcement.

Balding D, Krawczak M, Buckleton J, Curran J. 2013. Decision-making in familial database searching: KI alone or not alone?. *Forensic Sci Int-Gen* **7**:52-54.

Balding DJ, Nichols A.1994. DNA profile match probability calculation : how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**:125–140.

Bieber F, Brenner C, Lazer D. 2006. Finding Criminals Through DNA of Their Relatives. *Science* **312**:1315-1316.

Brenner CH, Weir BS.2003. Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical Population Biology*. **63(3)**:173-178.

Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM. 1999. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *J Forensic Sci* **44(6):**1277-1286.

Budowle B, Shea B, Niezgoda S, Chakraborty R. 2001. CODIS STR loci data from 41 sample populations. *J Forensic Sci* **46(3):**453-489.

Butler JM. 2009. Fundamentals of Forensic DNA Typing. Academic Press.

Butler JM. 2014. Probabilistic Strategies for Familial DNA Searching. *J R Stat Soc Ser C* **63(3):**361-384.

Casella G, and Berger RL.2001. Statistical Inference, Second Edition. Monterey, CA: Duxbury Press

Coalition Colorado Criminal Justice Reform. 2010 colorado quick facts: http://ccjrc.org/pdf/2010_Colorado_Quick_Facts.pdf

CODIS–NDIS Statistics, The Federal Bureau of Investigation; http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-statistics.

CODIS Popstats Program, Version 7.0 with Service Pack 4 provided by the Federal Bureau of Investigation.

Correctional populations in the United States, 1996. U.S. Department of Justice; 1999. NCJ 170013

Curran JM, Buckleton JS.2008. Effectiveness of familial searches. *Science and Justice*.**48(4)**:164-167.

Doumas D, Margolin, G, John, RS. 1994. The intergenerational transmission of aggression across three generations. *J Fam Violence* ***9***:157-175.

Dudbridge F.2007 Unphased. 3.0.12 ed. Cambridge, UK.

Durose MR.2008. Census of Publically Funded Forensic Crime Laboratories, 2005. Washington D.C.: Bureau of Justice Statistics

1996. National Research Council. Committee on DNA Technology in Forensic Science. The Evaluation of Forensic DNA Evidence. National Academy Press.

FBI-CODIS and NDIS Fact Sheet, The Federal Bureau of Investigation; http://www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet.

Ge J, Chakraborty R, Eisenberg A, Budowle B. 2011. Comparison of Familial DNA Database Searching Strategies. *J Forensic Sci* **56(6):**1448-1456.

Hill CR, Duewer DL, Kline MC, Coble MD, Butler JM. 2013. U.S. Population Data for 29 Autosomal STR Loci. *Forensic Sci Int Genet* **7(3):**e82-83

2009. SWGDAM Recommendations for the FBI Director on the "Interim plan for the release of information in the event of a 'partial Match' as NDIS". *Forensic Science Communications.* **11(4).**

JMP® 11, Statistical Discovery Program, SAS

Leclair B, Fregeau CJ, Bowen KL, Fourney RM.2005. Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair flight 111 disaster. *Journal of Forensic Sciences*. **49(5)**:939-53.

Li N, He G, Chen C. 2006.Distribution of STR locus DXS8027 polymorphism in the Han population in Qinba mountain areas. *Yi chuan Hereditas / Zhongguo yi chuan xue hui bian ji*. **28(3)**:273-278.

Morrissey MB, Wilson AJ.2010. Pedantics: an R Package for Pedigree-Based Genetic Simulation and Pedigree Manipulation, Characterization and Viewing. *Mol Ecol Resour* **10(4):**711–719.

Myers SP, Timken MD, Piucci ML, Sims GA, Greenwald MA, Weigand JJ, et al. 2010. Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach. *Forensic Science International: Genetics*.

Pham-Hoai P, Crispino F, Hampikian G. 2014. The First Successful Use of a Low Stringency Familial Match in a French Criminal Investigation. *J Forensic Sci* **59(3)**:816-819

Powers D. 2007. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *School of Informatics and Engineering Technical Report* **SIE-07-001**:1-24

Presciuttini S, Toni C, Tempestini E, Verdiani S, Casarino L, Spinetti I, De Stefano F, Domenici R, Bailey-Wilson J. 2002. Inferring relationships between pairs of individuals from locus heterozygosities. *BMC Genet* **3**:23.

Reid TM, Baird ML, Reid JP, Lee SC, Lee RF. 2008.Use of sibling pairs to determine the familial searching efficiency of forensic databases. *Forensic Science International: Genetics*.**2(4)**:340-342.

Slooten K, Meester R. 2014. Probabilistic Strategies for familial DNA searching. *J R Stat Soc: Series C* **63(3):**361-384.

VI. Dissemination of Research Findings:

International Symposium on Human Identification Conference, September 2014, Phoenix, Arizona, Poster Session:"*A Comparative Analysis of Likelihood Ratio Statistical Ranking to Expected Match Ratios and Estimated Kinship Ratios for Familial DNA Searching Applications."*

Webinar presentation\demonstration of web system, September 2015, *Familial DNA Database Search System: Hardware/Software Integration.*

International Symposium on Human Identification Conference, October 2015, Grapevine, Texas, Poster Session:"*Familial DNA Search System.*"

Manuscript pending submission.- Mountain B, Santorico S, LaBerge G. A Comparative Analysis of Likelihood Ratio Statistical Ranking to Expected Match Ratios and Estimated Kinship Ratios for Familial DNA searching Applications.