Title

Can simple population genetic models reconcile partial match frequencies observed in large forensic databases?

Journal

Journal of Genetics, 87(2)

ISSN

0022-1333

Author

Mueller, Laurence D

Publication Date

2008-08-01

DOI

10.1007/s12041-008-0016-4

PERSPECTIVES

Can simple population genetic models reconcile partial match frequencies observed in large forensic databases?

LAURENCE D. MUELLER*

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525, USA

A recent study of partial matches in the Arizona offender database of DNA profiles has revealed a large number of nine and ten locus matches. I use simple models that incorporate the product rule, population substructure, and relatedness to predict the expected number of matches in large databases. I find that there is a relatively narrow window of parameter values that can plausibly describe the Arizona results. Further research could help determine if the Arizona samples are congruent with some of the models presented here or whether fundamental assumptions for predicting these match frequencies requires adjustments.

Introduction

In 1994, the United States DNA identification Act gave the Federal Bureau of Investigation (FBI) authority to establish a national DNA index for law enforcement. The law also allows database samples to be used for 'a population statistics database'. The FBI implemented the Combined DNA Index System (CODIS) by establishing three levels of operation: the National DNA Index System (NDIS), State DNA Index System (SDIS) and the Local DNA Index System (LDIS). States have established criteria that determine whose DNA samples will be entered into their local and state databases; usually some type of criminal offense is required. A few states, like California, are permitted to collect DNA samples from people who have been arrested even if the charges are later dismissed or the person is found not guilty of the charged offense. Most states use the 13 core set of short tandem repeat loci to develop genetic profiles for these offender databases. The names of these loci are: D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, TH01, TPOX, CSF1PO and D16S539

*For correspondence. E-mail: ldmuelle@uci.edu.

Keywords. DNA typing; offender database; population substructure.

Most forensic DNA databases that are used to estimate allele frequencies and test for independence within and between loci consist of a few hundred people per racial group (Budowle *et al*. 1999; Cherni *et al*. 2005). Offender databases, on the contrary, are quite large. The NDIS consisted of 3,866,259 profiles as of November 2006. In individual states, the size of these databases range from a high of 602,338 in California to a low of 451 in Rhode Island.

Access to databases

The Federal DNA Identification Act of 1994 (42 U.S.C. §14132) specifies that these databases be '(3) maintained by federal, state and local criminal agencies . . . pursuant to rules that allow disclosures of stored DNA samples and DNA analyses only - (D) if personally identifiable information is removed, for population statistics databases for identification research and protocol development purposes, or for quality control purposes'. 'Identification research' and 'quality control' could mean many things including research to verify the accuracy of definitions of uniqueness or research to determine the reliability of statistical models used to determine DNA profile frequencies, etc.

In recent years there have been several examples of offender database samples being used for this type of research. (McElfresh and Kim 2000) utilized offender profiles from Virginia and North Carolina to estimate the minimum number of loci required to narrow a database search to a single individual. Their results were presented at an annual meeting sponsored by the Promega Corporation. Troyer *et al*. (2001) presented observations of 9-locus matches between unrelated people in the Arizona offender database. More recently Frank *et al*. (2006), used samples from the Illinois offender database to construct a Y-STR database. Frank listed all the observed Y-STR profiles in their *Journal of Forensic Science* paper albeit with all 'personally identifiable information' removed.

Outside the United States there is also precedent for information in offender databases to be made available to outside scientists for review. The offender database from the Victoria Police Forensic Services Centre has been available to outside scientists to review. This database and a false match found during a search of this database are discussed in the 2006 report of the State's coroner inquest into the death of Jaidyn Raymond Leskie (http://darwin.bio.uci.edu/~mueller/pdf/leskie_decision.pdf). These data were analysed by Weir (2004). Weir took 14768 9-locus profiles and compared all possible pairs of profiles, a total of 109,039,528 comparisons, and determined for each pair the number of loci that matched and the number that showed a partial match. He compared the number of matches at 1 and 5 loci to the number expected under different levels of population substructure. Weir could only examine at most 5 locus matches, since there were very few matches at 6 or more loci to do rigorous statistical analysis.

Roadblocks to database access

More recently, in 2005, informal requests to the Arizona Department of Public Safety (DPS) for information about additional databases searches conducted by Kathryn Troyer were rebuffed. As the result of a court order in November 2005 the Arizona DPS reported the results of a search of their offender database consisting of 65493 profiles at the 13 CODIS core set of STR loci. There were 122 pairs of individual who matched at 9 loci out of 13, 20 pairs matched at 10 loci, 1 matched at 11 loci and 1 matched at 12 loci.

Subsequently, in California, Maryland, Illinois, the District of Columbia, and several other states additional requests from the defense bar for information similar to that provided by the Arizona DPS have been made. In some instances the requests have been for copies of all profiles in the State Offender database. The requests have been uniformly resisted by state and federal officials. A number of common reasons for this resistance appear frequently in the documents filed by state and federal officials which I review below.

(i) Release of even the numbers of matches at 9 or more loci would be in violation of the Federal DNA Identification Act or the local state equivalent. Clearly, the legal system will have to interpret what is meant by the wording in the enabling legislation. However, the scientific interpretations of those words do not seem to preclude giving scientists access to anonymous profiles for the study of a large variety of statistical and population problems. The use of database information by McElfresh and Kim (2000), Troyer *et al*. (2001), and Frank *et al*. (2006) discussed previously are all consistent with this interpretation.

(ii) Doing such searches would tie up forensic lab computers for an extraordinary amount of time and prevent important database searches. There are many possible solutions to this problem like using a second computer, giving the profiles to outside scientists who can conduct the search, etc.

(iii) Laboratory personnel do not have the technical expertise to carry out such searches. Making the anonymous profiles available to outside scientists would again relieve the forensic scientists from carrying out these investigations.

(iv) Doing these searches or providing outside scientists with the genetic profiles in the database would violate the local labs' memorandum of understanding with the FBI. The penalty for such violations would be removal from the CODIS system. To date, three different state laboratories have turned over search results to the defense as a result of court ordered discovery requests. Additionally, as mentioned previously, several laboratories have utilized offender database profiles or offender database samples for material in scientific publications or presentations. None of these laboratories has been removed from CODIS.

(v) Other large databases are publicly available which scientists could use for any conceivable research. Examples of such databases are the FBI databases used to estimate allele frequencies, the Australian offender database discussed previously, and a database of 17000 profiles produced by Orchid Biosciences (Einum and Scapetta 2004). The FBI databases altogether number only a few thousand individuals and thus could not be expected to show multiple matches at 9 loci or more. In fact even the largest Australian database does not show a large number of matches at more than five loci. The paper by Einum and Scarpetta (2004) was apparently never available to the public nor is it at the time this article was written. The original paper listed an illegitimate URL to download the database. The correct site (http://www.orchidbio.com/technology/publications.asp) does not have the raw data and it is apparently against Orchid Biosciences policy to send, via e-mail, individual copies of the database to interested scientists (Dr David Einum, personal communication). In any case, the Orchid Biosciences database is about the same size as the Australian database and much smaller than most offender databases and would have limited utility to study matches at a large number of loci.

(vi) Worthwhile research can not be accomplished with offender databases and thus should not be attempted. On the face of it, these arguments are wrong since scientists like Weir have already undertaken research programmes with these types of databases. This argument presupposes that there are scientists with infinite wisdom who can foresee all possible avenues of research. The history of science has shown that all scientists with this view have been proven wrong. At the very least it will be difficult for the legal system to determine the merits of these types of arguments and therefore access to these databases should not be barred to outside scientists based on this type of speculation.

What can be studied?

These large databases offer new ways of testing the predictions of rarity provided by simple models like the product rule. One such method is to study the frequency of partial matches. A partial match at 9 loci, for instance, would be a pair of individual who match at 9 CODIS loci out of the 13.

Table 1. The proportion of people in the simulated databases from each of five ethic groups.

Ethic group	African American	Caucasian	Hispanic	Navajo	Apache
Proportion	0.138	0.453	0.357	0.0425	0.0095

In a very small database of 100–200 people, the chance of finding pairs of individual who match more than 5 or 6 loci is very small. With a large database, like the Arizona offender database, we not only see matches between pairs of individual at almost all loci, but more importantly we see many matches at 9 and 10 loci that will permit some rigorous statistical tests.

Another avenue of research is to use offender databases as a means of investigating conditions that have been proposed for establishing uniqueness of a genetic profile (Budowle *et al*. 2001). If matching profiles between nonrelatives are found that exceed such set limits then they serve as counterexamples against a claim of uniqueness.

The goal of this study is to see if simple population genetic models can plausibly explain the observations in Arizona. If a plausible explanation can be developed then further research could focus on whether the parameters values used with these plausible explanations are realized in the Arizona populations. The simple population genetic models considered here will take into account, (i) population substructure, (ii) the presence of relatives, and (iii) variable ethnicity of the population. However, we assume that there is independence within and between loci among the smallest population sampling units. This study involves the use of extensive computer simulations that are described in detail.

Methods

Simulated databases consisted of either 65493 unrelated individual or S pairs of relatives plus 65493-2S unrelated individual. The relatives were either full sibs or parent offspring pairs. The parents of sibs were assumed to come from the same subpopulation. The relatives and unrelated individual were divided into five ethnic/racial populations: African Americans, Caucasians, Hispanics, Navajo and Apache. The proportions of each ethnic group are shown in table 1. These numbers were derived from statistics on the Arizona prison population since no direct estimate from the database exists (http://azcorrections.gov/reports/annual2003.pdf). The Arizona Department of Corrections report did not distinguish between different tribes of Native Americans so this group has been divided into the two tribes for which STR data exists, Navajo and Apache. The relative proportions of Navajo and Apache were derived from census data (http://cals.arizona.edu/edrp/tribes.html). The allele frequencies for all populations listed in table 1 came from published databases (Budowle *et al*. 2001).

Let the frequency of the mth ($m = 1, \ldots, l_n$) allele in the ith population ($i = 1, \ldots, 5$), at the nth locus ($n = 1, \ldots, 13$) be x^i_{mn}. Let the vector of allele frequencies at locus n and population i be, $X^i_n = \left(x^1_{ln}, \ldots, x^i_{l_n{}^n}\right)$. The simulations also allowed for population substructure. This was done by assuming each population consisted of four equally sized subgroups. Allele frequencies at locus n within subgroups were chosen from a Dirichlet distribution, $\widetilde{X}^i_n$, with shape parameter λ, whose kth element is, $\dfrac{(1-\theta)x^i_{kn}}{\theta}$, and θ is the inbreeding coefficient that measures population substructure (Evett and Weir 1998; Balding 2005). The random allele frequency vector within a subpopulation was generated with the R function *rdiric* in the VGAM module (version 2.40, www.r-project.org).

An outline of the procedure used to generate genetic profiles for the simulated databases is shown in figure 1. This method was used to generate a sample of 65493 13-locus

Table 2. The P values for all simulations.

θ	Number of sib pairs	P value
0.005	0	$< 2 \times 10^{-7*}$
	200	0.040
	2300	0.0020
	2400	$< 2 \times 10^{-7}$
	2500	0.096
	2600	0.036
	2700	0.015
0.01	0	4.8×10^{-6}
	1600	0.0013
	1800	0.12
	2000	0.071
	2100	0.049
	2300	0.015
	2500	0.082
	3000	0.009
0.015	0	$< 2 \times 10^{-7}$
	1000	0.021
	1400	9.2×10^{-6}
	1600	0.10
	1800	0.026
	2000	0.079
	2200	0.039
0.01	Parent–offspring pairs	
	2000	$< 2 \times 10^{-7}$
	2500	$< 2 \times 10^{-7}$
	4000	$< 2 \times 10^{-7}$
	6000	$< 2 \times 10^{-7}$
	19,000	$< 2 \times 10^{-7}$
	32,000	$< 2 \times 10^{-7}$

*The minimum P value that could be accurately determined was 2×10^{-7}

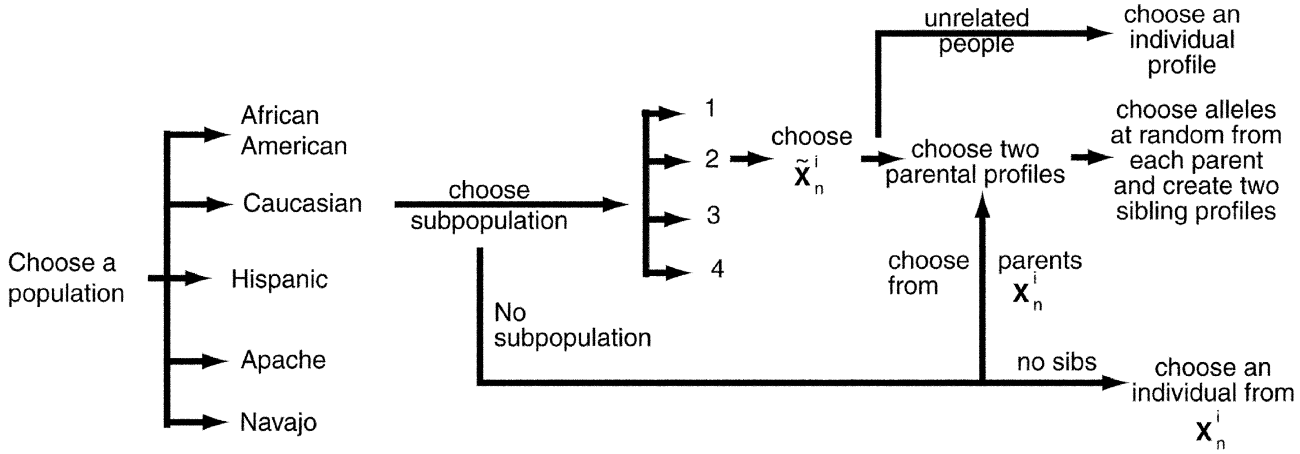

Figure 1. The sampling procedure used to generate genetic profiles in the simulated databases.

profiles. With this single database all possible pairs of individual were compared and the number of matching loci was determined for each pair. Ideally, this process could be repeated thousands of times and the results would then be used to construct an empirical distribution for the number of 9 and 10-locus matches for a particular model (Efron and Tibshirani 1993). These simulations unfortunately took too long to follow this ideal method. Instead I was forced to do a smaller set of simulations and then assume that the bivariate random vector of 9-locus and 10-locus matches had a bivariate normal distribution (BVN). For models that predict very small number of 10-locus matches, the BVN assumption is probably not a good one since the number of matches is constrained to be a positive number and the BVN is designed for continuous random variables. However, for models that yield predictions close to the Arizona observations the BVN assumption would appear to be a reasonable first approximation.

One simulated database produced one vector, Y, consisting of the number of matches at 9 (y_1) and 10 (y_2) loci. This process was repeated 10 times and from these 10 vectors I estimated the mean ($\overline{Y}$) and covariance matrix ($\sum$) for the vector Y. Under the assumption that Y has a bivariate normal distribution, e.g. $Y \sim BVN(\overline{Y}, \sum)$, we can then compute a 95% ellipsoid that encircles an area equal to 95% of the expected Y values for a particular model, using the R *ellipse* function. If the Arizona observation ($Y = 12, 220$) does not fall within a particular ellipsoid we conclude that the model associated with the ellipsoid is unlikely to be an accurate characterization of the Arizona population. Additionally, P values corresponding to the probability of observing the Arizona results or a more extreme results were calculated (table 2). This testing protocol requires that any specified population conditions simultaneously explain the number of matches at 9 and 10 loci. If some combination of θ, and relatives can correctly predict the number of 9-locus matches but not the number of 10-locus matches then it is an unsatisfactory explanation of the Arizona observations.

Results

No relatives and no population substructure

The simplest model assumes no population substructure within the populations that gave rise to the Arizona database and no relatives within the Arizona sample. This model (figure 2) predicts far too few matches at both 9 and 10 loci. The P value associated with the Arizona observation is $< 2\times10^{-7}$ (table 2). Even after allowing for the possibility that the MVN assumption is not precise, a sound conclusion is that the number of multi-locus matches in Arizona can not be predicted from a model of independence in the absence of substructure and relatives.

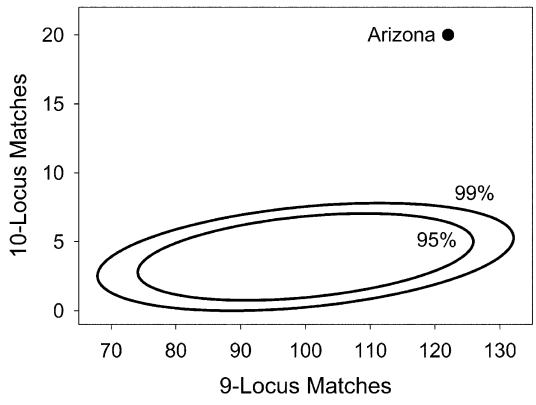

Figure 2. 95% and 99% confidence ellipsoids for simulations with no population substructure and no relatives. The mean for this model is at the centre of the ellipsoids ($Y = 100, 3.9$). The point labelled Arizona represents the combination of 9-locus and 10-locus matches seen in the Arizona offender database.

Population substructure and no relatives

In these simulations, I let the measure of population substructure θ vary over the range used in most forensic calculations, 0–0.03. The results (figure 3) show that as θ increases the number of matches at 9 and 10 loci increase but the num-

ber of matches at 9 loci increase faster than the number of matches at 10 loci. By the time $\theta = 0.03$ the predicted number of 9-locus matches is 154 but only 6, 10-locus matches are expected. If much higher θ values were used to boost the number of 10-locus matches there would continue to be too many 9-locus matches. Consequently, none of the models with substructure alone can adequately describe the Arizona results.

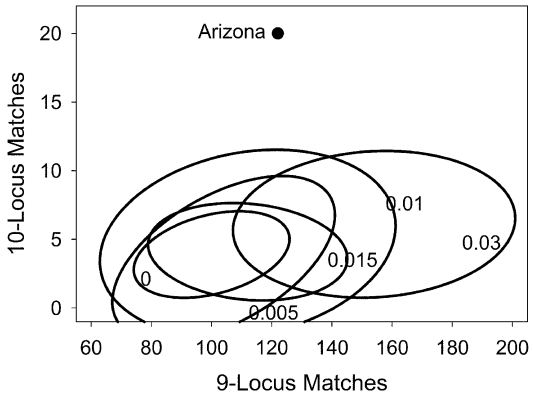

Figure 3. 95% confidence ellipsoids for simulations in which θ alone varied. The values of θ are placed on each ellipsoid.

At this point it is worth commenting on several other studies which have looked at this problem. Weir (2007) and Myers (2006) both derive analytic solutions for the expected number of matches in a large database. These studies looked at both differing values of θ and, in the case of Meyer's work, different numbers of siblings in the database. The difference between these studies and the present study is that Weir (2007) and Myers (2006) only computed the expected value for the number of matches, there is no estimate of the covariance matrix for number of matches at 9 and 10 loci. Weir (2007), for instance, derives expressions for the probability that two individuals will match at both the alleles at a locus, one of two alleles, and neither allele. Assuming independence between loci these formulae may be used to compute the expected number of matches and partial matches at any number of loci. However, to estimate the variance or place confidence intervals on these predictions would require estimates of the variance and covariance of the probability of matches at two, one or no alleles. Without these estimates no formal statistical test can be done to determine if the observations in Arizona represent a significant departure from the expectations.

Additionally, the numbers of matches that result from modest increases in θ are much greater in the Weir (2007) and Myers (2006) studies than here (figure 3). For instance Weir (2007) predicts 20 10-locus matches and 538 9-locus matches using allele frequencies from a Caucasian database and $\theta = 0.03$. Thus, while this example predicts exactly the number of 10-locus matches seen in Arizona, it also predicts too many 9-locus matches by a factor of five. The most likely reason for this difference is the assumption Myers (2006) and Weir (2007) relied on for computing the effects of population substructure. They used a calculation premised on the assumption that every pair of individuals from the same ethnic group (Myers (2006)) or the population (Weir (2007)) came from the same subpopulation. In these database comparisons, all possible pairs of individuals are examined. Accordingly, some of these pairs will be individual from the same subpopulation but many will not. In fact, the more subpopulations that are represented in the Arizona database, the less likely it is for two randomly chosen people to be from the same subpopulation and, therefore, the less accurate the method used by Myers (2006) and Weir (2007).

To demonstrate this effect, I have simulated databases of single locus profiles of 100,000 individual from one ethnic group and $\theta = 0.03$. The number of subpopulations was set to 2, 10 and 1000. There were 10-replicate simulations at each level of subpopulation. The theoretically expected number of matches was determined from the appendix of Weir's paper (2004). These results (figure 4) show that as the number of subpopulations increases, the expected number of matches decreases relative to Weir's prediction, since it is less likely that any pair of individual come from the same subpopulation. The sampling variance of the mean number of matches also decreases with increasing subpopulations as expected.

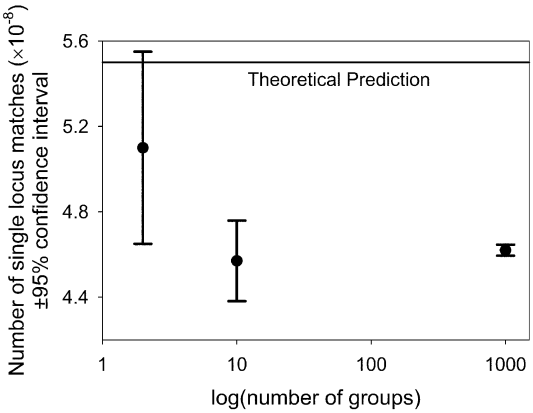

Figure 4. The number of single locus matches in simulated databases of 100,000 people. The subpopulations were assumed to be drawn from a single population with θ equal to 0.03. Each point is the mean of 10 independent simulations.

Full sibs and no population substructure

Adding pairs of full sibs to the Arizona database increases both the number of 9-locus and 10-locus matches (figure 5), but as in the substructure-only-simulations, the number of 9-locus matches quickly exceeds the number in Arizona well before the number of 10-locus matches is even close to 20. Consequently, no models that add sibs alone can adequately explain the Arizona observations. I next consider models that incorporate both the addition of full siblings and population substructure.

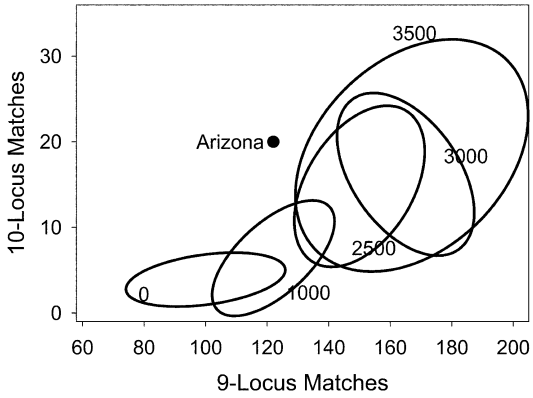

Figure 5. 95% confidence ellipsoids for simulations in which $\theta = 0$ and the number of full sibs varied. The number of pairs of full sibs is placed on each ellipsoid.

Full sibs and population substructure

In these simulations the chances of sibs matching is enhanced relative to the chances in the previous section, since the parents of each sib are assumed to come from the same subpopulation. Thus, the parents are more likely to share alleles and hence this further increases the chances that their progeny will share alleles and ultimately match at a large number of loci. Results are shown for three different values of θ: (i) 0.005 (figure 6), (ii) 0.01 (figure 7), and (iii) 0.015 (figure 8).

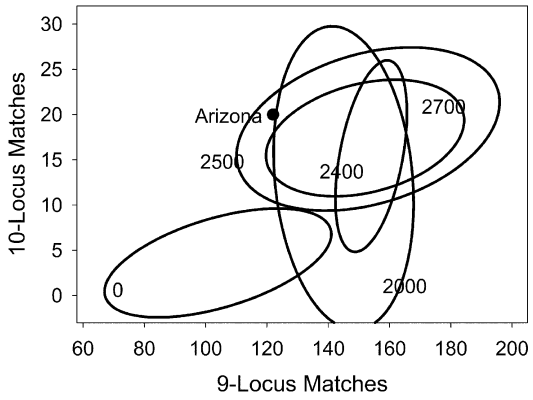

Figure 6. 95% confidence ellipsoids for simulations in which θ was set to 0.005 and the number of full sibs varied. The number on each ellipsoid corresponds to the number of pairs of sibs present in the simulated databases.

For each level of population substructure, there is now at least one set of parameter values that would make the probability of observing the Arizona results greater than 5%. For $\theta = 0.005$, around 2500 pairs of sibs will explain the Arizona observations but just a few hundred more or less sibs causes the Arizona results to fall below the critical 5% level (figure 6). With $\theta = 0.01$, this range is expanded to about 1800–2500 pairs of sibs, although even within this range we find some nonsignificant results. When $\theta = 0.015$, the bottom of the range is lowered to about 1600 pairs of sibs and continues to about 2000.

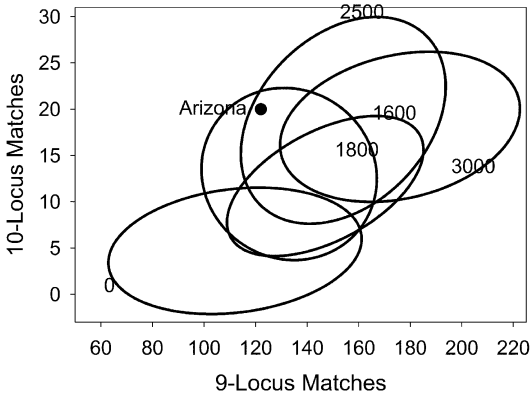

Figure 7. 95% confidence ellipsoids for simulations in which θ was set to 0.01 and the number of full sibs varied. The number on each ellipsoid corresponds to the number of pairs of sibs present in the simulated databases.

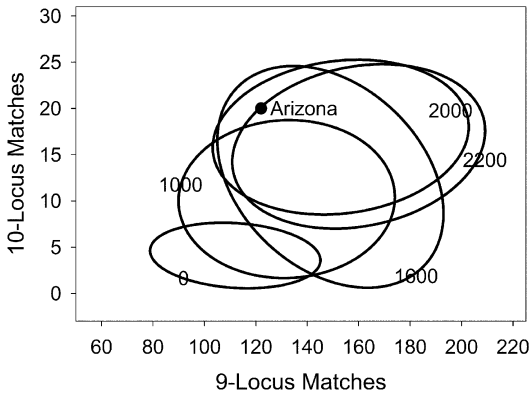

Figure 8. 95% confidence ellipsoids for simulations in which θ was set to 0.015 and the number of full sibs varied. The number on each ellipsoid corresponds to the number of pairs of sibs present in the simulated databases.

The general findings are that acceptable parameter values require fewer pairs of siblings as θ increases. The range of sibling pairs that produce an adequate description of the Arizona observations is relatively narrow. Thus, if the true number of sibling pairs was much less than 1000, or much greater than 3000, then none of these models would produce reliable predictions of the observed number of matches. The claim that there is a relatively narrow parameter range that explains the Arizona results can be put into perspective as follows. If 150 9-locus matches and 15 10-locus matches had actually been observed in Arizona, then virtually all simulations in figures 6–8 would have been consistent with this result.

Parent–offspring pairs and population substructure

Although siblings are the genetically closest relatives one is likely to encounter in a population, it is worth evaluating the effects of other relative types to assess the possible contribution they might make to the elevated numbers of matches seen in Arizona. Simulations with $\theta = 0.01$ were carried out as described above but using parent–offspring pairs rather than sibling pairs. These results (figure 9) show that even with a database composed almost entirely of parent–offspring pairs the number of matches at 10 loci is far below the Arizona value. From these results it is reasonable to conclude that the only relatives that would possibly contribute to explaining the Arizona observations are full sibs. The presence of parents and offspring, or more distant relatives, will make very minor contributions to the increasing number of matches above the number predicted with no relatives.

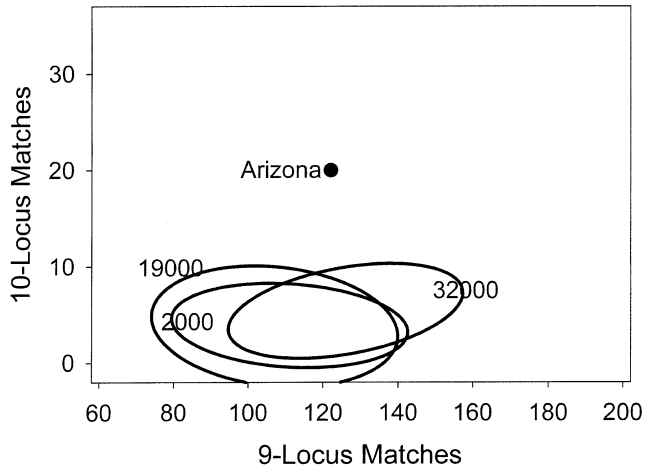

Figure 9. 95% confidence ellipsoids for simulations in which θ was set to 0.01 and the number of parent–offspring pairs varied. The number on each ellipsoid corresponds to the number of parent–offspring pairs present in the simulated databases.

Discussion

More information is needed before we can decisively conclude that the models utilized here are adequate or not. However, this work has permitted some relatively strong conclusions. To explain the Arizona observations will require the presence of a large number of siblings in the Arizona database. More remote relatives, even as close as parents and offspring, are unlikely to help much at explaining these observations. Not any number of siblings will work. The results from this study suggest that if the numbers were much less than about 1000 pairs or much more than about 3000 these models would not work.

Is there any way to verify these predictions? One solution is to determine for each person in the Arizona database whether or not they also have a full sibling in the database. This would clearly be a very tedious solution. However, this method could be streamlined by taking a random sample of, say, several hundred people from the Arizona database and determining what fraction of these people has siblings in the database. If that fraction were between 3.1% and 9.2% then that would be within the suitable range cited above.

An additional method for studying these problems would be to get the profiles from two different states, say Arizona and Maryland. The number of matches within databases could be compared to the number between databases. This latter number would not be expected to be inflated by numerous full sibs and thus should be close to the numbers predicted by substructure only.

It is clear from these simulations that, even for the best models, the probability of the Arizona observations is only 9%–12%. The study of additional offender databases would help add to the empirical foundation of this study and help assess whether Arizona is the norm or, for some reason, an odd outlier. Ultimately, if the simple models examined here cannot adequately explain the number of matches observed in the Arizona offender database, some modification of the underlying probability models may be required.

The product rule with some minor modification is the most common method for computing the frequency of DNA profiles in forensic laboratories. This method relies critically on the assumption that there is statistical independence between loci. The empirical support for this method comes mainly from tests of independence between pairs of loci (Budowle *et al*. 1999). However, recent research on finite populations, with mutation and a monogamous mating system shows that departures from the product rule get worse as one looks at more loci (Dr Yun Song, personal communication). Thus, rigorous testing of the product rule predictions at many loci may yield different results than prior work at only two loci. Perhaps the most important qu1ality control issue in forensic DNA typing is determining the adequacy of the methods for computing profile frequencies. In this respect offender databases can serve a useful and unique purpose, as apparently intended by the DNA Identification Act. The tremendous size of these databases makes them a unique resource which would cost many millions of dollars to recreate. There is certainly much more that can be learned from additional scientific research with offender databases.

Acknowledgements

I thank D. Balding, J. Gilder, D. Krane and S. Zabell for helpful comments on this research, and B. Barlow for her path breaking field work on offender databases.

References

Balding D. J. 2005 *Weight-of-evidence for forensic DNA profiles*. John Wiley, New York.

Budowle B., Moretti T. R., Baumstark A. L., Defenbaugh D. A. and Keys K. M. 1999 Population data on thirteen CODIS core short tandem repeat loci in African American, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *J. Forensic Sci*. **44**, 1277–1286.

Budowle B., Chakraborty R., Carmody G. and Monson K. L. 2000 Source attribution of a forensic DNA profile. *Forensic Sci. Comm*. 2. (http://www.fbi.gov/hq/lab/fsc/backissu/july2000/source.htm)

Budowle B., Shea B., Niezgoda S. and Chakraborty R. 2001 CODIS STR Loci Data from 41 sample populations. *J. Forensic Sci*. **46**, 453–489.

Cherni L., Yaacoubi B. L., Pereira L., Alves C., El Kill H. K., El Gaaied A. B. A. and Amorim A. 2005 Data for 15 STR markers (Powerplex 16 system) from two Tunisian populations: Kersa (Berber) and Zriba (Arab). *Forensic Sci. Int*. **147**, 101–106.

Efron B. and Tibshirani R. J. 1993 *An introduction to the Bootstrap*, CRC Press, Boca Raton.

Einum D. D. and Scapetta M. 2004 Genetic analysis of large data sets of North American Black, Caucasian, and Hispanic populations at 13 CODIS STR loci. *J. Forensic Sci*. **49**, 1381–1385.

Evett I. W. and Weir B. S. 1998 *Interpreting DNA evidence: statistical genetics for forensic scientists*. Sinauer Associates, Sunderland, MA.

Frank W. E., Ellinger E. R. and Krishack P. A. 2006 Y chromosome STR haplotypes and allele frequencies in Illinois Caucasian, African American, and Hispanic Males. *J. Forensic Sci*. **51**, 1207–1215.

McElfresh K. C. and Kim Y. K. 2000 Finding the needle in the haystack: how many loci does it take to find a single individual in a DNA database of 2 million individuals? The eleventh international symposium on human identification (http://www.promega.com/geneticidproc/ussympllproc/abstracts.html).

Myers S. P. 2006 Felon-to-felon STR partial profile matches in the Arizona database: don't panic. Presentation at the California association of criminalists DNA workshop.

Troyer K., Gilboy T. and Koeneman B. 2001 A nine STR locus match between two apparently unrelated individuals using Ampflstr Profiler Plus™ and Cofiler™. The twelfth international symposium on human identification. (http://www.promega.com/geneticidproc/ussympl2proc/abstracts.html).

Weir B. S. 2004 Matching and partially-matching DNA profiles. *J. Forensic Sci*. **49**, 1009–1014.

Weir B. S. 2007 The rarity of DNA profiles. *Ann. Appl. Stat*. **1**, 358–370.

Received 18 October 2007, in revised form 13 February 2008; accepted 14 February 2008
Published on the Web: 8 July 2008