www.sciencemag.org/cgi/content/full/1122655/DC1

Supporting Online Material for

Finding Criminals Through DNA of Their Relatives

Frederick R. Bieber,* Charles H. Brenner, David Lazer

*Author for correspondence. E-mail: fbieber@partners.org

Published 11 May 2006 on *Science* Express
DOI: 10.1126/science.1122655

The main PDF file includes the following:

Finding Criminals through DNA of Their Relatives

Frederick R. Bieber, Departments of Pathology, Brigham and Women's Hospital and
Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. Telephone: 617-732-6507.
E-mail: fbieber@partners.org

Charles Brenner, Forensic Science Group, School of Public Health, University of California,
Berkeley, CA 94730-7360, USA. Telephone: 510-339-1911. E-mail: cbrenner@berkeley.edu

David Lazer, John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138, USA.
Telephone: 617-496-0102. E-mail: david_lazer@harvard.edu

Materials and Methods

We performed computer-modeling experiments to assess chances of identifying, in a DNA database, a biological relative of the owner of a given multiple locus STR profile. Model parameters considered were the nature of the genetic relationship, and the size of the database. In the final section we also evaluate impact of the incorporation of geographic information on the effectiveness of familial searching.

Forensic Mathematics of Familial Searching

We performed computer simulations using published data (*S1*) of U.S. Caucasians to predict the effect of the following situation:

1. There is a computerized database of 2,500,000 DNA profiles $\{D_i\}$ representing the catalogued convicted offenders of the United States.
2. Each DNA profile comes from one or another of 50 "states" of equal size, 50,000 convicted offenders each. St_i is the state corresponding to offender D_i.

A relative—son or brother—of someone with profile D_j in the offender databank (from state St_j) commits a crime in some state *St*, leaving his DNA profile *C*. We modeled the parentage and the sibling relationships as separate experiments, but of course in practice one would be quite satisfied to find either or any relative at all. In real life it will occasionally happen that uncle-nephew or even cousin relationships will chance into the sibling net if the genetic sharing between the pair is more than usual.

We considered, but have not modeled, the idea of looking for a pair of related two offenders who simultaneously resemble a crime stain profile. The confidence of a true, rather

than spurious, relationship would be much greater in such a circumstance, and indeed, this idea was used to good effect in screening for identifications of World Trade Center victims (*S2*, *S3*). In practice, genealogical information about offenders is not now available in a fashion that could be easily be integrated with the offender databases. There is, though, no cost and some possible benefit to noting coincidence of last names in the offender database.

We consider several experiments with various assumptions about the relationship between *St* and St_j

3. The profile *C* is compared with all the profiles D_i using kinship analysis. Certainly, some D_i who are not related to *C* will, nonetheless, appear related by chance.

4. The individuals D_i are examined in priority order according to a reasonable strategy.

We asked how many leads would need to be examined before the actual family member was found. That is, how far down the list of examined D_i will the proxy relative profile, D_j, be?

For computerized Monte Carlo simulations (*S4*) to evaluate the foregoing,

I. First simulate a database:

   1. A databank of 2,500,000 offender/arrestee DNA profiles, each of 13 "CODIS" loci, are generated by simulation, alleles being chosen according to frequencies from standard population data (*S1*). The simulation assumes random mating, a slight simplification that, in principle, biases toward making true relatives easier to find. The bias is small, however.

   2. The DNA profiles are divided into 50 "state" databases of 50,000 profiles each by randomly labeling 50,000 of them as from "Alabama," the next 50,000 as from "Alaska," etc.

II. Then, repeatedly simulate a crime investigation strategy using family searching:

   1. A DNA profile, *C*, of a relative representing a novice (previously not catalogued) criminal—either sibling or child (depending on the experiment) of some offender databanked person D_j—is generated in accordance with Mendelian inheritance and population STR frequency data (*S1*).

   2. The "crime" is assigned to a state *St*, where *St* is more or less likely to be St_j, depending on the geographical component of the experiment.

III. For each "crime," we record the difficulty of finding the true relative using family searching:

   1. The "crime" scene DNA profile, *C*, is compared with that from every D_i and the kinship likelihood ratios L_i (parentage—evaluating father-son relationship) or S_i (sibling), depending on the experiment—are calculated. One of these—the one with index *j*—represents a true relationship, which is typically a large number. The rest represent false relationships and so are mostly small, although some of them may also be large. Figure S1 compares the distributions and illustrates the idea that they overlap somewhat.

2. Further, a geographical likelihood ratio, G_i, is computed whose value depends on the relationship between St_i and St.

3. An overall likelihood ratio for individual i is obtained as $L_i \times G_i$ (or $S_i \times G_i$) and these are sorted by size, largest first.

4. We record the position in the sorted list of individual j. This is the number, k, representing the number of leads pursued to find the true relative of the perpetrator.

IV. The method described above for the Monte Carlo simulation requires the evaluation of millions of kinship likelihood ratios per crime, hence, trillions of computations in order to simulate millions of crimes. This was reduced to a more feasible computation, as follows:

1. Generate 2,500,000 families. For each family, two parents are generated at random, and then two children are generated by selecting random alleles from each parent.

2. Use the families generated above to compute four likelihood ratios (LRs) within each family f:

   a. L_f, the paternity index between true father and son (one parent and one child)

   b. S_f, the sibling index between the two true siblings

   c. NL_f, the paternity index computed between the unrelated parents

   d. NS_f, the sibling index computed between the unrelated parents

   The definition of the relationship LRs is as follows. Let E represent the statement that two given people have the respective genotypes of the two people mentioned in any one of points *a-d* above, let "relationship" mean that a parentage or sibling relationship, as the case may be, exists between two people. Then

   L = Prob(E | relationship) / Prob(E | no relationship) and the same for each of the LRs.

Details of the calculation are below.

The above four distributions are shown in Fig. S1. The only purpose of the families was to compute these LR distributions used as described below; these families then have no further purpose for the simulations.

3. To simulate a crime by a son of a convicted offender, of the necessary 2,500,000 kinship indices L_i between C and D_i, one—representing the true-relative index—is chosen from among the set $\{L_f\}$ and the rest are the NL_i. These likelihood ratios are then rank ordered. The same process is followed to generate a set of likelihood ratios for a potential sibling relationship.

   Thus, the simulation is rather abstract. There are no longer any particular profiles that can be identified as C or as D_i. Instead, we only model likelihood ratios. In particular the true-relative index arises by taking C as the child of some family f;

the false-relative indices are calculated from some altogether different pairs of people, not including *C*.

As a check that this method is not biased, we calculated 100 *k*-values using the direct procedure described above in section III. Their distribution is indistinguishable from the distribution of several million *k*-values using the method described here.

V. Computation of likelihood ratios

1. A unified method for evaluating two-person kinship indices is given in (*S3*). We refine the method to account for mutation, which may be important in calculating the indices $\{NL_i\}$ as follows.

2. To compute a kinship likelihood comparing *C* with *D* at a particular locus, denote the genotype of *C* as *ab*; that of *D* by *cd*. Define variables u_h for the four "mating combinations" 1, *ac*; 2, *ad*; 3, *bc*; and 4, *bd*. If the two alleles in the *i*-th pair are the same type, then u_i is the reciprocal of the frequency of that allele. If they are different, then u_i is the probability that one would mutate to the other during meiosis between C and D. We incorporate a mutation rate of 1% for the simulations, a generously high figure for forensic STR loci. In real life, mutation must be included in the LR evaluation to avoid overlooking a few percent of real fathers, but the higher the mutation rate the more the false-parentage indices $\{NL_i\}$ are inflated, so the less conspicuous the true-parentage index *L* becomes. Hence, our mutation model is conservative, biased against finding the true parent or child.

   Let *U* be the average $(u_1 + u_2 + u_3 + u_4)/4$. *U* is the likelihood ratio by which the genotype information at the locus supports a father-son relationship between *C* and *D*. Multiplying across loci gives overall likelihood ratios L_i or NL_i as above.

   Also, let *W* be the average between-individual product $(u_1u_4 + u_2u_3)/2$. Then the likelihood ratio by which the genotypes at this locus support a sibling relationship between *C* and *D* is $¼ + ½U + ¼W$. (The coefficients are the probabilities for siblings to share 0, 1, or 2 alleles identically by descent.) Multiplying these across loci gives S_i or NS_i as above.

   We evaluated the magnitude and the rank-order of each of these likelihood ratios, which are usually much higher for close relatives than for unrelated pairs.

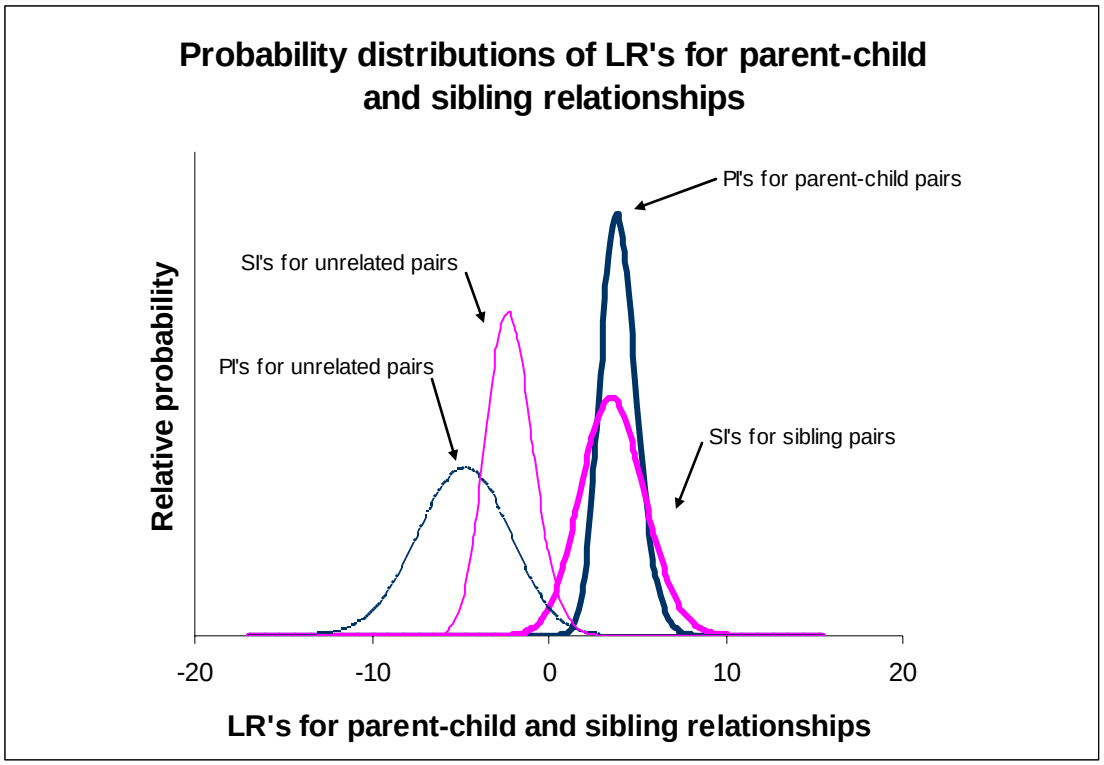

Fig. S1. Results of Monte Carlo simulations showing distributions of paternity and sibling likelihood ratios [paternity indices (PI) and sibling indices (SI), respectively] between related and unrelated pairs obtained by generating 2,500,000 pairs of each type.

VI. Geographical modeling experiments

A crime is committed in state *St*; a databank person *D* comes under consideration as a possible proxy relative and *D* is from a possibly different state *Su*. For a person unrelated to the crime, the probability to come from *Su* is 1/50, since "states" in the model are all the same size. For a proxy relative however, the probability *x* to come from *Su* is larger for *Su* proximate to *St*. Thus, the geographical information about *D* supplies some information as to whether *D* is a true relative; in particular, the "geographical" supporting likelihood ratio is *G*=X/Y, where

X = Prob(*D* is from *Su* | *D* is a true relative of the crime committed in *St*) = x;

Y = Prob(*D* is from *Su* | *D* is unrelated to the crime committed in *St*) = 1/50.

i.e., $G = 50x$.

1. A preliminary experiment ("50k database") assumes all crime is intrastate. The crime state *St* is chosen to be the same as St_j, the state of the true relative. Therefore $G_i = 50$ for each of the 50,000 offenders in *St* and $G_i = 0$ otherwise.

2. At the other extreme ("2.5m database"), assume crime is geographically random. Then $G_i \equiv 1$ and the effect is to model a single very large state. On average there will be 50 times as many false leads impeding the search for the true relative as in the preceding case.

3. Finally we introduce a realistic model for *G*, namely

   a) x = 85% that the true relative will be in the same state as the crime. $G = 50 \times 85\% = 85/2$.

   b) 12% that the true relative will be in one of the 4 neighboring states, hence x = 3% for each one. $G = 3/2$.

   c) 3% that the true relative will be in one of the 45 "remote" states, hence, x = 1/15% for each one. $G = 1/30$.

We considered several additional experiments, each evaluated using the realistic *G*:

4. All simulated crimes occur in the state of the relative. It is still a bit harder to find the relative than in the first case because, although the *G* ratios depress the LR's of the out-of-state (false) leads, a few of them are still large enough to dilute the list of leads.

5. All simulated crimes occur in a state neighboring the state of the relative. In this unrealistic and difficult situation, the LR for the true lead is always depressed by 3%/85% compared with false leads from the crime state.

6. The simulated crimes are distributed realistically according to the same model as that for *G*—85% in the state of the true relative, etc. Consequently, catching the perpetrators is not much harder than in experiment VI.1 or VI.4. Figure S2 shows

experiments VI.1 (50k database), VI.2 (2.5m database) and this experiment, VI.6 ("national").

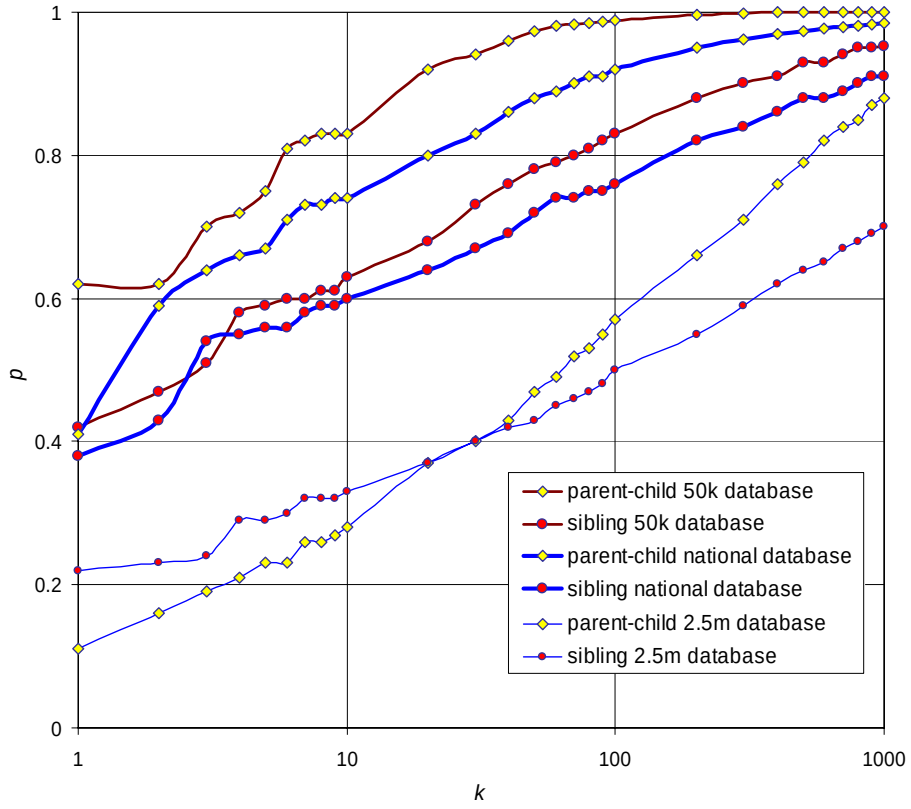

Fig. S2. Probability of finding a true relative in the convicted offender database of the crime scene DNA profile, assuming such a relative exists, within the first *k* leads investigated. "Leads" are defined by computing likelihood ratios for a parentage or sibling relationship between crime stain and each convicted offender, possibly modified by a "geographic" factor.

Three pairs of simulations are shown: one pair modeling a "state" database of 50,000 offenders; a pair of larger homogeneous databases of 2,500,000 offenders—comparable in size to the combined U.S. National DNA Index System (NDIS); and a realistic "national" model representing 2,500,000 offenders comprised of 50 "states" and with LRs calculated according to modeling assumptions that criminals preferentially offend near the state in which their relatives reside.

VII. Prediction and practice

In practice, a computerized family search of a crime scene DNA profile against a DNA database of convicted offenders will produce a collection of leads, each of which with an associated LR. We are not suggesting that investigators adopt a policy of following up on the first 10 leads or any other number of leads. Rather, the investigator will naturally make a judgment depending on the sizes of the LRs. We have discussed incorporating geographic information as well as genetic into the LRs, and the investigator may roughly adjust the LR further if any additional factors seem relevant. Then, probably through a rule of thumb that in effect implements Bayes' Theorem, the LR is in effect interpreted as an approximate posterior probability. The underlying mathematical thinking is roughly as follows:

1. There is some probability e (for relative *e*xists) that the perpetrator of the crime, not himself in the offender database, has a close relative who is. As a starting point we might guess that this probability is related to the 40% or so chance that an offender has a relative in jail—perhaps it is much less than that, but we assume that it is substantial in the sense that a worthwhile fraction of the time kinship searching in the database will be effective.

2. Any given offender in the database has some small prior probability to be the relevant relative; summed over the entire database these probabilities total e. Taking the simple approach of distributing the probability, e is distributed equally over N offenders, the prior probability is e/N for each one to be the relevant relative.

3. On that basis, a likelihood ratio L supporting a particular offender as a relative of the crime stain corresponds to posterior odds Le/N—essentially this is Bayes' Theorem stated in terms of odds. Probability = Odds/(Odds + 1), so odds of 1 corresponds to a probability of 50%, which surely indicates a worthwhile suspect. Whether odds of 1/10 or 1/100—for such small numbers odds and probability are approximately equal—are worthy of interest depends on available resources, the cost of pursuing a lead, and the importance of the case.

Large LRs—substantial compared with the size of the database—as a rule may serve to identify the most worthwhile suspects. Hence, in practice, the criterion for pursuing a lead would be closely related to the size of the LR.

References

S1. B. Budowle, T. R. Moretti, A. L. Baumstark, D. A. Defenbaugh, and K. M. Keys, 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 (1999).

S2. L. Biesecker *et al.*, DNA identifications after the 9/11 World Trade Center attack. *Science* **310**, 1122–1123 (2005).

S3. C. H. Brenner and B. S. Weir, Issues and strategies in the identification of World Trade Center victims. *Theor. Popul. Biol.* **63**, 173–178 (2003).

S4. N. Metropolis, S. Ulam, The Monte Carlo method. *J. Am. Stat. Assoc.* **44** (247), 335–341 (1949).