A manual for *DNAMIX* version 2

John D. Storey
Program in Statistical Genetics, Statistics Department
North Carolina State University, Raleigh, NC 27695
E-mail: storey@statgen.ncsu.edu

August 13, 1998

1 Introduction

Two computer programs have recently been written by JDS in order to perform calculations of likelihood ratios as they pertain to mixed DNA samples encountered in forensic science. *DNAMIX* v.1 was written in 1997 and performs calculations for the formulas and methods described in [4]. The formulas in [4] are based on the assumption of independence of all the alleles in the mixture. This assumption implies Hardy-Weinberg and linkage equilibrium, as well as independence between individuals. Hence, the low-level dependence among individuals within the same population is ignored. A more recent treatment of mixed DNA samples was given in [1], in which the coancestry coefficient (often denoted by θ) of the population was taken into account. See [5] for a thorough treatment of the coancestry coefficient in the context of forensic science. A general introduction to the applications of statistical genetics to forensic science is given in [2].

This document is a manual for the computer program *DNAMIX* v.2, which performs calculations for the formulas and methods presented in [1]. Up-to-date information on *DNAMIX* v.1 or *DNAMIX* v.2 is available at the website:

`http://statgen.ncsu.edu/storey/`

Either program can be downloaded at the anonymous ftp site:

`statgen.ncsu.edu`

in the directory `/pub/storey/`. The source code written in FORTRAN 90 and the object code for both PC's and Sun workstations should be available for *DNAMIX* v.2. The source code written in both C and FORTRAN 77 should be available for *DNAMIX* v.1.

2 Likelihood Ratios

The likelihood ratio method for assigning weight to evidence is described in [1], [2], [4]. Here, we will give a brief overview of the method so that one understands the input and output of the computer program.

Denote the evidentiary profile as E. We want to compare two alternative explanations for E. Let C be the explanation for, say, the prosecution and let $\overline{C}$ be the explanation for, say, the defense. We then form a likelihood ratio L by defining

$$L = \frac{\Pr(E|C)}{\Pr(E|\overline{C})} .$$

Thus, L gives us a measure of how much more likely the evidence E is under explanation C than under explanation $\overline{C}$. In the context of mixed DNA samples, E represents the alleles present in the mixture at

a particular locus. C and $\overline{C}$ are explanations as to which alleles come from a given number of known contributors, known non-contributors, and unknown contributors.

For example, suppose that at a particular locus, alleles A_1, A_2, A_3 are present in the mixture, and that the victim has genotype (A_1, A_2) while the defendant has genotype (A_2, A_3). In this case $E = \{A_1, A_2, A_3\}$. The prosecution may say that C is the event that there are two known contributors with genotypes (A_1, A_2) and (A_2, A_3), no known non-contributors, and no unknown contributors. The defense may say that $\overline{C}$ is the event that there is one known contributor with genotype (A_1, A_2), one known non-contributor with genotype (A_2, A_3), and two unknown contributors.

3 Input and Output

Since *DNAMIX* is an interactive program, the input procedure should not be too difficult to understand. The input can roughly be divided into three parts: the evidentiary profile, the numerator explanation C of the likelihood ratio L, and the denominator explanation $\overline{C}$ of L. The items required for entering the evidentiary profile are:

- the number of the alleles in the sample
- the names of the alleles in the sample
- the number of databases to be used
- the names of the databases
- the allele frequencies and the coancestry coefficient of each database

The names of the databases and alleles should be kept as short as possible without any spaces between words. For example, if one of the databases is a sample of the U.S. Caucasian population, then an appropriate name would be `US_Cauc`. I have set the maximum length of all character strings to eight characters. Therefore, any character string entered by the user is truncated after eight characters. For more information about the value of the coancestry coefficient θ for a given database, see [3] and [5]. The coancestry coefficient may be set to zero if one does not want to include population structure in the calculations (see section 5). After all of the above items are entered, the user has an option of creating a summary file. This is highly recommended since the file is automatically written and all the results will be neatly displayed. If this option is selected, a name of the file must be given as well as a title for the summary, such as a locus name. It is important to remember that if a file already exists under the same name, then this file will automatically be replaced by the *DNAMIX* file without any warning.

The user is then prompted to enter the numerator explanation of the likelihood ratio. The following items are required:

- number of known contributors
- genotypes of the known contributors
- number of known non-contributors
- genotypes of the known non-contributors
- lower bound on the number of unknown contributors
- upper bound on the number of unknown contributors

In order to enter the genotypes, the alleles are enumerated and the user selects the two alleles by entering their corresponding numbers. If the contributor is a homozygote, then the corresponding number is entered twice. A nice feature of this program is that probabilities can be calculated over a range of unknown contributors. This is practical in that one may be interested in the likelihood ratios under several different combinations

of unknown contributors. If only one number of unknown contributors is desired, then the user should enter that number as the both the lower and upper bounds. The probability of the evidence under the numerator explanation is then displayed for each database and for each number of unknown contributors.

The input and output for the denominator explanation is similar to that of the numerator and will not be discussed. After the denominator probabilities are displayed, the likelihood ratio(s) are displayed for each database. The database name, number of numerator unknown contributors, number of denominator unknown contributors, and likelihood ratio are displayed for each possible unknown contributor/database combination. Typically, there will be several likelihood ratios displayed since more than one database and a variable number of unknown contributors can be considered. The user is then given the option of performing more calculations under the same evidentiary profile and databases, but under different numerator and denominator explanations.

4 Examples

Suppose that we have a mixed sample containing the alleles A, B, and C from the HBGG locus. The victim has genotype AB at this locus, and the suspect has genotype BC. Suppose we have the following three databases available where θ is the estimated coancestry coefficient. (Note that these databases are not real and were made up for the example only.)

HBGG	African American	Caucasian	Hispanic
A	0.42000	0.55825	0.37500
B	0.23000	0.43689	0.58000
C	0.35000	0.00485	.045000
θ	0.02000	0.03000	0.03000

The prosecution's hypothesis is that the victim and the suspect are known contributors with zero unknown contributors and zero known non-contributors. The defense's explanation is that the victim is a known contributor, the suspect is a known non-contributor, and there are one or two unknown contributors. To calculate the likelihood ratios of these explanations we would run the program as follows.

```
*****************************************************
*                                                   *
*                  DNAMIX version 2                 *
*                                                   *
*****************************************************
*                                                   *
* This program performs calculations for the methods *
* and formulas presented in:                        *
*                                                   *
* Curran, J.M., Triggs, C.M., Buckleton, J., and    *
* B.S. Weir. 1998. Interpreting DNA mixtures in     *
* structured populations, preprint.                 *
*                                                   *
*****************************************************

Press RETURN to continue.
```

```
 Enter the number of alleles in the mixture:
3

 Enter the name of allele 1:
A
 Enter the name of allele 2:
B
 Enter the name of allele 3:
C

 Enter the number of databases to be used:
3

 Enter the name for database number 1:
Afr_Amer
 Enter the frequency of allele A:
.42
 Enter the frequency of allele B:
.23
 Enter the frequency of allele C:
.35
 Enter the coancestry coefficient for database number 1:
.02

 Enter the name for database number 2:
Cauc
 Enter the frequency of allele A:
.55825
 Enter the frequency of allele B:
.43689
 Enter the frequency of allele C:
.00485
 Enter the coancestry coefficient for database number 2:
.03

 Enter the name for database number 3:
Hispanic
 Enter the frequency of allele A:
.375
 Enter the frequency of allele B:
.58
 Enter the frequency of allele C:
.045
 Enter the coancestry coefficient for database number 3:
.03

 Do you want a summary file? ( No=0, Yes=1 ):
1
```

```
 Enter the name of the file:
summary

 Enter a title for the summary (such as a locus name):
HBGG

 ************************************************************************
     THE FOLLOWING INPUT IS FOR THE NUMERATOR OF THE LIKELIHOOD RATIO
 ************************************************************************

 Enter the number of known contributors:
2

 Enter the numbers corresponding to the genotype for known contributor number 1.
  1 = A
  2 = B
  3 = C
 Enter the number corresponding to the first allele:
1
 Enter the number corresponding to the second allele:
2

 Enter the numbers corresponding to the genotype for known contributor number 2.
  1 = A
  2 = B
  3 = C
 Enter the number corresponding to the first allele:
2
 Enter the number corresponding to the second allele:
3

 Enter the number of individuals known NOT to have contributed to the sample:
0

 Enter the lower bound on the number of unknown contributors:
0
 Enter the upper bound on the number of unknown contributors:
0

 Unknown Contributors:  0
 Database       Numerator Probability
 -----------------------------------
 Afr_Amer           0.300468E-01
 Cauc               0.185034E-02
 Hispanic           0.199937E-01


 Press RETURN to continue.
```

```
 ************************************************************************
    THE FOLLOWING INPUT IS FOR THE DENOMINATOR OF THE LIKELIHOOD RATIO
 ************************************************************************

 Enter the number of known contributors:
1

 Enter the numbers corresponding to the genotype for known contributor number 1.
  1 = A
  2 = B
  3 = C
 Enter the number corresponding to the first allele:
1
 Enter the number corresponding to the second allele:
2

 Enter the number of individuals known NOT to have contributed to the sample:
1

 Enter the numbers corresponding to the genotype for known non-contributor number 1.
  1 = A
  2 = B
  3 = C
 Enter the number corresponding to the first allele:
2
 Enter the number corresponding to the second allele:
3

 Enter the lower bound on the number of unknown contributors:
1
 Enter the upper bound on the number of unknown contributors:
2

 Unknown Contributors:  1
 Database     Denominator Probability
 -----------------------------------
 Afr_Amer           0.169302E-01
 Cauc               0.114421E-03
 Hispanic           0.257688E-02


 Unknown Contributors:  2
 Database     Denominator Probability
 -----------------------------------
 Afr_Amer           0.241070E-01
 Cauc               0.216317E-03
 Hispanic           0.470977E-02
```

```
 Press RETURN to continue.


 ************************************************************************
        THE FOLLOWING ARE THE LIKELIHOOD RATIOS FOR EACH DATABASE
 ************************************************************************

 Press RETURN to see the results for the next database or to continue.


            Numerator Unknown    Denominator Unknown
 Database      Contributors         Contributors         Likelihood Ratio
 ------------------------------------------------------------------------
 Afr_Amer           0                    1                        1.77
 Afr_Amer           0                    2                        1.25

 Cauc               0                    1                       16.17
 Cauc               0                    2                        8.55

 Hispanic           0                    1                        7.76
 Hispanic           0                    2                        4.25


 Do you want to compute another likelihood ratio? ( No=0, Yes=1 ):
0

            Program written by John Storey on May 1, 1998.
             Questions and/or comments should be sent to:
                      storey@statgen.ncsu.edu
```

The input and output are simple and straightforward. The program will warn the user if nonsensical input is performed such as giving the coancestry coefficient a value of 2. In most instances, the user will have to enter another value for a variable if it does not make sense. It is the case, however, that if a character is entered when the program asks for a number the program will crash. Be careful to enter integers when an integer is requested and to enter real numbers when a real number is requested. Numbers can be entered when a character string is requested. For example, we could have named our first database `1data` or `111`. Character strings are read until the first blank character is encountered. So if we had entered `Afr Amer` instead of `Afr_Amer` for our first database, then the output would have been so that the first database is named `Afr`. Also, note that character strings in this program have a maximum length of eight characters. Thus, `African_American` would appear as `African_` for example. It should be obvious throughout the program as to what type of data (such as an integer, real number, or character string) should be entered.

Going back to our example, note how we chose to save our results to a file. When we view the file `summary`, we should see the following.

Locus: HBGG

Database: Afr_Amer		Database: Cauc		Database: Hispanic	
Allele	Freq	Allele	Freq	Allele	Freq
A	0.420000	A	0.558250	A	0.375000
B	0.230000	B	0.436890	B	0.580000
C	0.350000	C	0.004850	C	0.045000
THETA =	0.020000	THETA =	0.030000	THETA =	0.030000

****

NUMERATOR HYPOTHESIS AND PROBABILITIES

Number of known contributors: 2

Individual	Genotype
1	A, B
2	B, C

Number of individuals known NOT to have contributed: 0

Unknown Contributors: 0

Database	Numerator Probability
Afr_Amer	0.300468E-01
Cauc	0.185034E-02
Hispanic	0.199937E-01

DENOMINATOR HYPOTHESIS AND PROBABILITIES

Number of known contributors: 1

Individual	Genotype
1	A, B

```
Number of individuals known NOT to have contributed: 1
Individual   Genotype
------------------------------
 1           B, C
------------------------------


Unknown Contributors:  1
Database       Denominator Probability
--------------------------------------
Afr_Amer           0.169302E-01
Cauc               0.114421E-03
Hispanic           0.257688E-02
--------------------------------------

Unknown Contributors:  2
Database       Denominator Probability
--------------------------------------
Afr_Amer           0.241070E-01
Cauc               0.216317E-03
Hispanic           0.470977E-02
--------------------------------------


 ------------------
| LIKELIHOOD RATIOS |
 ------------------

           Numerator Unknown    Denominator Unknown
Database      Contributors         Contributors         Likelihood Ratio
-------------------------------------------------------------------------
Afr_Amer           0                    1                          1.77
Afr_Amer           0                    2                          1.25

Cauc               0                    1                         16.17
Cauc               0                    2                          8.55

Hispanic           0                    1                          7.76
Hispanic           0                    2                          4.25
-------------------------------------------------------------------------


*************************************************************************
```

If we had computed more likelihood ratios (see the last line of the example program execution), the results would have been appended to the end of this file.

5 *DNAMIX* version 2 versus *DNAMIX* version 1

If population structure is to be ignored in the calculations, then *DNAMIX* version 2 can still be used. In other words, the likelihood ratios in *DNAMIX* version 2 are equal to *DNAMIX* version 1 when $\theta = 0$. If you are going to use *DNAMIX* version 2 in place of *DNAMIX* version 1, then there are two important things to remember. The first is that the numerator and denominator probabilities will not be the same, although the likelihood ratios will. This is because in [1] the genotypes of known contributors and known non-contributors are also included in the calculations of the probabilities. In [4], only the unknown alleles and unknown contributors are included in the calculations of the probabilities. When $\theta = 0$, however, there is no effect in including known contributors and known non-contributors once the likelihood ratio is formed. The second thing to keep in mind when using *DNAMIX* version 2 in place of *DNAMIX* version 1 is that if a profiled individual is identified as a known contributor or known non-contributor for the numerator of the likelihood ratio, then he/she also has to be identified as a known contributor or known non-contributor for the denominator of the likelihood ratio (and vice versa). I recommend using *DNAMIX* version 2 whether you are performing calculations according to [1] or [4] because it is more efficient than *DNAMIX* version 1 in several ways.

References

[1] Curran JM, Triggs CM, Buckleton J, Weir BS. 1998. Interpreting DNA mixtures in structured populations, submitted.

[2] Evett IW, Weir BS. 1998. Interpreting DNA evidence: Statistical genetics for forensic science. Sunderland, MA: Sinauer.

[3] National Research Council. 1996. The evaluation of forensic DNA evidence. Washington, DC: National Academy Press.

[4] Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KAJ, Buckleton J. 1997. Interpreting DNA Mixtures. *Journal of Forensic Sciences* **42**:213-222.

[5] Weir BS. 1998. The coancestry coefficient in forensic science. Proc 8th Int Symp Hum Identification. Madison, WI: Promega.