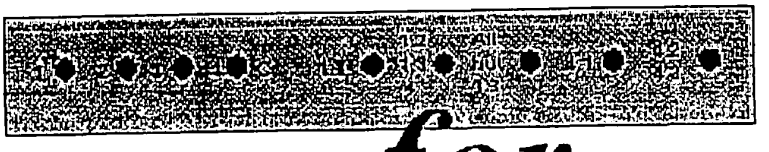
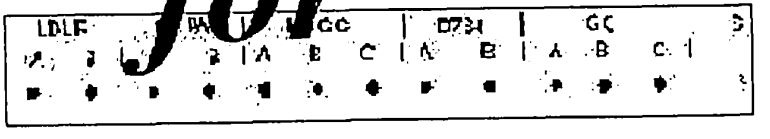
Dnatype



for



Windows 95/NT

DNATYPE

User Manual and Guide

by Ranajit Chakraborty, Ph.D.
Allen King Professor

Yixi Zhong

David Stivers, Ph.D.

Human Genetics Center
School of Public Health
University of Texas Health Science Center
Houston, Texas

DNATYPE is a collection of computer programs for analyzing forensic population genetic data. Running in this Windows 95/NT interface, these programs can be used to create, edit and check population database files, perform tests for independence within, between and among loci in a database, search database files for complete or partial matches, and calculate the probability of a chance match for a user-specified profile following NRC II guidelines.

The Windows 95 interface was written by Snehit Mathew Cherian and R.E. Gaensslen, Forensic Science Program, University of Illinois at Chicago, under an interagency agreement between the Board of Trustees of the University of Illinois and the National Institute of Justice, U. S. Department of Justice, Washington DC.

Contents

1 Introduction to DNATYPE

DNATYPE is a collection of computer programs for analyzing forensic population genetic data and databases. Most of the programs are written in BASIC (one is written in FORTRAN), and run in DOS windows that pop up when programs are run. Each program has an alphabetic-letter designation that serves simply as a shorthand label -- except for NRC, which does calculations of probability of a chance duplicate in a population according to the recommendations of the NRC 1996 report.

Use of most of the programs in DNATYPE (all except NRC) requires population databases, and they must be in a format suitable for the programs. Some example databases are included; their main purposes is to illustrate how to use the different programs. The TWGDAM RFLP databases for six loci for five different U.S. populations are also included. Databases can contain RFLP and/or any PCR-based locus data for up to 16 loci.

Several types of tests can be done with the programs in DNATYPE, including:

1. Check a database file for entry errors and accuracy of format, and search for duplicates. In the case of RFLP database files, users can set the "match window" to be used in searching for duplicates.
2. There are several tests for independence (under HWE assumptions) within a locus. These tests can be run on any specified locus within any database.
3. The pairwise comparison of loci for independence can be done between any two loci in any database.
4. An independence test across all loci in a database can be done on any database.
5. Two databases can be compared with one another for genetic distance (similarity).
6. Any database can be searched for a complete or partial match to a user-specified DNA profile.
7. The NRC program will calculate probabilities of a chance match (and their reciprocals, i.e., 1 in N values) for any user-specified DNA profiles with any combination of loci in any population (for which allele or bin frequencies are available), using the recommendations in the NRC 1996 report. Users can choose up to three values of theta for each profile analysis. The program does these calculations for unrelated persons and for various relatives (such as full siblings, half siblings, etc.).

The programs are written in BASIC or FORTRAN, and run in DOS windows that pop up when a program is run. The programs request specific input data within the DOS window (such as name of database file, name(s) of locus(i) to be analyzed, etc.) before they execute.

Detailed information on each program's requirements, the background of the statistical tests it performs, it's output, and the interpretation of the results, is found in this manual as well as in easily accessible Help files. Most of the information is in both locations.

Click on:	To:
• File	Read introductory material; Exit the program
• Database	Edit, Create, See examples of Databases, or Check databases
• Tests	Run any of the statistical tests on a database or databases
• NRC	Run the NRC program
• Results	View results from a program in WordPad
• Help	Get detailed information about any program or feature in DNATYPE
• Getting Started	Get advice and tutoring on navigating and using DNATYPE
• References	See full citations of references in Help files or the User Manual

The software is written in modules, so that one or more specific tests can be done in a session. The modules have been conveniently grouped in this Windows interface. For faster operation, the interface along with all the programs and files, should be run from the user's hard drive, and the software and data to be analyzed should reside in the same directory.

In Help and in this manual, general guidance is given for running each test or routine, and brief explanations of the required input, and the output, are presented with illustrations. No mathematical details are given; these can be found in the references cited in each section.

2 Getting Started

Installation of DNATYPE

DNATYPE is furnished on a CD ROM, and should be installed to a user's hard disk. The program has an install shield typical of Windows 95 programs. Place the CD ROM in the CD ROM drive. From the Windows 95 Start menu, select Run. Browse to the CD ROM drive letter, and select Setup.exe. Click OK to continue.

DNATYPE will be installed by default to a folder called "DNATYPE" within the Program Files folder. A user can choose a different installation folder name during setup.

All the programs in DNATYPE will look for the database files within the folder that contains the programs. User created database files should, therefore, be saved in the DNATYPE folder (or in the same folder as the programs, if the user has chosen a different folder name).

Using DNATYPE

Detailed documentation on how to use the DNATYPE programs has been included in this manual, and in the Help files built into the program.

Prior to using one of the programs, a user can read through the relevant manual and/or Help file sections, and practice running the program using the examples and example database files provided.

It is generally recommended that user created database files be checked for errors and for possible duplicates using the CDupl and CError programs (see § 4, Databases). Errors in database file structure can cause problems in running the programs.

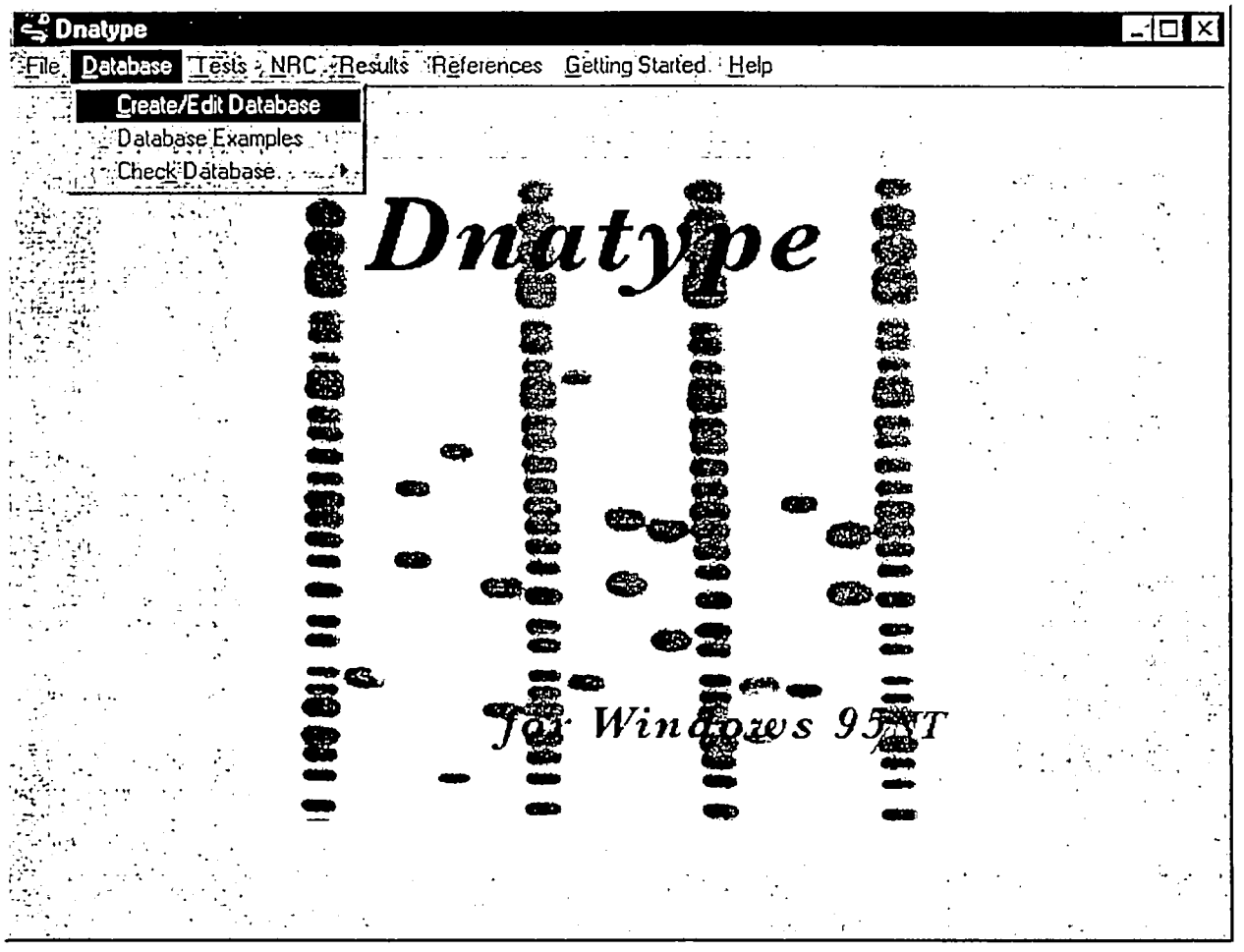
Some programs will run very slowly on large database files. Database files should not contain data on more than 16 loci.

A brief tutorial is provided below to help introduce users to some of the features and operations of DNATYPE.

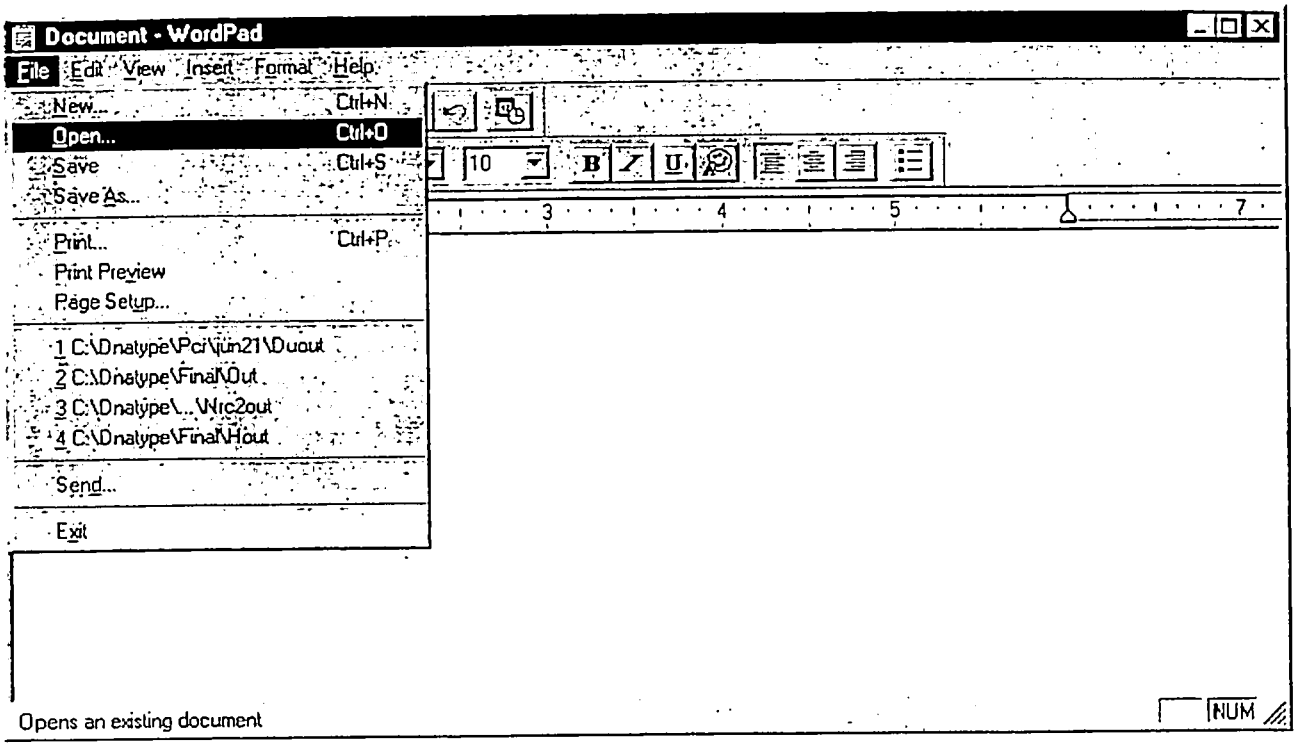
Tutorial for viewing a sample database, running program H on this database and finally viewing the output.

Viewing a sample database

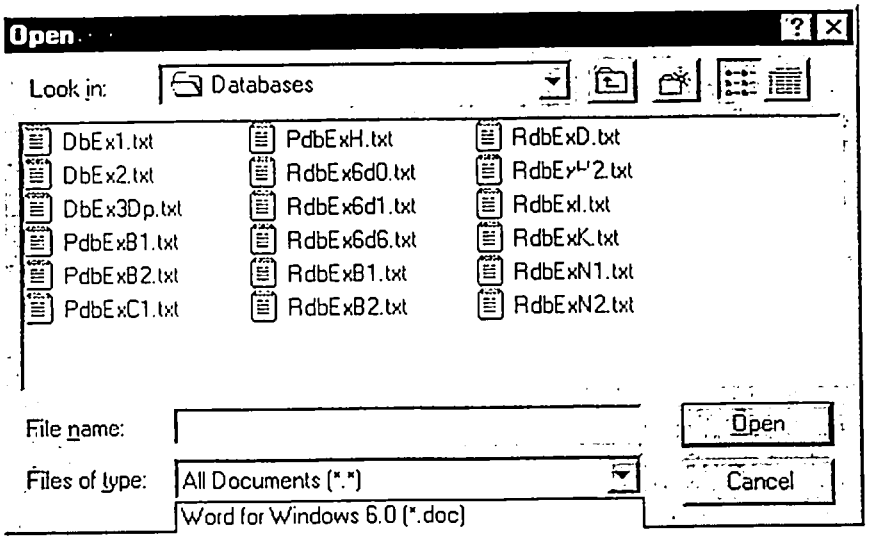
1. Select the "Database" option from the main menu.
2. Select the "Create/Edit Database" option from the drop-down menu. This will open WordPad.



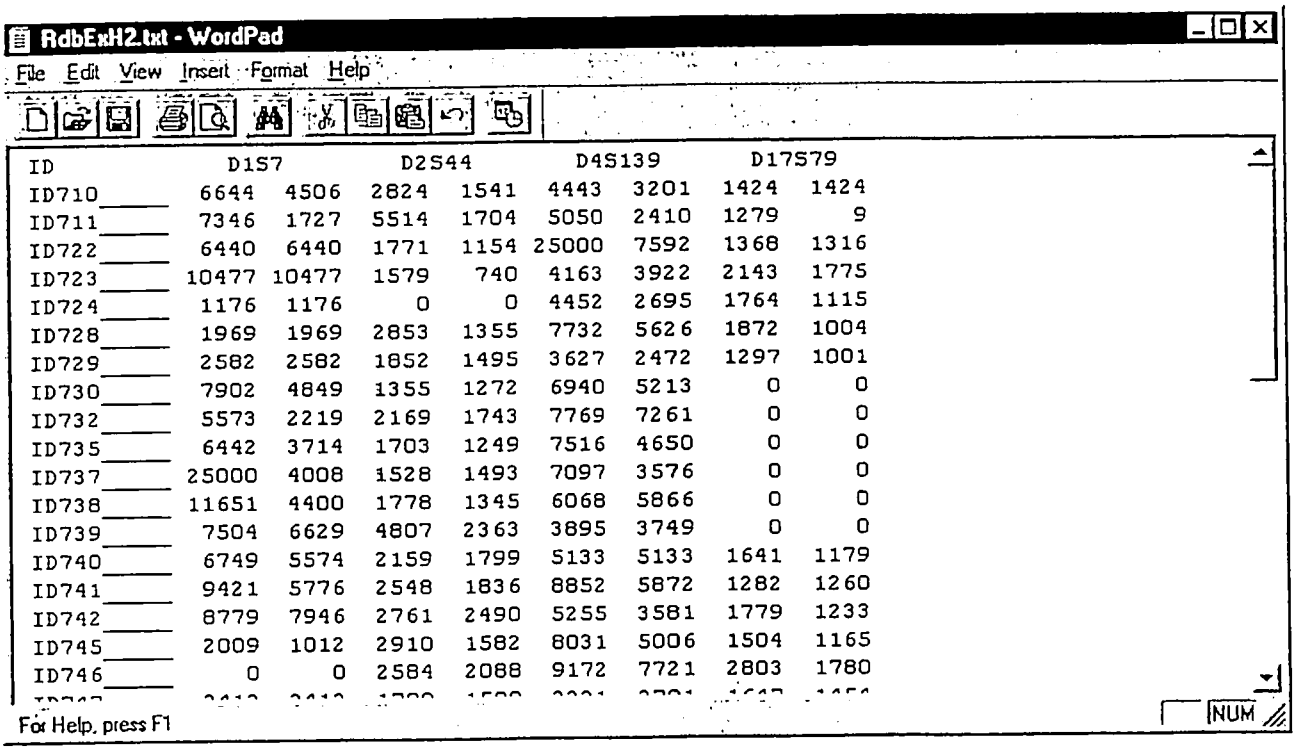
3. Select the file Open option from the WordPad file menu.



4. Change the file type from *.Doc to all files.



5. Select file RdbExH2.txt the list of files displayed.
6. Click Open. This opens RdbExH2.txt in WordPad.

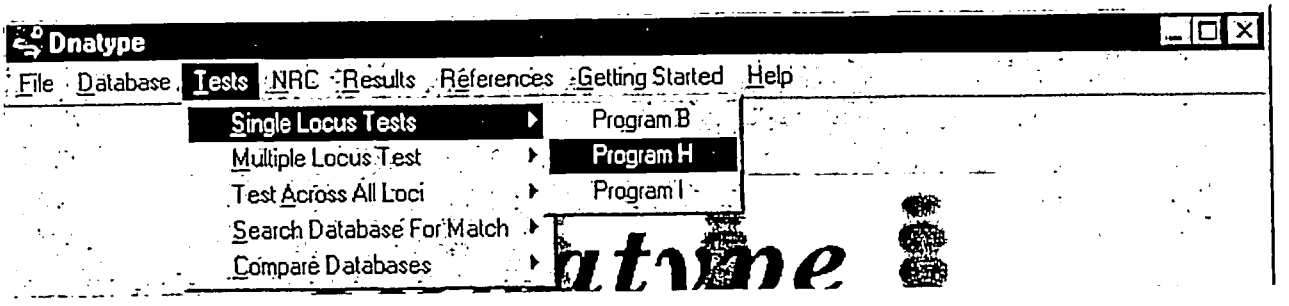


ID	D1S7		D2S44		D4S139		D17S79	
ID710	6644	4506	2824	1541	4443	3201	1424	1424
ID711	7346	1727	5514	1704	5050	2410	1279	9
ID722	6440	6440	1771	1154	25000	7592	1368	1316
ID723	10477	10477	1579	740	4163	3922	2143	1775
ID724	1176	1176	0	0	4452	2695	1764	1115
ID728	1969	1969	2853	1355	7732	5626	1872	1004
ID729	2582	2582	1852	1495	3627	2472	1297	1001
ID730	7902	4849	1355	1272	6940	5213	0	0
ID732	5573	2219	2169	1743	7769	7261	0	0
ID735	6442	3714	1703	1249	7516	4650	0	0
ID737	25000	4008	1528	1493	7097	3576	0	0
ID738	11651	4400	1778	1345	6068	5866	0	0
ID739	7504	6629	4807	2363	3895	3749	0	0
ID740	6749	5574	2159	1799	5133	5133	1641	1179
ID741	9421	5776	2548	1836	8852	5872	1282	1260
ID742	8779	7946	2761	2490	5255	3581	1779	1233
ID745	2009	1012	2910	1582	8031	5006	1504	1165
ID746	0	0	2584	2088	9172	7721	2803	1780

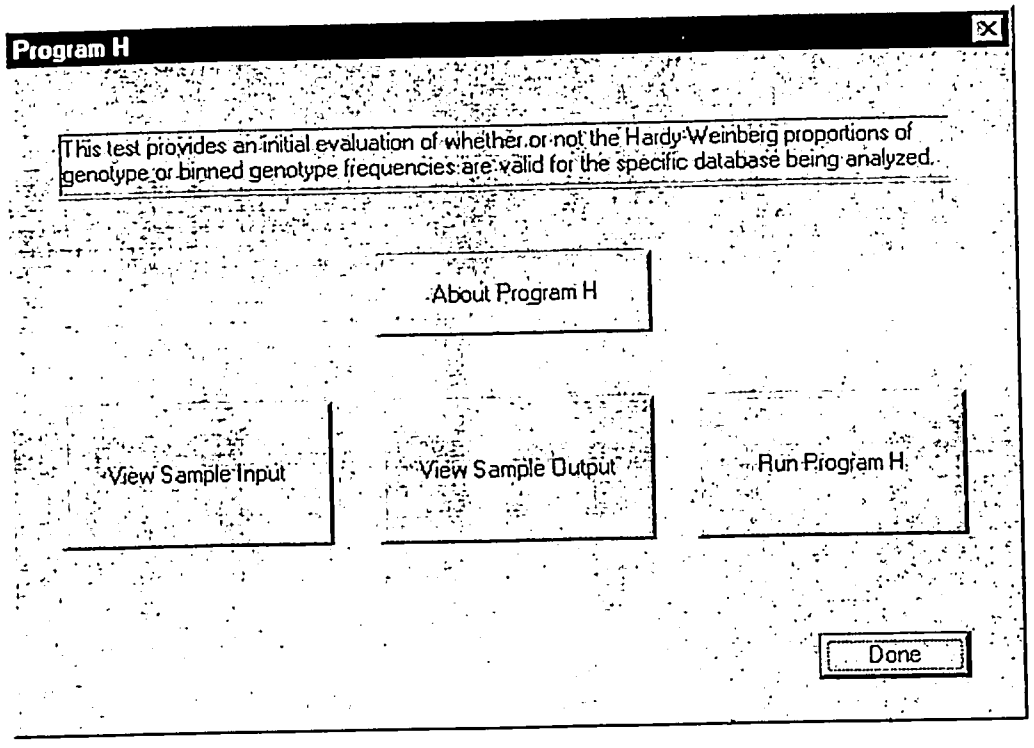
7. After viewing the file, you can either close it or minimize it, proceeding to your next operation.

Running Program H on the Example Database

1. Select the "Tests" option from the main menu.
2. Select the "Single Locus Tests" option from the drop-down menu.
3. Select the "Program H" option from the pop-up menu.

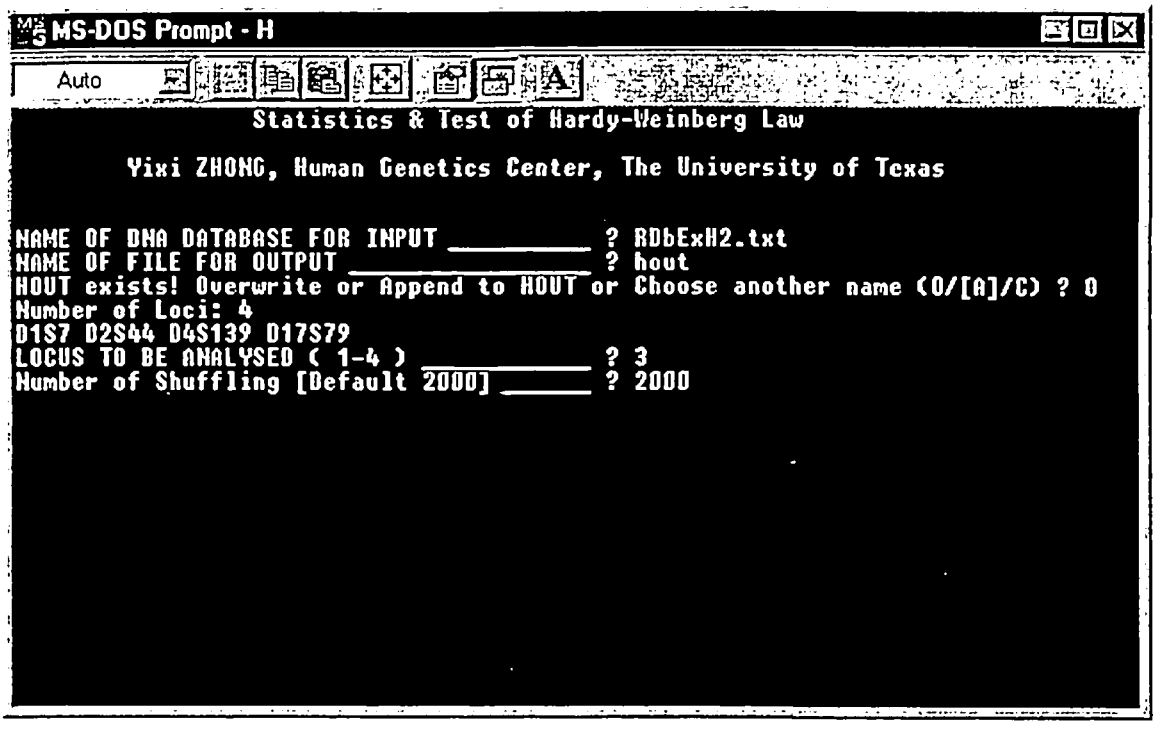


4. This will open a window for Program H.



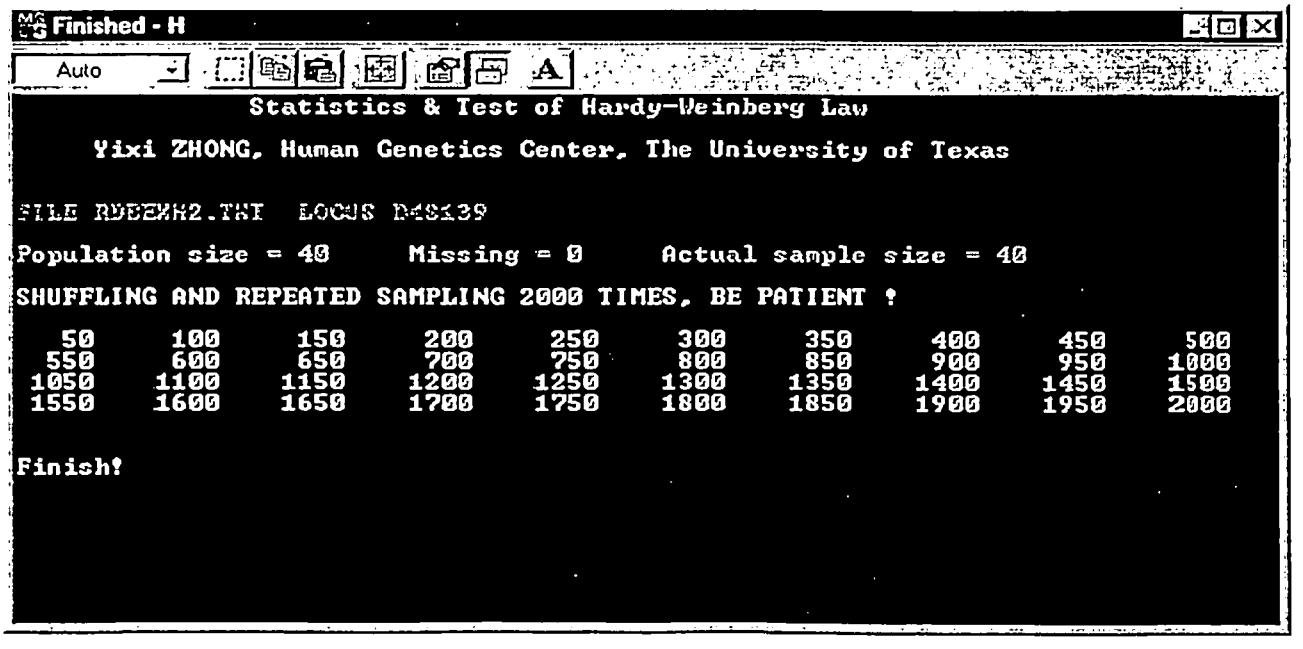
4.1 To view a brief/detailed description of Program H, Click on "About Program H"
4.2 To view a sample input screen and input data, Click on the "View Sample Input"
4.3 To see the output of the program, Click on the "View Sample Output"
4.4 To actually run the test, Click on the "Run Program H" button.

5. This will open a DOS window.



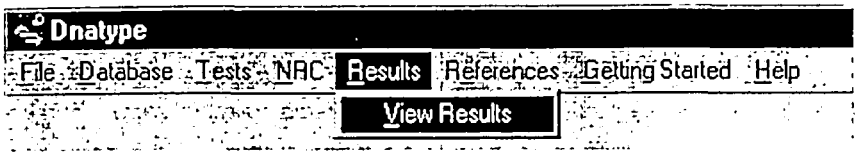
5.1 For the name of the Database, type: RdbExH2.txt
5.3 For the name of the output file, type: hout
5.2 For the locus to be analyzed, type: 3 (corresponding to D4S139)
5.4 For the number of shufflings, hit enter (default 2000) or type: 2000

6. This will perform the test and shut the DOS window.

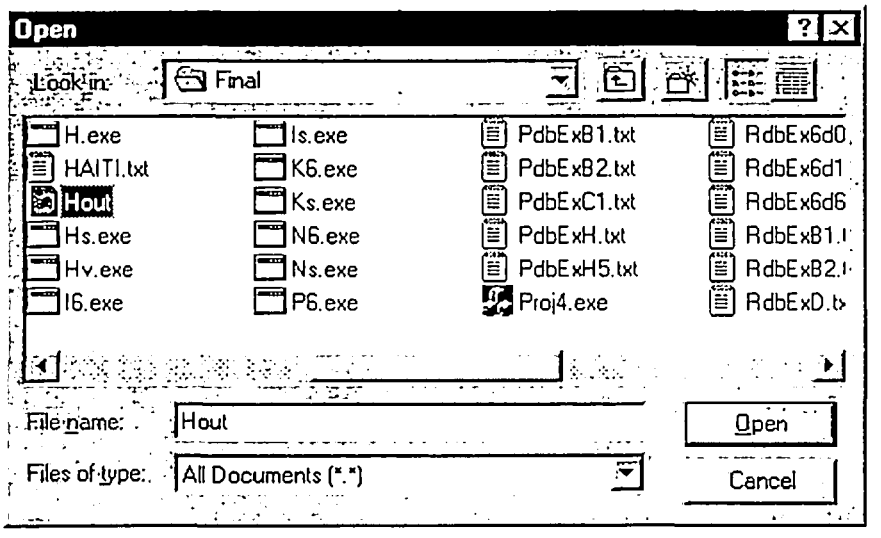


Viewing the output of the Program

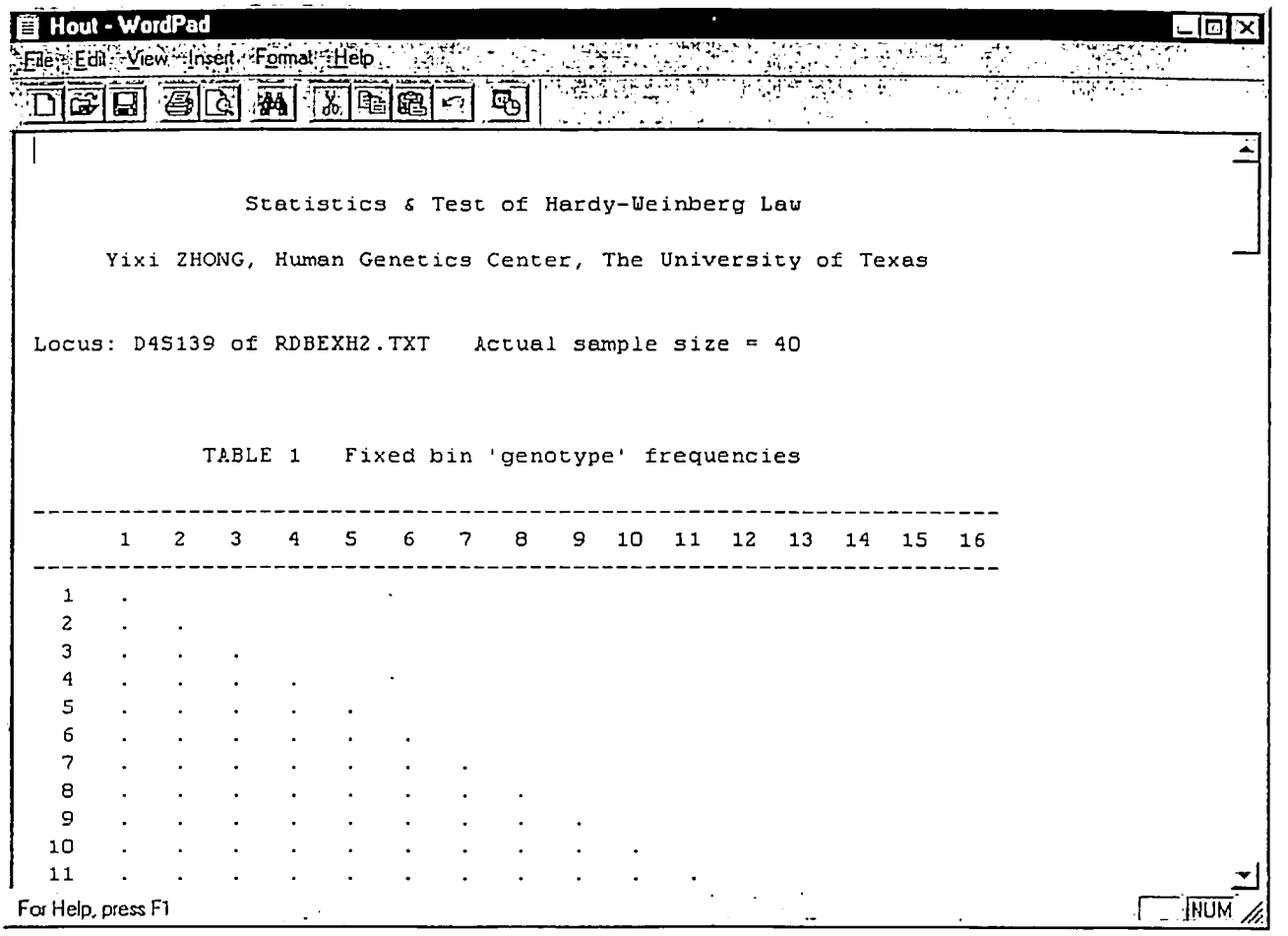
1. Close any windows if open and return to the main screen with the main menu.
2. Select the "Results" option from the main menu.
3. Select the "View Results" option from the drop-down menu.



4. This will open WordPad.
5. Select the file Open option from the WordPad file menu.
6. Change the file type from "*.Doc" to "all files".
7. Select file Hout from the list of files displayed.



8. Click Open. This opens Hout in WordPad.



9. After viewing the file, you can print the file, close it or minimize it, proceeding to your next operation. Look for more information on Program H and on how to interpret the output results by viewing the help files on Program H.

3 RFLP Population Genetics

The programs in DNATYPE consist of software that can be used for the analysis of population data on VNTR fragment sizes scored by Southern blot RFLP analysis.

Genetic variation at variable number of tandem repeats (VNTR) loci are caused by differences in the number of repeat units of short nucleotide sequences that occur tandemly at specific regions of the genome. One method of detecting such polymorphism is the Southern blot, or Restriction Fragment Length Polymorphism (RFLP) analysis, whereby DNA extracted from individuals is digested with specific restriction enzymes (e.g., *Hae*III) which can cut genomic DNA into fragments at specific sites. Electrophoresis of the resulting DNA fragments results in differential migration dependent on the size of the fragments. Immobilization of the DNA on a nylon membrane preserves the order of the restricted DNA fragments relative to their positions on the gel. The DNA on the membranes can then be hybridized with locus-specific probes to detect fragments at specific chromosomal locations. Details of methods of DNA extraction, restriction enzyme digestion, electrophoretic conditions, and probes used for DNA typing following this procedure are described elsewhere (Budowle and Baechtel 1990).

In principle, the alleles at all VNTR loci are truly discrete, and can be represented by the numbers of the repeat units at any such locus (Wyman and White 1980; Jeffreys et al. 1985; Nakamura et al. 1987). However, when variation is detected by the Southern blot RFLP analysis (Southern 1975), the fragment sizes are estimated in base-pair units by comparing their electrophoretic mobility relative to DNA fragments of known molecular sizes run in other lanes of the same gel, i.e., relative fragment sizes in the DNA of a specific individual can be calibrated against the standard sizes of a molecular ladder.

A computer algorithm for this calibration is available for sizing DNA fragments using a digitized method (Monson & Budowle 1989.). Forensic laboratories around the world have generated databases of DNA profiles; these databases typically consist of fragment sizes for many individuals at one or more loci.

The purpose of the DNATYPE programs is to conduct a statistical analysis of such data. In particular, this software: (i) generates fixed bin frequency counts of fragments at each RFLP locus analyzed, from which fixed bin frequency estimates with their respective standard errors are evaluated; (ii) checks the validity of the assumption of independence of genotypes in individuals' profiles within a locus by various statistical procedures; (iii) tests the assumption of independence of genotypes between all possible pairs of loci in the database; (iv) provides a generalized test of the independence of genotypes across all loci in the database; and (v) computes genetic distances between populations to examine the genetic proximity of the sample populations from which such databases are constructed. In addition, the software also performs several utility functions generally meaningful for forensic inferences. For example, search for the presence of any specific DNA profile in a database can be done (with a specified window of measurement-size variation in the case of RFLP loci). Users can also determine if two or more individuals in a database have nearly similar RFLP profiles simply revealed by their estimated fragment sizes.

These tasks are important for several reasons. Since VNTR loci are hypervariable in most populations studied thus far, it can be postulated that any specific multi-locus DNA profile is very rare in any population.

Usually, the frequency is so rare that it cannot be meaningfully obtained by directly counting the number of observations in any database; in particular, the absence of a specific multi-locus profile in a database cannot predict its frequency by counting in the population with any precision (Chakraborty 1992). Therefore, the frequency of a given profile has to be estimated by alternative procedures. The biology of VNTR polymorphism indicates that alleles at such loci are transmitted by principles of Mendelian genetics. Each person inherits one allele from each parent, giving the genotype of the individual. Furthermore, when the probes used are not related to any functional gene, genotypes can be assumed on a biological level to be formed by independent combinations of alleles. This should enable computation of genotype frequencies from allele frequencies using the assumption of Hardy-Weinberg proportions (i.e., independence of alleles within loci). Computed in this fashion, the genotype frequencies for each locus may be multiplied over loci (i.e., product rule), when the different loci segregate independently of each other (which, at the biological level, occurs when the loci are on separate chromosomes, or far apart on the same chromosome). This second assumption is technically known as the assumption of linkage equilibrium (or gametic phase equilibrium).

In principle, both of these two assumptions are well-founded, based on theoretical as well as empirical data. However, there are some factors that might cause these assumptions to be violated. Some factors that may cause apparent or real deviation from the independence assumption both within as well as across loci are the presence of individuals from heterogeneous populations, extensive inbreeding (consanguinity), nonrandom sampling, nondetectability of alleles of certain fragment sizes, and incomplete resolution of alleles of very similar fragment size. Therefore, tests of independence are important elements in checking validity of DNA typing databases for forensic calculations. It is also significant that the statistics generated from the calculations done by these programs are based on the biological rationale of Southern blot RFLP analysis. For example, if a database exhibits significant evidence of allelic non-independence within any locus, other calculations allow the investigator to judge whether or not this observation can be explained by nondetectability of DNA fragments -- a common phenomenon in Southern blot RFLP analysis (Jeffreys et al.1991; Budowle et al.1991; Devlin and Risch 1992; Steinberger et al.1993; Chakraborty et al.1994). Failure of independence also can be checked to examine which combination of fragment lengths (grouped by bins) are responsible for such departures, so that where appropriate, caution may be exercised when using the assumption of independence in specific forensic casework.

4 Databases

The Database Menu choices do the following:

Create / Edit opens WordPad

Database / Examples opens a file showing the names of database files supplied with this software and a description of their content and purpose

Check Database opens a menu to two database checking utility programs (see below)

There are population databases included with this software. Some are actual U.S. population databases. Some are examples, to show the file structure and/or to run with particular programs to generate illustrative results. A section of the Help file below called Database / Examples gives a list of the included databases and a brief description of each.

The databases containing RFLP data exclusively are all named Rdb*****, where the Rdb signifies a RFLP database file and the ***** signifies up to 5 descriptive characters (such as ExC, indicating that the database is an example that can be run with program C to yield particular results). Keep in mind that the programs in DNATYPE are basically DOS based, so that filenames longer than eight characters will not be recognized. Database files that contain data for the PCR-based loci HLA-DQA1, LDLR, GYPA, HBGG, D7S8, GC or D1S80 or any combination of them exclusively, are named Pdb*****, where ***** signifies five descriptive characters. Database files that contain data for the STR loci exclusively are named Sdb*****. Finally, databases that contain data for combinations of RFLP, HLA-DQA1, PM loci, D1S80, and/or STR loci, are named Db******.

All the database files are, and must be, in plain text format (prepared in a text processor and saved as plain text files). The database files included have the extension .txt, because most text processor programs in Windows 95 (e.g Wordpad) will assign this extension by default. The .txt extension is not required by the programs. However, if the text processor assigns the .txt extension by default, or the user does so, then the whole name should be typed when a program asks for the name of the input file (e.g. filename.txt).

We suggest that database files for the programs be constructed using a text editor (such a WordPad. Although it is possible to prepare database files in Excel, there are a number of complications that make this method less than desirable, and actually impossible in some cases (see below).

A number of the programs use the same database input file(s). With programs that analyze single loci, or do pairwise comparisons, users must specify which locus (loci) is (are) to be analyzed. Some of the programs request locus(i) specification by number as well as by name. The "locus number" refers to its order in the rows of the database file. Thus, if D1S7 data is entered first, D1S7 is "locus 1" for purposes of the database file. For this reason, we have maintained a consistent locus order in all the database files furnished with this software. The database files included here have data for up to six RFLP loci, in the order D1S7, D2S44, D4S139, D10S28, D14S13, D17S79. Database files can contain data for up to 16 loci. Users

creating different database files, or entering data for additional loci should keep track of the locus order in which data were entered.
See Database / Examples for a list and brief description of the database files included with this software.

NOTE: USERS SHOULD NOT ADD RFLP TYPING DATA TO THE U.S. TWGDAM DATABASE FILES PROVIDED. THE DATA FROM A USER'S LABORATORY COULD ALREADY BE IN THOSE DATABASE FILES. THE EFFECT OF ADDING DATA COULD BE TO INTRODUCE DUPLICATES AND CAUSE SUBSEQUENT ANALYSIS OF THE DATABASE TO BE INCORRECT AND MISLEADING.

Database Create / Edit

Database / Create/Edit opens WordPad, allowing a user to view, and/or edit any database that is on the disk or drive. WordPad can also be used to create a new database file.

Users can create their own databases which can then be analyzed by these programs, based on their own DNA typing data. All database files that serve as input for these programs are straightforward ASCII text files that can be created in any text processor. It doesn't matter if the text processor is DOS or Windows based. The files must be saved in plain ASCII text format. Some Windows-based text processors will automatically assign a .txt extension to a saved text file. If a database file has a .txt extension, users must remember to give the full name of the file to the data analysis program when prompted to do so (i.e., filename.txt, not simply filename).

We have followed a consistent nomenclature for the database files provided with the software. All RFLP database filenames start with "Rdb," for example. Remember that filenames are limited to eight characters. It is recommended that user-created filename extensions be avoided because of possible conflicts in Windows 95 (A number of file extensions are default-assigned by windows. The details of how Windows does this are not necessarily obvious, and may depend on what programs are actually running on a particular computer). Some example and actual database files have been included. See Database / Examples for more information about the database files and the nomenclature. Users can look at these example files to see the database file structure, and can also use them as "templates" for user-created database files.

Database Structures

Database files have specific structure/syntax requirements. The first row should have column headers, consisting of "ID #" then the names / designators for the loci. Actual data begins in the second row. Column 1 is an alphameric identifier field of up to ten characters, followed by a comma. Columns 2 and 3 are allele bandsizes or names for the two alleles of the first locus; Columns 4 and 5 are allele bandsizes or names for the two alleles of the second locus; etc. The order of loci in the database file is up to the user. The only punctuation in a row consists of the comma after the "ID" designator. The alleles or bandsizes are separated by spaces -- one space is enough, but there can be more than one. A few conventions are important: Missing data (e.g. locus not typed, three-banded RFLP pattern) is entered as 0 0 (zero zero). Single-banded patterns or homozygotes are entered in duplicate, e.g. 2456 2456, or A A.

With RFLP data, bands detected in lanes that are outside the sizing ladder range should be recorded as 9 (for small size fragments), or 25000 (for large unmeasurable sizes). The last row of a database file must consist of at least 2n+1 "-1" (not in quotes) separated by commas, where n is the number of loci in the database file (this row is an end signal). The maximum number of loci that a database file can contain is sixteen.

Special note on RFLP data: Occasionally DNA typing for some loci show multiple bands (more than two bands per locus). Such observations may be caused by presence of restriction site polymorphisms within the VNTR region. These types of profiles cannot be directly interpreted as single locus genotypes. In preparing database files for use with this software, data on such loci should not be recorded in the datafile. It is advised that such loci be recorded as untyped, (place zeroes at the alleles of the locus) and that the investigator keep records of these DNA samples in separate files. Generally, these multibanded profiles do not occur often and can be reported individually. Moreover, the profiles can be examined if there is a suggestion that any peculiar characteristic with respect to these profiles exists at other loci.

An example of part of a RFLP database file is shown below for four loci (D1S7, D2S44, D4S139, and D17S79 - as listed in the top row of the file).

```
ID #          D1S7          D2S44         D4S139        D17S79
ID710_____,  6644  4506   2824  1541   4443  3201   1424  1424
ID711_____,  7346  1727   5514  1704   5050  2410   1279     9
ID722_____,  6440  3299   1771  1154  25000  7592   1368  1316
ID723_____, 11801 10477   1579   740   4163  3922   2143  1775
ID724_____,  5856  1176      0     0   4452  2695   1764  1115
ID728_____,  2458  1969   2853  1355   7732  5626   1872  1004
ID729_____,  5915  2582   1852  1495   3627  2472   1297  1001
ID730_____,  7902  4849   1355  1272   6940  5213      0     0
ID732_____,  5573  2219   2169  1743   7769  7261      0     0
-1,-1,-1,-1,-1,-1,-1,-1,-1,-1
```

An example of part of a mixed locus database file is shown below for two RFLP, two STR, HLA-DQA1, LDLR and D1S80.

```
ID      D2S44       D4S139      THO1     TPOX     HLA-DQ       LDLR    D1S80
10LE  3620  1164  8115  5286  8   8    6   9    1.2  1.3     A   B   29  29
10-1V 3134  2996  6953  4735  7   9    9   11   1.1  1.2     A   B   24  24
100A  5711  1231  5638  4143  7   8    9   11   1.1  4.1     B   B   24  29
101VS 1730  1228  5070  3155  9   9    6   10   1.1  4.2/4.3 B   B   24  26
109A  1131  1131  6909  2206  0   0    0   0    1.1  4.1     A   B   19  24
10AS  290   2169  9037  8174  6   8    9   10   1.1  1.3     A   B   24  29
10H   99999 1234  18528 7669  8   9.3  8   11   2    2       B   B   24  29
10V   2902  2020  7186  5231  7   9    8   8    1.2  1.3     A   B   18  24
12E   0     0     4334  3418  9   9.3  6   8    1.3  2       A   B   18  18
121A  3932  1501  16939 7333  7   8    9   11   1.3  4.2/4.3 B   B   22  24
12A   1721  1328  7074  2211  0   0    0   0    1.1  1.2     A   A   26  31
138S  4373  3454  0     0     0   0    0   0    2    4.1     A   B   21  24
13AB  1230  1230  14997 3217  7   8    8   11   1.3  1.3     A   B   24  24
-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1
```

Creating Database Files Using Microsoft Excel

It is possible to create database files using Excel, but there are some problems that users should be aware of. Excel is capable of saving files essentially as space-delineated text files, and that is what is wanted for the DNATYPE programs. Using Excel to prepare database files has the advantage that the data "cells" are divided, and it may be easier to keep track of data entry than when using a plain text processor. Using Excel may also help to keep columns of data lined up. Although the programs don't necessarily require that columns be lined up, it is easier for a user to keep track of the data when they are lined up.

Generally, column A is the "ID" column, Columns B and C have the alleles of the first locus, columns D and E for the second locus, etc. Row 1 is a "label" row. Typically, "ID" would be placed in A1, the name of the first locus in B1, the name of the second locus in D1, etc. The required series of -1,-1,-1, ... etc. that serves as a file end signal must be entered in column A in the row immediately following the last data profile. Finally, the file must be saved as a "space-delineated text file", one having the .prn extension. Excel provides this file type as an option under its Save As command. Files of the .prn type are really text files, and can be opened, viewed and edited in text processors, such as WordPad.

There are several potential problems that can be encountered when using Excel to create database files.

1. Problems can be encountered when the letters ID are used in Cell A1 of an Excel .prn file. This problem is a property of Excel. The Excel problem can be avoided by placing the letters in quotes (i.e., "ID"). However, because the quotation marks may cause problems with the operation of the DNATYPE programs, it is recommended that the quotes be removed after the database file is completed and saved. This step can be done later in a text processor.

2. The DNATYPE programs will accept files of the .prn type. Users must be sure to provide the full filename to the DNATYPE program, however. Thus, if your database filename is MYFILE.PRN, you have to type in MYFILE.PRN when queried for a database filename; you cannot just type MYFILE. A user can also change the name of a .prn database file prepared in Excel by opening it in WordPad or another text processor, then using the Save As command to rename the file with a .txt extension.

3. Trying to enter -1,-1,-1,-1, ... etc into the last row of column A in an Excel sheet only works if the series is preceded by a single quote ('). Otherwise, Excel tries to treat the series as a formula. Once the database file has been saved in the .prn format, it should be opened in a text processor, and the single quote removed before trying to run the file with DNATYPE programs. Users can also use this opportunity to check that there are 2n+1 "-1," (no quotes) present as an end signal (where n is the number of loci).

Thus, Excel may provide a convenient way to enter a lot of data into a file, but the resulting .prn file should be opened in a text processor like WordPad and made to conform to the requirements of the DNATYPE programs as described above.

Database Examples

This section provides the names we have given to the actual and example databases provided with this software, along with a brief description of each. Databases constructed with text editors have the .txt extension.

Database nomenclature: In DNATYPE, database files containing only RFLP locus data are named Rdb***** (where ***** can be any five character description). Database files that contain data only on the PCR-based loci HLA-DQA1, LDLR, GYPA, HBGG, D7S8, GC or D1S80 or any combination of them, are named Pdb*****. Database files that contain data only on the STR loci are named Sdb*****. Finally, databases that contain data for combinations of RFLP, HLA-DQA1, PM loci, D1S80, and/or STR loci, are named Db******.

RdbExH.txt A six locus database with a total of 224 profiles. The operation of program H is illustrated by running this database with program H for locus 1 (D1S7). There are no entry errors or duplicates. See the Help File for program H for more details.

RdbExH2.txt A four locus database with a total of 40 profiles. The operation of program H is illustrated for null allele calculations by running this database with program H for locus 1 (D1S7). There are no entry errors or duplicates. See the Help File for program H for more details.

RdbExB1.txt An example input file for program B, containing the same data as locus 1 (D1S7) in the RdbExH.txt database file. Program B requires different input, but outputs results very similar to program H. See the Help files for program B for more details.

RdbExB2.txt An example input file for program B, containing the same data as locus 2 (D2S44) in the RdbExH.txt database file. Program B requires different input, but outputs results very similar to program HR. See the Help files for program B for more details.

RdbExI.txt A six locus database with a total of 224 profiles. The operation of program I is illustrated by running this database with program I for locus 3 (D4S139). There are no entry errors or duplicates in the database. See the Help File for program I for more details.

RdbExK.txt A four locus database with a total of 40 profiles. There are no entry errors or duplicates. The operation of program K is illustrated by running this database with program K for the loci D2S44 and D4S139. See the Help File for program K for more details.

RdbExD.txt A four locus database with a total of 40 profiles. There are no entry errors or duplicates. The operation of program D is illustrated by running this database with program D. See the Help File for program D for more details.

RdbExN1.txt and RdbExN2.txt

These are six locus databases with a total of 224 and 329 profiles, respectively. There are no entry errors or duplicates. The operation of program N is illustrated by running these databases with program N. See the Help File for program N for more details.

RdbEx6d0.txt and RdbEx6d1.txt

These are six locus databases with 101 profiles. There are no entry errors or profile duplicates. The operation of programs S and T can be illustrated by running RdbEx6d0.txt with the programs. RdbEx6d1.txt is identical to RdbEx6d0.txt except that one record has an identical D1S7 profile to another record. See Help files for programs S and T.

RdbEx6d6.txt

This is a six locus database with 101 distinct profiles. There are no entry errors, but two profiles are duplicated. The operation of program CDupl can be illustrated by running this database with the program.

The following are the five filenames of the collected U.S. population RFLP six-locus databases from the TWGDAM laboratories (Ca Caucasian, Bl Black, Hs Hispanic, In Amerindian and Or Oriental):
RdbTwgCa.txt, RdbTwgBl.txt, RdbTwgHs.txt, RdbTwgIn.txt, RdbTwgOr.txt

NOTE: USERS SHOULD NOT ADD RFLP TYPING DATA TO THE U.S. TWGDAM DATABASE FILES PROVIDED. THE DATA FROM A USER'S LABORATORY COULD ALREADY BE IN THOSE DATABASE FILES. THE EFFECT OF ADDING DATA COULD BE TO INTRODUCE DUPLICATES AND CAUSE SUBSEQUENT ANALYSIS OF THE DATABASE TO BE INCORRECT AND MISLEADING

PdbExH.txt A data input file for several programs (including H) containing 210 profiles for the HLA-DQA1, five Polymarker and D1S80 loci. Not every individual is typed at every locus. Operation of programs H, D and others can be illustrated by running them with this database file.

PdbExB1.txt A data input file for program B containing data for the HLA-DQA1 locus identical to that in DbEx1.txt. Operation of program B can be illustrated by running it with this database file.

PdbExB2.txt A data input file for program B containing data for the GYPA locus identical to that in DbEx1.txt. Operation of program B can be illustrated by running it with this database file.

PdbExC1.txt A data input file containing 210 profiles for the HLA-DQA1, PM and D1S80 loci. Not every person is typed at every locus. Operation of programs S and T can be illustrated by running them with this database file.

SdbEx1.txt A data input file containing 100 profiles for the "CTT" STR loci (THO1, TPOX and CSF1PO). Not every person is typed at every locus.

DbEx1.txt A sixteen locus database with a total of 100 profiles for the six RFLP loci D2S44, D1S7, D17S79, D4S139, D10S28, D17S26, the three STR loci THO1, TPOX, CSF1PO, HLA-DQA1, five Polymarker loci and D1S80. There are no entry errors or duplicates.

DbEx2.txt A sixteen locus database with a total of 100 profiles for the six RFLP loci D2S44, D1S7, D17S79, D4S139, D10S28, D17S26, the three STR loci THO1, TPOX, CSF1PO, HLA-DQA1, five Polymarker loci and D1S80. There are several data entry errors (illegal characters such as %, <, >, letters in number fields) that are found when this database file is run with program CError. In addition, row 2 is blank in this database file (labels are in row 1; data begins in row 3). Program CError also finds the empty row defect.

DbEx3Dp.txt A sixteen locus database with a total of 100 profiles equivalent to DbEx1.txt, except that two complete profiles are duplicated. The operation of program CDupl can be illustrated by running it with this database file.

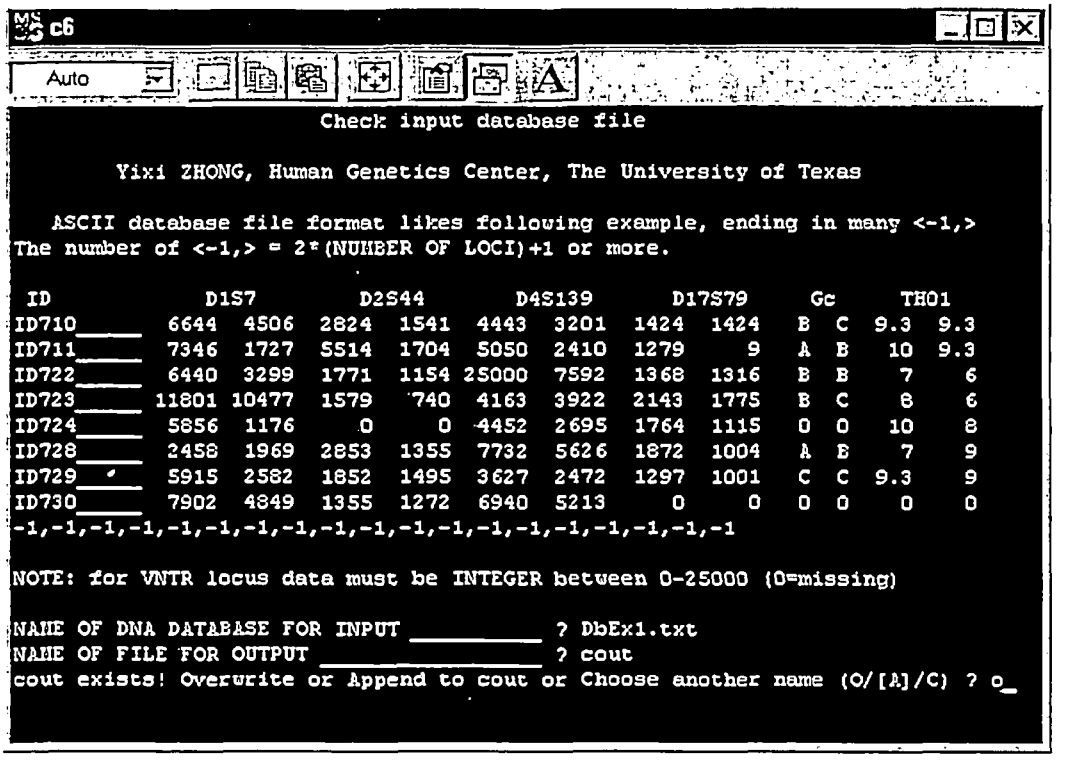
Database Checking

There are two database checking programs: CError and CDupl.

Program CError

Clicking on Database / CError will run a program whose purpose is to check the correctness of the data format in the database file. Generally, the program checks for correctness of the formatted data, and for blank rows, illegal characters in fields, etc.

Running CError with the database file DbEx1.txt (which contains no errors) is shown on the input / dialog screen below:



The user must enter the name of the database file (in this case DbEx1.txt), a name for the output file (in this case cout) -- and if the filename selected already exists (as it does here), decide whether to Overwrite / Append / Change the output filename.

The database file DbEx1.txt contains no errors, and yields the following output file.

```
                    Check input database file

          Yixi ZHONG, Human Genetics Center, The University of Texas

Input database name = DBEX1.TXT

No errors were found in the input file DBEX1.TXT
```

Running CError with the database file DbEx2.txt, however, yields the following output file.

```
                    Check input database file

          Yixi ZHONG, Human Genetics Center, The University of Texas

Input database name = DBEX2.TXT

 32 bands expected in line 1 But-1 bands found
Illegal character <>> found in line 4
1003 >5711 1231 6755 3344 2004 1527 5638 4143 2924 1369 1932 1501 7 8 9 11 11 12 1.1 4 B B A B
  B B  A A  B C  24 29
Illegal character <<> found in line 5
1004  1730 <1228 5588 768 1413 1413 5070 3155 1166 990  4172 1392 9 9 6 10 10 13 1.1 4 B B A A
```

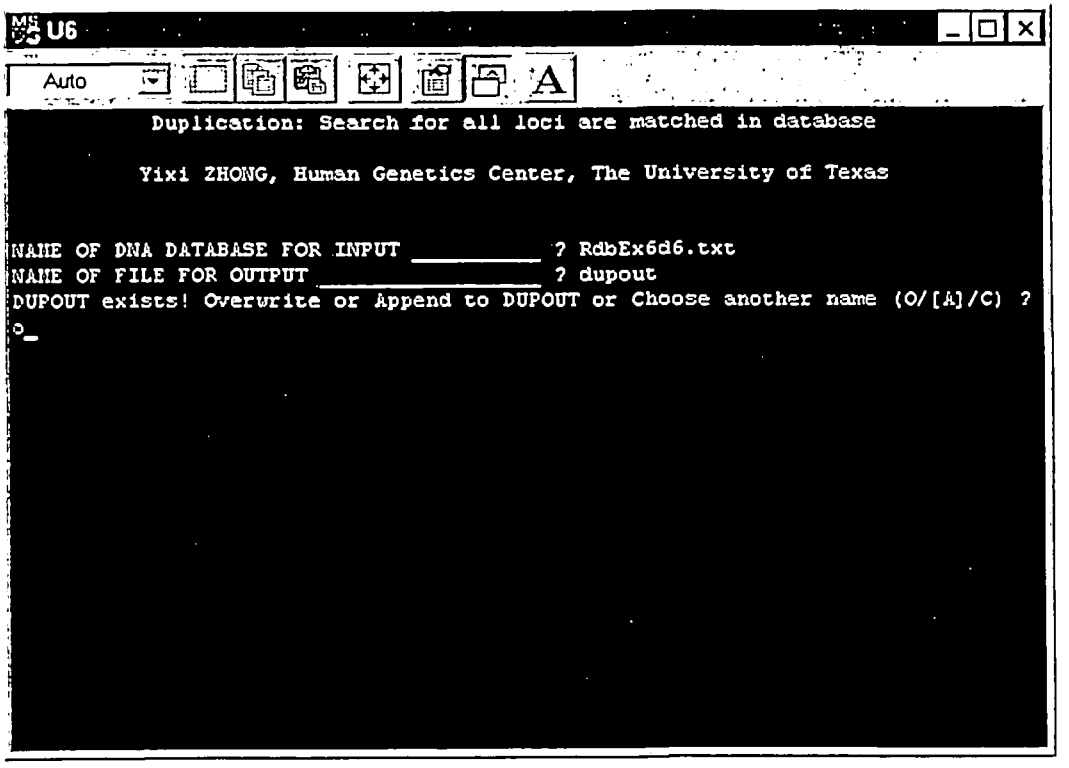
```
  A B  A A  C C  24 26
Illegal character <A> found in line 6
1005 11A1  1131 11532 8166 13%1   9 6909 2206 2188 2188 1857 1499 0 0 0 0  0 0   1.1 4 A B A B
  A B  A B  B C  19 24
Illegal character <%> found in line 6
1005 11A1  1131 11532 8166 13%1   9 6909 2206 2188 2188 1857 1499 0 0 0 0  0 0   1.1 4 A B A B
  A B  A B  B C  19 24
Illegal character <P> found in line 7
1006 2900   21P9 7294  6072 2165  1480 9037 8174 6822 1891 3034 1964 6 8 9 10 8 12 1.1 1.3 A B
  A A  A B A A  C C  24 29
```

This file has a blank row 2 (where data entry should begin) that is detected. In addition data entries within the RFLP loci contain several illegal characters that are detected as shown. The program reports the line number of the erroneous line, and reproduces it in the output file.

Program CDupl

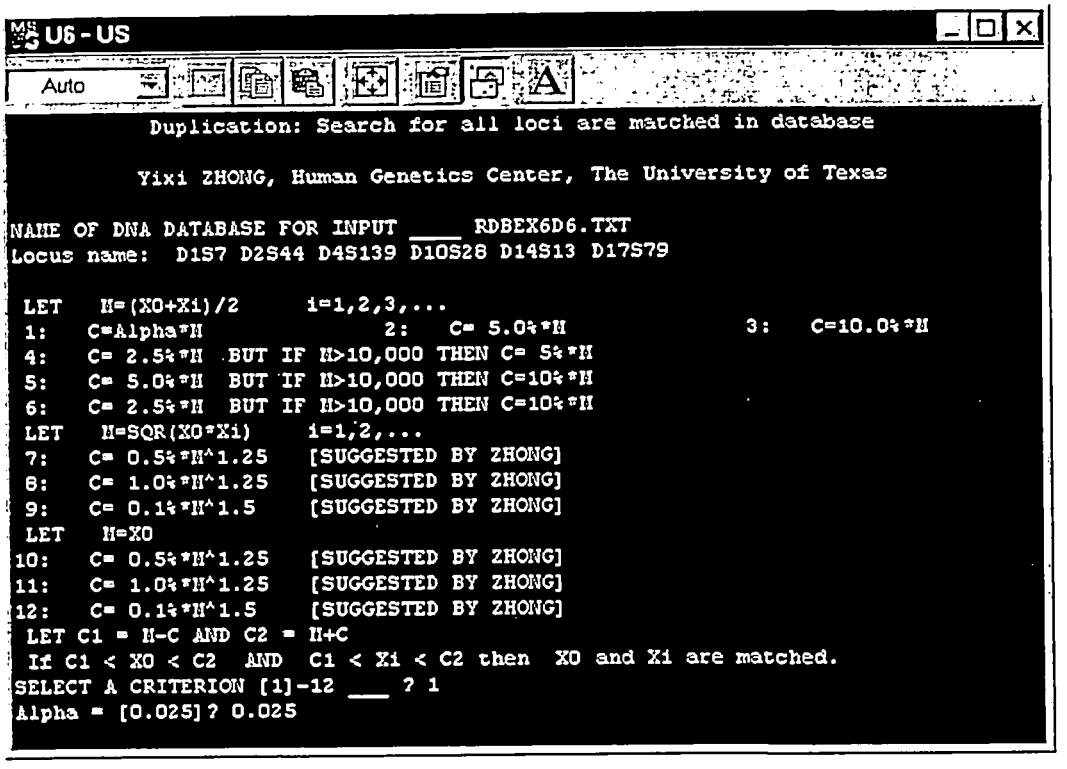
The CDupl program is a utility that checks database files for duplicates. If the database file contains any RFLP locus data, the program offers a series of criteria for user selection in searching for duplicates. If the simplest "window" criterion is chosen, the user can select a +/- value for the window (called alpha). 0.025 is the default alpha value, i.e. ± 2.5%). Should matches be found, the records are displayed on the screen output (and saved in the user-specified output file) showing the complete individual profiles and ID numbers that match each other.

The CDupl program input screen and dialog when run with RdbEx6dp6.txt is shown below:



The program requests the input filename. RdbEx6d6.txt was given. It next requests an output filename. "dupout" (no quotes) was given. Since "dupout" existed, the program gave the options of overwrite (O), append (A) or choose (C) a new name. Overwrite was selected. Next, the program

reiterates the name of the input database file, and lists the loci in the database file. Since the database file contains RFLP locus data, the match / window criteria dialog appears.



In this case, the user chose choice 1 (C=Alpha*M) and typed 0.025 for alpha (hitting <Return> to accept the default value of alpha -- 0.025 or 2.5% -- would have accomplished the same thing.

The program reports that ID # F0063 at line 63 matches ID # F0063D16 at line 64 (NL = 6 means "number of loci = 6), and that ID # F0097 matches ID # F0097D16, and gives all bandsizes and locus names. These are the duplicates that were introduced into this example database file to illustrate the operation of program CDupl.

The output file from running program CDupl with RdbEx6d6.txt is shown below:

```
          Duplication: Search for all loci are matched in database

         Yixi ZHONG, Human Genetics Center, The University of Texas

NAME OF DNA DATABASE FOR INPUT ____ RDBEX6D6.TXT

Locus name:  D1S7 D2S44 D4S139 D10S28 D14S13 D17S79

 LET   M=(X0+Xi)/2      i=1,2,3,...
 1:   C=Alpha*M
 2:   C= 5.0%*M
 3:   C=10.0%*M
 4:   C= 2.5%*M  BUT IF M>10,000 THEN C= 5%*M
 5:   C= 5.0%*M  BUT IF M>10,000 THEN C=10%*M
```

```
 6:   C= 2.5%*M  BUT IF M>10,000 THEN C=10%*M
 LET   M=SQR(X0*Xi)     i=1,2,...
 7:   C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
 8:   C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
 9:   C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET   M=X0
10:   C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
11:   C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
12:   C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET C1 = M-C AND C2 = M+C
 If C1 < X0 < C2  AND  C1 < Xi < C2 then X0 and Xi are matched.
SELECT A CRITERION 1-12 ___  1
   Alpha = .025

ID = F0063_____ Line    63  MATCH WITH  ID = F0063D16__ Line    64    NL = 6
 6469/ 1892  2177/ 1613  7207/ 6951  1853/ 1661  5018/ 1294  1762/ 1330
 6469/ 1892  2177/ 1613  7207/ 6951  1853/ 1661  5018/ 1294  1762/ 1330
ID = F0097_____ Line    98  MATCH WITH  ID = F0097D16__ Line    99    NL = 6
 4480/ 2235  1925/ 1617  9486/ 8577  3730/ 2290  8869/ 1811  1542/ 1423
 4480/ 2235  1925/ 1617  9486/ 8577  3730/ 2290  8869/ 1811  1542/ 1423
  D1S7        D2S44       D4S139      D10S28      D14S13      D17S79
Above pairs: All loci are matched; They maybe duplicated
```

A second example of CDupl is illustrated by running it with an example database file called DbEx3Dp.txt. This database file contains two complete profile duplicates, but is otherwise the same as DbEx1.txt. The output file from running CDupl with DbEx3Dp.txt is shown below.

```
          Duplication: Search for all loci are matched in database

          Yixi ZHONG, Human Genetics Center, The University of Texas

NAME OF DNA DATABASE FOR INPUT ____ DBEX3DP.TXT

Locus name:  D2S44 D1S7 D17S79 D4S139 D10S28 D17S26 THO1 TPOX CSF1PO HLA-DQ
LDLR GYPA HBGG D7S8 GC D1S80

 LET   M=(X0+Xi)/2      i=1,2,3,...
 1:   C=Alpha*M
 2:   C= 5.0%*M
 3:   C=10.0%*M
 4:   C= 2.5%*M  BUT IF M>10,000 THEN C= 5%*M
 5:   C= 5.0%*M  BUT IF M>10,000 THEN C=10%*M
 6:   C= 2.5%*M  BUT IF M>10,000 THEN C=10%*M
 LET   M=SQR(X0*Xi)     i=1,2,...
 7:   C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
 8:   C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
 9:   C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET   M=X0
10:   C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
11:   C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
12:   C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET C1 = M-C AND C2 = M+C
 If C1 < X0 < C2  AND  C1 < Xi < C2 then X0 and Xi are matched.
```

```
SELECT A CRITERION 1-12 ___  1
   Alpha = .025

ID = 1001       Line     1  MATCH WITH  ID = 1001A      Line     2    NL =16
 3620/ 1164  4794/ 4794  1518/ 1404  8115/ 5286  3720/ 1419  2165/ 1533
8.0/  8.0   6/  9  10/ 11   1.2/  1.3  A/B   A/A   B/B   A/B   B/C   29/ 29
 3620/ 1164  4794/ 4794  1518/ 1404  8115/ 5286  3720/ 1419  2165/ 1533
8.0/  8.0   6/  9  10/ 11   1.2/  1.3  A/B   A/A   B/B   A/B   B/C   29/ 29
ID = 1013       Line    11  MATCH WITH  ID = 1013A      Line    12    NL =16
 3932/ 1501  5418/ 3724  2243/ 1283 16939/ 7333  2999/ 2554  3034/ 2729
7.0/  8.0   9/ 11  11/ 11   1.3/  4.0  B/B   A/A   A/B   A/B   C/C   22/ 24
 3932/ 1501  5418/ 3724  2243/ 1283 16939/ 7333  2999/ 2554  3034/ 2729
7.0/  8.0   9/ 11  11/ 11   1.3/  4.0  B/B   A/A   A/B   A/B   C/C   22/ 24
  D2S44       D1S7        D17S79      D4S139      D10S28      D17S26
THO1        TPOX    CSF1PO  HLA-DQ      LDLR  GYPA  HBGG  D7S8  GC    D1S80
Above pairs: All loci are matched; They maybe duplicated
```

Note: Program CDupl does not identify profiles as duplicates unless they are duplicated at all the loci in the database, i.e., the program is not designed to find partial duplicates (profiles that are the same at some loci but not at others).

When duplicates or RFLP "matches" are found, they may warrant further investigation. It should be noted that some computer-detected matches may not truly be "forensic" matches. Budowle et al. (1991) clearly state that the determination of a forensic match with a series of RFLP loci is a two-step procedure. Two DNA profiles are considered a forensic match only when they pass the visual match test. Since computer matches of profiles fragment by fragment, e.g., two fragments match each other if windows of ±2.5%, opened around each fragment are found overlapping, i.e., two fragments x and y are match if for $x_1 < x_2$, $x_1 + cx_1 > x_2 - cx_2$, it may so happen that the profile at a locus (x_1, y_1) would be matched by the computer with a profile (x_2, y_2) where for $x_1 > y_1$ and $x_2 > y_2$, it is found that $x_1 < x_2$ and $x_1 + cx_1 > x_2 - cx_2$ (so that x_1 and x_2 are declared to be match), but $y_1 > y_2$ with $y_1 - cy_1 < y_2 + cy_2$ (so that y_1 is called a match of y_2). When the differences between x_1 and x_2 and/or y_1 and y_2 are large (but not outside the match window), the compared profiles may not pass the visual match test. This scenario is the one described in Figure 1b of Budowle et al. (1991), and there is evidence in database analyses that matches detected by a computer search alone do not necessarily equate to forensic matches. Nonetheless, a search for matches in a database is instructive, to weed out shared duplicate samples between analysts, or to weed out duplicate blood samples from the same person. It helps to evaluate if sampling could be the cause, e.g. blood donors are known to donate blood at frequent intervals, and sometimes under different names and identifications.

Certain Automatic Correction Features

As noted, program CError finds certain types of data entry errors and reports them to the user. However, certain data entry errors are automatically "corrected" by DNATYPE programs without notifying users of their existence. The most significant example of such an error -- one that could change the calculations slightly -- is one where the two alleles for a locus were entered as a value and as a zero, by mistake, e.g. 2345 0 for an RFLP locus entry. The program automatically

treats this entry as a single-banded pattern and "corrects" it to read 2345 2345 for purposes of calculation. If the user had intended to enter 0 0, the automatic correction used by the programs would cause a slight calculation error. RFLP band size entries greater than 25000 are automatically corrected to 99999, and band size entries smaller than the smallest ladder fragment are automatically corrected to 9. These latter errors, though they are not "corrected" in the actual database file, should not cause any calculation errors.

Program CError may also add the 2n+1 "-1" separated by commas as an end signal as the last line of a database file if a user has omitted it. Users should check database files to make sure that there are no blank rows between the last data entry row and the end signal row. Some programs can treat such blank rows as empty data rows and the calculations may be affected.

5 Tests / Programs

Check Database Programs

The two check database programs CError and CDupl are discussed in Section 4 Databases.

Program H Allele or Fixed-Bin Frequencies and Tests for HWE - Independence at a Single Locus

This routine generates a series of outputs: fixed bin fragment size frequencies and several tests performed to check fragment size independence within a locus. In other words, this function provides an initial evaluation of whether or not the Hardy-Weinberg proportions of binned genotype (for RFLP) frequencies are valid for the specific database being analyzed. If a "departure" is detected, it warrants further evaluation to determine effects. It is recommended that users assure that the data format is correct (using program CError) and that the database contains no duplicates (using program CDupl) prior to running this test.

Program H Results / Output

The results of this program are the most frequently used descriptors of a database analysis. In this version of the software, DNA fragment sizes are binned by following the fixed bin boundaries as defined in Budowle et al. (1991). This generates 31 bins, although simple changes in the source code can provide alternative bin boundaries, as well as the number of bins. The output of this routine consists of four tables of information (There are actually four examples of Test H example database input and output The first two, for D1S7 of RdbExH.txt and another one for D1S7 of RdbExH2.txt, show analyses of RFLP loci. The second two for HLA-DQA1 and for GYPA of the example database DbEx1.txt, show analyses of PCR-based loci.

Program H Information and Calculations

This routine generates a table of fixed bin fragment size (for RFLP loci) or allele frequencies (for PCR loci). Additionally, several tests are performed to check allele or fragment-size independence within a locus. This program provides an initial evaluation of whether or not the Hardy-Weinberg proportions of genotype or binned genotype frequencies are valid for the loci in the specific database being analyzed. If a test indicates significant deviation from HWE expectation, the source of deviation should be investigated. It is recommended that users ensure that the data format is correct using program C prior to running this test (or other tests).

Program H calculates several of the most commonly used statistics or descriptors of a population database. With RFLP loci, DNA fragment sizes are binned by following the fixed bin boundaries as defined in Budowle et al. (1991), which produces 31 bins. Simple changes in the source code can change the bin boundaries and change the number of bins. Four tables are produced as the output of program H. (Test H Example Output for an example). The table data differ according to whether RFLP loci or PCR/STR loci were analyzed.

RFLP Locus Analysis:

The first table (Table 1) is self-explanatory. For RFLP data it provides counts of binned profiles, with dots representing types not observed in the database, and the numbers representing the frequency (count) of individuals whose two fragments reside within the indicated bin or within the genotype. The header of the table indicates the locus and database name for which this table is created. At the bottom is the total number of individuals for which frequency counts are given in the table. This may be smaller than the sample size (actual sample size mentioned in the header), because for any database, some individuals may not be typed at that locus.

The next table (Table 2) lists the bin number and bin boundaries (the chosen bin boundaries of the analysis), chromosomal counts of fragments within each bin, frequencies in percents, and upper and lower 95% confidence interval estimates of these frequencies (in percent) calculated using Goodman's algorithm (Goodman 1965). The last column shows the count of number of individuals with single-banded profile (any fragment within the same bin) at the locus. The header identifies the locus and database for this table. The total number reflects the number of chromosomal counts used in the frequency calculations. Again, this number at any locus may be smaller than twice the number of individuals (stated in the header), because of possible missing data.

The binned counts are analogous to gene count estimators (Li, 1976) of allele frequencies. Since a single-banded genotype has the alleles listed twice (i.e., a heterozygous profile might be "8974,11760" whereas a single-banded profile that measured 8974 would be entered as "8974,8974"), it is counted twice in the appropriate bin. When fragment sizes are noted as 9 or 99999, the respective fragment is assumed to be present in the profile, but its frequency is counted in the first bin (if 9), or in the last (when it is 99999). Apart from true homozygosity, a type may be single-banded for at least two other reasons. One is incomplete separation of bands of nearly equal size (Devlin et al., 1990). However, allelic coalescence does not substantially affect the binned frequency counts, since alleles of similar size usually fall within the same bin (the rare exception is when the two alleles straddle a bin boundary). Second, a single-banded pattern may also be due to heterozygosity for a non-detectable allele; this has a more appreciable effect. A non-detectable allele may be due to a very large or very small allele, as observed at some RFLP loci. Such alleles are classified as non-detectable or null. The presence of these alleles cause the bin counts to overestimate the true bin frequencies (Chakraborty et al. 1992, 1994). If the frequency of null alleles in the population is high enough, an "excess of homozygosity" may be observed, and the genotype frequencies may show significant deviation from HWE. Other than these, there is no other assumption involved in the frequency computations.

This second table alone could be a summary descriptor of the data, when fixed bins are used for any frequency computation. The listing of the actual chromosomal counts along with relative frequencies is enough information to produce a different binning scheme (Budowle et al. 1991). This would usually be used to merge bins with small absolute counts (NRC 1996).

The next part of this output represents the results of four different tests on the binned genotype and allele counts to check for independence of binned fragment sizes, i.e. do the binned genotype frequencies (based on the fragment sizes grouped in 31 fixed bins) agree with their Hardy-Weinberg expectations (p^2 for homozygotes, and 2pq for heterozygotes)? The four tests are: (i) a chi-square test based on total counts of homozygotes and heterozygotes; (ii) a test based on number of distinct genotypes seen in the sample (Chakraborty 1993a,b); (iii) the likelihood ratio test based on contrasts of observed and expected frequency after binning (i.e., if a type's bands fall into different bins, it is a heterozygote; otherwise, it is considered a

homozygote) of each of the 496 binned genotypes in the data (Weir 1991) (if there are k bins, the possible number of binned genotypes is k(k+1)/2 or in this case 31 x 32 / 2 = 496); and (iv) the exact test of Guo and Thompson 1992. Details are given in the cited references.

Even when binned, there are usually many genotypes that are either absent in the database, or present with very low frequencies (1 or 2). This implies that large-sample approximations of the test statistics are not applicable (Chakraborty et al. 1991; Weir 1992 and Chakraborty et al. 1994 for detailed explanation). The method used by program H to determine significance is by a random permutation of the data or "reshuffling" (Chakraborty et al. 1991). Consult Shuffling Routine topic for more information.

PCR / STR Locus Analysis:

Program H output here is similar but not identical to that described above for RFLP loci. Table 1 is a graphical representation of genotype frequencies. Table 1A lists observed and expected genotype frequencies, with expected frequencies calculated using both biased and unbiased methods. The difference between these two estimates is discussed under the topic Chi-square test based on total counts of homozygotes and heterozygotes. (see Note 2 for an explanation of the omission of this Table for an RFLP locus).

Table 2 lists the alleles with number, percent and S.E. (%) and provides a column that counts the number of homozygotes observed for that allele in the database.

Tables 3 and 4 provide the same information for these loci that is provided for RFLP loci, namely observed and expected frequency contrasts using both biased and unbiased methods of estimating expected numbers, and numbers of distinct genotypes with S.E.

Shuffling routine: In running this program, the data on fragment sizes or genotypes are shuffled 2000 times (or any number of times desired) to generate random distributions of test criteria for this as well as other menus of this software. With data on n individuals, from the list of $2n$ observed fragment sizes or alleles, two fragments or alleles are randomly selected to form a simulated DNA profile. This process is repeated with the remaining $2n$-2 fragment sizes to obtain the simulated profile of the second individual. This resampling is continued until the list of all fragment sizes or alleles is exhausted. Once this process is complete, it constitutes a single replicate database of shuffled data. When this is treated as a simulated database, the fragment size or allele distribution remains unaltered, but any summary statistic (e.g., counts of total number of heterozygotes or homozygotes, likelihood ratio, exact test probability, etc.) will vary from replicate to replicate. The summary statistic may be computed for each run of this shuffling routine without any change in the fragment size distribution in the sample. Note that the original database file is unaffected by the shuffling procedure; all the shuffling is done in computer memory. Also note that because the shuffling process assorts the data randomly, repeated shufflings of the same data even for the same number of times will not yield identical results (although they will be very similar). Therefore, if a user were to rerun the examples presented in the Help files, the results will be very similar but some will not be identical to the output files shown.

Chi-square test based on total counts of homozygotes and heterozygotes: Table 3 lists the observed numbers and proportions in percents of homozygotes and heterozygotes along with their expectations (counts and frequencies in percent). Recall that heterozygosity and homozygosity for RFLP loci are defined at the level of bins in these computations. In other

words, individual profiles where the two fragments are within the same bin (even if the profile is a two-banded pattern) are counted operationally as homozygotes (and see Note 1). Likewise, when the fragment sizes are in different bins, the individual is defined as a heterozygote. Clearly, this may not reflect the actual heterozygosity at a VNTR locus. Coalescence of fragments of nearly equal size, as well as neglecting the possibility of nondetectable alleles, will lead to underestimation of the actual heterozygosity. Two different sets of numbers, labeled "biased" and "unbiased", are under the expected column. Under the assumption of independence, the probability of homozygosity at a locus Σp_i^2 estimated by replacing p_i with its gene count estimator is not an unbiased estimate of the population homozygosity (even when the Hardy-Weinberg assumption is appropriate). Nei (1978) provides an alternative estimator, which corrects for bias of this estimator. In large samples (say, number of individuals larger than 100), the bias correction does not appreciably change the estimate. However, for the sake of completeness, both estimates are used to check whether or not the observed number of homozygotes (and consequently, the number of heterozygotes) deviate significantly from that predicted under the independence assumption. The statistic of deviation, a goodness-of-fit chi-square shown in this table, should follow a χ^2 distribution with 1 degree of freedom, and hence, the level of significance (probability) is determined by the probability that a χ^2_1 variate exceeds the observed goodness-of-fit chi-square value, which is listed at the bottom of this table. When this probability is small (say, less than 0.05), one could infer that the data are at discrepancy with the prediction of the independence rule at the level of total homozygosity (or heterozygosity). The test based on unbiased estimators of homozygosity (or, heterozygosity, the complement of homozygosity) is preferred, since it allows one to examine whether or not the observed deviation is due to the sampling bias of the estimation of homozygosity and heterozygosity.

Three aspects of this test procedure are worth noting. First, even though the approximation of this goodness-of-fit test statistic by a χ^2 distribution with 1 d.f. is generally adequate, when heterozygosity is too high (say, above 95%), this large sample approximation may not be adequate for the tail probability estimation. Therefore, the significance levels of the goodness-of-fit test criteria (based on both biased and unbiased estimates of heterozygosity and homozygosity) are also judged by empirically determining the level of significance from the shuffling replications in Table 4. In other words, for each of the 2000 shufflings, the goodness of fit chi-square values are computed and departure from HWE expectations is tested. The χ^2 statistic of the reshuffled data is compared to the χ^2 statistics of the data being tested; the significance level (or P value) is the number of times the χ^2 of the reshuffled data is larger than or equal to χ^2 of the original data, divided by the number of reshufflings.

These probabilities are noted under Table 4. [Heading is number of shuffled = #; Biased and unbiased rows]. Their interpretations are again similar to ones of the large sample approximation, namely, small probabilities (say, less than 0.05) are indicators of deviation from strict predictions of the Hardy-Weinberg expectation (HWE) proportions. Second, even when departures from HWE are found, this test alone is not definitive proof of the fact that the independence rule fails, because not all single-banded patterns are truly homozygotes. Note that in DNA typing data on VNTR loci, the total counts of homozygotes (at bin level) largely consist of counts of single-banded patterns. Since at least a fraction of these include heterozygotes with nondetectable alleles, further examination of this test is needed to rule out the possibility that such a deviation is not an artifact of the presence of nondetectable alleles (the last part of the

results of this menu). Third, non-significance of this goodness-of-fit test criterion is not definitive proof of independence either. This is a test based on total homozygosity (and heterozygosity). It may be that at individual bin levels, genotypes deviate from HWE predictions, but the net effect is that the bins cancel each other and result in conformity of the total homozygosity and heterozygosity. Also, the converse is true. Therefore, likelihood ratio and the exact tests are performed to guard against this possibility. (See below)

Nevertheless, a check of conformity of observed and expected homozygosity is meaningful in the context of other applications of VNTR DNA typing database, e.g., in detecting relatedness of individuals (see Chakraborty and Jin 1992), and the goodness-of-fit test of total homozygosity can be instructive. See also Help files for program D for implications of comparison of observed and expected heterozygosities.

Likelihood ratio test: A likelihood ratio statistic (G-square) is a test criterion that determines the deviation of observed and expected frequencies of all possible genotypes (see Weir 1991 for the algebraic expression of this test statistic in terms of binned genotype frequencies and allele frequency estimates). In principle, when k alleles are present in a database, the large sample approximation for this test statistic is a χ^2 distribution with d.f. = k(k+1)/2-k (see Rao 1973). For example, if 25 of the 31 binned RFLP alleles occurred in the sample with non-zero frequency, the d.f. of the G-square statistic in a large sample should be (25 x 26 /2)-25 = 300. The level of significance (noted as probability on the same line of the G-square value) can be computed based on this large sample approximation. However, this is not generally applicable for the observed G-square value and for data structure such as the ones seen in VNTR binned genotype distributions.

Again, the RFLP binned genotype table (Table 1) is highly sparse; most genotypes are either unobserved in the sample, or even when they are present, the occurrence is small. For example, in one of the examples, there are 24 binned alleles, and thus 276 possible genotypes, i.e. (24x25)/2. Only 36 of these are observed in the database of 39 individuals, and in no case does the count exceed 2. Obviously, for such sparse data, the large sample approximation may not be adequate. Below Table 4, MAXLOGL, the empirical level of significance of obtaining a G-value greater than or equal to the one observed, is shown for 2000 replications of shuffling. When this probability is small (say, less than 0.05), there is an indication of deviation from the strict HWE predictions at the individual binned genotype level. As in the case of the goodness-of-fit test statistic, this deviation may also be due to the presence of nondetectable alleles (not addressed until later in the output data).

Exact Test: Guo and Thompson (1992) developed a Monte-Carlo test of HWE predictions for sparse data with a large number of alleles. The exact test determines a conditional probability of a given data structure (observed genotype table) for a given set of (binned) allele frequencies. For polymorphic loci, the probability level of observing any particular data structure (i.e., any particular assortment of alleles within genotypes) is low (as an indicator of deviation from HWE). By determining how often a lower probability is observed in the randomly shuffled data, an indication of deviation from HWE is obtained. This information is listed as output information on the first line [probability] below Table 4 (number of shuffled conditional probability $\leq$ observed conditional probability). If this proportion is small (say, less than 0.05),

an indication of significant deviation from HWE could be inferred. As in the case of the other two test criteria, this test does not address the possible existence of nondetectable alleles.

Test based on number of distinct genotypes: Sparse data can also be summarized into the number of distinct genotypes (of heterozygous and homozygous type) observed in a sample and can be contrasted with their expectations based on HWE (output data in Table 4). Chakraborty (1992, 1993a) gave the analytical details of the logic of this test procedure, which also evaluates the standard errors (S.E.) of these statistics. Even in sparse data, the observed number of (heterozygous and homozygous) genotypes follow a normal distribution. By examining the deviation of the observed from expected values (by more than 2 SE), the departure from HWE can be significant or not at the 5% level. This test does not contrast each genotype frequency (at the bin level) with its expectation; thus the test is not as informative as the exact test regarding forensic implications of the independence test. However, for some population genetic applications (mainly for evolutionary purposes) this test provides some information.

Given the multiplicity of tests (at least three tests), special attention should be given to the interpretation of these test results. Chakraborty and Zhong (1994) have shown that of the three tests described here, the exact test is the most powerful in detecting deviation from HWE. Therefore, although it is generally in conformity with the likelihood ratio test result (empirical level, obtained from shuffling), the exact test procedure results are most dependable. The chi-square goodness-of-fit test is the least powerful. Therefore, even though each of these tests emphasizes different features of a genotype distribution in a sample, a closer examination of the exact test and likelihood ratio test results is recommended for forensic implications of using DNA typing databases.

Thus far, all tests of independence assumed the absence of nondetectable alleles. However, the test results can provide indications of whether or not nondetectable alleles are present in the data. Chakraborty et al. (1992, 1994) discussed these issues and developed algorithms to revise the three test procedures (chi-square, likelihood ratio, and the exact test) by incorporating the presence of nondetectable alleles. These are implemented in the last part of the output of this program.

Program H, when run on a data set, can detect situations when "null" alleles may have to be invoked. This is done internally in the program by checking whether the estimated null allele frequency is significantly different from zero (shown in the output under the null-allele statistics as a T-statistic and its normal deviate Z). When Z has a one-sided probability below 0.05, revisions of the shuffling tests based on Chi-square, likelihood ratio, and the exact test are automatically run by the program. A user can see on the screen when this occurs, because the shuffling "counter" will re-set and start again after reaching its pre-set value (see Program H Input Examples).

One concluding comment on this program: Since the inference from this example would be that for this locus a trace amount of null alleles may indeed be present in the data, then how would the estimated binned frequency estimates be affected? Strictly speaking, all binned frequencies should be revised by a factor of $p_i(1 - r)$. However, since the binned frequencies reported in Table 2 are already conservative, no further revision of these estimates is needed. Of course, if one wants precise estimates, all binned frequencies for fragment sizes should be multiplied by the factor $(1 - r)$.

Note 1. Table 1 – the graphical representation -- for a RFLP locus, in the diagonal entries, includes single-banded as well as two-banded profiles when both fragments are within the same (fixed) bin. Thus, for a RFLP locus, this table gives the binned genotype counts of the data.

In all subsequent analyses for RFLP, therefore, homozygosity / heterozygosity refers to binned data (and, thus, homozygosity will generally be larger than the observed proportion of single-banded individuals).

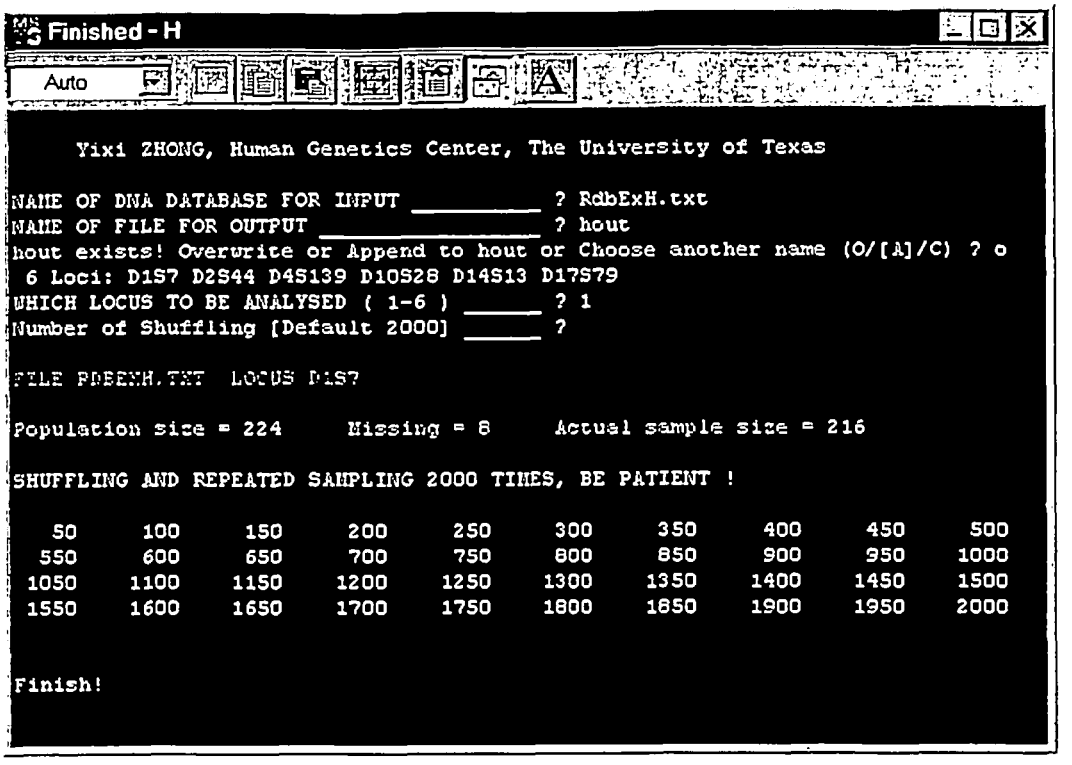
Note 2. The analogue of Table 1A for RFLP loci, although implicit in one or more of the HWE tests, is not included in the output file because of the large number of possible genotypes. Note that even when the observed frequency of a given binned genotype is zero, its expectation can be non-zero (under HWE) if all (binned) alleles relevant for that genotype are present in the population.

Program H Examples – Input and Output

To illustrate program H operation, four examples are given, using example databases (that are provided with the software). The first two examples are performed with RFLP loci; the second two, with PCR-based loci.

EXAMPLE 1

Below is shown an input / dialog screen (Example Input Screen / Dialog 1) for program H in which the first locus (D1S7) is analyzed in the example database RdbExH.txt. (The screen may switch after entering the first few items, but the screen dialog for the two screens is identical to that on the one pictured.)



Here, note that the name of the database for analysis (RdbExH.txt) was provided. The output filename chosen was "hout" (no quotes), and since it existed, the program asked whether to append to it, overwrite it, or whether the user wanted a new name. "O" for overwrite was entered. The six loci in this database are then listed. Number 1 (D1S7) was chosen. The number of shufflings is next requested. Here, the user accepted the default value of 2000 by hitting <Enter> (<Return>). The program then reports the database file and locus, total number of people (224), the number of missing data for the D1S7 locus (8), and then the actual sample size for D1S7 (216). The numbers 50, 100, 150 etc will appear as the program runs until shuffling is completed. Then the program terminates. The "Finish" message at the bottom of the screen indicates that the program has finished, and the user can open and view the output file (in WordPad).

The output file (Example Output File 1) resulting from the analysis of D1S7 in the database RdbExH.txt is shown below:

```
                Statistics & Test of Hardy-Weinberg Law

      Yixi ZHONG, Human Genetics Center, The University of Texas

Locus:  1 D1S7 of RDBEXH.TXT   Actual sample size = 216
```

TABLE 1 Fixed bin 'genotype' frequencies

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	.															
2	.	.														
3	.	.	.													
4	.	.	.	.												
5	.	.	.	.	.											
6	.	.	.	.	.	.										
7	.	.	.	.	.	.	.									
8	.	.	.	.	.	.	.	.								
9	.	.	.	.	.	.	.	.	.							
10	.	.	.	.	.	.	.	.	.	.						
11	.	.	.	.	.	.	.	.	.	.	.					
12	.	.	.	.	.	.	.	.	.	.	.	1				
13	.	.	.	.	.	1	.	.	1	1	.	1	.			
14	.	.	.	.	.	.	1	.	.	.	.	.	.	.		
15	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	
16	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.
17	.	.	.	.	.	.	.	.	1	.	.	.	1	.	1	.
18	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	2
19	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1
20	.	.	.	.	.	.	.	2	.	1	1	.	2	.	.	1
21	.	.	.	.	.	2	.	.	.	1	1	1	3	.	2	.
22	.	.	.	.	.	.	.	.	.	.	1	.	.	2	.	.
23	.	.	.	.	.	.	.	1	.	.	.	1	1	1	1	1
24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.
25	.	.	.	.	.	.	.	1	1	.	.	.	.	.	1	.
26	.	.	.	.	.	.	.	.	.	2	2	.	.	1	1	.
27	.	.	.	1	.	.	.	.	.	.	1	.	3	1	.	1
28	.	.	.	.	.	.	1	.	.	.	.	.	2	.	1	.
29	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
30	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
31	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1	.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Total number of individuals = 216 . = 'genotype' frequency is 0

TABLE 1 Fixed bin 'genotype' frequencies (continued)

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
17	1														
18	2	2													
19	2	2	.												
20	.	.	1	2											
21	.	2	3	.	.										
22	.	2	4	2	4	3									
23	.	3	.	.	2	4	2								

24	2	1	6	2	1	1	3	.							
25	1	1	1	.	4	2	1	3	.						
26	5	3	1	3	4	1	1	2	3	1					
27	1	5	2	1	2	1	2	1	.	1	.				
28	.	2	.	1	1	3	.	5	1	1	1	.			
29	.	.	.	.	.	1	.	.	.	1	.	.	.		
30	.	1	.	.	2	1	1	.	1	.	.	.	.	.	
31	.	1	1	.	1	1	.	2	1	1	2	.	.	.	1

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

. = 'genotype' frequency is 0

Locus: 1 D1S7 of RDBEXH.TXT Actual sample size = 216

TABLE 2 Fixed bin 'allele' frequencies and confidence intervals

Bin	Size Classes	Observed		95% Limit		SINGLE BAND
		Num.	Percent	LCL	UCL	
1	1 - 639	1	0.231	0.020	2.684	0
2	640 - 772	0	0.000	0.000	2.251	0
3	773 - 871	0	0.000	0.000	2.251	0
4	872 - 963	1	0.231	0.020	2.684	0
5	964 - 1077	0	0.000	0.000	2.251	0
6	1078 - 1196	3	0.694	0.136	3.473	0
7	1197 - 1352	2	0.463	0.068	3.088	0
8	1353 - 1507	4	0.926	0.218	3.843	0
9	1508 - 1637	4	0.926	0.218	3.843	0
10	1638 - 1788	5	1.157	0.312	4.202	0
11	1789 - 1924	6	1.389	0.414	4.552	0
12	1925 - 2088	8	1.852	0.641	5.230	1
13	2089 - 2351	18	4.167	2.028	8.369	0
14	2352 - 2522	8	1.852	0.641	5.230	0
15	2523 - 2692	12	2.778	1.156	6.525	0
16	2693 - 2862	8	1.852	0.641	5.230	0
17	2863 - 3033	18	4.167	2.028	8.369	1
18	3034 - 3329	33	7.639	4.495	12.690	2
19	3330 - 3674	25	5.787	3.141	10.424	0
20	3675 - 3979	21	4.861	2.495	9.259	1
21	3980 - 4323	36	8.333	5.020	13.522	0
22	4324 - 4821	36	8.333	5.020	13.522	2
23	4822 - 5219	27	6.250	3.472	10.997	1
24	5220 - 5685	31	7.176	4.150	12.130	0
25	5686 - 6368	22	5.093	2.654	9.553	0
26	6369 - 7241	35	8.102	4.844	13.246	1
27	7242 - 8452	26	6.019	3.306	10.711	0
28	8453 - 10093	19	4.398	2.182	8.667	0

29	10094 - 11368	2	0.463	0.068	3.088	0
30	11369 - 12829	6	1.389	0.414	4.552	0
31	12830 - 25000	15	3.472	1.580	7.459	1
Total		432	100.000			10

Locus: 1 D1S7 of RDBEXH.TXT Actual sample size = 216

Table 3 Observed and expected frequency contrasts

	Obs. No.	Percent (%)	Expected No. and Percent			
			Biased		Unbiased	
Homozygote	13	6.02	12.20	5.65	11.73	5.43
Heterozygote	203	93.98	203.80	94.35	204.27	94.57
Chi-Square			0.055732		0.146311	
Degrees of Freedom			1		1	
Probability			0.813373		0.702086	

Table 4 Numbers of distinct genotypes

	Obs.	Exp.	S.E.
Homozygote	8	7.888	1.872
Heterozygote	121	120.921	7.217
Total	129	128.809	7.456

OF SHUFFLED = 2000

PROBABILITY: # OF SHUFFLED CONDIT. PR.<= OBS. CONDIT.PR. = 310
PROBABILITY(1) = 0.15500

BIASED: # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 1764
PROBABILITY(2) = 0.88200

UNBIASED: # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 1529
PROBABILITY(3) = 0.76450

MAXLOGL.: # OF SHUFFLED G-SQUARE >= OBS. G-SQUARE = 326
PROBABILITY(4) = 0.16300
(Obs G-Square = 293.925 DF = 378; Approx Probability = 0.99950)

Locus: 1 D1S7 of RDBEXH.TXT Actual sample size = 216

```
Null Allele Freq. Computations:

Prop of single banded types = .0463
Freq. of null alleles:
                        Biased =     0.001969
                      Unbiased =     0.003128

                             T =     1.178090
                             Z =     0.503716
          1-SIDED PROBABILITY =     0.307231
                             r =     0.003298
                        SD(r) =     0.008136
               HETEROZYGOSITY =     0.937320
       NULL ALLELE CHI-SQUARE =     0.022886
       NULL ALLELE   G-SQUARE =   293.658081

05-16-1998     18:34:04
```

Table 1 data shown in this output are the binned genotype frequencies that are self explanatory. The dots indicate binned genotypes with no count of frequencies in the 216 individuals of the data set. Note that in this example, D1S7 data on 216 individuals in the sample are available, so that the actual sample size (noted on the top of Table 1) is the same as the total number of individuals listed below the Table 1. The Table 1 is split into two parts for accommodating all binned genotypes within the page format (with normal size fonts).

Table 2 gives estimated the binned allele frequencies (with the bin boundaries). Frequencies are tabulated in actual counts (out of the 2n chromosomes on which data is available, 2n = 432 in this example) as well as in percents. The data on actual counts should facilitate two purposes. First, if the count is below 5, a user should merge the bin with its nearest ones to increase the frequency (to make it above 5/2n, where n is the number of individuals) to be used for forensic calculations. Second, should any fragment size straddle any bin boundary, the count data readily helps to select one of the adjacent bins that contain a higher frequency. Other data presented in Table 2 contains upper and lower 95% confidence limits of estimated (binned) allele frequencies that incorporate the presence of multiple alleles (according to Goodman 1965) and the number of single banded profiles at the locus within each of the 31 bins. Data contained under the observed frequencies (counts and percent frequencies) are the most commonly used statistics of this menu for forensic applications.

Table 3 data shows the results of the homozygosity test of HWE. Along with the observed counts (and percent frequencies) of binned homozygotes and heterozygotes, their expectations (with and without bias-correction, see Nei, 1978 for the computations) are also given (in terms of counts and percent frequencies). The chi-square values listed below the expected columns signify the extent of agreement between observed and expected. The probability value listed in the last row of this table is the level of significance of the chi-square statistic, under the assumption that this statistic truly follows a chi-square distribution with a single degree of freedom. Although the inference is generally the same, it is preferred to use the unbiased expectation of the homozygosity and heterozygosity values, particularly when the sample size is

small. In the present example, the observed binned heterozygosity (93.98%) is statistically indistinguishable from the expected (94.57%, the unbiased estimate) under the HWE assumption (with the P-value of 0.702086).

Table 4 data is another alternative test of HWE based on summary data (on the number of distinct (binned) genotypes, grouped into homozygotes and heterozygotes). For each, the observed and expected (with standard errors) number of distinct homozygote and heterozygote genotypes are shown. Although generally this test is not very commonly used, it serves two purposes. First, it shows the sparseness of the data. In this example, of the possible 496 (binned) genotypes (31 x 32 / 2 = 496), in the 216 individuals sampled only 129 distinct genotypes are observed. Second, under the HWE assumption, in large samples the number of distinct genotypes follow a normal distributions, and hence, when the observed counts of distinct homozygotes and heterozygotes are within the plus/minus 2-SE limits of their expectations, the data may be considered consistent with the HWE predictions (as it is seen in this example).

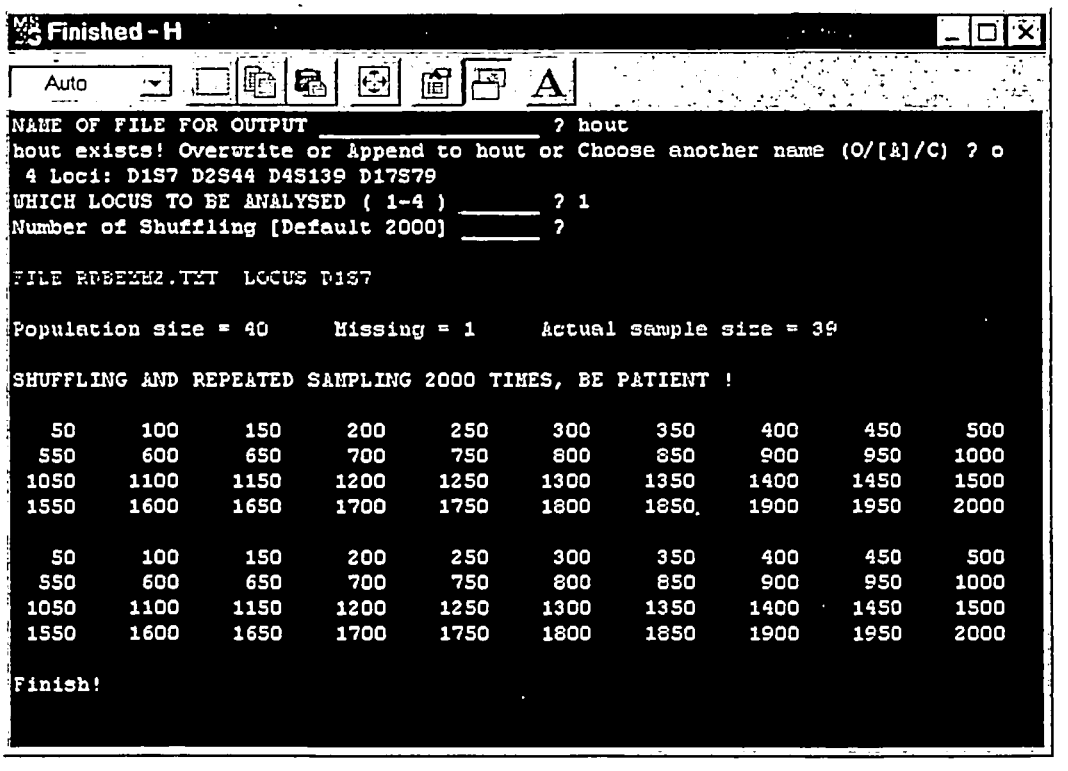
Since Table 4 shows that even the binned genotype distribution is sparse, below this table are the results of the shuffling tests based on 2,000 replications of allele shuffling (see notes on the Shuffling routine). Four empirical probabilities are listed here (the ones applicable to the exact test, Guo and Thompson 1992, called here "Shuffled conditional probability" or "PROBABILITY (1)"; chi-square test based on biased ("PROBABILITY (2)")and unbiased ("PROBABILITY (3)") estimate of (binned) heterozygosity and the likelihood ratio test, called here the Shuffled G-square ("PROBABILITY (4)"). Of the 2,000 replications, the number of times the shuffled values of the corresponding statistics that exceed the ones in the observed data are recorded along with the empirical levels of significance, called probability, in the output. When these probabilities exceed a nominal level of 1%, or 5%, the (binned) genotype data is regarded as being consistent with their HWE predictions. In this example, all four tests have P-value far exceeding the 5% level, indicating consistency with the HWE predictions. The last row of this section also records the observed G-square statistic (of the likelihood ratio test), its degrees of freedom and the probability level under the large sample assumption of its chi-square distribution. Although the multiplicity of tests may sometime lead to different conclusions, of these HWE tests, the exact test and the likelihood ratio test are most informative (powerful) since they contrast observed and expected frequencies of all (binned) genotypes at the locus analyzed.

In all of the analyses this menu conducted at this stage it was assumed that there is no nondetectable allele. However, some of the 10 single-banded individuals observed in 216 individuals may indeed have a non-zero frequency of nondetectable alleles. The last section of this output further examines this possibility. The proportion of single banded types is simply 10/216 = 0.463, readily interpretable. This yields two estimates of null allele frequency (biased and unbiased, see Chakraborty et al. 1994). In this example, the unbiased estimate of null allele frequency is 0.3128%. It is not significantly different from zero (judged from the T-value, its normal deviate transformation, Z = 0.503716, whose 1-sided probability of 0.307231 is obviously above the 5% nominal level of significance). The maximum likelihood estimate of null allele frequency (r) in this data is 0.3298% with a standard deviation (SD) of 0.8136%. Had the null alleles been incorporated (which amounts to a reduction of observed homozygosity), the

revised (binned) heterozygosity and tests of HWE would have to done under a somewhat different sampling assumption, namely, the sampling is conditional to having no null-homozygote in the data (see Chakraborty et al. 1994). Under this assumption, the revised (binned) heterozygosity and the corresponding chi-square statistic for the heterozygosity (Table 3) and likelihood ratio (G-square) tests of HWE are shown in the last three rows of this output. Since in this example, the null allele frequency estimate (and its associated Z-value) is not significant, the program completes its execution at this stage.

EXAMPLE 2

As a second example of the program H operation (Example Input screen / Dialog 2) in which null allele calculations are further considered in order to revise the shuffling test results is shown below. Here, D1S7 is analyzed in the example database RdbExH2.txt. (The screen may switch after entering the first few items, but the screen dialog for the two screens is identical to that on the one pictured.)



Note that the input screen is the same (except some lines have scrolled off), but for this data (RbdExH2.txt) when locus 1 (D1S7) is chosen by the user (with the default value of 2,000 shufflings), the program recognizes that of the 40 individuals in this sample D1S7 fragment size data are available for 39 individuals. Unlike in Example 1, during the execution of the program, the screen indicates that shufflings are being done two sets of times. This itself suggests that the program first performed all tasks described in the Example 1, and found null alleles to be a significant factor and, hence, decided to repeat the shuffling tests with null alleles incorporated in the analysis.

The output file (Example Output File 2) resulting from the analysis of D1S7 in the database RdbExH2.txt is shown below:

```
                Statistics & Test of Hardy-Weinberg Law

      Yixi ZHONG, Human Genetics Center, The University of Texas

Locus:  1 D1S7 of RDBEXH2.TXT   Actual sample size = 39

              TABLE 1   Fixed bin 'genotype' frequencies

-------------------------------------------------------------------------
      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
-------------------------------------------------------------------------
  1   .
  2   .   .
  3   .   .   .
  4   .   .   .   .
  5   .   .   .   .   .
  6   .   .   .   .   .   1
  7   .   .   .   .   .   .   .
  8   .   .   .   .   .   .   .   .
  9   .   .   .   .   .   .   .   .   .
 10   .   .   .   .   .   .   .   .   .   .
 11   .   .   .   .   .   .   .   .   .   .   .
 12   .   .   .   .   1   .   .   .   .   .   1   1
 13   .   .   .   .   .   .   .   .   .   .   .   .   .
 14   .   .   .   .   .   .   1   .   .   .   .   .   .   .
 15   .   .   .   .   .   .   .   .   .   .   .   .   1   .   1
 16   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 17   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 18   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 19   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 20   .   .   .   .   .   .   .   .   .   .   .   1   .   .   .   .
 21   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 22   .   .   .   .   .   .   .   .   .   .   1   .   .   .   .   .
 23   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 24   .   .   .   .   .   .   .   .   .   .   .   .   1   .   .   .
 25   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 26   .   .   .   .   .   .   1   .   .   .   .   1   .   .   .   .
 27   .   .   .   .   .   .   .   .   .   1   .   .   1   .   1   .
 28   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 29   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 30   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
 31   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
-------------------------------------------------------------------------
      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
-------------------------------------------------------------------------

Total number of individuals = 39     . = 'genotype' frequency is 0

      TABLE 1   Fixed bin 'genotype' frequencies (continued)

-------------------------------------------------------------
```

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
17	.														
18	.	.													
19	.	.	1												
20	.	.	.	.											
21	.	.	.	.	.										
22	.	2	.	.	2	.									
23	.	.	.	.	1	1	.								
24	1	.	.	.	.	.	.	.							
25	.	1	.	1	.	.	.	.	.						
26	.	.	.	1	.	1	.	1	.	1					
27	.	.	.	.	.	.	1	.	.	2	.				
28	.	.	.	.	.	.	.	.	1	.	1	.			
29	.	.	.	.	1	.	.	.	.	.	.	.	1		
30	.	.	.	.	.	1	.	.	.	.	.	.	.	.	
31	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

. = 'genotype' frequency is 0

Locus: 1 D1S7 of RDBEXH2.TXT Actual sample size = 39

TABLE 2 Fixed bin 'allele' frequencies and confidence intervals

Bin	Size Classes	Observed		95% Limit		SINGLE BAND
		Num.	Percent	LCL	UCL	
1	1 - 639	0	0.000	0.000	11.310	0
2	640 - 772	0	0.000	0.000	11.310	0
3	773 - 871	0	0.000	0.000	11.310	0
4	872 - 963	0	0.000	0.000	11.310	0
5	964 - 1077	1	1.282	0.108	13.476	0
6	1078 - 1196	2	2.564	0.377	15.481	1
7	1197 - 1352	2	2.564	0.377	15.481	0
8	1353 - 1507	0	0.000	0.000	11.310	0
9	1508 - 1637	0	0.000	0.000	11.310	0
10	1638 - 1788	1	1.282	0.108	13.476	0
11	1789 - 1924	2	2.564	0.377	15.481	0
12	1925 - 2088	6	7.692	2.318	22.636	1
13	2089 - 2351	3	3.846	0.755	17.377	0
14	2352 - 2522	1	1.282	0.108	13.476	0
15	2523 - 2692	4	5.128	1.215	19.191	1
16	2693 - 2862	0	0.000	0.000	11.310	0
17	2863 - 3033	1	1.282	0.108	13.476	0
18	3034 - 3329	3	3.846	0.755	17.377	0
19	3330 - 3674	2	2.564	0.377	15.481	1

20	3675 - 3979	3	3.846	0.755	17.377	0
21	3980 - 4323	5	6.410	1.740	20.940	0
22	4324 - 4821	8	10.256	3.602	25.901	0
23	4822 - 5219	3	3.846	0.755	17.377	0
24	5220 - 5685	3	3.846	0.755	17.377	0
25	5686 - 6368	3	3.846	0.755	17.377	0
26	6369 - 7241	9	11.538	4.297	27.480	1
27	7242 - 8452	8	10.256	3.602	25.901	0
28	8453 - 10093	2	2.564	0.377	15.481	0
29	10094 - 11368	3	3.846	0.755	17.377	1
30	11369 - 12829	1	1.282	0.108	13.476	0
31	12830 - 25000	2	2.564	0.377	15.481	0
Total		78	100.000			6

Locus: 1 D1S7 of RDBEXH2.TXT Actual sample size = 39

Table 3 Observed and expected frequency contrasts

	Obs. No.	Percent (%)	Expected No. and Percent			
			Biased		Unbiased	
Homozygote	6	15.38	2.42	6.21	1.95	5.00
Heterozygote	33	84.62	36.58	93.79	37.05	95.00
Chi-Square			5.630013		8.871163	
Degrees of Freedom			1		1	
Probability			0.017655		0.002897	

Table 4 Numbers of distinct genotypes

	Obs.	Exp.	S.E.
Homozygote	6	2.106	1.271
Heterozygote	30	32.129	5.035
Total	36	34.235	5.193

OF SHUFFLED = 2000

PROBABILITY: # OF SHUFFLED CONDIT. PR.<= OBS. CONDIT.PR. = 92
PROBABILITY(1) = 0.04600

BIASED: # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 17
PROBABILITY(2) = 0.00850

UNBIASED: # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 17
PROBABILITY(3) = 0.00850

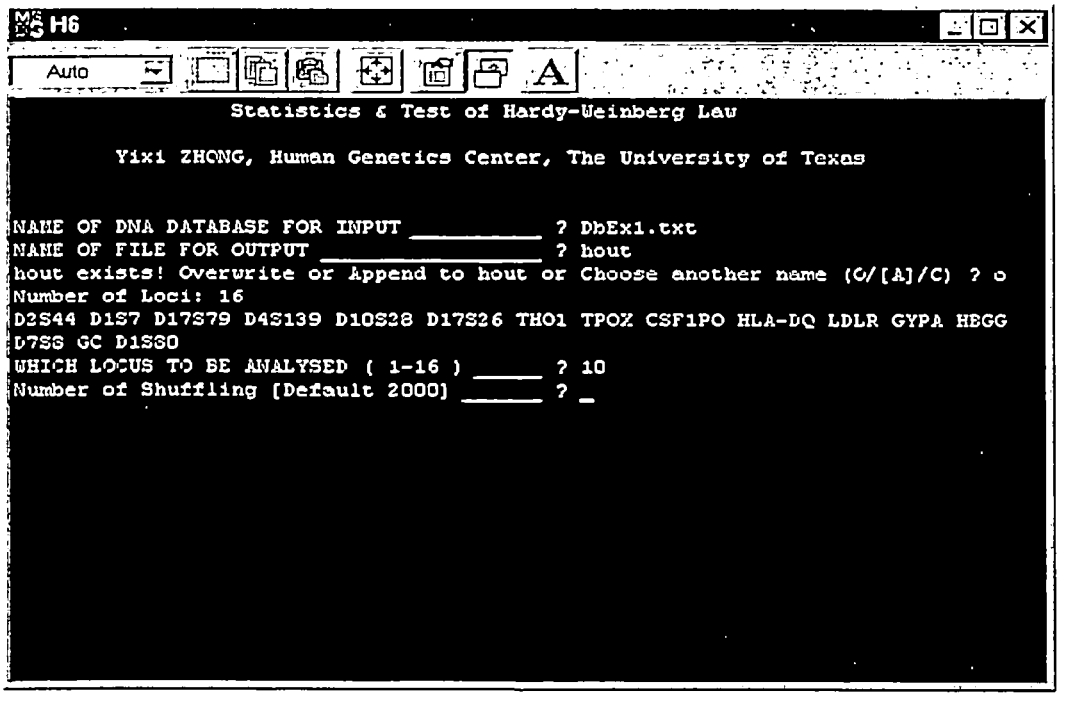
```
MAXLOGL.:     # OF SHUFFLED   G-SQUARE >= OBS.    G-SQUARE = 281
          PROBABILITY(4)                                    = 0.14050
(Obs G-Square = 138.904 DF = 276;  Approx Probability = 1.00000)


Locus:  1 D1S7 of RDBEXH2.TXT   Actual sample size = 39


Null Allele Freq. Computations:

Prop of single banded types = .1538
Freq. of null alleles:
                      Biased =    0.051410
                    Unbiased =    0.057842

                           T =    3.722222
                           Z =    3.544802
         1-SIDED PROBABILITY =    0.000197
                           r =    0.059179
                       SD(r) =    0.035094
              HETEROZYGOSITY =    0.833068
      NULL ALLELE CHI-SQUARE =    0.048023
      NULL ALLELE   G-SQUARE =  132.247284

NULL ALLELES:  # OF SAMPLING = 2000

PROBABILITY: # OF SAMPLING CONDIT. PR.<= OBS. CONDIT.PR. = 1595
          PROBABILITY(5)                                    = 0.79750

HOMO-EXPECT  # OF SAMPLING CHI-SQUARE >= OBS. CHI-SQUARE = 1669
          PROBABILITY(6)                                    = 0.83450

MAXLOGL.:     # OF SAMPLING   G-SQUARE >= OBS.    G-SQUARE = 1500
          PROBABILITY(7)                                    = 0.75000


05-16-1998     18:34:47
```

Except for the last section of this output, all entries in it are similar to the ones of the previous example. However, Table 3 notes that there is a deficiency of (binned) heterozygosity (and a corresponding excess homozygosity) and this is significant (chi-square = 8.871163 with a P-value of 0.002897 for the unbiased heterozygosity test) even with its large sample approximation. Table 4 data confirms this, showing a deficiency of the number of distinct (binned) heterozygote genotypes. The shuffling tests also failed for the exact test (P = 0.046), i.e. PROBABILITY (1), and with the chi-square test (P = 0.0085), both biased and unbiased, i.e., PROBABILITY (2) and PROBABILITY (3). The likelihood ratio test, i.e., PROBABILITY (4), however, still did not show significant deviation from HWE (G-square = 138.904 with an empirical level of significance of p = 0.1405). The null-allele calculations from the occurrence of 15.38% single-banded individuals resulted in a maximum likelihood estimate of 5.9179% null alleles (parameter r) with a S.D. of 3.5094% (parameter SD(r)). The z-value (3.544802) for

null alleles is significant, as judged from its 1-sided probability of 0.000197. This prompted the program to examine whether or not the significant departures from HWE can be explained due to the presence of null alleles, which is done from the second set of replications incorporating 5.9179% null alleles in the data and performing a conditional sampling of no null-homozygote in the sample of 39 individuals. The results on the empirical levels of significance (with 2,000 such shufflings) are shown in the last section for the exact test (P = 0.7975), i.e. PROBABILITY (5), chi-square test (based only on the unbiased estimate of (binned) heterozygosity, P = 0.8345), i.e. PROBABILITY (6), and the likelihood ratio test (G-square = 132.247284 with P = 0.7500), i.e. PROBABILITY (7). Clearly, all test results now show that the data is consistent with the HWE predictions invoking a null allele frequency of 5.9179%.

EXAMPLE 3

For completeness, the execution of this program for PCR loci including the input / dialog screens and their outputs are shown with two other examples. Here, HLA-DQA1 in the example database DbEx1.txt is analyzed. (The screen may switch after entering the first few items, but the screen dialog for the two screens is identical to that on the one pictured.)



The output file (Example Output File 3) resulting from the analysis of HLA-DQA1 in the database DbEx1.txt is shown below:

```
Statistics & Test of Hardy-Weinberg Law for autosomal locus

          Yixi ZHONG, Human Genetics Center, The University of Texas

Locus: HLA-DQ of DbEx1.txt   Actual sample size = 100
```

TABLE 1 Genotype frequencies

	1.1	1.2	1.3	2.0	3.0	4.0
1.1	7					
1.2	5	1				
1.3	8	6	3			
2.0	4	2	3	2		
3.0	2	5	1	.	.	
4.0	14	10	11	5	2	9
	1.1	1.2	1.3	2.0	3.0	4.0

Total number of individuals = 100 . = 'genotype' frequency is 0

Locus: HLA-DQ of DbEx1.txt Actual sample size = 100

TABLE 1A Genotype frequencies and its expects

Obs. genotype	Obs. freq #(%)	Expected freq. Unbiased (%)	Biased (%)
1.1/ 1.1	7(7.00)	5.432(5.43)	5.523(5.52)
1.1/ 1.2	5(5.00)	7.085(7.09)	7.050(7.05)
1.1/ 1.3	8(8.00)	8.266(8.27)	8.225(8.22)
1.1/ 2.0	4(4.00)	4.251(4.25)	4.230(4.23)
1.1/ 3.0	2(2.00)	2.362(2.36)	2.350(2.35)
1.1/ 4.0	14(14.00)	14.171(14.17)	14.100(14.10)
1.2/ 1.2	1(1.00)	2.186(2.19)	2.250(2.25)
1.2/ 1.3	6(6.00)	5.276(5.28)	5.250(5.25)
1.2/ 2.0	2(2.00)	2.714(2.71)	2.700(2.70)
1.2/ 3.0	5(5.00)	1.508(1.51)	1.500(1.50)
1.2/ 4.0	10(10.00)	9.045(9.05)	9.000(9.00)
1.3/ 1.3	3(3.00)	2.990(2.99)	3.063(3.06)
1.3/ 2.0	3(3.00)	3.166(3.17)	3.150(3.15)
1.3/ 3.0	1(1.00)	1.759(1.76)	1.750(1.75)
1.3/ 4.0	11(11.00)	10.553(10.55)	10.500(10.50)
2.0/ 2.0	2(2.00)	0.769(0.77)	0.810(0.81)
2.0/ 4.0	5(5.00)	5.427(5.43)	5.400(5.40)
3.0/ 4.0	2(2.00)	3.015(3.02)	3.000(3.00)
4.0/ 4.0	9(9.00)	8.894(8.89)	9.000(9.00)
Others	0(0.00)	1.131(1.13)	1.150(1.15)
Total	100(100.00)	100.000(100.00)	100.000(100.00)

Locus: HLA-DQ of DbEx1.txt Actual sample size = 100

TABLE 2 Allele frequencies and standard errors

Allele Classes	Observed			SINGLE BAND
	Num.	Percent	SE (%)	
1.1	47	23.500	2.998	7
1.2	30	15.000	2.525	1
1.3	35	17.500	2.687	3
2.0	18	9.000	2.024	2
3.0	10	5.000	1.541	0
4.0	60	30.000	3.240	9
Total	200	100.000		22

Locus: HLA-DQ of exam Actual sample size = 100

Table 3 Observed and expected frequency contrasts

	Obs. No.	Percent (%)	Expected No. and Percent					
			Biased		SD	Unbiased		SD
Homozygote	22	22.00	20.90	20.90%	4.07	20.50	20.50%	4.04
Heterozygote	78	78.00	79.10	79.10	4.07	79.50	79.50	4.04
Chi-Square			0.073872			0.138533		
Degrees of Freedom			1			1		
Probability			0.785781			0.709743		

Maximum Log Likelihood Ratio Methods:
G-Square 11.787 Degrees of Freedom 15 Probability 0.695108

Table 4 Numbers of distinct genotypes

	Obs.	Exp.	S.E.
Homozygote	5	4.6272	0.7422
Heterozygote	14	13.9349	0.8826
Total	19	18.5622	1.1760

OF SHUFFLED = 2000

```
PROBABILITY: # OF SHUFFLED CONDIT. PR.<= OBS. CONDIT.PR. = 1573
          PROBABILITY(1)                                 = 0.78650

BIASED:      # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 1607
          PROBABILITY(2)                                 = 0.80350

UNBIASED:    # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 1411
          PROBABILITY(3)                                 = 0.70550

MAXLOGL.:    # OF SHUFFLED   G-SQUARE >= OBS.   G-SQUARE = 1599
          PROBABILITY(4)                                 = 0.79950


Locus: HLA-DQ of DbEx1.txt   Actual sample size = 100


Null Allele Freq. Computations:

Prop of single banded types = .2200
Freq. of null alleles: .
                      Biased =    0.007033
                    Unbiased =    0.009540

                           T =    1.058190
                           Z =    0.260232
        1-SIDED PROBABILITY =    0.397342
                           r =    0.005819
                      SD(r) =    0.023237
              HETEROZYGOSITY =    0.781897
      NULL ALLELE CHI-SQUARE =    0.002110
      NULL ALLELE   G-SQUARE =   11.716675

06-16-1998     10:56:08
```

In this example, the user chose the DbEx1.txt file with locus HLA-DQA1 in which all 100 individuals had the genotype data on the HLA-DQA1 locus. As before, the Table 1 shows the genotype counts (which indicate that although there are 21 possible genotypes in this data, with 4.0 allele not subdivided in this typing protocol, there are two genotypes 2,3 and 3,3 not encountered in this sample).

Since the number of genotypes is not very large for such a locus, Table 1A (not reproduced in analysis of RFLP loci - see Note 2) lists the observed and expected genotype frequencies in terms of their counts and percentages (in parentheses) in later of which both biased and unbiased estimates are given. This table helps the user to identify which specific genotypes, if any, fail to meet the HWE expectations. In this example no such discrepancy worth further evaluation is found, since for all genotypes the observed frequencies are in good agreement with their expectations (a further support of which is obtained by the exact and likelihood ratio tests indicated below).

Table 2 results for a PCR locus are also similar to that of the RFLP loci, with the exception that in this table only the standard errors of the estimated allele frequencies are given (and not their 95% confidence intervals). The last column (single-band) in the context of a PCR locus refers to the homozygotes for each allele in the data. As before, this is the table that gives most useful results for the purposes of applications, since allele counts (frequencies) reflect their relative frequencies in the sample, and provides readily usable data identifying whether or not any allele frequency has to be increased to meet the minimum allele frequency requirement.

Table 3 results are exactly parallel to a RFLP locus, showing the results of the homozygosity test. When the number of alleles are not so large (as in this case), the large-sample approximation of the chi-square statistic appears adequate (since the P-values of this table are in good agreement with their corresponding shuffling results).

Table 4 data confirms that the HWE assumption is supported for this data, since the observed number of distinct genotypes are within the plus/minus 2 S.E. limits of their corresponding expectations. For small number of possible alleles (say, for the polymarker and HLA-DQA1 loci), data in this table is generally not very informative.

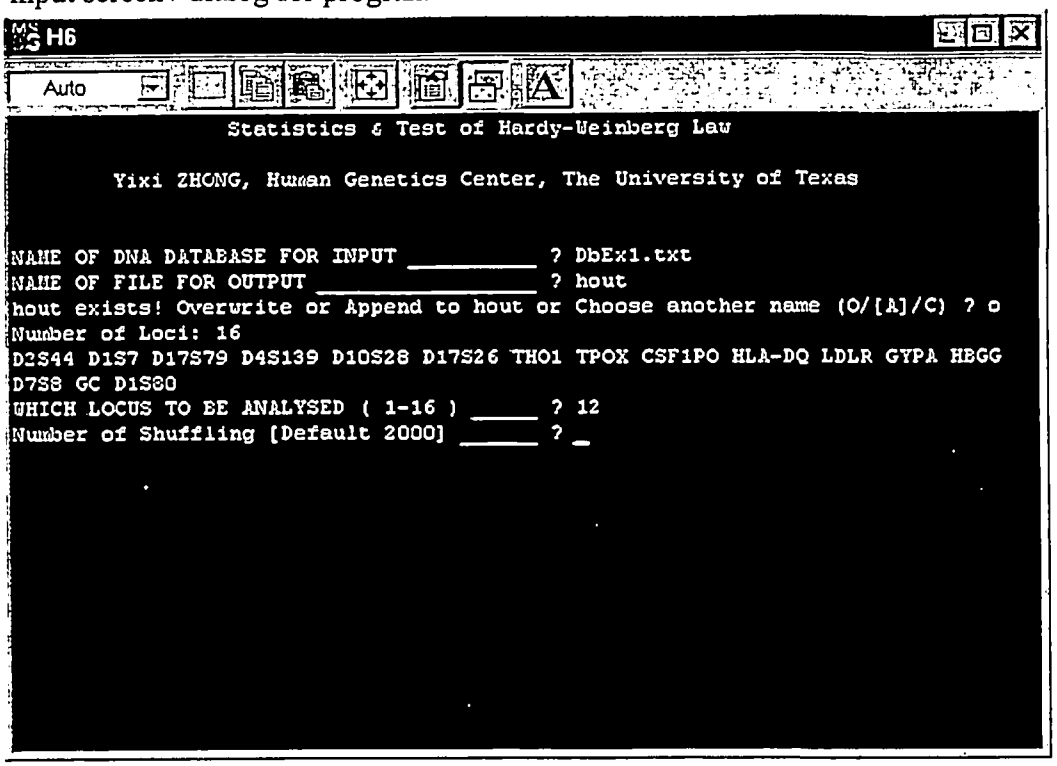
The shuffling results on the four HWE tests (exact, heterozygosity - biased and unbiased, and likelihood ratio) are shown in the next section. For this example, all empirical P-values are well above the nominal level of significance, say 5%), suggesting no significant departure from HWE.

The preliminary computations for checking whether or not null alleles are to be invoked in the analysis are similar to the ones of the analysis of a RFLP locus, shown in the last section. The maximum likelihood estimate of null-allele frequency (which is theoretically possible for a PCR locus due to "allele drop out" in the reverse dot-blot protocol of HLA-DQA1 and Polymarker loci, or due to differential amplification of alleles in STR loci) is 0.5819% in this example, which is not significantly different from zero (as judged from its corresponding Z-value of 0.260232, 1-sided P = 0.397342). Thus, the program completes its execution at this stage.

EXAMPLE 4

Example 4 of program H is another application to a PCR locus (GYPA) on the DbEx1.txt data where the number of alleles is still more limited (alleles A and B for GYPA). This example is chosen to indicate that the allele nomenclature can be alphabetic as well.

The input screen / dialog for program H for GYPA with the database DbEx1.txt is shown below.



With datafile DbEx1.txt chosen by the user and the locus GYPA selected, the program prompts that GYPA genotype data is available for all 100 individuals in the sample, and one set of 2,000 (default number) of shufflings are being done to complete the execution of the program producing the output file ("hout" (without quotes) name is given by the user with the overwrite option).

The output file produced during this execution is shown below.

```
Statistics & Test of Hardy-Weinberg Law for autosomal locus

          Yixi ZHONG, Human Genetics Center, The University of Texas

Locus: GYPA of DbEx1.txt    Actual sample size = 100



TABLE 1  Genotype frequencies

-----------------
        A     B
-----------------
 A      56
 B      34    10
-----------------
```

```
      A    B
----------------
Total number of individuals = 100       . = 'genotype' frequency is 0
```

Locus: GYPA of DbEx1.txt Actual sample size = 100

TABLE 1A Genotype frequencies and its expects

Obs. genotype	Obs. freq #(%)	Expected freq. Unbiased (%)	Expected freq. Biased (%)
A / A	56(56.00)	53.191(53.19)	53.290(53.29)
A / B	34(34.00)	39.618(39.62)	39.420(39.42)
B / B	10(10.00)	7.191(7.19)	7.290(7.29)
Others	0(0.00)	-0.000(-0.00)	0.000(0.00)
Total	100(100.00)	100.000(100.00)	100.000(100.00)

Locus: GYPA of DbEx1.txt Actual sample size = 100

TABLE 2 Allele frequencies and standard errors

Allele Classes	Observed Num.	Observed Percent	Observed SE (%)	SINGLE BAND
A	146	73.000	3.139	56
B	54	27.000	3.139	10
Total	200	100.000		66

Locus: GYPA of DbEx1.txt Actual sample size = 100

Table 3 Observed and expected frequency contrasts

	Obs. No.	Obs. Percent (%)	Expected No. and Percent Biased		SD	Unbiased		SD
Homozygote	66	66.00	60.58	60.58%	4.89	60.38	60.38%	4.89
Heterozygote	34	34.00	39.42	39.42	4.89	39.62	39.62	4.89

Chi-Square	1.230134	1.319401
Degrees of Freedom	1	1
Probability	0.267381	0.250700

Maximum Log Likelihood Ratio Methods:
G-Square 1.819 Degrees of Freedom 1 Probability 0.177421

Table 4 Numbers of distinct genotypes

	Obs.	Exp.	S.E.
Homozygote	2	1.9995	0.0227
Heterozygote	1	1.0000	0.0000
Total	3	2.9995	0.0261

OF SHUFFLED = 2000

PROBABILITY: # OF SHUFFLED CONDIT. PR.<= OBS. CONDIT.PR. = 437
PROBABILITY(1) = 0.21850

BIASED: # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 437
PROBABILITY(2) = 0.21850

UNBIASED: # OF SHUFFLED CHI-SQUARE >= OBS. CHI-SQUARE = 437
PROBABILITY(3) = 0.21850

MAXLOGL.: # OF SHUFFLED G-SQUARE >= OBS. G-SQUARE = 437
PROBABILITY(4) = 0.21850

Locus: GYPA of DbEx1.txt Actual sample size = 100

Null Allele Freq. Computations:

Prop of single banded types = .6600
Freq. of null alleles:

Biased =	0.073822
Unbiased =	0.076314
T =	1.137494
Z =	1.374936
1-SIDED PROBABILITY =	0.084576
r =	0.068747
SD(r) =	0.052346
HETEROZYGOSITY =	0.343486
NULL ALLELE CHI-SQUARE =	0.005390
NULL ALLELE G-SQUARE =	0.007141

06-16-1998 14:30:51

These results are exactly similar to the Example 3 results (output for the HLA-DQA1 Locus from the same example database). No genotype is missing in the sample (Table 1); the observed frequencies of all three genotypes agree with their HWE expectations (Table 1A); both alleles are rather common (frequency of A is 73% and that of B 27% - Table 2); and the none of the four tests (exact, chi-square - biased and unbiased, and likelihood ratio) detects any significant deviation from HWE (since all four empirical probabilities are far greater than any nominal level of significance, say 5%). The good agreement of the large-sample approximation of the P-values (see Table 3) with the shuffling P-values (for the chi-square and G-square test statistics) shows that for a 2-allele locus, a sample of 100 individuals is adequate for using the large sample approximation for the HWE tests.

Even though the maximum likelihood estimate of "null" allele is 6.8747%, its large standard deviation (5.2346) indicates that there is no need to invoke nondetectability of alleles for this locus in this database (Z-score = 0.068747, with 1-sided P = 0.084576).

Program B Independence at a Single Locus - Same as Program H but Uses Binned Genotype or Genotype Input Data

Program B generates results almost identical to program H, but the input data and its file format are different. The principal application of program B is to situations where, for some reason, the full population data are not available, and the only data available are binned genotype or genotype frequencies. It is unlikely that most users will find any practical need for program B with RFLP database files. However, it could be useful for data from PCR-based loci.

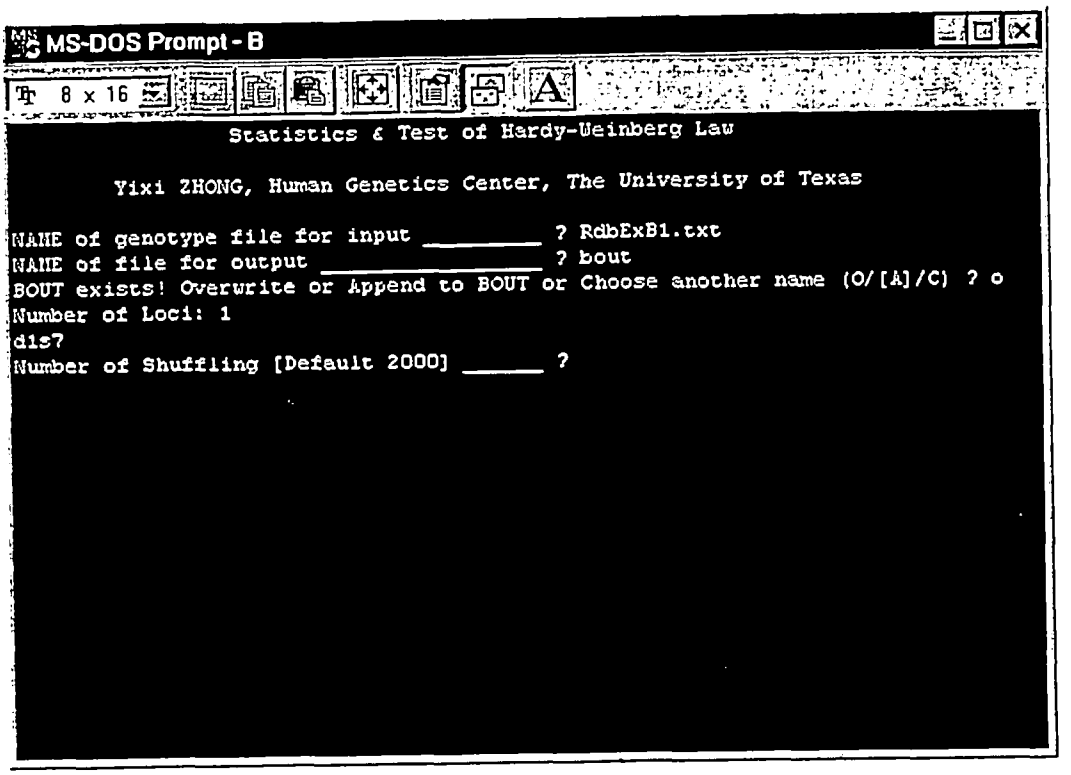
Program B input data for RFLP loci are in the format:

```
dls7 of RdbExH.txt
 1  18   1
 4  27   1
 6  13   1
 6  21   2
 7  14   1
 7  28   1
 8  20   2
 8  23   1
etc. etc.
```

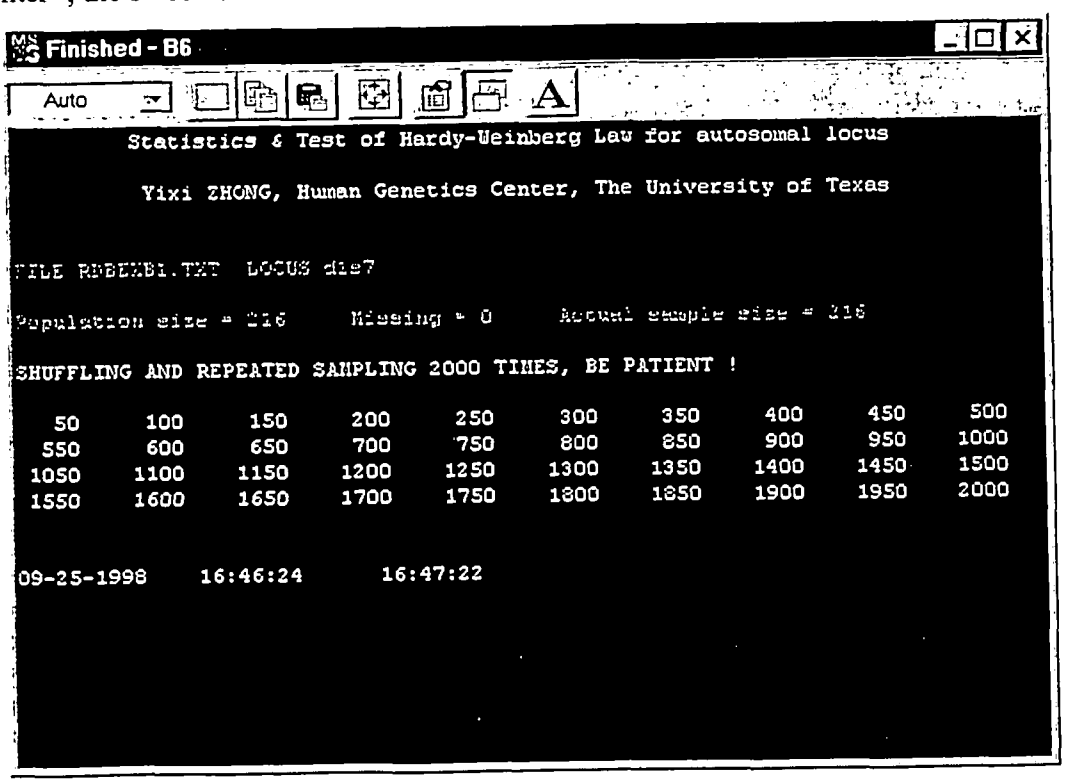
The first row is used for labels. As with H, B analyzes one locus for independence. The data shown above represent the first several lines of an example input file for program B called RdbExB1.txt. The data entries mean that there is one person with bands in bin 1 and bin 18, a second person with bands in bin 4 and bin 27, a third with bands in bin 6 and bin 13, etc. If, for some reason a user had data in this form only (and not the complete population bandsize data for the locus, program B would be required to analyze the data for independence.

Program B will run input files prepared in text processors as text files. See the Database or Help files on Database / Create / Edit for further discussion of this point. Example file RdbExB1.txt run with program B will give output that is almost identical to running RdbExH.txt with program H for the same locus, i.e. locus 1 or D1S7. Similarly, running RdbExB2.txt with program B will give output results almost identical to running RdbExH.txt with program H for the same locus, i.e. locus 2 or D2S44. The reason the output files are almost identical and not completely identical in all the output tables lies in the difference in input data file information and structure, and in the fact that the shuffling routine is random and thus slightly different with each iteration.

A program B input screen and running dialog is shown below. The first screen requests the database filename (RdbExB1.txt in this case), output filename (bout in this case) and offers the Append, Change, Overwrite option - if the filename already exists. The locus name is reported (from the database file), and the number of shufflings is requested (default is 2000).



After accepting the default shuffling number by hitting <Enter>, or entering a different number and hitting <Enter>, the screen switches.



The program reports the database filename, locus, total population size, number of missing data (if any) and actual population size. Next, it echoes the numbers of shufflings as it executes, until completion. The screens are very similar to those for program H, except that with H a user has to specify a locus number and name because the database file contains multiple loci, whereas here the database file contains data for only one locus.

Since program B output files are so similar to program H output files, the output files are not illustrated nor further discussed. The output from program H is fully discussed in the Help file for program H.

A program B input data file for HLA-DQA1 data, called PdbExB1.txt, is shown below:

```
HLADQA1
1.1 1.1 7
1.1 1.2 5
1.2 1.2 1
1.3 1.1 8
1.3 1.2 6
1.3 1.3 3
2 1.1 4
2 1.2 2
2 1.3 3
2 2 2
3 1.1 2
3 1.2 5
3 1.3 1
4 1.1 14
4 1.2 10
4 1.3 11
4 2 5
4 3 2
4 4 9
-1,-1,-1,-1,-1
```

The first line should contain only the locus name. Then each genotype and the number of people who were observed to have it , in space delineated ASCII format, is entered on a separate line. The database file PdbExB1.txt contains the same HLA-DQA1 data as DbEx1.txt does for the HLA-DQA1 locus. Thus, running PdbExB1.txt with B produces nearly identical output results as does running DbEx1.txt with H for HLA-DQA1. The shuffled probabilities will not be quite identical because of the randomness in the shuffling routine (just as repeated runs of DbEx1.txt with H for HLA-DQA1 will not give identical shuffled probabilities).

A final example of program B input data file for GYPA, called PdbExB2.txt, is shown below:

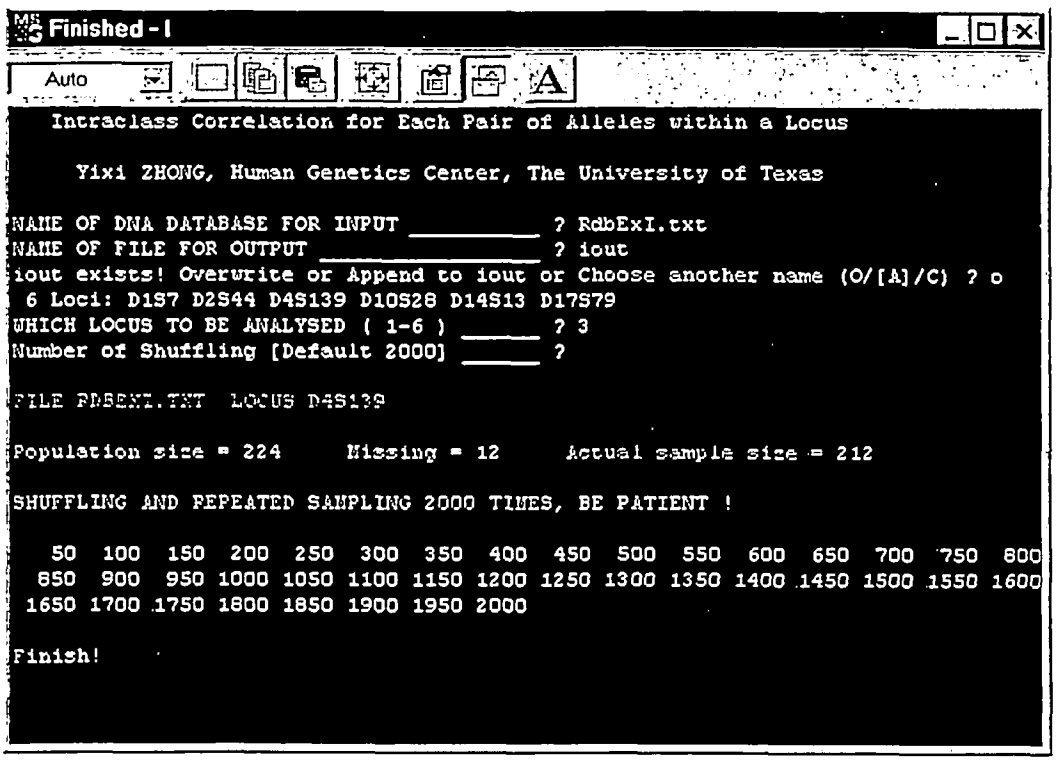
```
GYPA
A A 56
B A 34
B B 10
-1,-1,-1,-1,-1
```

This file contains the same GYPA data as DbEx1.txt does for the GYPA locus. Thus, running B with PdbExB2.txt produces nearly identical output results as does running DbEx1.txt with H for GYPA.

Program I Karlin's Intraclass Correlation Within a Locus - Independence at a Single Locus

This program generates intraclass correlation coefficients of fragment sizes within loci and tests independence based on these correlations. Karlin's nonparametric correlation coefficients (Karlin et al.1981; Chakraborty et al. 1993) are specifically computed in this routine, which are analytically, as well as empirically, shown to be almost identical to the Analysis of Variance (ANOVA)-based measure of intraclass correlation (Weir 1992).

An example input screen for program I for the third locus (D4S139) of RdbExI.txt is shown below. (The screen may switch after entering the first few items, but the screen dialog for the two screens is identical to that on the one pictured.)



The filename "iout" (no quotes) was chosen for output. Since it already existed, the program offered the choices of append, overwrite or choose new. Overwrite was selected. Locus 3 (D4S139) was selected for analysis, and the default number of shufflings (2000) was accepted by pressing <Enter>.

The output from executing program I on locus 3 (D4S139) is shown below:

```
        Intraclass Correlation for Each Pair of Alleles within a Locus

     Yixi ZHONG, Human Genetics Center, The University of Texas

 File name: RDBEXI.TXT       LOCUS  3 D4S139
Population size = 224      Missing = 12      Actual sample size = 212

RHO                              =  -0.002733
KARLIN-RHO (Rk)                  =  -0.005094
T-VALUE OF OBSERVED RHO          =  -0.005350
T-VALUE OF OBSERVED KARLIN-RHO =  -0.039566

NUMBER OF SHUFFLED = 2000

NUMBER OF SHUFFLED RHO DEPARTURE >= OBSERVED RHO DEPARTURE  =   1014
                                    ONE SIDED PROBABILITY   = 0.5070
NUMBER OF SHUFFLED  Rk DEPARTURE >= OBSERVED  Rk DEPARTURE  =   1014
                                    ONE SIDED PROBABILITY   = 0.5070
NUMBER OF SHUFFLED |RHO| >= OBSERVED |RHO|                  =   1943
                                    TWO SIDED PROBABILITY   = 0.9715
NUMBER OF SHUFFLED  |Rk| >= OBSERVED  |Rk|                  =   1885
                                    TWO SIDED PROBABILITY   = 0.9425

05-16-1998    19:49:20
```

The quantity called "RHO" in this output is the ANOVA-based estimate of the intra-class correlation (equation 3 of Chakraborty et al. 1993) while the quantity called "KARLIN-RHO" is the nonparametric estimate proposed by Karlin et al. (1981) (equation 2 of Chakraborty et al. 1993). As mentioned before, their differences are usually very small (equation 4 of Chakraborty et al. 1993 shows that one can be obtained from the other, with the sample size in the database
known), and for the purpose of checking whether or not the observed intraclass correlation is significant, either one of them would suffice. In the present example, both of them indicate that there is no significant intraclass correlation, the two-sided probabilities are 0.9715 and 0.9425 for RHO and KARLIN-RHO, respectively.

Since, under population substructure, we expect significant positive intraclass correlation, the corresponding test statistic (from either RHO or KARLIN-RHO) is the normal deviate (equation 9 of Chakraborty et al. 1993) which, in this output, is called the T-VALUE. While in large samples, the normal deviate would follow a standard normal distribution (and hence, T-value greater than 1.6495 would signify significant positive intraclass correlation), this program avoids making the large sample approximation.

As in program H, alleles (fragment sizes) are shuffled for a number of replications (in this output 2000 replications of shufflings were done) and the number of occurrences in the shuffled data with RHO greater than the observed RHO in the real data were computed through one-sided (shuffled RHO > Observed RHO) or two-sided (Shuffled absolute value of RHO > Observed absolute value of RHO) tests. The proportion of these cases (divided by number of replications) are the empirical levels of significance, called ONE-SIDED and TWO-SIDED probabilities, each of which are computed in the program for the ANOVA-based estimator (RHO) as well as for the non-parametric estimator (KARLIN-RHO).

Inference of the presence of significant intraclass correlation is to be judged from the two-sided probabilities, a value of which smaller than 0.05 would be regarded as significant departure (at 5% level) from the independence assumption. As stated earlier, this example shows that the two-sided probabilities (0.9715 for RHO and 0.9425 for KARLIN-RHO) far exceed a nominal 5% level of significance, indicating the absence of any significant intraclass correlation. In addition, in this example the one-sided probabilities for RHO and KARLIN-RHO are exactly the same (0.5070, which is = 1014/2000), each far greater than 0.05. Thus, the one-sided test as well shows that there is no indication of a positive significant departure from the null value (RHO = 0) of intraclass correlation at the D4S139 locus in the data.

NOTE: When RHO or KARLIN-RHO is negative, one-sided probabilities would generally be larger than 0.5. Nevertheless, since the empirical distributions of these two statistics are not exactly standard normal, their two-sided probabilities need not be identical, nor can they be theoretically predicted from the one-sided probabilities, as is usually possible for the strict standard normal deviate. In some situations, the two-sided probabilities may be smaller than a nominal level of significance (say, 5%), in which case if the estimates (RHO and KARLIN-RHO) are negative, that significant (negative) intraclass correlation cannot be ascribed to population substructure, since under substructuring, we expect a positive intraclass correlation. Usually such aberrant results are due to chance occurrence, or due to technical problems in the data that may require additional investigation.

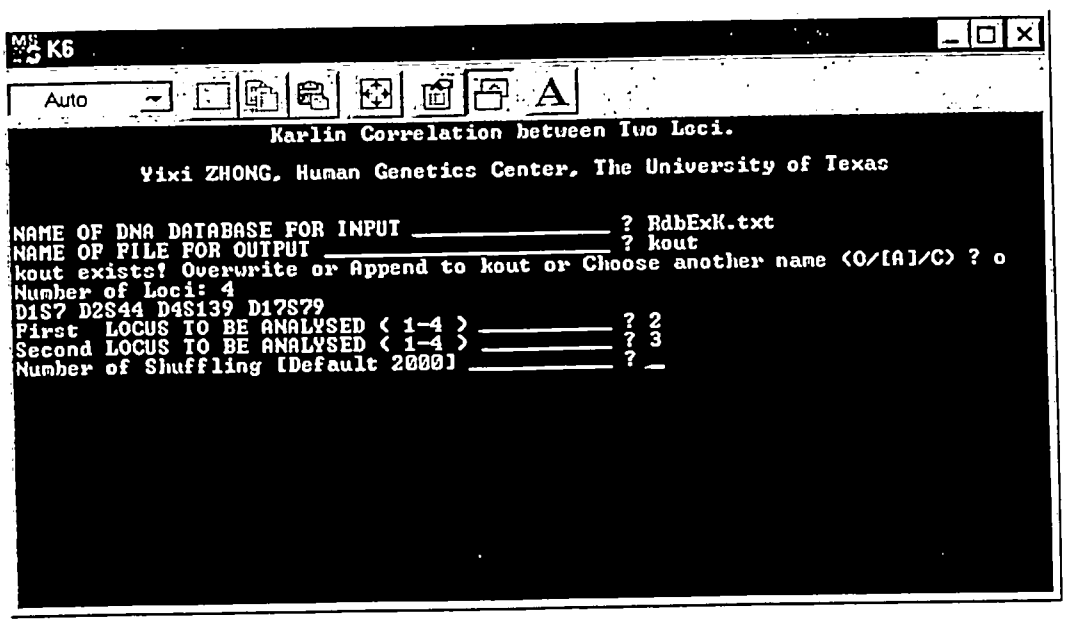
Although this correlation test will run even on loci where the allele designations are not expressed in terms of repeat sizes (as is the case with any of the Polymarker or HLA-DQA1 loci), results of program I from these databases should be interpreted with caution, since by chance, significant intraclass correlation is occasionally expected for such loci even when the Hardy-Weinberg expectations of genotype proportions show a good fit to the observed data. This can happen specifically when one of the alleles is comparatively rarer than the other(s), so that homozygotes of the rarer alleles are uncommon in the data. For the HLA-DQA1 or Polymarker loci, alleles with labels A, B and C (or 1.1, 1.2, etc.) are given nominal scores (1, 2, etc. for A, B, etc.; and 11, 12, 13, 20, etc. for 1.1, 1.2, 1.3, 2, etc.) so that the intraclass correlation concept is dependent on the nominal scale of the alleles.

Program K Pairwise Comparison of Loci for Independence - Karlin's Interclass Correlation Between Two Loci

This program performs a test of independence of fragment sizes between pairs of loci. Karlin's nonparametric interclass correlations (Karlin et al. 1981) are used, which are analytically, as well as empirically, identical to ANOVA-based estimates of the interclass correlation test (Weir 1992; Chakraborty et al. 1993).

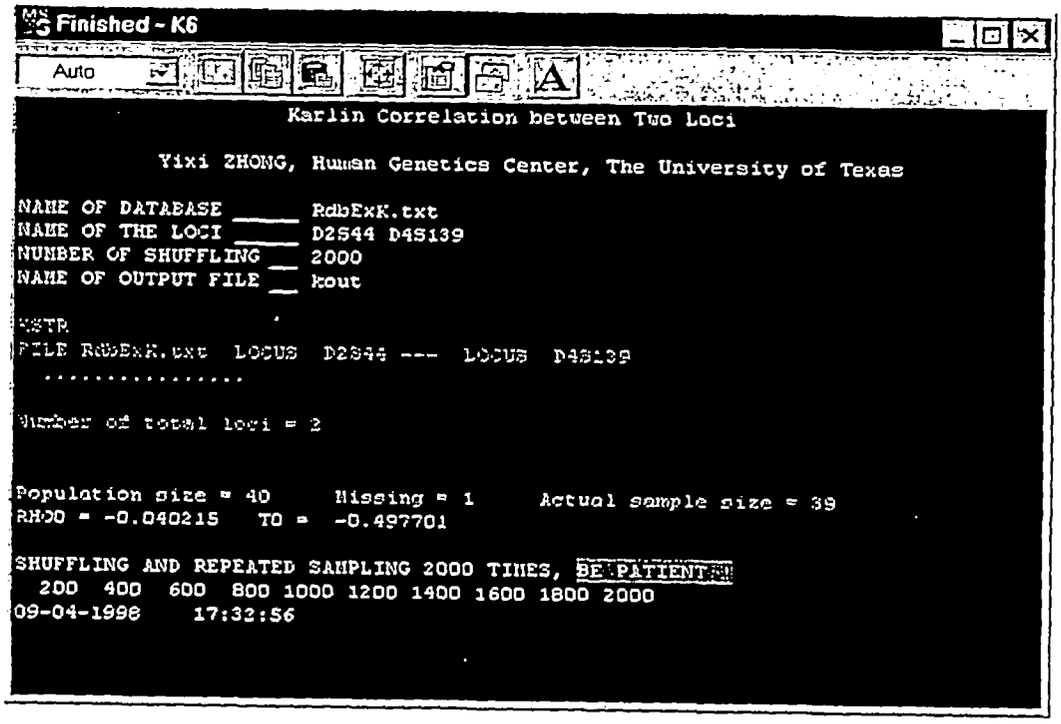
Program K is a test for independence of fragment sizes across a pair of loci. Therefore, this menu is to be executed for all pairs of loci available in the database (for n loci there will be n(n-1)/2 pair-wise comparisons). Interclass correlations are computed based on fragment size data available for the chosen pair of loci. Total sample size, as well as the number of individuals typed at both loci are recorded in the output. Only Karlin's nonparametric estimate (see Chakraborty et al. 1993) is computed; its sampling properties are virtually identical to Weir's estimate (Weir 1992) for large samples. Both one-sided and two-sided probabilities are determined by shuffling. Two-sided probabilities lower than 0.05 are indicative of a lack of independence of fragment sizes for the pair of loci. Only positive correlations are expected if the lack of independence is the result of linkage disequilibrium due to population substructuring in the data. In the strict sense, the fragment sizes across the two loci may not be truly independent when negative correlations are observed. However, negative significant correlations are of no concern for forensic applications of the database, since in such situations, the multiplication of fragment size frequencies across the pair of loci would produce a conservative estimate.

A program K input / dialog is shown below:



In this first input / dialog screen for program K, with RdbExK.txt selected as the database file, kout selected a the output file, and "overwrite" selected as the option because kout already exists. The program reported the number and names of the loci, and asks for the 1st, then the 2nd, locus to be included in the analysis. Number 2 (D2S44) and number 3 (D4139) were selected. The default number of shufflings (2000) was accepted by hitting <Enter>.

The screen then changes to the one below, as the program executes and finishes.



The program shows that the database has 40 persons total and a missing record for one or both of the selected loci, and thus an actual sample size of 39. The shuffling dialog follows.

The results (output) of program K when run with RdbExK.txt for D2S44 and D4S139 is shown below:

```
                    Karlin Correlation between Two Loci

      Yixi ZHONG, Human Genetics Center, The University of Texas

File name: RDBEXK.TXT  The 1st locus = 2 D2S44  The 2nd locus = 3 D4S139

Population size = 40      Missing = 1      Actual sample size = 39


Rho of observed data          =     -0.040215

T-value of observed data      =     -0.497701

NUMBER OF SHUFFLED = 2000

NUMBER OF SHUFFLED RHO DEPARTURE >= OBSERVED RHO DEPARTURE  =     699
                                    ONE SIDED PROBABILITY    = 0.3495
```

```
NUMBER OF SHUFFLED |RHO| >= OBSERVED |RHO|                    =   1310
                                TWO SIDED PROBABILITY     = 0.6550
```

Unlike the intraclass correlation routine (Program I), this program output lists only the non-parametric estimate of interclass correlation, called Rho (equation 6 of Chakraborty et al. 1993) and its associated normal-deviate test criterion (called T-value). One- and two-sided shuffling test results on the Rho are also listed. As in the intraclass correlation test, when the two-sided probability (empirical) is below a nominal level of significance (say, 5%), we would infer the presence of a significant interclass correlation. In the example give, the two-sided probability (0.6550) far exceeds the nominal level of 5%, indicating that there is no interclass correlation between D2S44 and D4S139 loci in this data.

Should one observe a significantly smaller (than 5%) value of the two-sided probability, it is instructive again to check the sign of Rho, since negative significant interclass correlations cannot be readily ascribed to the presence of population substructure in the data (see Chakraborty et al. 1993).

As seen in the example, the one-sided empirical level of significance (699/2000 = 0.3495) is not exactly one-half of the two-sided empirical level of significance (1310/2000 = 0.6550), because the empirical distribution of the T-value of Rho is not exactly a standard normal (which in turn justifies why the testing should be done through the shuffling algorithm).

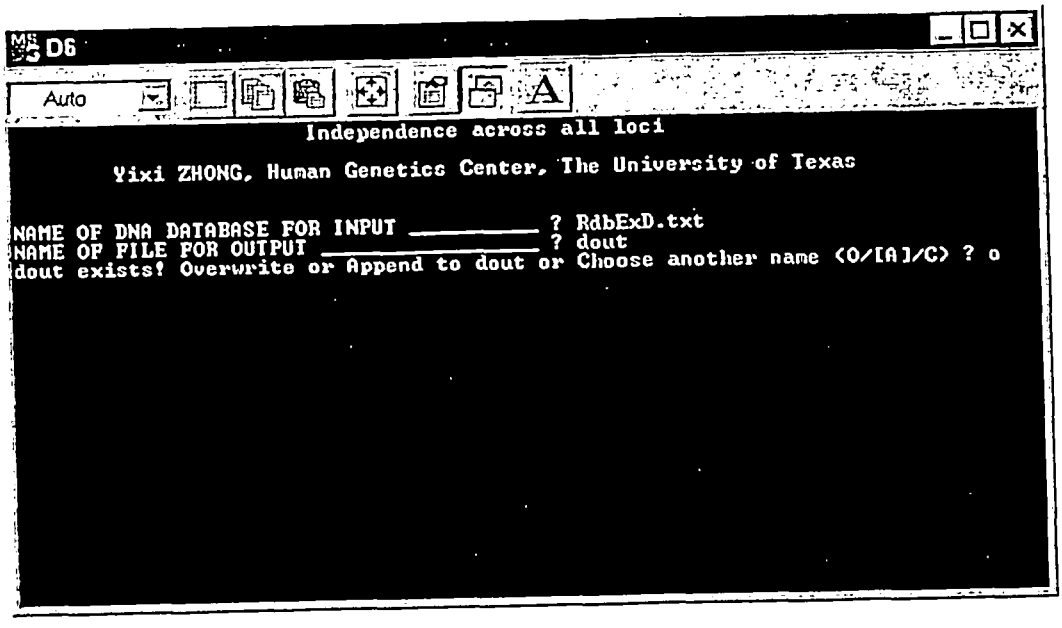
NOTE: As in the case of the program I, when one or both loci of the interclass correlation test have only a few alleles (PM and HLA-DQA1 loci, for example), the nominal scales used for the allelic designation may affect the result, and may cause departure from independence due to the presence of one or more rare alleles. Thus, a significant test result with respect to this test must be assessed in reference to actual genotype frequency tables to decide whether one or more rare alleles are the source of such significant departures from independence.

Program D Test of Independence Across All Loci

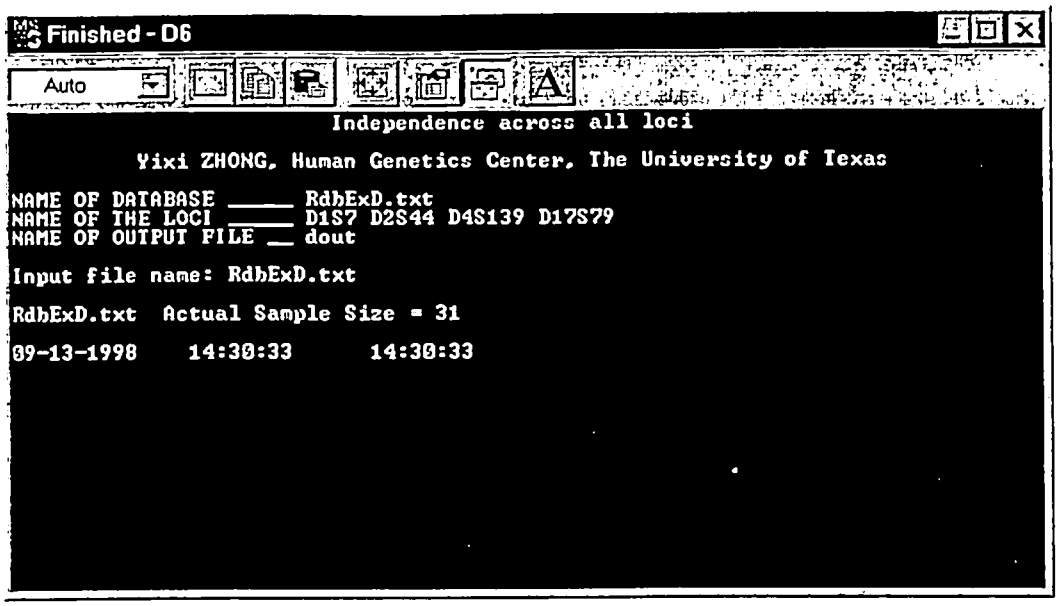
This program provides a global test of independence of loci with data where individuals are typed for all loci. Thus, individuals with one or more untyped loci in the data set are excluded from analysis by this program. The test done in this program is based on the distribution of the number of heterozygous loci across individuals (the theory for which is described in Brown et al. 1980 and in Chakraborty 1984).

Both RFLP and PCR-based loci (such as the PM loci, HLA-DQA1 and STRs) can be used in this program. The fragment sizes of the RFLP loci are first grouped into bins and binned genotypes are considered for the RFLP group of loci (from one of the tables of the output of the program H). With K loci in the database, the variable considered in this test is a random variable that takes on values of 0 (for individuals who are homozygous for all loci) to K (for individuals who are heterozygous for all loci). The variance of the distribution of the number of heterozygous loci provides information about linkage disequilibrium across loci (Brown et al. 1980; Chakraborty 1984). Under linkage equilibrium (independence of alleles across loci), the expected variance of the number of heterozygous loci, as well as its 95% confidence limits, can be computed (see Brown et al. 1980 for their computations), which can be done from the locus-specific heterozygosities. Since there are two ways of obtaining heterozygosity (observed proportions and the ones under the HWE assumption (see note numbers 2 and 3 below), the expected variance and its 95% confidence intervals are computed using both alternative estimates of locus-specific heterozygosities.

Two examples of the operation of this program menu are shown for illustration. The first of these is run on an example database called RdbExD.txt. The first input screen is shown below.



Note that the program requests an input database – RdbExD.txt was given, and an output file name. The name "dout" (without the quotes) was given. The program reported that this output file already exists, and gave the option of appending, overwriting or creating a new name. The option of overwrite was selected. Hitting <Enter> yields the next screen as the program executes and finishes.



The program finishes its execution and, on the screen, it reports that it ran the database file RdbExD.txt that had four loci (D1S7, D2S44, D4S139 and D17S79), and that the file contained 31 individuals with genotype data on all four loci available. The output may be viewed (in WordPad) by opening the file name chosen for output (dout, in this example).

Output of program D with the RdbExD.txt example database is shown below:

```
                    Number of Heterozygous Loci (Fixed Bin)
      Yixi ZHONG, Human Genetics Center, The University of Texas

FILE IS RDBEXD.TXT  Actual Sample Size = 31
TOTAL 4 LOCI: D1S7 D2S44 D4S139 D17S79

                                TABLE
Distribution of Number of Heterozygous Loci

-------------------------------------------------------------------
 Heterozygosity at Loci     Observed        Expected Freq. Based on
                           ------------    --------------------------
          x                Number  Freq.    Obs. Het.   Expected Het.
-------------------------------------------------------------------
          0                  0    0.00000    0.00015      0.00004
          1                  0    0.00000    0.00687      0.00196
          2                  0    0.00000    0.08674      0.03753
          3                 18    0.58065    0.38598      0.28856
          4                 13    0.41935    0.52027      0.67191
-------------------------------------------------------------------
Mean                             3.41935    3.41935      3.63035
Variance                         0.25161    0.45994      0.32025
95% C.I. of Variance:
    Lower Limit                                0.21942      0.12091
    Upper Limit                                0.70046      0.51959
```

The top part of this output is self-explanatory, showing that, of the 31 individuals in the database typed for all four loci (which are listed), the observed counts and proportions (frequencies) of individuals who are heterozygous at 0, 1, ..., 4 loci are given along with their expectations under the the hypothesis of global linkage equilibrium (i.e., allelic independence across all loci), using two alternative estimates of locus-specific heterozygosities. The expectations appear under the column titled "Expected Freq. Based on", and it shows observed proportions of locus-specific heterozygosities (under the sub-column Obs. Het.) and under the HWE assumption using the bias-corrected estimate of the expected heterozygosity at each locus (under the sub-column Expected Het.). These columns may be imported in any graphic software to show the agreements of the observed and expected distributions of the number of heterozygous loci in the data.

At the bottom part of the output, the test results of global linkage equilibrium are shown. For the purpose of completeness, the observed and expected mean number of heterozygous loci in the data are shown (not relevant for the global tests).

Inference with regard to global independence of alleles across loci is made through the comparison of the observed variance of the number of heterozygous loci (shown in the row labeled as variance) and its expectation under the independence assumption. In particular, the user should check to see if the observed variance is within the 95% confidence interval. While, in general, any of the two expected columns will give the same inference, since the last column in its strict sense tests for HWE for all loci along with the global linkage equilibrium, it is more desirable to examine if the observed variance is within the 95% confidence limits under the Obs. Het. column where the range (95% limit) of variance is computed under the allelic independence (across loci) without invoking the HWE assumption for the individual loci.

In this example, the observed variance (0.25161) is between the 95% confidence limits of either of the two expected columns, indicating that there is no evidence of global disequilibria between loci; thus, the use of the product rule is appropriate for multilocus genotype probability calculations.

A second example of program D operation can be illustrated with a database called PdbExH.txt that contains data for HLA-DQA1, PM and D1S80. The input screens and dialog work the same way as illustrated above.

The output from the program for PdbExH.txt is shown below.

```
Independence across all loci

          Yixi ZHONG, Human Genetics Center, The University of Texas


INPUT FILE NAME: PdbExH.txt

 HLA-DQA1 LDLR GYPA HBGG D7S8 GC D1S80

Actual Sample Size = 191
```

TABLE

Distribution of Number of Heterozygous Loci

Heterozygosity at Loci	Observed		Expected Freq. Based on	
			Obs. Het.	Expected Het.
x	Number	Freq.		
0	0	0.00000	0.00050	0.00077
1	1	0.00524	0.00993	0.01360
2	16	0.08377	0.06972	0.08290
3	45	0.23560	0.21329	0.22945
4	53	0.27749	0.32657	0.32474
5	55	0.28796	0.26063	0.24341
6	13	0.06806	0.10333	0.09159
7	8	0.04188	0.01603	0.01354
Mean		4.13089	4.13089	4.02806
Variance		1.54594	1.39053	1.42664
95% C.I. of Variance:				
Lower Limit			1.12723	1.15651
Upper Limit			1.65384	1.69677

This output indicates that the database used has 191 individuals each of whom were typed for all 7 loci as listed in the top two lines of the output file.

Below the table of the comparison of the observed and expected distributions of the number of heterozygous loci in these 191 individuals, the global test of linkage disequilibrium again indicates that the observed variance (1.54594) is within the 95% confidence limits (1.12723 to 1.65384) based on the observed heterozygosities as well as that (1.15545 to 1.69524) based on the (bias-corrected) expected heterozygosities (under HWE). Thus, in this data, there is no evidence of a significant departure from the global independence of alleles across the seven PCR loci.

NOTES for Program D:

1. Since individuals with data missing for some loci are excluded from analysis in this program, it may be preferred to split the data into groups of loci (say, RFLP and PCR data) where large sets of individuals are typed for one group of loci only in cases where the database file have a large number of missing data for one or more loci.

2. Since RFLP loci are rescored into binned genotypes for this test, individuals who are called homozygotes for any of the RFLP loci are not necessarily single-banded at those specific loci. They simply have both fragments within the same bin (of the fixed bin categories). Thus, heterozygosity for a RFLP locus is simply the proportion of individuals having two fragments in any two different bins (observed heterozygosity), or the complement of sum of squares of binned allele frequencies (the expected heterozygosity under the HWE assumption); both of which are computed in program H.

3. For the PCR loci, homozygotes refer to the individuals with two copies of the same allele and the heterozygotes have two different alleles at the locus. Of course, locus-specific heterozygosities are either obtained by the observed ones (the observed proportion) or their expectations under the HWE assumption (the complement of sum of squares of allele frequencies, with the bias correction - see documentation of the H program for further details).

4. It is possible that the observed variance may be within the 95% limits of the expected column under the heading of "Obs. Het." but not for the column under "Expected Het.". Should such a situation be encountered, it is advisable to go back to the H program results of the same database for each individual locus separately and see if any of the individual loci is out of HWE for some reason or the other. In such cases, for casework analysis, the user may have to invoke adjustments for computing single-locus genotype frequencies (such as the 2p rule for RFLP loci, the theta-adjustment for a PCR locus, or using the minimum threshold allele frequency for certain loci for rare genotypes). But still, the individual locus-specific genotypes may be multiplied, because the "Obs. Het." column showed no evidence of departure from allelic independence across loci without invoking the HWE assumption for each individual locus.

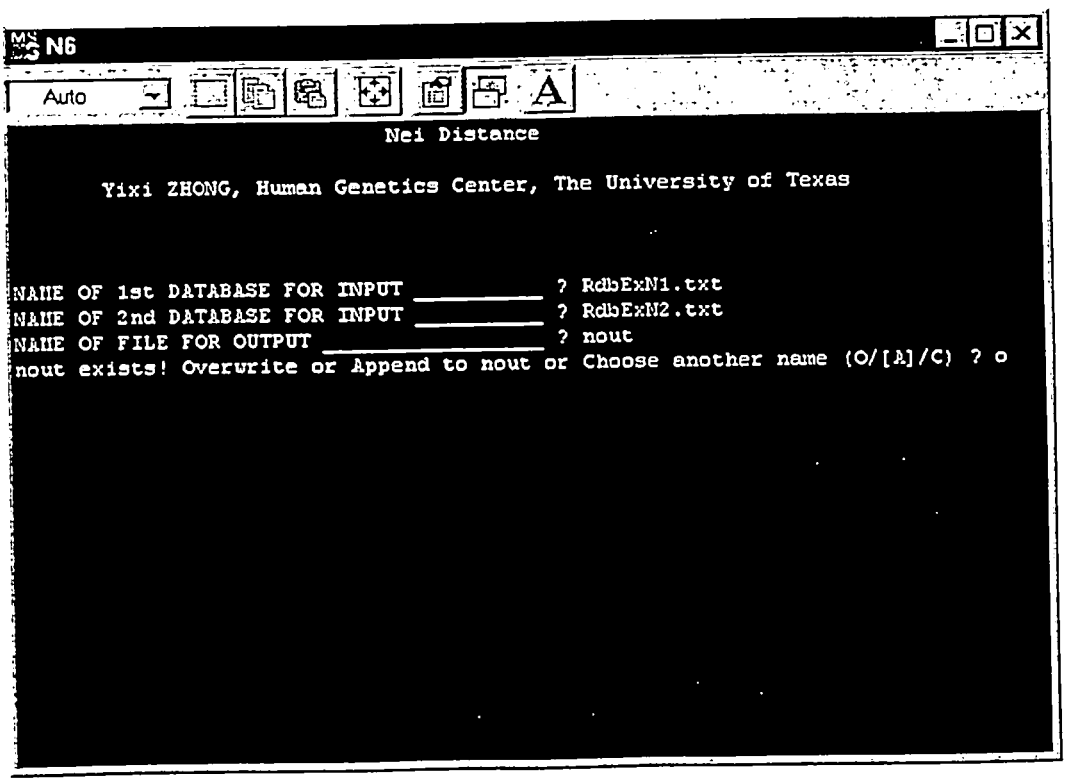
5. In this test of global independence, since the actual genotype data are summarized in the form of heterozygotes and homozygotes, some may claim that it does not actually test whether each individual multilocus genotype satisfies the mutual independence of loci. In principle, this criticism is valid. However, with large number of alleles segregating, tests based on individual multilocus genotypes are either too crude with reference to the actual level of significance (even when suffling is done to obtain the level for significance empirically), or there are complications that are subject to multiple testing problems. Thus, in spite of the data summary, this test serves the purpose of approximately determining the adequacy of the prodcut rule, when this program is run, particularly if done in combination with the programs S and/or T which perform complete and partial match tests with respect to all loci in the dataset in all pairwise comparisons of individuals in the database.

Program N Nei's Genetic Distance Between Populations (Across Same Sets of Loci in Two Databases) - Compare Databases for Genetic Distance

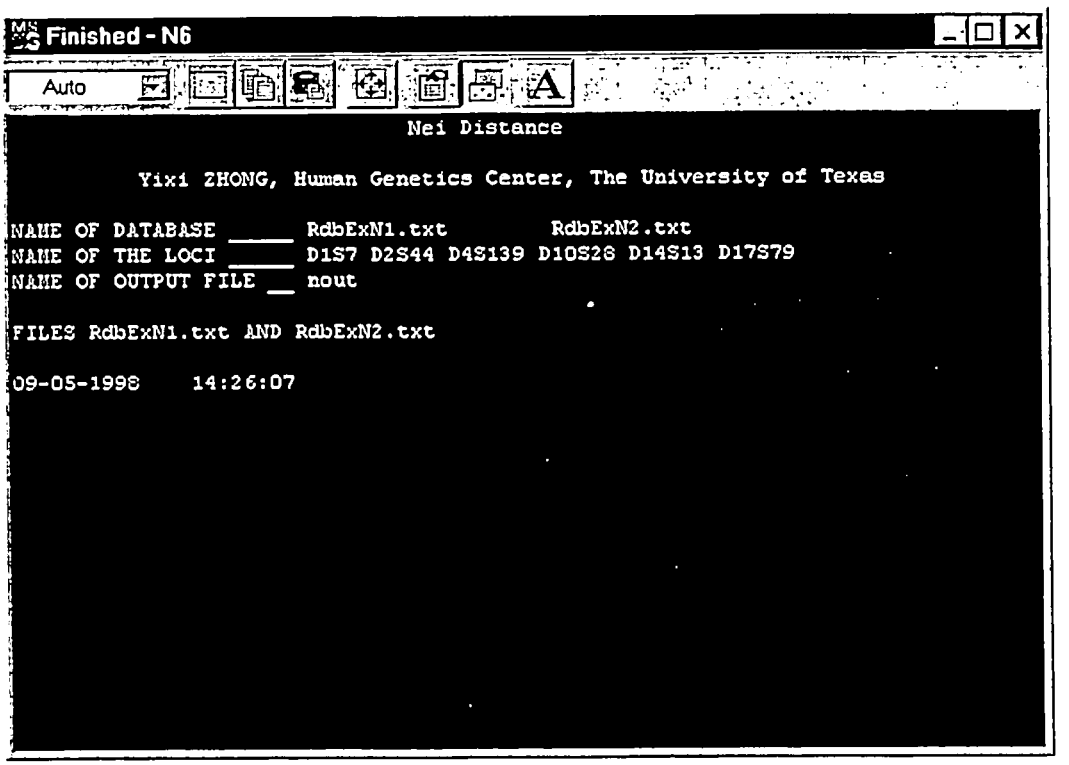
Program N performs comparison of allele frequencies in two databases each containing genotype data on the same set of loci. Although the tasks performed in this menu are generally beyond the need of most forensic analysts, it is useful in certain situations when the analyst needs to assess whether or not either of the two databases can be used for any specific purpose (e.g., if two alternative sources of data from the same population (or two similar populations) can be interchangeably used).

Three specific tasks are performed in this program: (i) computations of locus-specific heterozygosities and their standard errors for each of the two populations (the theory of which is given in Nei, 1978 with appropriate bias-correction of estimation); (ii) a 2 X C contingency table chi-square analysis to check the significance of allele frequency differences at each locus between the two populations; and (iii) computations of Nei's minimum (Dm) and standard (Ds) genetic distances and their standard errors (intra-locus as well as inter-locus, called SE. INTER and SE. INTRA separately for Dm and Ds, respectively) for each locus, each pair, triplet, etc. of loci in the databases. In all computations, Nei's bias-corrected estimation method is used (see Nei, 1978) since for loci with as many alleles segregating as most of the loci used for forensic purposes, genetic distances without bias-correction can take absurd values that are not meaningful for any biological inference.

The input / dialog screen for program N using two databases called RdbExN1.txt and RdbExN2.txt is shown below:



The filename nout is specified for the output data. Since it already exists, the choices of overwrite, append or change are given. Overwrite is selected. The program then runs, ending with the screen below.



The program requested the names of the two databases to be compared. In this case, RdbExN1.txt and RdbExN2.txt, databases with 224 and 329 people, respectively, from the same ethnic group but different locations, are given. The program requested an output filename -- "nout" (no quotes) was given. Since it existed, the user was prompted to choose append, overwrite or change the name. Overwrite was chosen. The program then reported the names of the six loci in the databases and the names of the input files as it executes and finishes. The user could then retrieve for viewing or printing the "nout" (no quotes) output file (by opening it in WordPad).

The program N output file from running it with RdbExN1.txt and RdbExN2.txt is shown below:

```
HETEROZYGOSITY AND DISTANCES BETWEEN TWO POPULATIONS WITH ITS S.E.

          Yixi ZHONG, Human Genetics Center, The University of Texas

  (A)RdbExN1.txt and (B)RdbExN2.txt:D1S7 D2S44 D4S139 D10S28 D14S13 D17S79

                HETEROZYGOSITY AND HETEROGENEITY CHI-SQUARE
```

LOCUS	POPULATION (A) EXPECTED H	SE(H)	POPULATION (B) EXPECTED H	SE(H)	CHI-SQUARE	DF	p-Value
1	0.94539	0.00233	0.94804	0.00204	105.48	29	0.000000
2	0.92867	0.00335	0.92987	0.00248	133.05	24	0.000000
3	0.88982	0.00524	0.90721	0.00376	70.18	20	0.000000
4	0.93611	0.00299	0.94226	0.00257	63.51	25	0.000034
5	0.94715	0.00331	0.91554	0.00602	210.59	29	0.000000
6	0.83225	0.00753	0.81337	0.00574	101.29	17	0.000000

NEI'S MINIMUM, STANDARD DISTANCES AND THEIR INTER-LOCUS & INTRA-LOCUS S.E.

LOCUS	Dm	SE.DmINTER	SE.DmINTRA	Ds	SE.DsINTER	SE.DsINTRA
100000	0.0022825	0.0000000	0.0014308	0.0434709	0.0000000	0.0274085
010000	0.0036469	0.0000000	0.0017921	0.0529034	0.0000000	0.0260356
001000	0.0016351	0.0000000	0.0017389	0.0125596	0.0000000	0.0163635
000100	0.0010464	0.0000000	0.0014628	0.0160750	0.0000000	0.0240244
000010	0.0087169	0.0000000	0.0027952	0.1085587	0.0000000	0.0385539
000001	0.0068294	0.0000000	0.0031843	0.0378842	0.0000000	0.0184960
110000	0.0029647	0.0006822	0.0011466	0.0488732	0.0046164	0.0189403
101000	0.0019588	0.0003237	0.0011260	0.0235383	0.0141640	0.0143440
011000	0.0026410	0.0010059	0.0012486	0.0296934	0.0197255	0.0144211
100100	0.0016644	0.0006181	0.0010231	0.0288640	0.0136368	0.0181076
010100	0.0023467	0.0013003	0.0011567	0.0359394	0.0182981	0.0178388
001100	0.0013407	0.0002943	0.0011362	0.0140234	0.0017194	0.0136629
100010	0.0054997	0.0032172	0.0015701	0.0874375	0.0303859	0.0258374
010010	0.0061819	0.0025350	0.0016602	0.0869041	0.0281918	0.0242158
001010	0.0051760	0.0035409	0.0016460	0.0619023	0.0504573	0.0198673
000110	0.0048817	0.0038353	0.0015774	0.0735471	0.0452550	0.0244811
100001	0.0045559	0.0022734	0.0017455	0.0397182	0.0025688	0.0155787
010001	0.0052381	0.0015912	0.0018270	0.0425401	0.0064419	0.0151958
001001	0.0042322	0.0025972	0.0018141	0.0308417	0.0105409	0.0132846
000101	0.0039379	0.0028915	0.0017521	0.0332926	0.0073399	0.0150768
000011	0.0077732	0.0009438	0.0021186	0.0600249	0.0306073	0.0175566
111000	0.0025215	0.0005929	0.0009593	0.0330115	0.0138257	0.0127916
110100	0.0023253	0.0007510	0.0009067	0.0381059	0.0114265	0.0149560
101100	0.0016546	0.0003570	0.0008951	0.0214464	0.0086142	0.0123354
011100	0.0021095	0.0007873	0.0009647	0.0261218	0.0127808	0.0123664
110010	0.0048821	0.0019574	0.0012052	0.0764637	0.0233551	0.0192549
101010	0.0042115	0.0022605	0.0011965	0.0578778	0.0345044	0.0165418
011010	0.0046663	0.0021069	0.0012494	0.0595124	0.0319587	0.0160514
100110	0.0040153	0.0023778	0.0011547	0.0662344	0.0319617	0.0193244
010110	0.0044701	0.0022522	0.0012094	0.0674949	0.0290338	0.0185692
001110	0.0037995	0.0024646	0.0012008	0.0504610	0.0351317	0.0160221
110001	0.0042529	0.0013471	0.0013081	0.0429718	0.0045381	0.0134164
101001	0.0035823	0.0016343	0.0013001	0.0329089	0.0079162	0.0119914
011001	0.0040371	0.0015121	0.0013489	0.0352784	0.0084594	0.0118402
100101	0.0033861	0.0017582	0.0012617	0.0353461	0.0053724	0.0133092
010101	0.0038409	0.0016722	0.0013120	0.0378622	0.0064026	0.0130819
001101	0.0031703	0.0018374	0.0013040	0.0283915	0.0089361	0.0117179
100011	0.0059429	0.0019096	0.0014907	0.0582462	0.0223638	0.0152976
010011	0.0063977	0.0014794	0.0015335	0.0595053	0.0215086	0.0149196
001011	0.0057271	0.0021173	0.0015267	0.0495970	0.0216894	0.0135368
000111	0.0055309	0.0023075	0.0014942	0.0530066	0.0221418	0.0149332

111100	0.0021527	0.0005584	0.0008071	0.0293912	0.0105465	0.0112920
111010	0.0040704	0.0016046	0.0010030	0.0567873	0.0252073	0.0140718
110110	0.0039232	0.0016839	0.0009751	0.0629988	0.0232499	0.0158232
101110	0.0034202	0.0017835	0.0009690	0.0492837	0.0274303	0.0140225
011110	0.0037613	0.0017432	0.0010059	0.0510936	0.0258409	0.0137283
111001	0.0035985	0.0011557	0.0010731	0.0363952	0.0071175	0.0108954
110101	0.0034513	0.0012450	0.0010470	0.0388047	0.0052617	0.0118664
101101	0.0029483	0.0013181	0.0010414	0.0304425	0.0073526	0.0107920
011101	0.0032894	0.0013047	0.0010757	0.0325752	0.0077179	0.0106895
110011	0.0053689	0.0014672	0.0012045	0.0578282	0.0175210	0.0133969
101011	0.0048660	0.0017272	0.0011996	0.0490834	0.0180953	0.0123083
011011	0.0052071	0.0015849	0.0012296	0.0503782	0.0175511	0.0121074
100111	0.0047188	0.0018226	0.0011763	0.0521848	0.0182389	0.0133921
010111	0.0050599	0.0016983	0.0012069	0.0534866	0.0176179	0.0131346
001111	0.0045569	0.0019002	0.0012020	0.0451432	0.0180938	0.0120837
111110	0.0034656	0.0013823	0.0008541	0.0500299	0.0215340	0.0123833
111101	0.0030880	0.0010305	0.0009069	0.0338376	0.0067249	0.0099727
111011	0.0046222	0.0013599	0.0010244	0.0497941	0.0152638	0.0111869
110111	0.0045044	0.0014280	0.0010070	0.0526065	0.0151636	0.0120184
101111	0.0041020	0.0015406	0.0010033	0.0451213	0.0157672	0.0111597
011111	0.0043749	0.0014831	0.0010262	0.0463947	0.0153463	0.0110118
111111	0.0040262	0.0012602	0.0008878	0.0462223	0.0136915	0.0102895

In the output, the table on the top part lists unbiased estimate of heterozygosities for all loci in the two datsets along with the standard errors of the estimates. For RFLP loci, these are at the level of binned alleles (31 possible fixed bins - see note 1 below). Although there is no formal test for checking if the heterozygosities in the two datasets are statistically different, approximate equality of them may be judged by visually examining to see whether the heterozygosity differences are smaller than their summed standard errors. The last three columns of the top table show the results of the 2 X C contingency table analysis of allele frequency differences between the two populations. The chi-square statistic has (k – 1) degrees of freedom, where k is the number of alleles at the locus for which at least one of the two populations has a non-zero frequency. The P-value of the chi-square statistic is based on shuffling of alleles across populations (keeping the sample sizes the same as in the observed data sets).

The larger table in the bottom gives the detailed sets of computations of bias-corrected estimates of Nei's minimum (Dm) and standard (Ds) distances along with their standard errors (inter-locus and intra-locus). The first set of entries in each row (consisting of 1's and 0's) show which loci are considered for computations (1 indicating that the locus is used, and 0 not). For example, and entry with 1 in the first column and 0's elsewhere means that for this row computations of genetic distances are done with allele frequency data on the first locus alone. When a single-locus computation is done, there is no inter-locus variance of the estimates of genetic distances, and hence the corresponding standard errors (the ones under the SE. INTER - columns) are zero.

Overall, an user should first look at the last line of this table to examine whether the two populations have any siginificant genetic distance between them with data on the loci. Thus, for the row with 1 for all columns if Dm (or Ds) is larger than 2-times the corresponding INTER-locus SE, the two populations would be judged being significantly dissimilar.

If so, the user may be interested in knowing which locus (or which sets of loci) causes this significant genetic distance. This can be done first by examining the results of the 2 X C

contingency table P-values and by checking all rows of the distance calculations. Generally, even when the two populations are genetically not dissimilar, the contigency table analysis may show significant differences of allele frequencies. This is likely to occur particularly for loci with larger number of alleles, since rare alleles (in one or both populations) contribute to significant P-value (i.e., P smaller than the nominal levels for significance, 1% or 5%). This can be checked by examining the significance of Dm or Ds for the same locus in the table shown in the bottom part of the output.

NOTES for Program N:

1. As in the other routines, the loci included in this analysis may be RFLP and/or PCR loci, or combinations thereof. For RFLP loci, alleles are binned (fixed bin categories are used in the program), so that there are 31 binned alleles for each RFLP locus. Bins with no allele in both databases are excluded from analysis and the number of alleles are accordingly reduced for analysis.

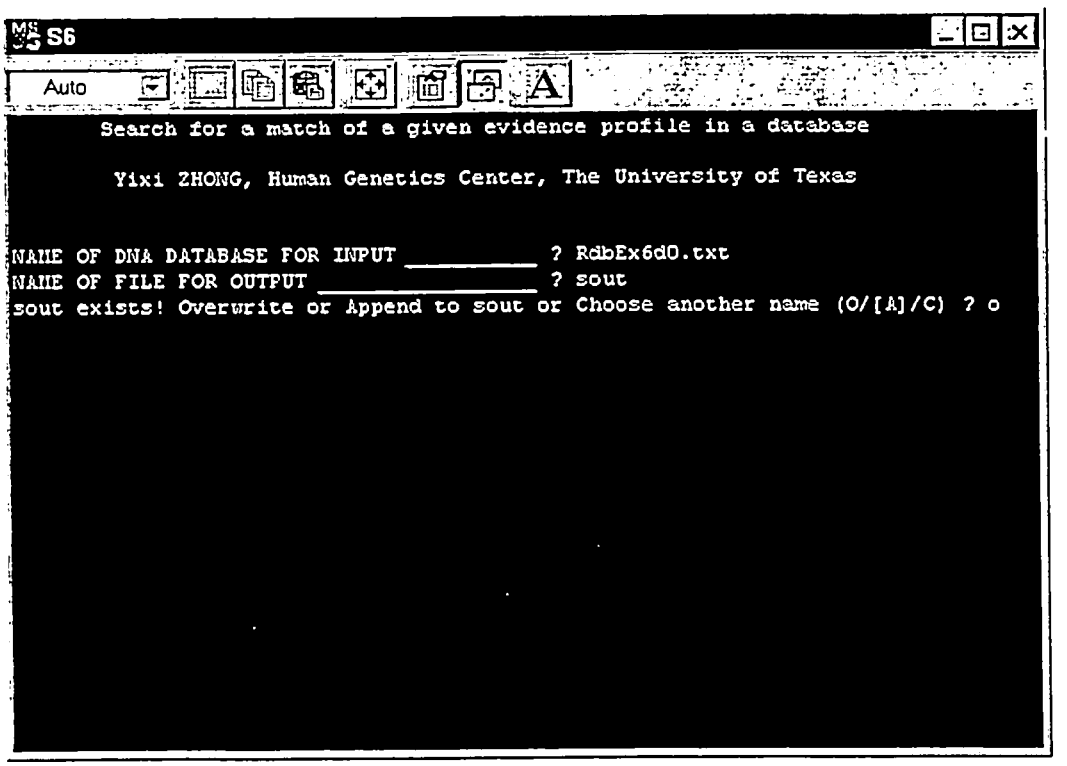
2. Significant P-values in the 2 X C contingency table analysis (top table, last column, obtained by allele shuffling across the two populations) that result from mainly the dissimilarities of rare allele frequencies in the populations would be indicated by small Dm and Ds values (judged with respect to the INTRA-locus SE of the corresponding distance estimate) for the same locus.

3. If genetic distances are not significant between populations, the two datasets may be pooled should there be need of bigger population samples for the population of interest.

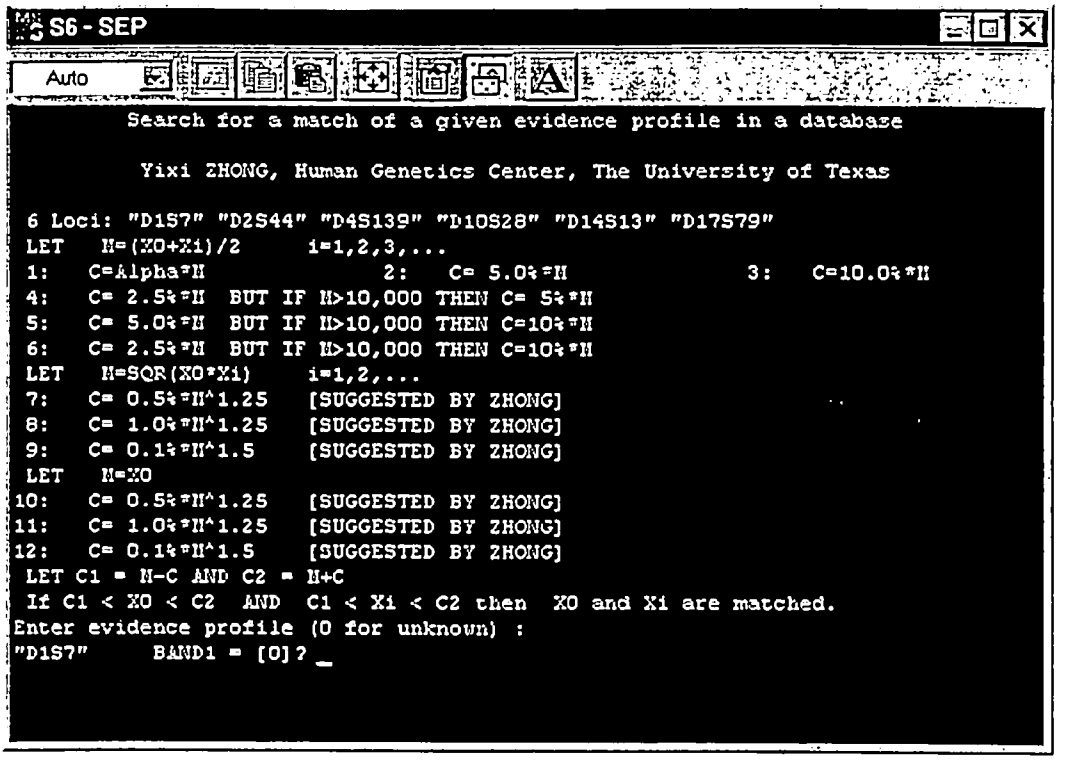
Program S Search Database for a Complete Match to User Specified Profile

Program S is a utility routine that searches for a match in the database with a specific user-specified multiple locus profile. This allows the investigator to examine how often any given multiple locus profile is found (if at all) in a database.

The program prompts for an input database file (RdbEx6d0.txt was entered), and an output filename (sout was entered). If an output filename that already exists is given, the user is prompted to enter O for Overwrite, A for Append or C for Choose another name (O was entered). See the screen below.

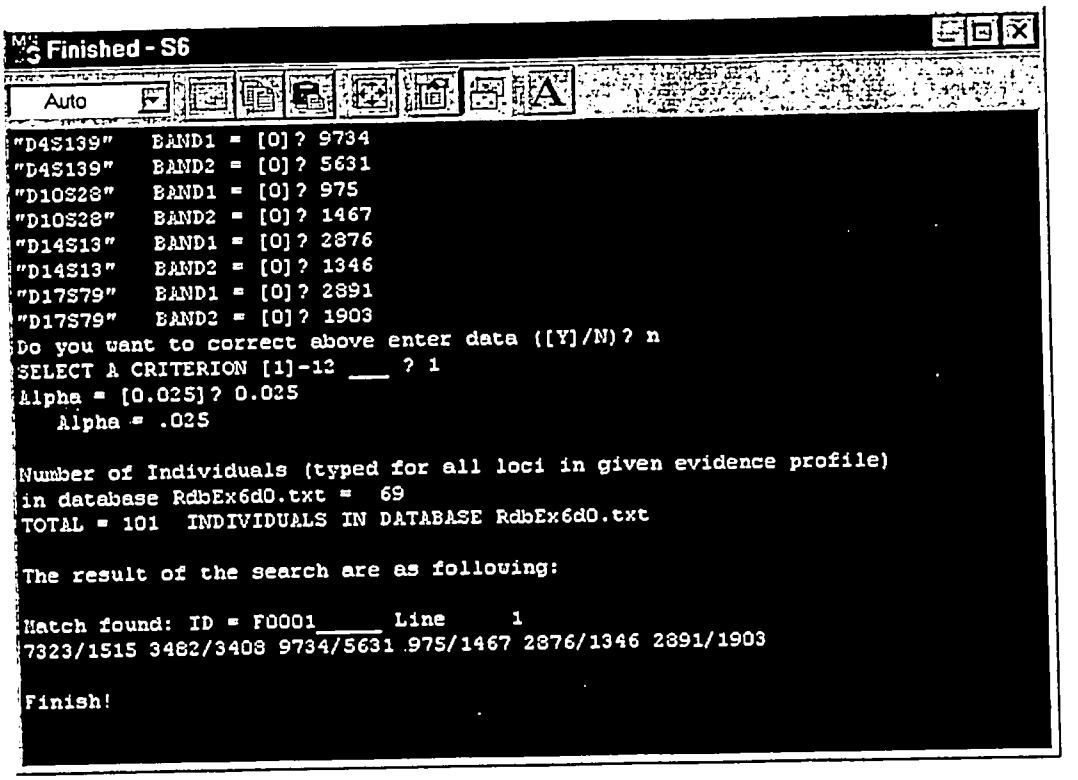


Once the O, A or C has been entered, another screen will appear listing possible match criteria if the specified database file contains any RFLP locus data. The user has control over the criteria the program uses to determine a "match" with RFLP data.



Of the 12 choices listed, a typical criterion for matching might be number 1 with alpha = 0.025 (2.5 %). The program next prompts for band1 and band2 bandsizes for each locus in sequence. The six locus profile 7323 / 1515 for D1S7, 3482 / 3408 for D2S44, 9734 / 5631 for D4S139, 975 / 1467 for D10s28, 2876 / 1346 for D14S13 and 2891 / 1903 for D17S79 was entered.

On the screen illustrated, only the first of these twelve prompts is visible. Zero is entered if the bandsize is unknown (hitting <Return> will result in entering zero). In this illustration, the user entered the bandsizes for a profile that did exist in the database. Criterion 1 was selected to use in searching for matches, and alpha was set at 0.025.
The program reported back the single match it found -- see the next screen image.



Most of the user-entered bandsizes can be seen at the top of this screen. The program provides an opportunity to correct entered data as well -- choosing Yes will take the user back through the bandsize entry dialog. Next the user must select a criterion, specifying it by a number (from the screen menu as noted above), and specifying an alpha value if a criterion involving alpha was selected.

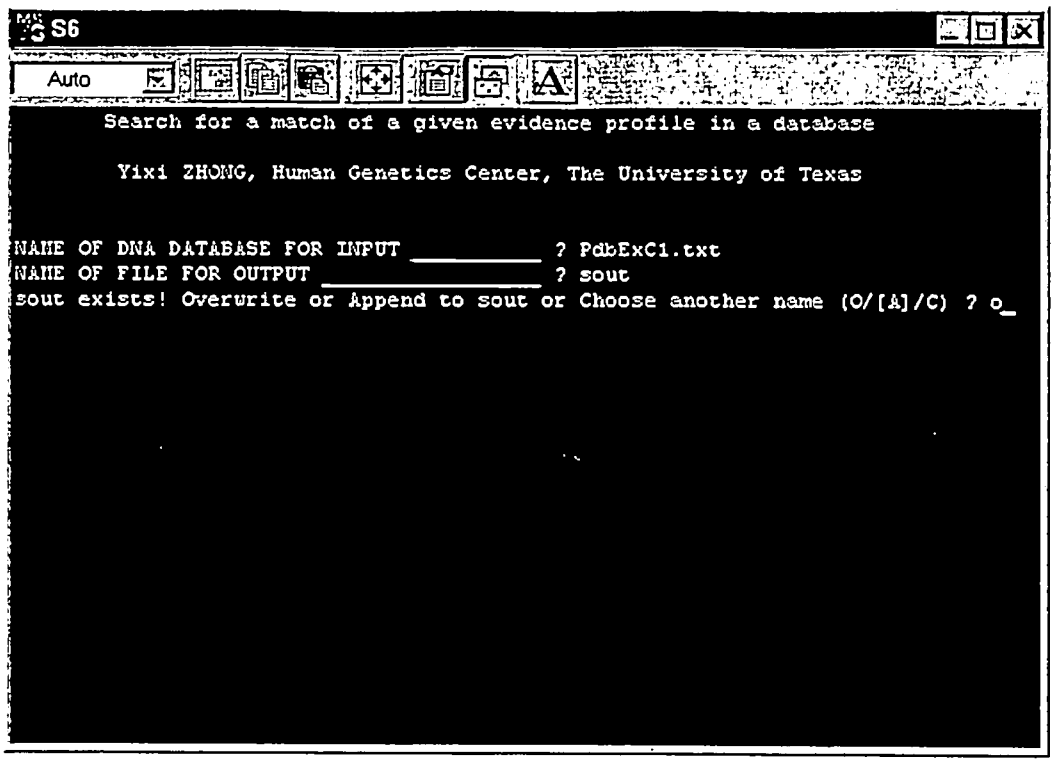
The program reports the number of individuals in the database file who have been typed for all the loci for which values were entered, i.e., in this case that number consists of the people typed for all six loci -- no one with a 0 0 entry at any locus is counted. With criterion 1 and alpha = 0.025, the program reports that the entered data match the first database record. And this is the data that was in fact entered.

Much of the input screen information is repeated in the output file, along with a record of the inputed profile and the ID # and line of the database record that matched. The output file produced by program S with RdbEx6d0.txt, and the entered profile described, is shown below.

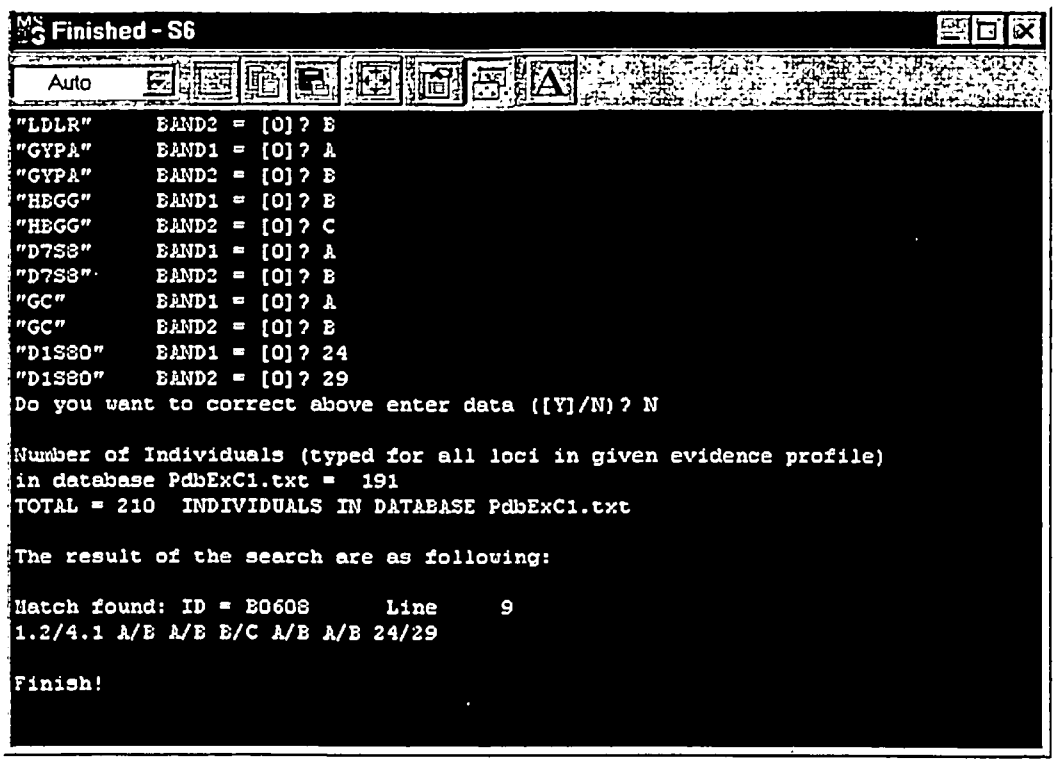
```
Search for a match of a given evidence profile in a database

          Yixi ZHONG, Human Genetics Center, The University of Texas

NAME OF DNA DATABASE FOR INPUT ____ RdbEx6d0.txt

 LET    M=(X0+Xi)/2      i=1,2,3,...
 1:    C=Alpha*M                  2:    C= 5.0%*M                  3:    C=10.0%*M
 4:    C= 2.5%*M   BUT IF M>10,000 THEN C= 5%*M
 5:    C= 5.0%*M   BUT IF M>10,000 THEN C=10%*M
```

```
 6:    C= 2.5%*M  BUT IF M>10,000 THEN C=10%*M
 LET    M=SQR(X0*Xi)     i=1,2,...
 7:    C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
 8:    C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
 9:    C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET    M=X0
10:    C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
11:    C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
12:    C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET C1 = M-C AND C2 = M+C
 If C1 < X0 < C2  AND  C1 < Xi < C2 then X0 and Xi are matched.

The given evidence profile:
 6 loci: "D1S7" "D2S44" "D4S139" "D10S28" "D14S13" "D17S79"
7323/1515 3482/3408 9734/5631 975/1467 2876/1346 2891/1903

SELECT A CRITERION 1-12 ___  1
   Alpha = .025

Number of Individuals (typed for all loci in given evidence profile)
in database RdbEx6d0.txt =  69
TOTAL = 101  INDIVIDUALS IN DATABASE RdbEx6d0.txt

The result of the search are as following:

Match found: ID = F0001_____ Line     1
7323/1515 3482/3408 9734/5631 975/1467 2876/1346 2891/1903

06-25-1998    17:32:16
```

Running program S with database files containing PCR-based locus data is very similar. If the database file contains no RFLP loci, the program will not bring up the "match" criteria. The rest of the input dialog is very similar -- see the screen below.



The database file PdbExC1.txt was entered, along with an output filename (sout), and the program was told to "overwrite" (o). This database file has 210 individuals typed for HLA-DQA1, all five Polymarker loci and D1S80. The profile HLA-DQA1 1.2/4.1, LDLR AB, GYPA AB, HBGG BC, D7S8 AB, GC AB, and D1S80 24/29 was entered, one allele at a time as prompted by the program. This profile does exist in the database. Most of that entry dialog is visible on the screen shown below.



N (for No) was entered at the "Do you want to correct the above data?" prompt. The program then reports that 191 individuals in the database had been typed for all the loci (out of a total of 210), and that the entered profile matched ID # BO608 at line 9 -- this is the profile that was entered.

The output file produced by program S with PdbExC1.txt, and the entered profile described, is shown below.

```
Search for a match of a given evidence profile in a database

          Yixi ZHONG, Human Genetics Center, The University of Texas

NAME OF DNA DATABASE FOR INPUT ____ PdbExC1.txt


The given evidence profile:
 7 loci: "HLA-DQA1" "LDLR" "GYPA" "HBGG" "D7S8" "GC" "D1S80"
1.2/4.1 A/B A/B B/C A/B A/B 24/29


Number of Individuals (typed for all loci in given evidence profile)
in database PdbExC1.txt =  191
TOTAL = 210  INDIVIDUALS IN DATABASE PdbExC1.txt

The result of the search are as following:

Match found: ID = B0608      Line     9
1.2/4.1 A/B A/B B/C A/B A/B 24/29

06-25-1998    18:10:00
```

Program S finds only those profiles in the specified database file that exactly match the user-entered profile at all loci. To search for partial matches (matches at some loci but not others) in a database file, program T is used.

Program T Search Database for All Partial Matches to User Specified Profile

Program T serves the same purpose as program S, but the analysis is more detailed, since it also lists partial matches in the database with any given profile, by considering each locus individually, in pairs, in triplets, etc.

The input / dialog screens for program T are identical to those for program S, but the output file is different.

An output file is shown below from program T with RdbEx6d0.txt, and inputing a random set of bandsizes for all the loci except D1S7, where the bandsizes from one of the existing records were entered. The data that were entered are reproduced in the output file.

The 1 0 0 0 0 0 designator means that locus 1 matched but that none of the remaining five matched. That number is 1 out of 96 in this example. An exact two-band match at D1S7 was intentionally entered in the profile (and 96 of the 101 total individuals were typed for D1S7). If the first two loci matched but none of the others, there would be a nonzero entry for the 1 1 0 0 0 0 deignator, and so forth. The program reports the proportion of matches found for that locus or loci. In this illustration, there was one match at D2S44 (0 1 0 0 0 0 shows 1/93) and three matches at D17S79 (0 0 0 0 0 1 shows 3/98). These "matches" were found on the basis of the user-selected match criterion (number 1) and a +/- 2.5% window. No two- or multiple-locus matches were detected.

```
                PROPORTIONS MATCHED OF ALL LOCUS COMBINATIONS

          Yixi ZHONG, Human Genetics Center, The University of Texas

NAME OF DNA DATABASE FOR INPUT ____ RdbEx6d0.txt

 LET   M=(X0+Xi)/2      i=1,2,...
 1:   C=Alpha*M                2:   C= 5.0%*M                3:   C=10.0%*M
 4:   C= 2.5%*M  BUT IF M>10,000 THEN C= 5%*M
 5:   C= 5.0%*M  BUT IF M>10,000 THEN C=10%*M
 6:   C= 2.5%*M  BUT IF M>10,000 THEN C=10%*M
 LET   M=SQR(X0*Xi)     i=1,2,...
 7:   C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
 8:   C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
 9:   C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET   M=X0
10:   C= 0.5%*M^1.25    [SUGGESTED BY ZHONG]
11:   C= 1.0%*M^1.25    [SUGGESTED BY ZHONG]
12:   C= 0.1%*M^1.5     [SUGGESTED BY ZHONG]
 LET C1 = M-C AND C2 = M+C
 WHEN C1 < X0 < C2  AND  C1 < Xi < C2 THEN  WE GET A MATCH.

Evidence profile:
 6 loci: "D1S7" "D2S44" "D4S139" "D10S28" "D14S13" "D17S79"
 7323/1515 2220/1500 2970/19410 0/0 3540/1580 1990/1300

SELECT A CRITERION 1-12 ___  1
  Alpha = .025
TOTAL = 101  INDIVIDUALS IN DATABASE RdbEx6d0.txt
```

L O C U S		PROPORTION	PRODUCT
1 0 0 0 0 0 =	1/	96 = 0.0104167	0.01041667
0 1 0 0 0 0 =	1/	93 = 0.0107527	0.01075269
0 0 1 0 0 0 =	0/	97 = 0.0000000	0.00000000
0 0 0 1 0 0 =	0/	0 = 0.0000000	0.00000000
0 0 0 0 1 0 =	0/	91 = 0.0000000	0.00000000
0 0 0 0 0 1 =	3/	98 = 0.0306122	0.03061225
1 1 0 0 0 0 =	0/	89 = 0.0000000	0.00011201
1 0 1 0 0 0 =	0/	93 = 0.0000000	0.00000000
0 1 1 0 0 0 =	0/	90 = 0.0000000	0.00000000
1 0 0 1 0 0 =	0/	0 = 0.0000000	0.00000000
0 1 0 1 0 0 =	0/	0 = 0.0000000	0.00000000
0 0 1 1 0 0 =	0/	0 = 0.0000000	0.00000000
1 0 0 0 1 0 =	0/	86 = 0.0000000	0.00000000
0 1 0 0 1 0 =	0/	83 = 0.0000000	0.00000000
0 0 1 0 1 0 =	0/	87 = 0.0000000	0.00000000
0 0 0 1 1 0 =	0/	0 = 0.0000000	0.00000000
1 0 0 0 0 1 =	0/	94 = 0.0000000	0.00031888
0 1 0 0 0 1 =	0/	92 = 0.0000000	0.00032916
0 0 1 0 0 1 =	0/	95 = 0.0000000	0.00000000
0 0 0 1 0 1 =	0/	0 = 0.0000000	0.00000000
0 0 0 0 1 1 =	0/	88 = 0.0000000	0.00000000
1 1 1 0 0 0 =	0/	86 = 0.0000000	0.00000000
1 1 0 1 0 0 =	0/	0 = 0.0000000	0.00000000
1 0 1 1 0 0 =	0/	0 = 0.0000000	0.00000000
0 1 1 1 0 0 =	0/	0 = 0.0000000	0.00000000
1 1 0 0 1 0 =	0/	79 = 0.0000000	0.00000000
1 0 1 0 1 0 =	0/	83 = 0.0000000	0.00000000
0 1 1 0 1 0 =	0/	80 = 0.0000000	0.00000000
1 0 0 1 1 0 =	0/	0 = 0.0000000	0.00000000
0 1 0 1 1 0 =	0/	0 = 0.0000000	0.00000000
0 0 1 1 1 0 =	0/	0 = 0.0000000	0.00000000
1 1 0 0 0 1 =	0/	88 = 0.0000000	0.00000343
1 0 1 0 0 1 =	0/	91 = 0.0000000	0.00000000
0 1 1 0 0 1 =	0/	89 = 0.0000000	0.00000000
1 0 0 1 0 1 =	0/	0 = 0.0000000	0.00000000
0 1 0 1 0 1 =	0/	0 = 0.0000000	0.00000000
0 0 1 1 0 1 =	0/	0 = 0.0000000	0.00000000
1 0 0 0 1 1 =	0/	84 = 0.0000000	0.00000000
0 1 0 0 1 1 =	0/	82 = 0.0000000	0.00000000
0 0 1 0 1 1 =	0/	85 = 0.0000000	0.00000000
0 0 0 1 1 1 =	0/	0 = 0.0000000	0.00000000
1 1 1 1 0 0 =	0/	0 = 0.0000000	0.00000000
1 1 1 0 1 0 =	0/	76 = 0.0000000	0.00000000
1 1 0 1 1 0 =	0/	0 = 0.0000000	0.00000000
1 0 1 1 1 0 =	0/	0 = 0.0000000	0.00000000
0 1 1 1 1 0 =	0/	0 = 0.0000000	0.00000000
1 1 1 0 0 1 =	0/	85 = 0.0000000	0.00000000
1 1 0 1 0 1 =	0/	0 = 0.0000000	0.00000000
1 0 1 1 0 1 =	0/	0 = 0.0000000	0.00000000
0 1 1 1 0 1 =	0/	0 = 0.0000000	0.00000000
1 1 0 0 1 1 =	0/	78 = 0.0000000	0.00000000
1 0 1 0 1 1 =	0/	81 = 0.0000000	0.00000000
0 1 1 0 1 1 =	0/	79 = 0.0000000	0.00000000

```
1 0 0 1 1 1 =     0/     0 = 0.0000000    0.00000000
0 1 0 1 1 1 =     0/     0 = 0.0000000    0.00000000
0 0 1 1 1 1 =     0/     0 = 0.0000000    0.00000000
1 1 1 1 1 0 =     0/     0 = 0.0000000    0.00000000
1 1 1 1 0 1 =     0/     0 = 0.0000000    0.00000000
1 1 1 0 1 1 =     0/    75 = 0.0000000    0.00000000
1 1 0 1 1 1 =     0/     0 = 0.0000000    0.00000000
1 0 1 1 1 1 =     0/     0 = 0.0000000    0.00000000
0 1 1 1 1 1 =     0/     0 = 0.0000000    0.00000000
1 1 1 1 1 1 =     0/     0 = 0.0000000    0.00000000
```

Program T runs similarly with database files containing PCR-based locus data. The output file from program T with PdbExC1.txt and the profile HLA-DQA1 1.2/1.3, LDLR AA, GYPA AB, HBGG BC, D7S8 BB, GC CC, and D1S80 24/24 is shown below. The 1 0 0 0 0 0 0 designator (profiles in the database that have HLA-DQA1 types of 1.2/1.3 but no other matches to the entered profile) shows 10 out of 206 (206 of the 210 people are typed for HLA-DQA1). The 0 0 1 1 0 0 0 = 18 / 206 = 0.0873786 entry indicates that 18 profiles in the database matched the entered types for both loci 3 and 4 (GYPA and HBGG) and that 206 people (out of the 210) were typed for both these loci; and so forth.

```
                PROPORTIONS MATCHED OF ALL LOCUS COMBINATIONS

          Yixi ZHONG, Human Genetics Center, The University of Texas

NAME OF DNA DATABASE FOR INPUT ____ PdbExC1.txt


Evidence profile:
 7 loci: "HLA-DQA1" "LDLR" "GYPA" "HBGG" "D7S8" "GC" "D1S80"
 1.2/1.3 A/A A/B B/C B/B C/C 24/24

TOTAL = 210  INDIVIDUALS IN DATABASE PdbExC1.txt

L O C U S                          PROPORTION      PRODUCT

1 0 0 0 0 0 0 =   10/  206 = 0.0485437   0.04854369
0 1 0 0 0 0 0 =    6/  206 = 0.0291262   0.02912621
0 0 1 0 0 0 0 =  122/  206 = 0.5922330   0.59223300
0 0 0 1 0 0 0 =   33/  206 = 0.1601942   0.16019417
0 0 0 0 1 0 0 =   25/  206 = 0.1213592   0.12135922
0 0 0 0 0 1 0 =    7/  206 = 0.0339806   0.03398058
0 0 0 0 0 0 1 =    9/  193 = 0.0466321   0.04663213
1 1 0 0 0 0 0 =    0/  206 = 0.0000000   0.00141389
1 0 1 0 0 0 0 =    6/  206 = 0.0291262   0.02874917
0 1 1 0 0 0 0 =    2/  206 = 0.0097087   0.01724951
1 0 0 1 0 0 0 =    1/  206 = 0.0048544   0.00777642
0 1 0 1 0 0 0 =    1/  206 = 0.0048544   0.00466585
0 0 1 1 0 0 0 =   18/  206 = 0.0873786   0.09487227
1 0 0 0 1 0 0 =    0/  206 = 0.0000000   0.00589122
0 1 0 0 1 0 0 =    1/  206 = 0.0048544   0.00353473
0 0 1 0 1 0 0 =   15/  206 = 0.0728155   0.07187293
0 0 0 1 1 0 0 =    4/  206 = 0.0194175   0.01944104
1 0 0 0 0 1 0 =    1/  206 = 0.0048544   0.00164954
```

0	1	0	0	0	1	0	=	0/	206	= 0.0000000	0.00098973
0	0	1	0	0	1	0	=	2/	206	= 0.0097087	0.02012442
0	0	0	1	0	1	0	=	0/	206	= 0.0000000	0.00544349
0	0	0	0	1	1	0	=	1/	206	= 0.0048544	0.00412386
1	0	0	0	0	0	1	=	1/	191	= 0.0052356	0.00226370
0	1	0	0	0	0	1	=	0/	191	= 0.0000000	0.00135822
0	0	1	0	0	0	1	=	2/	191	= 0.0104712	0.02761708
0	0	0	1	0	0	1	=	3/	191	= 0.0157068	0.00747019
0	0	0	0	1	0	1	=	2/	191	= 0.0104712	0.00565924
0	0	0	0	0	1	1	=	3/	191	= 0.0157068	0.00158459
1	1	1	0	0	0	0	=	0/	206	= 0.0000000	0.00083735
1	1	0	1	0	0	0	=	0/	206	= 0.0000000	0.00022650
1	0	1	1	0	0	0	=	0/	206	= 0.0000000	0.00460545
0	1	1	1	0	0	0	=	1/	206	= 0.0048544	0.00276327
1	1	0	0	1	0	0	=	0/	206	= 0.0000000	0.00017159
1	0	1	0	1	0	0	=	0/	206	= 0.0000000	0.00348898
0	1	1	0	1	0	0	=	0/	206	= 0.0000000	0.00209339
1	0	0	1	1	0	0	=	0/	206	= 0.0000000	0.00094374
0	1	0	1	1	0	0	=	0/	206	= 0.0000000	0.00056624
0	0	1	1	1	0	0	=	3/	206	= 0.0145631	0.01151363
1	1	0	0	0	1	0	=	0/	206	= 0.0000000	0.00004804
1	0	1	0	0	1	0	=	1/	206	= 0.0048544	0.00097691
0	1	1	0	0	1	0	=	0/	206	= 0.0000000	0.00058615
1	0	0	1	0	1	0	=	0/	206	= 0.0000000	0.00026425
0	1	0	1	0	1	0	=	0/	206	= 0.0000000	0.00015855
0	0	1	1	0	1	0	=	0/	206	= 0.0000000	0.00322382
1	0	0	0	1	1	0	=	0/	206	= 0.0000000	0.00020019
0	1	0	0	1	1	0	=	0/	206	= 0.0000000	0.00012011
0	0	1	0	1	1	0	=	0/	206	= 0.0000000	0.00244228
0	0	0	1	1	1	0	=	0/	206	= 0.0000000	0.00066062
1	1	0	0	0	0	1	=	0/	191	= 0.0000000	0.00006593
1	0	1	0	0	0	1	=	0/	191	= 0.0000000	0.00134064
0	1	1	0	0	0	1	=	0/	191	= 0.0000000	0.00080438
1	0	0	1	0	0	1	=	0/	191	= 0.0000000	0.00036263
0	1	0	1	0	0	1	=	0/	191	= 0.0000000	0.00021758
0	0	1	1	0	0	1	=	0/	191	= 0.0000000	0.00442410
1	0	0	0	1	0	1	=	0/	191	= 0.0000000	0.00027472
0	1	0	0	1	0	1	=	0/	191	= 0.0000000	0.00016483
0	0	1	0	1	0	1	=	1/	191	= 0.0052356	0.00335159
0	0	0	1	1	0	1	=	1/	191	= 0.0052356	0.00090658
1	0	0	0	0	1	1	=	0/	191	= 0.0000000	0.00007692
0	1	0	0	0	1	1	=	0/	191	= 0.0000000	0.00004615
0	0	1	0	0	1	1	=	0/	191	= 0.0000000	0.00093844
0	0	0	1	0	1	1	=	0/	191	= 0.0000000	0.00025384
0	0	0	0	1	1	1	=	0/	191	= 0.0000000	0.00019230
1	1	1	1	0	0	0	=	0/	206	= 0.0000000	0.00013414
1	1	1	0	1	0	0	=	0/	206	= 0.0000000	0.00010162
1	1	0	1	1	0	0	=	0/	206	= 0.0000000	0.00002749
1	0	1	1	1	0	0	=	0/	206	= 0.0000000	0.00055891
0	1	1	1	1	0	0	=	0/	206	= 0.0000000	0.00033535
1	1	1	0	0	1	0	=	0/	206	= 0.0000000	0.00002845
1	1	0	1	0	1	0	=	0/	206	= 0.0000000	0.00000770
1	0	1	1	0	1	0	=	0/	206	= 0.0000000	0.00015650
0	1	1	1	0	1	0	=	0/	206	= 0.0000000	0.00009390
1	1	0	0	1	1	0	=	0/	206	= 0.0000000	0.00000583

1	0	1	0	1	1	0	=	0/	206	=	0.0000000	0.00011856
0	1	1	0	1	1	0	=	0/	206	=	0.0000000	0.00007113
1	0	0	1	1	1	0	=	0/	206	=	0.0000000	0.00003207
0	1	0	1	1	1	0	=	0/	206	=	0.0000000	0.00001924
0	0	1	1	1	1	0	=	0/	206	=	0.0000000	0.00039124
1	1	1	0	0	0	1	=	0/	191	=	0.0000000	0.00003905
1	1	0	1	0	0	1	=	0/	191	=	0.0000000	0.00001056
1	0	1	1	0	0	1	=	0/	191	=	0.0000000	0.00021476
0	1	1	1	0	0	1	=	0/	191	=	0.0000000	0.00012886
1	1	0	0	1	0	1	=	0/	191	=	0.0000000	0.00000800
1	0	1	0	1	0	1	=	0/	191	=	0.0000000	0.00016270
0	1	1	0	1	0	1	=	0/	191	=	0.0000000	0.00009762
1	0	0	1	1	0	1	=	0/	191	=	0.0000000	0.00004401
0	1	0	1	1	0	1	=	0/	191	=	0.0000000	0.00002641
0	0	1	1	1	0	1	=	0/	191	=	0.0000000	0.00053690
1	1	0	0	0	1	1	=	0/	191	=	0.0000000	0.00000224
1	0	1	0	0	1	1	=	0/	191	=	0.0000000	0.00004556
0	1	1	0	0	1	1	=	0/	191	=	0.0000000	0.00002733
1	0	0	1	0	1	1	=	0/	191	=	0.0000000	0.00001232
0	1	0	1	0	1	1	=	0/	191	=	0.0000000	0.00000739
0	0	1	1	0	1	1	=	0/	191	=	0.0000000	0.00015033
1	0	0	0	1	1	1	=	0/	191	=	0.0000000	0.00000934
0	1	0	0	1	1	1	=	0/	191	=	0.0000000	0.00000560
0	0	1	0	1	1	1	=	0/	191	=	0.0000000	0.00011389
0	0	0	1	1	1	1	=	0/	191	=	0.0000000	0.00003081
1	1	1	1	1	0	0	=	0/	206	=	0.0000000	0.00001628
1	1	1	1	0	1	0	=	0/	206	=	0.0000000	0.00000456
1	1	1	0	1	1	0	=	0/	206	=	0.0000000	0.00000345
1	1	0	1	1	1	0	=	0/	206	=	0.0000000	0.00000093
1	0	1	1	1	1	0	=	0/	206	=	0.0000000	0.00001899
0	1	1	1	1	1	0	=	0/	206	=	0.0000000	0.00001140
1	1	1	1	0	0	1	=	0/	191	=	0.0000000	0.00000626
1	1	1	0	1	0	1	=	0/	191	=	0.0000000	0.00000474
1	1	0	1	1	0	1	=	0/	191	=	0.0000000	0.00000128
1	0	1	1	1	0	1	=	0/	191	=	0.0000000	0.00002606
0	1	1	1	1	0	1	=	0/	191	=	0.0000000	0.00001564
1	1	1	0	0	1	1	=	0/	191	=	0.0000000	0.00000133
1	1	0	1	0	1	1	=	0/	191	=	0.0000000	0.00000036
1	0	1	1	0	1	1	=	0/	191	=	0.0000000	0.00000730
0	1	1	1	0	1	1	=	0/	191	=	0.0000000	0.00000438
1	1	0	0	1	1	1	=	0/	191	=	0.0000000	0.00000027
1	0	1	0	1	1	1	=	0/	191	=	0.0000000	0.00000553
0	1	1	0	1	1	1	=	0/	191	=	0.0000000	0.00000332
1	0	0	1	1	1	1	=	0/	191	=	0.0000000	0.00000150
0	1	0	1	1	1	1	=	0/	191	=	0.0000000	0.00000090
0	0	1	1	1	1	1	=	0/	191	=	0.0000000	0.00001824
1	1	1	1	1	1	0	=	0/	206	=	0.0000000	0.00000055
1	1	1	1	1	0	1	=	0/	191	=	0.0000000	0.00000076
1	1	1	1	0	1	1	=	0/	191	=	0.0000000	0.00000021
1	1	1	0	1	1	1	=	0/	191	=	0.0000000	0.00000016
1	1	0	1	1	1	1	=	0/	191	=	0.0000000	0.00000004
1	0	1	1	1	1	1	=	0/	191	=	0.0000000	0.00000089
0	1	1	1	1	1	1	=	0/	191	=	0.0000000	0.00000053
1	1	1	1	1	1	1	=	0/	191	=	0.0000000	0.00000003

6 NRC Program

The "NRC" program calculates the probabilities and reciprocal probabilities of chance duplicates (estimated frequencies) for a user-specified DNA profile, in accordance with the NRC II report's Recommendation 4.1 and Equation 4.10. These are:

Recommendation 4.1. In general, calculation of a profile frequency should be made with the product rule. If the race of the person who left the evidence-sample DNA is known, the database for the person's race should be used; if the race is not known, calculations for all racial groups to which possible suspects belong should be made. For systems such as VNTRs, in which a heterozygous locus can be mistaken for a homozygous one, if an upper bound on the frequency of the genotype at an apparently homozygous locus (single band) is desired, then twice the allele (bin) frequency, 2p, should be used instead of p^2. For systems in which exact genotypes can be determined, $p^2 + p(1-p)\theta$ should be used for the frequency at such a locus instead of p^2. A conservative value of θ for the U.S. population is 0.01; for some small, isolated populations, a value of 0.03 may be more appropriate. For both kinds of systems, $2p_ip_j$ should be used for heterozygotes.

Recommendation 4.10.

$$\text{Homozygote: } P(A_iA_i|A_iA_i) = \frac{\left[2\theta + (1-\theta)p_i\right]\left[3\theta + (1-\theta)p_i\right]}{(1+\theta)(1+2\theta)}, \qquad (4.10a)$$

$$\text{Heterozygote: } P(A_iA_j|A_iA_j) = \frac{2\left[\theta + (1-\theta)p_i\right]\left[\theta + (1-\theta)p_j\right]}{(1+\theta)(1+2\theta)}. \qquad (4.10b)$$

Further information about these recommendations and equations may be found in the NRC II report itself (National Research Council 1996).

The program is designed to do computations for a user-specified DNA profile. The profile can consist of any combination of loci, i.e., RFLP, PCR dot blot, D1S80 and/or STR. There are two methods by which the program obtains the allele (or band) frequencies that it use in the calculations: (1) users can input the allele frequencies and population size from the keyboard; or (2) the program will query a user-specified database and compute the frequencies of the specified alleles or bands and the population size.

Keyboard Entry

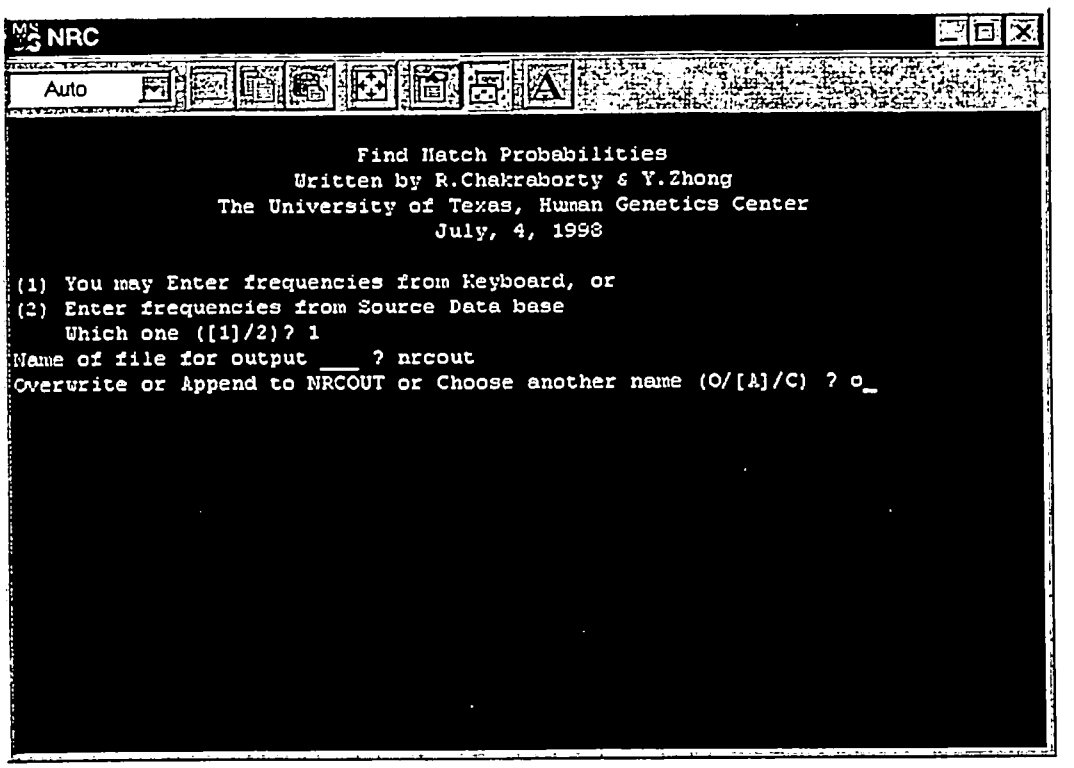
With keyboard entry, the allele frequencies and population size (number of people typed at a locus on which the frequencies are based) are obtained independently in advance of running NRC (say, for example, using literature values, or by running program H). The program will query as to whether each locus is RFLP or PCR.

When the program executes, it will specify that a user can (1) enter frequencies from the keyboard; of (2) enter frequencies from a source database; and it will ask

Which one ([1]/2)?

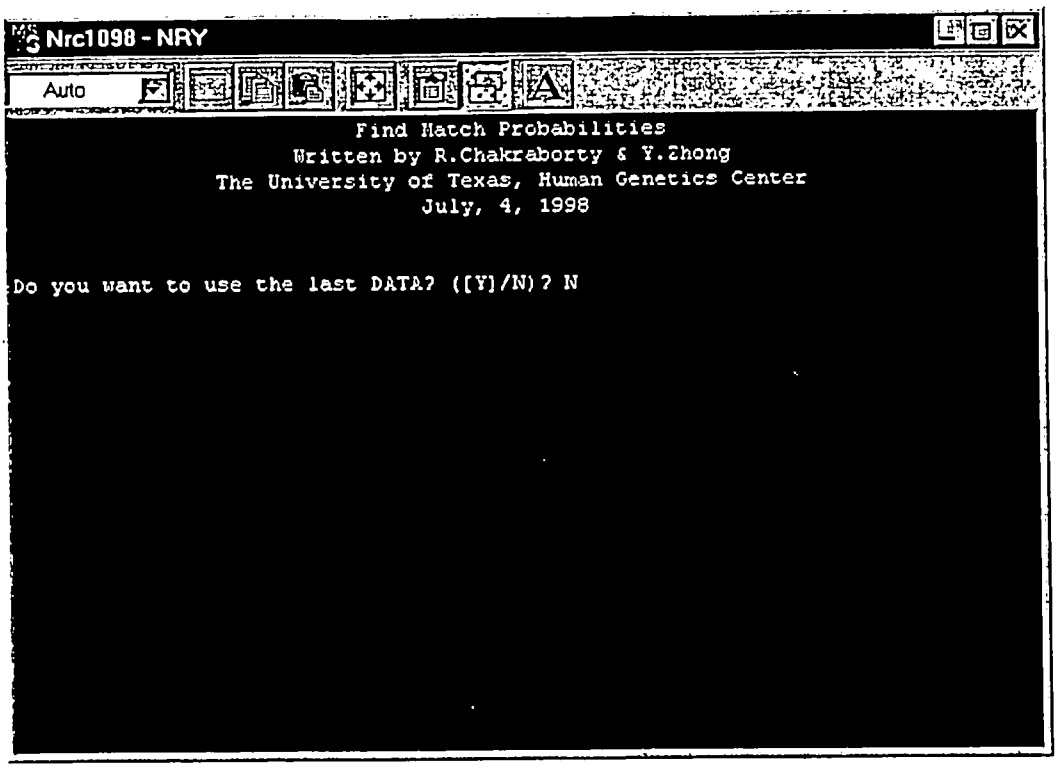
Keyboard entry is the default, and hitting <Enter> at this point will select option 1. A user can also type "1" (no quotes) and get the same results. With either option, a name for the output file is requested - if a user chooses an output file that already exists, the program prompts for (O/[A]/C), meaning overwrite, append or change, with "append" as the default.

An example opening screen is shown below:



Here, keyboard input (choice 1) was entered (the user could also have just hit <Enter> at this point), "nrcout" (no quotes) was entered as the output file name, and since it already existed, the (O/[A]/C) choices appeared. Overwrite (o) was selected.

Hitting <Enter> after selecting the overwrite option brings up another screen, shown below.



This option: Do you want to use the last DATA? ([Y]/N)? is asking whether the user wants to use the data entered into the program the last time it was used before this time. [Y] for Yes is the default. Typing Y or hitting <Enter> selects Yes for this question. In the example screen, N (for No) was selected. The "use the last data" option is useful if you are running the program several times in a row, and some of the data does not change from run to run. If you select Yes for the "use the last data" option, the program presents your previously entered data as the default values for the variables requested, and these can be accepted by hitting <Enter>.

The input screen dialog continues as shown in the next screen below. A user is asked for a Profile ID. Any label up to 8 characters can be entered, and this label will appear in the program's output.

Next, a user is asked to enter:

Number of loci in the profile (the default is [3])

Name of locus 1:

Type (1=VNTR, 2=PCR) of locus 1:

In this question, enter 1 for an RFLP locus, and enter 2 for any PCR or STR locus

Number of alleles for locus 1 [2]:

The default value is 2. For heterozygotes, users can type 2 or hit <Enter>. For homozygotes, users should type 1.

Name of allele 1:

This value will be a fragment size or an allele designator

Name of allele 2:

This question appears if the user specified that the locus has 2 alleles.

If the user specified that the locus has 1 allele, the question does not appear.

Name of locus 2:

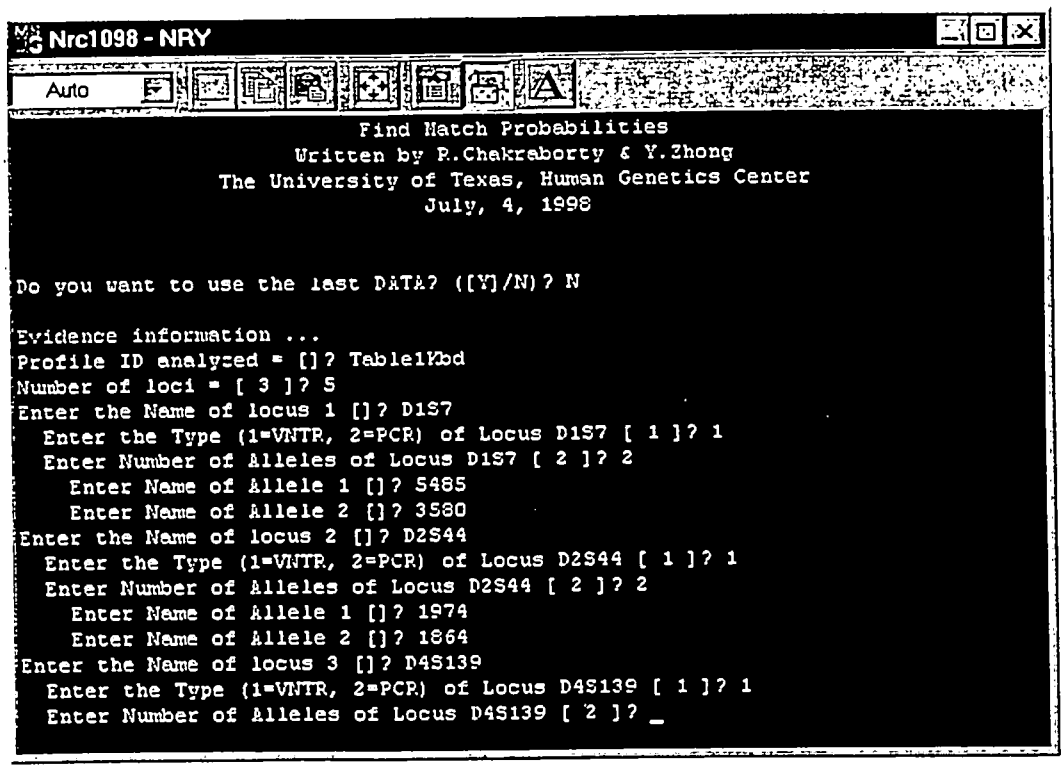
Type (1=VNTR, 2=PCR) of locus 2:

...

...

etc.

The same information is requested for all the loci in the profile. If the user specified that the profile had five loci, for example, this dialog will be repeated for all five loci.



Once the number of loci, their names, types, numbers and names of alleles have been entered, the program will query for additional information (see the next screen below):

Name of the population:

This is a label for the population from which the frequencies were derived. Any eight character label can be entered.

Enter 1st Theta [0]:

Enter 2nd Theta [0.01]:

Enter 3rd Theta [0.03]:

A user can enter three theta values to be used in the calculations. The default values are 0, 0.01 and 0.03, and these can be accepted by hitting <Enter> after each question.

Number of individuals of population [population label], locus 1 [locus 1 name]:

where [population label] is the name of the population entered by the user above, and [locus 1 name] is the name of locus 1 entered by the user above.

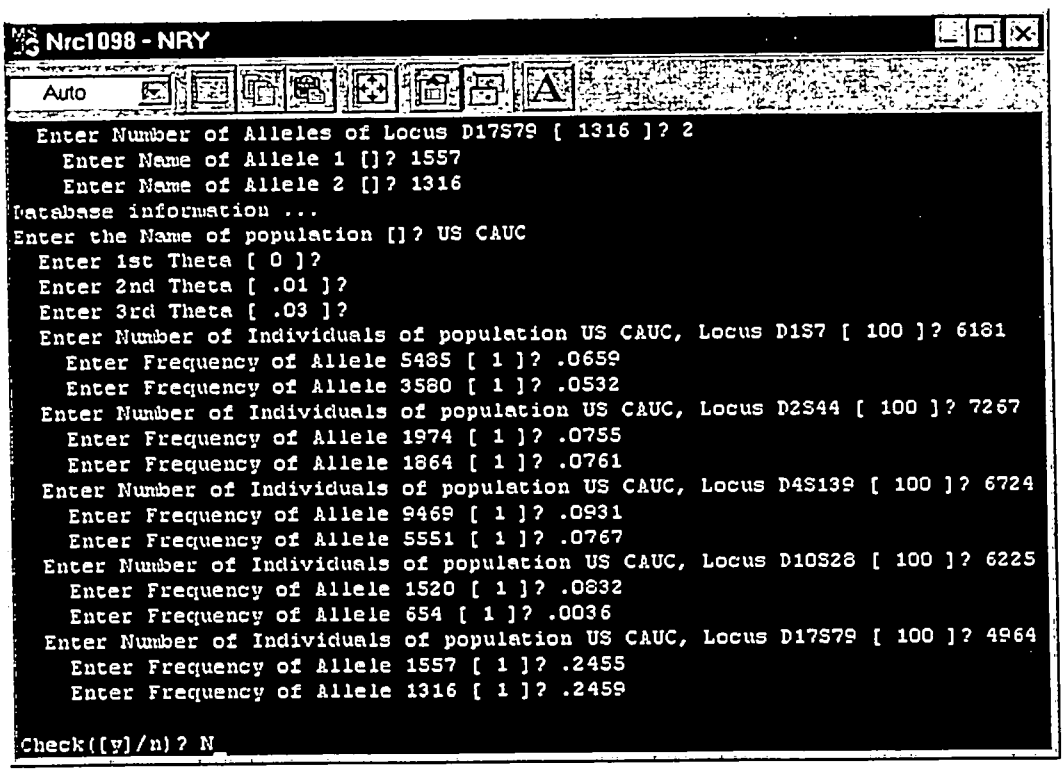
In this example, the query reads: Number of Individuals of population US CAUC, Locus D1S7 [100]? because US CAUC was entered for the name of the population, and D1S7 was entered as the name of the first locus.
NOTE that the default number of individuals is 100. If the population size is any number other than 100, it should be typed in. In this case, 6181 was entered.

Frequency of Allele [Allele 1 Name] [1]:

Here, the frequency of allele 1 for the first locus is entered. Note that the program presents a default value of 1. [Allele 1 Name] is the name for allele 1 of the first locus entered by the user above. In this case, the question reads: Enter Frequency of Allele 5485 [1]? because 5485 was the first D1S7 fragment size entered above.

Frequency of Allele [Allele 2 Name] [1]:

If the user specified that the locus has two alleles, the frequency of the second allele is requested. Again, the program uses the user specified locus 1, allele 2 name, 3580 in this case.



This dialog, requesting number of people who were typed for the locus in the database (from which the frequencies are derived), and the allele frequency(-ies), continues for all the loci in the profile.

At this point, the program says "Check [Y]/N". Accepting the default (or choosing Y) allows (actually forces) the user back through all the user-entered data to make any changes or corrections. Choosing to "check the data" will cause the program to pace back through every question previously asked, so that any entry item can be corrected. Note that your originally entered value will be presented as the default in every case, however.

Thus for data entry questions that do not require changes, hitting <Enter> accepts the previously entered value, and moves on to the next question. Once a user is satisfied with all the user-entered data, and chooses "N" for the "Check" option, the program executes.

All the data entered in the example above, illustrating keyboard input of data for a profile consisting of five RFLP loci, are given in Table 1 below.

Table 1. Example Profile Data for Five RFLP Loci

Locus Name	Number of People Typed	Locus Type	No of alleles	Allele Name	Allele Frequency
D1S7	6181	1	2	5485	0.0659
				3580	0.0532
D2S44	7267	1	2	1974	0.0755
				1864	0.0761
D4S139	6724	1	2	9469	0.0931
				5551	0.0767
D10S28	6225	1	2	1520	0.0832
				654	0.0036
D17S79	4964	1	2	1557	0.2455
				1316	0.2459

Running the program, using keyboard input as illustrated above, with the data from Table 1 gives the following output data file:

```
                Find Match Probabilities
          Written by R.Chakraborty and Y.Zhong
                The University of Texas
                  10-17-1998  15:22:28


Case Number = 1    Table1Kbd
  US CAUC


 Data used:
-------------------------------------------------------------
 Type        Locus         Allele         Freq.          n
-------------------------------------------------------------
 VNTR        D1S7          5485           6.59000E-2    6181
 VNTR        D1S7          3580           5.32000E-2    6181
 VNTR        D2S44         1974           7.55000E-2    7267
 VNTR        D2S44         1864           7.61000E-2    7267
 VNTR        D4S139        9469           9.31000E-2    6724
 VNTR        D4S139        5551           7.67000E-2    6724
 VNTR        D10S28        1520           8.32000E-2    6225
 VNTR        D10S28        654            3.60000E-3    6225
 VNTR        D17S79        1557           2.45500E-1    4964
 VNTR        D17S79        1316           2.45900E-1    4964
-------------------------------------------------------------
```

Frequency estimates:

Locus	Recommendation 4.1			Recommendation 4.10		
	th = 0.0	th = 0.01	th = 0.03	th = 0.0	th = 0.01	th = 0.03
D1S7	7.0118E-03	7.0118E-03	7.0118E-03	7.0118E-03	9.1540E-03	1.4040E-02
D2S44	1.1491E-02	1.1491E-02	1.1491E-02	1.1491E-02	1.4040E-02	1.9633E-02
D4S139	1.4282E-02	1.4282E-02	1.4282E-02	1.4282E-02	1.7045E-02	2.3008E-02
D10S28	5.9904E-04	5.9904E-04	5.9904E-04	5.9904E-04	2.4323E-03	6.7919E-03
D17S79	1.2074E-01	1.2074E-01	1.2074E-01	1.2074E-01	1.2450E-01	1.3189E-01
Combined Prob. and 95% CI:						
Estim.=	8.3226E-11	8.3226E-11	8.3226E-11	8.3226E-11	6.6339E-10	5.6812E-09
or 1 in	1.2015E+10	1.2015E+10	1.2015E+10	1.2015E+10	1.5074E+09	1.7602E+08
L 95% CI	5.9674E-11	5.9674E-11	5.9674E-11	5.9674E-11	5.6450E-10	5.0399E-09
or 1 in	1.6758E+10	1.6758E+10	1.6758E+10	1.6758E+10	1.7715E+09	1.9842E+08
U 95% CI	1.1607E-10	1.1607E-10	1.1607E-10	1.1607E-10	7.7960E-10	6.4041E-09
or 1 in	8.6152E+09	8.6152E+09	8.6152E+09	8.6152E+09	1.2827E+09	1.5615E+08

Use B.S. Weir and W.G. Hill's Formulas (JFSS 1993:33(4):218-255)
(1-5 use EXACT V(Pi); While Unrelated use approx. V(Pi))
For Unrelated individual, homozygoties at a VNTR locus, Pi=2p (Not Pi=p^2)

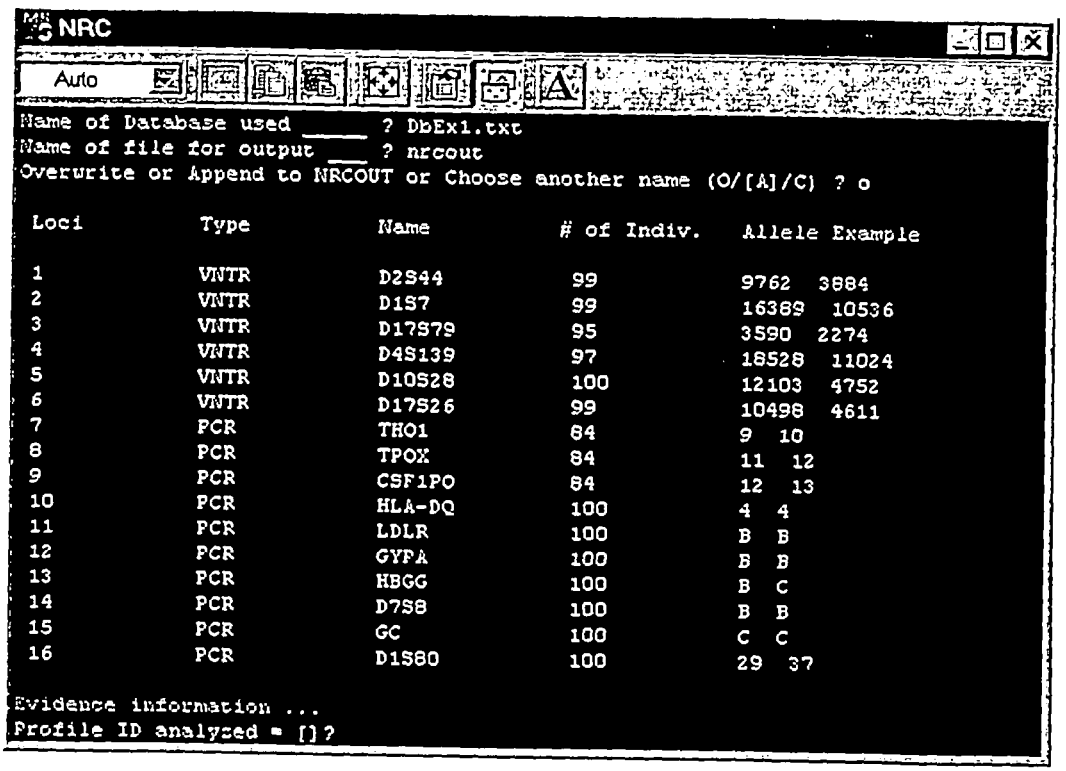
(1) Parent or Offspring	Probability =	4.0865E-06
	95% CI	(3.7202E-06 ; 4.4889E-06)
	Or 1 in	2.4471E+05
	95% CI	(2.2277E+05 ; 2.6880E+05)
(2) Full Sibling	Probability =	2.6550E-03
	95% CI	(2.6169E-03 ; 2.6937E-03)
	Or 1 in	3.7665E+02
	95% CI	(3.7124E+02 ; 3.8213E+02)
(3) Half Sibling	Probability =	2.9035E-07
	95% CI	(2.6195E-07 ; 3.2182E-07)
	Or 1 in	3.4441E+06
	95% CI	(3.1073E+06 ; 3.8175E+06)
(4) Uncle/Aunt or Nephew/Niece	Probability =	2.9035E-07
	95% CI	(2.6195E-07 ; 3.2182E-07)
	Or 1 in	3.4441E+06
	95% CI	(3.1073E+06 ; 3.8175E+06)
(5) First Cousin	Probability =	3.0459E-08
	95% CI	(2.7135E-08 ; 3.4191E-08)
	Or 1 in	3.2831E+07
	95% CI	(2.9247E+07 ; 3.6853E+07)
(6) Unrelated	Probability =	8.3226E-11
	95% CI	(5.9674E-11 ; 1.1607E-10)
	Or 1 in	1.2015E+10
	95% CI	(8.6152E+09 ; 1.6758E+10)

Note that in this example, and in general, the output file repeats the input data so a user can verify the data that was actually used for the calculations.

Database Entry:

Choosing option 2 (database frequency input) gives a slightly different screen dialog. It asks for the name of the database to be used, then for the name of the output file (with the usual options). Next, it will delineate the loci, their names, types (VNTR or PCR; VNTR in this context means RFLP), names, number of people typed for each locus and examples of the allele names. This screen output is a handy reference to the loci that are in the selected database. Obviously, the loci that are in the profile to be analyzed must all be present in the selected database, if database entry is to be used.

An example initial screen where database input was selected, and a small example database called DbEx1.txt was specified, is shown below:



The profile data shown in Table 2 below is used to illustrate the operation of the NRC program using database input mode, and the relationship of database input mode to keyboard input mode.

In database entry mode, a user must still input the locus and allele names in the profile to the program. With the database input option, the program will go and calculate both the numbers of people typed for a given locus, and the frequencies of the alleles.

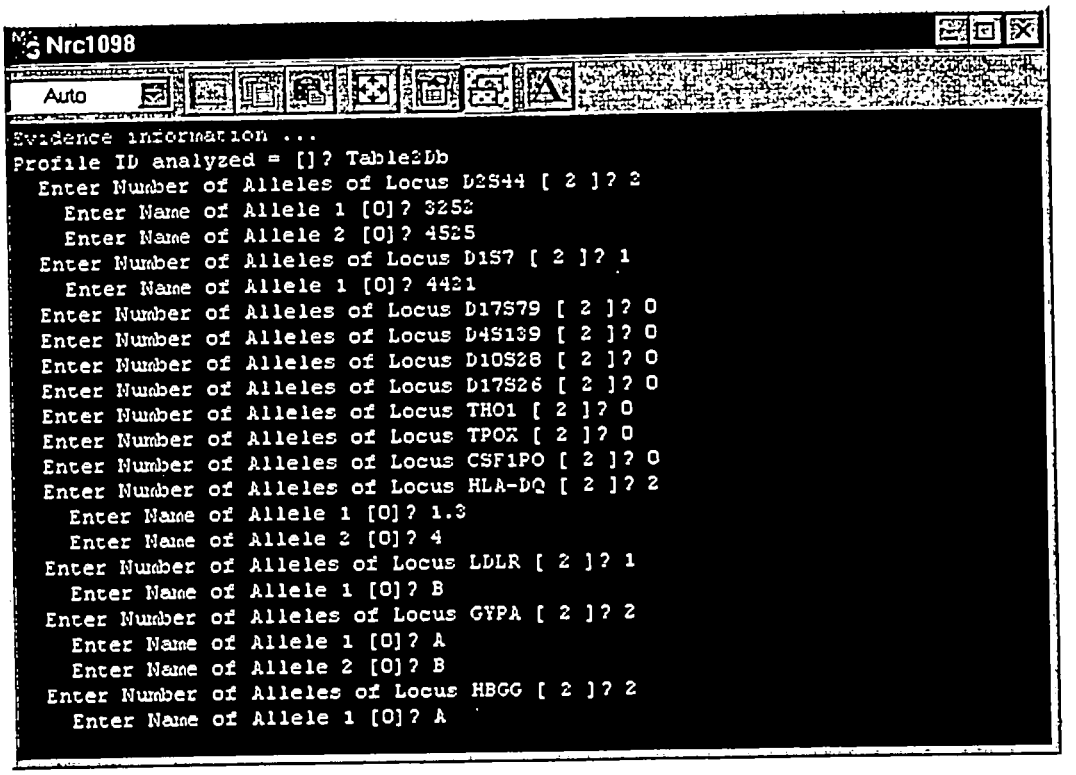
Table 2. Example Profile Data

Locus Name	Locus Type	No of alleles	Allele Name	Database Input *		Keyboard Input **	
				Number Typed	Allele Frequency	Number Typed	Allele Frequency
D1S7	1	1	4421	[99]	[0.0808]	99	0.08081
D2S44	1	2	3252	[99]	[0.455]	99	0.04545
			4525		[0.0303]		0.0303
HLA-DQA1	2	2	1.3	[100]	[0.1375]	100	0.175
			4		[0.13]		0.30
LDLR	2	1	B	[100]	[0.58]	100	0.58
GYPA	2	2	A	[100]	[0.73]	100	0.73
			B		[0.27]		0.27
HBGG	2	2	A	[100]	[0.365]	100	0.365
			B		[0.63]		0.63
D7S8	2	1	A	[100]	[0.76]	100	0.76
GC	2	2	A	[100]	[0.275]	100	0.275
			B		[0.14]		0.14

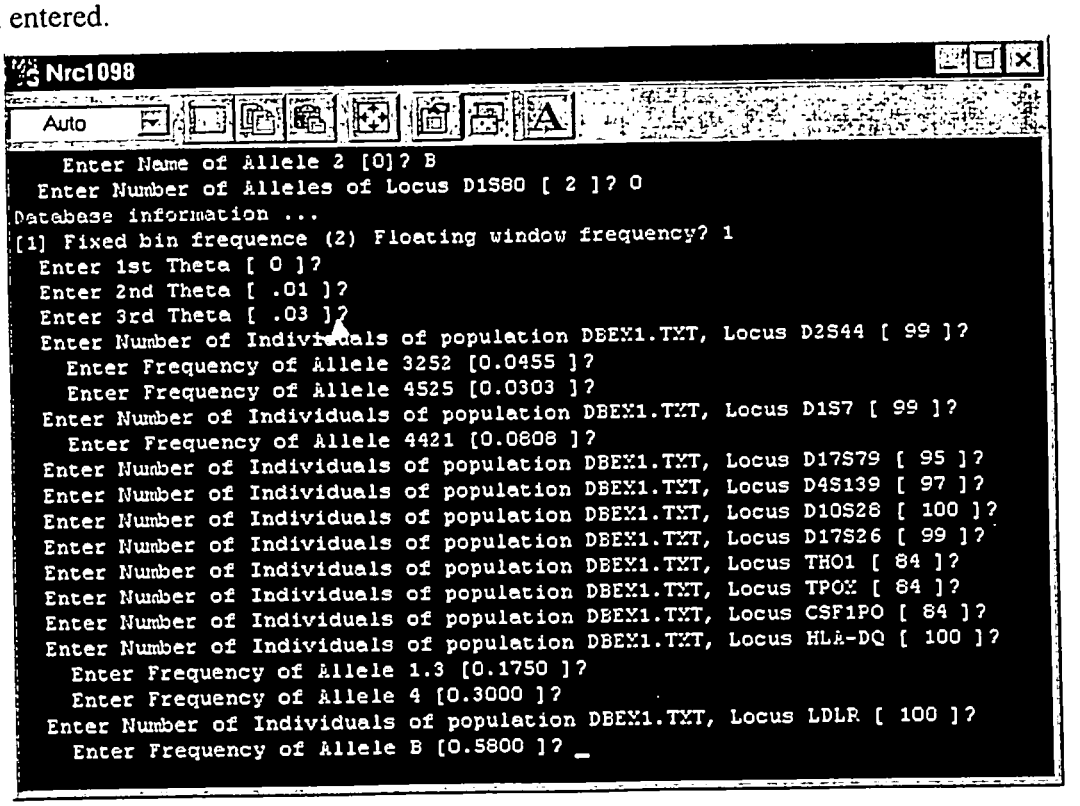
* Values are accepted by pressing <Enter> when the program presents them as the default values. The program derives them from the database file.

** Values for keyboard input derived from running Program H on DbEx1.txt for the loci represented in the profile

As can be seen in the screen immediately below, the program queries for "number of alleles" at every locus that the database has. If your profile doesn't have data for a particular locus, you specify zero for the number of alleles at that locus.



This dialog continues until the numbers and names of all alleles for the loci in the profile have been entered.



Next, you are prompted for a choice of fixed or floating bin frequency computation if any of the loci in the specified profile are RFLP. In this example two of the loci are RFLP, so the choice appears.
In this example, "fixed bin" was selected. In fixed bin mode, the program computes the frequency of the fragment size based on a division of all possible fragment sizes into 31 bins, as explained elsewhere (in the discussion of Program H and how it handles RFLP databases). The "floating bin" option will be discussed below.
Next, the program offers the choice of three values of theta, one at a time, with the defaults equal to 0, 0.01 and 0.03 (which were all accepted in the example).
Finally, the program requests the number of individuals typed at each locus, and the frequencies of the allele(s) -- but it presents its calculated results as the defaults. This feature is the principal advantage to using database input mode. In database mode, the program has computed the number of individuals typed and the allele (or bin) frequencies. Users should simply accept each default value by pressing <Enter> as was done in the example.
Note that the program queries for number of people typed even at loci that are not present in the profile, and shows the correct answer as the default. Users should just accept these values by pressing <Enter> as well.

Running NRC in database input mode, with DbEx1.txt as the selected database, gives the following results:

```
                    Find Match Probabilities
            Written by R.Chakraborty and Y.Zhong
                  The University of Texas
                    10-17-1998  18:13:01


Case Number = 1    TABLE2DB   DBEX1.TXT
```

Data used:

Type	Locus	Allele	Freq.	n
VNTR	D2S44	3252	4.54546E-2	99
VNTR	D2S44	4525	3.03030E-2	99
VNTR	D1S7	4421	8.08081E-2	99
PCR	HLA-DQ	1.3	1.75000E-1	100
PCR	HLA-DQ	4	3.00000E-1	100
PCR	LDLR	B	5.80000E-1	100
PCR	GYPA	A	7.30000E-1	100
PCR	GYPA	B	2.70000E-1	100
PCR	HBGG	A	3.65000E-1	100
PCR	HBGG	B	6.30000E-1	100
PCR	D7S8	A	7.60000E-1	100
PCR	GC	A	2.75000E-1	100
PCR	GC	B	1.40000E-1	100

Frequency estimates:

Locus	Recommendation 4.1 th = 0.0	th = 0.01	th = 0.03	Recommendation 4.10 th = 0.0	th = 0.01	th = 0.03
D2S44	2.7548E-03	2.7548E-03	2.7548E-03	2.7548E-03	4.2710E-03	8.0611E-03
D1S7	1.6162E-01	1.6162E-01	1.6162E-01	6.5299E-03	1.0678E-02	2.1342E-02
HLA-DQ	1.0500E-01	1.0500E-01	1.0500E-01	1.0500E-01	1.0922E-01	1.1746E-01
LDLR	3.3640E-01	3.3884E-01	3.4371E-01	3.3640E-01	3.4849E-01	3.7215E-01
GYPA	3.9420E-01	3.9420E-01	3.9420E-01	3.9420E-01	3.9444E-01	3.9467E-01
HBGG	4.5990E-01	4.5990E-01	4.5990E-01	4.5990E-01	4.5685E-01	4.5102E-01
D7S8	5.7760E-01	5.7942E-01	5.8307E-01	5.7760E-01	5.8661E-01	6.0400E-01
GC	7.7000E-02	7.7000E-02	7.7000E-02	7.7000E-02	8.1426E-02	9.0129E-02
Combined Prob. and 95% CI:						
Estim.=	1.2680E-07	1.2812E-07	1.3078E-07	5.1233E-09	1.4940E-08	7.2872E-08
or 1 in	7.8864E+06	7.8050E+06	7.6463E+06	1.9519E+08	6.6934E+07	1.3723E+07
L 95% CI	3.6597E-08	3.6986E-08	3.7768E-08	1.1600E-09	4.5219E-09	2.9945E-08
or 1 in	2.7325E+07	2.7038E+07	2.6477E+07	8.6205E+08	2.2114E+08	3.3395E+07
U 95% CI	4.3934E-07	4.4383E-07	4.5287E-07	2.2627E-08	4.9362E-08	1.7734E-07
or 1 in	2.2762E+06	2.2531E+06	2.2081E+06	4.4195E+07	2.0259E+07	5.6390E+06

Use B.S. Weir and W.G. Hill's Formulas (JFSS 1993:33(4):218-255)
(1-5 use EXACT V(Pi); While Unrelated use approx. V(Pi))
For Unrelated individual, homozygoties at a VNTR locus, Pi=2p (Not Pi=p^2)

TABLE2DB DBEX1.TXT

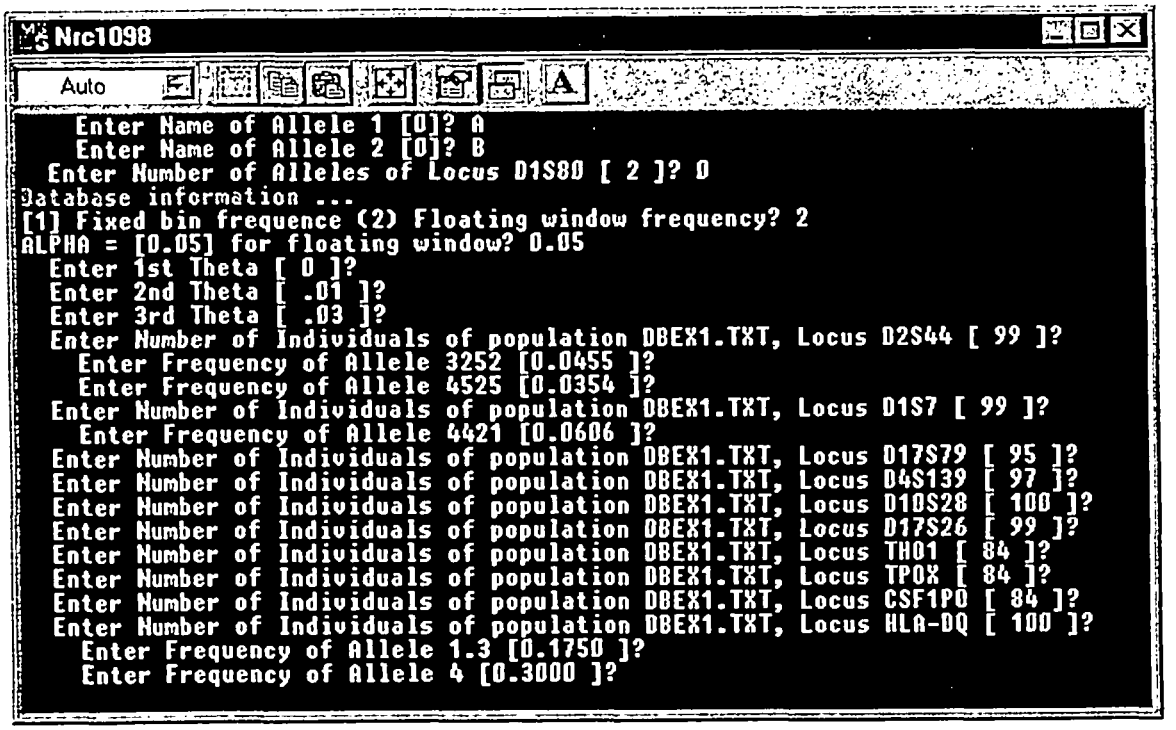
(1) Parent or Offspring	Probability =	1.6540E-05	
	95% CI	(8.0099E-06 ;	3.4154E-05)
	Or 1 in	6.0459E+04	
	95% CI	(2.9279E+04 ;	1.2485E+05)
(2) Full Sibling	Probability =	2.0597E-03	
	95% CI	(5.2588E-04 ;	8.0671E-03)
	Or 1 in	4.8551E+02	
	95% CI	(1.2396E+02 ;	1.9016E+03)
(3) Half Sibling	Probability =	1.4175E-06	
	95% CI	(1.0216E-07 ;	1.9669E-05)
	Or 1 in	7.0545E+05	
	95% CI	(5.0841E+04 ;	9.7887E+06)
(4) Uncle/Aunt or Nephew/Niece	Probability =	1.4175E-06	
	95% CI	(1.0216E-07 ;	1.9669E-05)
	Or 1 in	7.0545E+05	
	95% CI	(5.0841E+04 ;	9.7887E+06)
(5) First Cousin	Probability =	2.1430E-07	
	95% CI	(7.3712E-09 ;	6.2300E-06)
	Or 1 in	4.6665E+06	
	95% CI	(1.6051E+05 ;	1.3566E+08)
(6) Unrelated	Probability =	1.2680E-07	
	95% CI	(3.6597E-08 ;	4.3934E-07)
	Or 1 in	7.8864E+06	
	95% CI	(2.2762E+06 ;	2.7325E+07)

Table 2, which shows the profile that was used for the above example, also shows that if a user employed Program H to compute the frequencies for the alleles ands bins in the profile from DbEx1.txt, the same results would be obtained as Program NRC obtained. Furthermore, if a user took the values in Table 2 computed from Program H (shown under the "Keyboard Input" heading), then ran Program NRC using keyboard input and the values in the "keyboard input" columns, exactly the same results (output file) would be obtained as shown above.

Users should also note that the order in which the program presents the loci for data entry corresponds to the order of the loci in the specified database.

Database Entry Using Floating Bins:

As noted above, running Program NRC in database entry mode with databases containing RFLP data permits a choice between fixed bin and floating bin methods of computing binned allele frequencies. In the example above, with the profile in Table 2 and specifying DbEx1.txt as the database, the "fixed bin" method was selected.
Users can also select "floating bin," and specify the window size the program will use in computing the binned allele frequencies for RFLP locus bandsizes. The relevant portion of the input dialog screen is shown below.



Here, data are being entered in database entry mode using the profile from Table 2 and DbEx1.txt as the database. Note under the "Database information ..." line that the program offers the choice of: [1] Fixed bin frequency (2) Floating window frequency? indicating by the 1 in square brackets that "fixed bin" is the default -- it can be selected by hitting <Enter> or by entering 1. Here, 2 (for "floating window") was entered.

Next the program offers a choice of window size for the floating bin analysis: ALPHA = [0.05] for floating window? This line indicates that 0.05 is the default value and can be accepted by hitting <Enter>, or as has been done in the example screen, 0.05 can be entered from the keyboard. (Any other desired window value could be entered as well). A value of 0.05 indicates a "5% window," i.e. the program will set a window around bandsizes in the database that are within ± 5% of the profile bandsize value, count the number of times all the bandsizes within the window were observed in the population, and express the result as a fraction of the total binned alleles in the database.

Running Program NRC with the profile in Table 2, specifying DbEx1.txt as the database, selecting the "floating window" and using 0.05 for Alpha (5% window) yields the following results:

```
                    Find Match Probabilities
                Written by R.Chakraborty and Y.Zhong
                      The University of Texas
                        10-17-1998  22:37:52


Case Munber = 1    TBL2   DBEX1.TXT

 Data used:
```

Type	Locus	Allele	Freq.	n
VNTR	D2S44	3252	4.54546E-2	99
VNTR	D2S44	4525	3.53535E-2	99
VNTR	D1S7	4421	6.06061E-2	99
PCR	HLA-DQ	1.3	1.75000E-1	100
PCR	HLA-DQ	4	3.00000E-1	100
PCR	LDLR	B	5.80000E-1	100
PCR	GYPA	A	7.30000E-1	100
PCR	GYPA	B	2.70000E-1	100
PCR	HBGG	A	3.65000E-1	100
PCR	HBGG	B	6.30000E-1	100
PCR	D7S8	A	7.60000E-1	100
PCR	GC	A	2.75000E-1	100
PCR	GC	B	1.40000E-1	100

```
 Frequency estimates:
```

Locus	Recommendation 4.1			Recommendation 4.10		
	th = 0.0	th = 0.01	th = 0.03	th = 0.0	th = 0.01	th = 0.03
D2S44	3.2140E-03	3.2140E-03	3.2140E-03	3.2140E-03	4.8049E-03	8.7260E-03
D1S7	1.2121E-01	1.2121E-01	1.2121E-01	3.6731E-03	6.9889E-03	1.6188E-02
HLA-DQ	1.0500E-01	1.0500E-01	1.0500E-01	1.0500E-01	1.0922E-01	1.1746E-01
LDLR	3.3640E-01	3.3884E-01	3.4371E-01	3.3640E-01	3.4849E-01	3.7215E-01
GYPA	3.9420E-01	3.9420E-01	3.9420E-01	3.9420E-01	3.9444E-01	3.9467E-01
HBGG	4.5990E-01	4.5990E-01	4.5990E-01	4.5990E-01	4.5685E-01	4.5102E-01
D7S8	5.7760E-01	5.7942E-01	5.8307E-01	5.7760E-01	5.8661E-01	6.0400E-01
GC	7.7000E-02	7.7000E-02	7.7000E-02	7.7000E-02	8.1426E-02	9.0129E-02

```
 Combined Prob. and 95% CI:
 ------------------------------------------------------------------------------
  Estim.=  1.1095E-07 1.1211E-07 1.1443E-07   3.3621E-09 1.1001E-08 5.9832E-08
  or 1 in  9.0130E+06 8.9201E+06 8.7386E+06   2.9743E+08 9.0897E+07 1.6713E+07

  L 95% CI 3.2170E-08 3.2512E-08 3.3200E-08   7.0621E-10 3.2488E-09 2.4574E-08
  or 1 in  3.1085E+07 3.0758E+07 3.0121E+07   1.4160E+09 3.0781E+08 4.0693E+07

  U 95% CI 3.8265E-07 3.8656E-07 3.9444E-07   1.6007E-08 3.7254E-08 1.4568E-07
  or 1 in  2.6133E+06 2.5869E+06 2.5352E+06   6.2474E+07 2.6842E+07 6.8645E+06
 ------------------------------------------------------------------------------

Use B.S. Weir and W.G. Hill's Formulas (JFSS 1993:33(4):218-255)
   (1-5 use EXACT V(Pi); While Unrelated use approx. V(Pi))
   For Unrelated individual, homozygoties at a VNTR locus, Pi=2p (Not Pi=p^2)

TBL2  DBEX1.TXT

(1) Parent or Offspring          Probability = 1.3232E-05
                                    95% CI   ( 6.1389E-06 ;  2.8521E-05 )
                                 Or 1 in       7.5574E+04
                                    95% CI   ( 3.5062E+04 ;  1.6289E+05 )

(2) Full Sibling                 Probability = 1.9935E-03
                                    95% CI   ( 5.1475E-04 ;  7.7207E-03 )
                                 Or 1 in       5.0162E+02
                                    95% CI   ( 1.2952E+02 ;  1.9427E+03 )

(3) Half Sibling                 Probability = 1.1199E-06
                                    95% CI   ( 7.6690E-08 ;  1.6354E-05 )
                                 Or 1 in       8.9294E+05
                                    95% CI   ( 6.1148E+04 ;  1.3039E+07 )

(4) Uncle/Aunt or Nephew/Niece   Probability = 1.1199E-06
                                    95% CI   ( 7.6690E-08 ;  1.6354E-05 )
                                 Or 1 in       8.9294E+05
                                    95% CI   ( 6.1148E+04 ;  1.3039E+07 )

(5) First Cousin                 Probability = 1.6581E-07
                                    95% CI   ( 4.8807E-09 ;  5.6331E-06 )
                                 Or 1 in       6.0309E+06
                                    95% CI   ( 1.7752E+05 ;  2.0489E+08 )

(6) Unrelated                    Probability = 1.1095E-07
                                    95% CI   ( 3.2170E-08 ;  3.8265E-07 )
                                 Or 1 in       9.0130E+06
                                    95% CI   ( 2.6133E+06 ;  3.1085E+07 )
```

To illustrate the difference in the calculations depending on the window size specification, the following results are obtained by running Program NRC with the profile in Table 2, specifying DbEx1.txt as the database, selecting the "floating window" and using 0.025 for Alpha (2.5% window):

```
                        Find Match Probabilities
                 Written by R.Chakraborty and Y.Zhong
                       The University of Texas
                         10-17-1998  22:44:20

Case Number = 1    TBL2 2.5%  DBEX1.TXT
```

Data used:

Type	Locus	Allele	Freq.	n
VNTR	D2S44	3252	3.03030E-2	99
VNTR	D2S44	4525	5.05051E-3	99
VNTR	D1S7	4421	5.05050E-2	99
PCR	HLA-DQ	1.3	1.75000E-1	100
PCR	HLA-DQ	4	3.00000E-1	100
PCR	LDLR	B	5.80000E-1	100
PCR	GYPA	A	7.30000E-1	100
PCR	GYPA	B	2.70000E-1	100
PCR	HBGG	A	3.65000E-1	100
PCR	HBGG	B	6.30000E-1	100
PCR	D7S8	A	7.60000E-1	100
PCR	GC	A	2.75000E-1	100
PCR	GC	B	1.40000E-1	100

Frequency estimates:

Locus	Recommendation 4.1			Recommendation 4.10		
	th = 0.0	th = 0.01	th = 0.03	th = 0.0	th = 0.01	th = 0.03
D2S44	3.0609E-04	3.0609E-04	3.0609E-04	3.0609E-04	1.1648E-03	3.7970E-03
D1S7	1.0101E-01	1.0101E-01	1.0101E-01	2.5508E-03	5.4358E-03	1.3875E-02
HLA-DQ	1.0500E-01	1.0500E-01	1.0500E-01	1.0500E-01	1.0922E-01	1.1746E-01
LDLR	3.3640E-01	3.3884E-01	3.4371E-01	3.3640E-01	3.4849E-01	3.7215E-01
GYPA	3.9420E-01	3.9420E-01	3.9420E-01	3.9420E-01	3.9444E-01	3.9467E-01
HBGG	4.5990E-01	4.5990E-01	4.5990E-01	4.5990E-01	4.5685E-01	4.5102E-01
D7S8	5.7760E-01	5.7942E-01	5.8307E-01	5.7760E-01	5.8661E-01	6.0400E-01
GC	7.7000E-02	7.7000E-02	7.7000E-02	7.7000E-02	8.1426E-02	9.0129E-02

Combined Prob. and 95% CI:

Estim.=	8.8056E-09	8.8974E-09	9.0821E-09	2.2236E-10	2.0743E-09	2.2315E-08
or 1 in	1.1356E+08	1.1239E+08	1.1011E+08	4.4971E+09	4.8208E+08	4.4813E+07
L 95% CI	9.1981E-10	9.2949E-10	9.4899E-10	1.8447E-11	5.5529E-10	9.5044E-09
or 1 in	1.0872E+09	1.0759E+09	1.0537E+09	5.4210E+10	1.8009E+09	1.0521E+08
U 95% CI	8.4299E-08	8.5168E-08	8.6918E-08	2.6805E-09	7.7489E-09	5.2392E-08
or 1 in	1.1863E+07	1.1741E+07	1.1505E+07	3.7307E+08	1.2905E+08	1.9087E+07

Use B.S. Weir and W.G. Hill's Formulas (JFSS 1993:33(4):218-255)
(1-5 use EXACT V(Pi); While Unrelated use approx. V(Pi))
For Unrelated individual, homozygoties at a VNTR locus, Pi=2p (Not Pi=p^2)

TBL2 2.5% DBEX1.TXT

(1) Parent or Offspring
Probability = 4.8242E-06
95% CI (1.8086E-06 ; 1.2868E-05)
Or 1 in 2.0729E+05
95% CI (7.7713E+04 ; 5.5291E+05)

(2) Full Sibling
Probability = 1.8685E-03
95% CI (4.8500E-04 ; 7.1985E-03)
Or 1 in 5.3519E+02
95% CI (1.3892E+02 ; 2.0618E+03)

```
(3) Half Sibling                Probability = 3.8110E-07
                                   95% CI   ( 2.3862E-08 ;  6.0864E-06 )
                                Or 1 in       2.6240E+06
                                   95% CI   ( 1.6430E+05 ;  4.1907E+07 )

(4) Uncle/Aunt or Nephew/Niece  Probability = 3.8110E-07
                                   95% CI   ( 2.3862E-08 ;  6.0864E-06 )
                                Or 1 in       2.6240E+06
                                   95% CI   ( 1.6430E+05 ;  4.1907E+07 )

(5) First Cousin                Probability = 5.0024E-08
                                   95% CI   ( 1.2880E-09 ;  1.9429E-06 )
                                Or 1 in       1.9990E+07
                                   95% CI   ( 5.1469E+05 ;  7.7641E+08 )

(6) Unrelated                   Probability = 8.8056E-09
                                   95% CI   ( 9.1981E-10 ;  8.4299E-08 )
                                Or 1 in       1.1356E+08
                                   95% CI   ( 1.1863E+07 ;  1.0872E+09 )
```

The probabilities differ in the calculations where a 2.5% window was employed as against those where a 5% window was employed.
It is important to note that in the profile illustrated the differences lie only in the binned alleles (fragment sizes) for the RFLP loci. The PCR locus alleles are obviously not affected by "window" sizes as they are discrete. Thus only RFLP bandsize frequency calculations are affected by the "window" size option.

As noted above, the "keyboard" input option can be selected, and the user must then get the allele or binned fragment size frequencies from some other source. They could be obtained by running Program H, for example. Note, however, that program H sets up fixed bins, where NRC offers an option of fixed or floating bins, and a user-specified % match window in the case of floating bins. Values gleaned from program H, therefore, could differ from those calculated by NRC even with the same locus and database, depending on the options selected.

Output data for the NRC program is placed in a user-specified file that can be viewed in WordPad or any other text processor.

The NRC program calculates the probabilities and reciprocal probabilities of chance duplicates (estimated frequencies) for a user-specified DNA profile in accordance with the NRC 1996 recommendations. Values are separately calculated for 4.1 and for 4.10 and for all three user-selected values of theta. 95% confidence intervals are computed and printed as well. In addition the program calculates probabilities and reciprocal probabilities of chance duplicates among relatives (siblings, half siblings, etc.) as well as among unrelated persons using the formulas of Weir and Hill, 1993.

Note that NRC is independent of the type of locus; that is, a user can calculate a frequency for any combination of RFLP and/or PCR-based loci provided the appropriate data is supplied to the program.

Users should also note that it can sometimes be desirable to use scientifically estimated minimum allele frequencies in calculations of the probability of a chance match for a profile in order to be maximally conservative.

Consult Budowle, Monson, and Chakraborty (1996) for detailed information on minimum allele frequencies. These values can be calculated where it is deemed warranted, and inputted to NRC using the manual (keyboard) entry mode. Calculating minimum allele frequencies for HLA-DQA1 and PM loci from any database is trivial, because the value depends solely on the population size. The calculations are more complicated with other loci, and cannot be done unless a user has the complete database, i.e. these calculations cannot be done on loci other than the blot-dot loci from summary data such as one finds in most population survey papers.

7 Results

Clicking on Results / View Results opens WordPad so that the results file from a program can be viewed on the screen. In WordPad, a user should click on File then on Open (or on the File/Open icon). From the filenames in the directory, click on the name of the desired results file, then click OK. Be sure that WordPad has opened the correct directory -- the one in which all the DNATYPE programs are located, so that the output file can be found in the directory list.

8 References

Brown, A.D.H., Feldman, M.W., and Nevo, E. (1980) Multilocus structure of natural populations of Hordeum spontaneum. Genetics 96:523-536.

Budowle, B., and Baechtel, F.S. (1990) Modifications to improve the effectiveness of restriction fragment length polymorphism typing. Appl Theor Electrophoresis 1:181-187.

Budowle, B., Giusti, A.M., Waye, J.S., Baechtel, F.S., Fourney, R.M., Adams, D.E., Presley, L.A., Deadman, H.A., and Monson, K.L. (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. Am J Hum Genet 48:841-855.

Budowle, B., Monson, K.L. and Chakraborty, R. (1996) Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci. Int J Legal Med 108:173-176.

Chakraborty, R. (1984) Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. Genetics 108:719-731.

Chakraborty, R. (1992) Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum Biol 64(2):141-159

Chakraborty, R. (1993) A class of population genetic questions formulated as the generalized occupancy problem. Genetics 134:953-958.

Chakraborty, R., de Andrade, M., Daiger, S.P., and Budowle, B. (1992) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet 56:45-57.

Chakraborty, R., Fornage, M., Guegue, R., and Boerwinkle, E. (1991) Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In: DNA Fingerprinting: Approaches and Applications. T. Burke, G. Dolf, A. J. Jeffreys, and R. Wolff (Eds.) Birkhauser, Basel, pp. 127-143.

Chakraborty, R. and Jin, L. (1992) Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Human Genetics 88(3): 267-272

Chakraborty, R., and Li, Z. (1995) Correlation of DNA fragment sizes within loci in the Presence of Non-detectable Alleles. Genetica 96:27-36.

Chakraborty, R., Srinivasan, M. R., and de Andrade, M. (1993) Intraclass and interclass correlations of allele sizes within and between loci in DNA typing data. Genetics 133:411-419.

Chakraborty, R., Srinivasan, M.R., and Daiger, S.P. (1993) Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implication in DNA forensics. Am J Hum Genet 52:60-70.

Chakraborty, R., and Zhong, Y. (1994) Statistical power of an exact test of Hardy-Weinberg proportions of genotype data at a multiallelic locus. Hum Hered 44:1-9.

Chakraborty, R., Zhong, Y., Jin, L., and Budowle, B. (1994) Nondetectability of restriction fragments and independence of DNA fragment sizes within and between loci in RFLP typing of DNA. Am J Hum Genet 55:391-401.

Devlin, B., and Risch, N. (1992) A note on Hardy-Weinberg equilibrium of VNTR data by using the Federal Bureau of Investigation's fixed-bin method. Am J Human Genet 51:549-553.

Devlin, B., Risch, N., and Roeder, K. (1990) No excess of homozygosity at loci used for DNA fingerprinting. Science, 249:1416-1420.

Goodman, L.A. (1965) On simultaneous confidence interval intervals for multinomial distributions. Technometrics 7:247-254.

Guo, S. W., and Thompson, E.A. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361-372.

Jeffreys, A. J., Brookfield, J. F. Y., and Semeonoff, R. (1985) Positive identification of an immigration test-case using human DNA fingerprints. Nature, 317:818-819.

Jeffreys, A.J., Turner, M., and Debenham, P. (1991) The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework. Am J Hum Genet 48:824-840.

Karlin, S., Cameron, E.C., and Williams, P.T. (1981) Sibling and parent-offspring correlation estimation with variable family size. Proc. Nat. Acad. Sci. USA 78:2664-2668.

Li, C.C. (1976) First Course in Population Genetics. Boxwood, Pacific Grove, CA.

Monson, K.L. and Budowle, B. (1989) A system for semi-automated analysis of DNA autoradiograms, In: Proc. Intl. Symp. Forensic Aspects of DNA Analysis, Washington DC: U.S. Government Printing Office

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., and White, R. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 237:1616-1622

National Research Council, The Evaluation of Forensic DNA Evidence, Committee / Commission on DNA Forensic Science: An Update, Washington DC: National Academy Press, 1996

Nei, M. (1978) Estimates of average heterozygosity and genetic distance from a small number of individuals. Genetics, 89:583-590.

Rao, C.R. (1973) Linear Statistical Inference and Its Applications. Wiley, New York.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503-517.

Steinberger, E.M., Thompson, L.D., and Hartmann, J.M. (1993) On the use of excess homozygosity for subpopulation detection. Am J Hum Genet 52:1275-1277.

Weir, B. S. (1991) Genetic Data Analysis. Sunderland MA, Sinauer Associates

Weir, B.S. (1992) Independence of VNTR alleles defined as fixed bins. Genetics 130:873-887.

Weir, B.S. and Hill, W.G. (1993) Population genetics of DNA profiles. J Forensic Sci Soc 33(4): 218-225.

Wyman, A.R., and White, R. (1980) A highly polymorphic locus in human DNA. Proc Nat Acad Sci USA 77:6754-6758.

9 Troubleshooting

The programs in DNATYPE should run predictably, and as indicated, using the example database files included with the program.

Some of the programs can terminate and either give arcane error messages, or close the DOS window, if there is something wrong with the database filename entry or with the database file itself.

Make sure that the database filename is correctly typed, and that the complete filename was typed. For example, for a database file named RdbExC.txt, a user must type the whole name, including the extension, not just RdbExC.

If file names were completely and correctly typed, database file format is the next place to look if a program should terminate unexpectedly when it is executed. Check to be sure that there are no blank rows in the data file. For example, a blank row after the first (labels) row may be treated as an error, or may be treated as a blank data entry row. Similarly, insure that there are no blank rows after the last data entry row and the end signal row (-1,-1,-1, etc.).