

# GPMDNA Program User Manual

## Table of Contents

1 Introduction.....	2
2 Glossary.....	3
3 Running the program.....	4
4 The GPMDNA Input Language.....	6
5 Viewing, Entering and Editing Profiles.....	7
5.1 Select Profiles.....	7
5.2 New.....	7
5.3 Edit.....	7
5.4 Delete.....	8
5.5 Clear.....	8
5.6 Save.....	8
5.7 Dataset.....	8
5.8 Metadata.....	8
6 Import/Export.....	9
7 Database Management.....	10
8 Searching for matching profiles.....	11
8.1 Selecting Crime and Reference Datasets.....	11
8.2 Match parameters.....	11
8.2.1 Frequency database.....	12
8.2.2 Deltas.....	12
8.2.3 Relationships to match.....	12
8.2.4 LR threshold.....	12
8.3 Processors.....	13
8.4 Database.....	13
9 Results.....	16
9.1 Save Results.....	16
9.2 Clear Results.....	16
10 Message window.....	17

# 1 Introduction

The GPMDNA Program looks for matches between forensic DNA profiles.

DNA profiles may be imported or entered manually, and are stored in a database. The software supports *probabilistic* data entry. I.e. when the identity of an allele can not be determined with certainty, a probability distribution may be entered. This is useful when interpreting poor quality EPGs (partial or low-template samples), and mixtures.

Probabilistic interpretation of EPGs should be performed by an SO.

The user selects two sets of profiles from the database (The “crime” set and the “reference” set) for comparison. One-to-one, one-to-many and many-to-many comparisons are supported.

The software is able to detect a possible *identity match* between two profiles (i.e. the samples may have come from the same person), or a possible *familial match* (i.e. the samples may have come from people with a particular familial relationship, e.g. siblings, or parent-child).

The strength of a match is reported as a *Likelihood Ratio (LR)*. LRs are calculated relative to the null hypothesis that the profiles come from unrelated people.

**WARNING! The LR is a statistical measure of the strength of the genetic evidence in favour of the tested hypothesis, compared to the null hypothesis. THE LR DOES NOT HAVE A SIMPLE INTERPRETATION AS THE PROBABILITY OR THE ODDS THAT THE TESTED HYPOTHESIS IS TRUE. The interpretation of LRs should be undertaken by trained individuals.**

When the software is run on a computer with one or more NVIDIA CUDA-capable graphics cards, the software can make use of hardware acceleration to perform one-to-many and many-to-many comparisons many times faster than is possible on the CPU.

## 2 Glossary

<b>DNA</b>	Deoxyribonucleic acid
<b>EPG</b>	Electropherogram
<b>Identifiler</b>	AmpFlSTR® Identifiler® PCR Amplification Kit (Applied Biosystems)
<b>LR</b>	Likelihood Ratio
<b>MRGS</b>	Military Grid Reference System
<b>PCR</b>	Polymerase Chain Reaction
<b>PMF</b>	Probability Mass Function
<b>Powerplex 16</b>	PowerPlex® 16 System (Promega)
<b>SGM</b>	Second Generation Multiplex PCR Amplification Kit (Applied Biosystems)
<b>SGM+</b>	AmpFlSTR® SGM Plus® PCR Amplification Kit (Applied Biosystems)
<b>SO</b>	Scientific Officer
<b>STR</b>	Short Tandem Repeat

### 3 Running the program

The GPMDNA software runs on desktop and laptop computers under the CentOS Linux operating system. To make use of hardware acceleration an NVIDIA CUDA graphics card is required. For the greatest processing power a workstation with multiple GPGPUs is used. However all functionality is available on low-powered laptops.

The GPMDNA GUI is started by running a script named `gmatch.sh` at the command prompt.

```
> gmatch.sh
```

A number of defaults and options are controlled by environment variables set within this script.

Environment variable	Values	Notes
POPDATA	Path to default frequency database	May be set within the Match Parameters tab of the GUI
CRIME_DELTA	Default value of C-Set delta (error) value	May be set within the Match Parameters tab of the GUI
REF_DELTA	Default value of R-Set delta (error) value	May be set within the Match Parameters tab of the GUI
UNKNOWN_ALLELES	IGNORE	Ignore alleles not in the frequency database
	ADD_AS_RARE	Add allele at an appropriate low frequency (for this match only)
PROC	CPU	Default choice of processor. May be set within the Processors tab of the GUI
	GPU	
PRINT_PROFS	true	Whether to print profile details in the message are of the GUI
	false	

*Table 1: GPMDNA environment variables*

The appearance of the GUI at start-up is illustrated in Figure 1.



## 4 The GPMDNA Input Language

When entering data the user is required to enter a value for each allele at each locus in the sample. This is illustrated in Figure 2.

Sample ID	Profile ID	N in mix	Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01
S001	P001	1	1	11	11				
			2	12	F				

Figure 2: Data entry area

A column is provided for every locus in all the commonly used forensic genetics STR PCR amplification kits, including Identifiler, Powerplex 16, Powerplex ESI 17, NGM Select, GlobalFiler.

In accordance with convention, an allele is represent in the format *repeats.variant*, and an unknown allele (ie for which there is no evidence) is represented as 'F'.

In addition the GPMDNA program provides an extended notation the *GPMDNA input language*, in which probabilistic statements may be made. This is summarised in Table 2.

Example	Comment
12.3	Allele name in the format <i>repeats.variant</i>
F	Unknown (not measured). The background distribution is used.
12.3@0.6	Allele with associated certainty: <i>allele@p-value</i> . If <i>p-value</i> <1 background is added to normalize, so 12.3@0.6 is equivalent to 12.3@0.6/F@0.4.
12.3@0.6/13@0.3	A list of alleles separated by '/' (OR). In this case F @ 0.1 would be added automatically in order to normalise the PMF.
(11/12/12.3)@B	The listed alleles at relative background frequencies, normalised to 1.
(11/12/12.3)@B@0.6	The listed alleles at relative background frequencies, normalised to the indicated p-value. In this case F @ 0.4 would be added automatically.
D	Indicates all the alleles observed in the sample at this locus {S}, at background frequencies, i.e. {S}@B.
D@0.8	Equivalent to {S}@B@p-value In this case F @ 0.2 would be added automatically.

Table 2: GPMDNA input language

## 5 Viewing, Entering and Editing Profiles

Profiles are viewed, entered and edited in the *input area* at the top of the GUI main window.

### 5.1 Select Profiles

The *Select Profiles From...* button in the input area pops up the *Select Profiles* dialog (Figure 3), which allows one or more profiles to be selected by key for display in the input area. The select profiles dialog is populated with either all the profiles in the database, or the profiles selected in the C-Set or R-Set tabs (see Section 8.1 ). This is controlled by the selector to the right of the button.

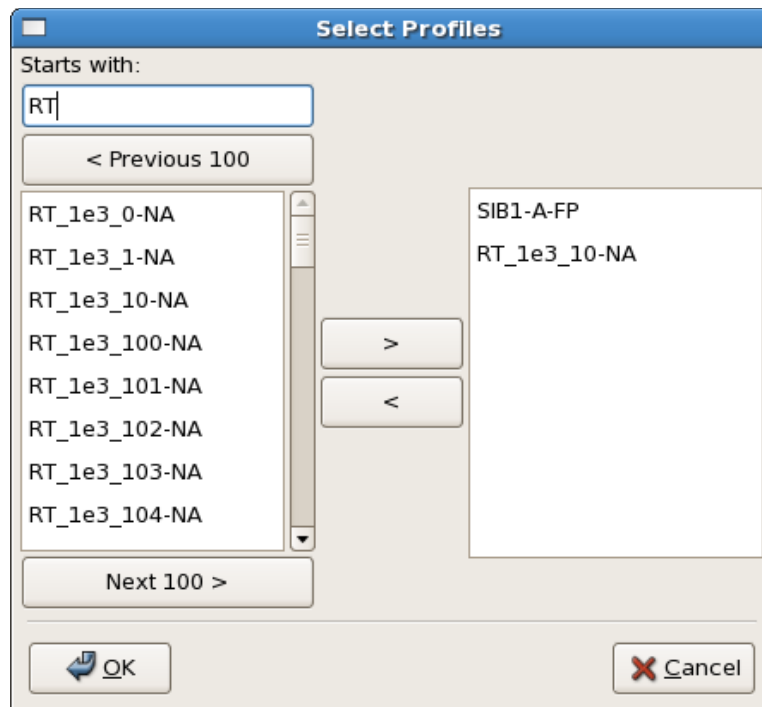


Figure 3: Select Profiles dialog

### 5.2 New

The *New sample* option in the *Edit* menu, and the *New Sample* button in the input area, initialize the input area for entry of a new sample.

The user must enter a Sample ID and a Profile ID, which must together be unique in the profile database.

A sample may consist of several profiles, and a profile may have one or more components. (A profile with more than one component is a mixture).

The number of alleles to be entered (at each locus) is equal to twice the number of components.

### 5.3 Edit

When the input area is being used to view a single profile, the *Edit profile* option in the *Edit* menu and the *Edit* button in the input area put the input area into edit mode, in which the profile (and its

metadata) may be edited. When viewing multiple profiles the edit function is not available.

#### **5.4 Delete**

The *Delete profile* option in the *Edit* menu, and the *Delete* button in the input area, delete the currently displayed profile from the database.

#### **5.5 Clear**

The *Clear profile* option in the *Edit* menu, and the *Clear* button in the input area, clear the current entry in the input area. (The profile stored in the database is not affected).

#### **5.6 Save**

The *Save profile* option in the *Edit* menu, and the *Save* button in the input area, allow an edited profile, (or the profiles of a newly entered sample) to be saved to the database.

NB: When a new sample is saved, each profile is saved separately in the database. Each profile can be recalled and edited independently, but they can not again be edited as a single sample.

#### **5.7 Dataset**

Each profile is assigned a *dataset* when it is first saved to the database. The dataset serves as a means of grouping profiles, and its meaning is a matter for the user. The dataset of the currently viewed profile is displayed at the top right of the input area. When in edit mode, the dataset may be edited.

#### **5.8 Metadata**

Each profile may have values assigned to a number of *metadata* fields. The metadata for the current profile in the input area may be viewed or edited via the Metadata dialog, which is displayed by the Metadata button at the top right of the input area.



## 6 Import/Export

The *Import* option in the *File* menu pops up the *Import Profiles* dialog (Figure 4). This dialog allows profiles to be imported in a number of common formats.

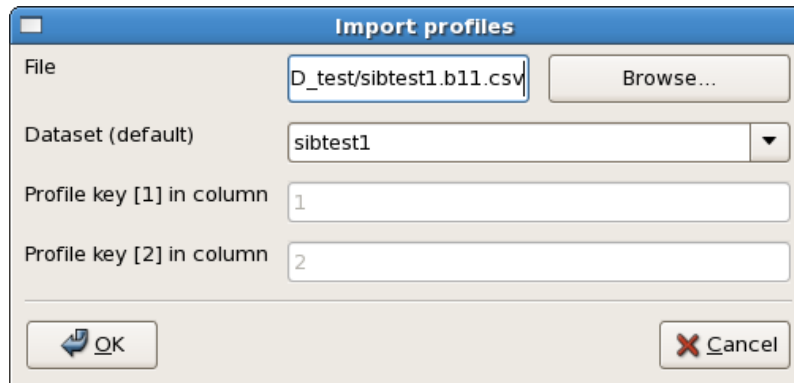


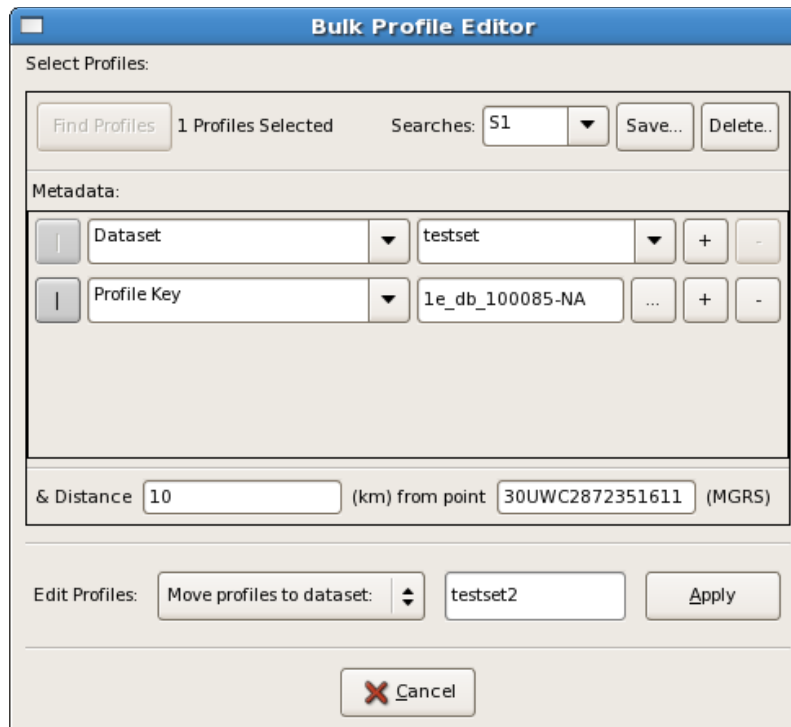
Figure 4: Import profiles dialog

## 7 Database Management

The *bulk editor* option in the *Edit* menu pops up the *Bulk Profile Editor* dialog (Figure 5). In the top part of this dialog is a *Select Profiles Panel* allowing a set of profiles to be selected based on dataset, profile key, metadata fields and within a given radius of a fixed location. Locations are specified in MRGS.

The *Find Profiles* button searches the database for profiles matching the specified criteria and displays the number of profiles found. The selected profiles may be either moved to a different dataset, or deleted.

Search parameters may be saved and recalled by name using the controls at the top right of the panel.



The **Bulk Profile Editor** dialog box is designed for selecting and managing profiles. It features a **Select Profiles:** section at the top with a **Find Profiles** button, a status indicator showing **1 Profiles Selected**, and search controls including a **Searches:** dropdown set to **S1**, **Save...**, and **Delete..** buttons. Below this is a **Metadata:** section with two rows of controls. The first row includes a **Dataset** dropdown, a text field containing **testset**, and **+** and **-** buttons. The second row includes a **Profile Key** dropdown, a text field containing **1e\_db\_100085-NA**, an ellipsis button, and **+** and **-** buttons. Further down, there is a section for distance and location with a text field for **& Distance** set to **10**, the unit **(km)**, the text **from point**, a text field for the point **30UWC2872351611**, and the unit **(MGRS)**. At the bottom, the **Edit Profiles:** section contains a **Move profiles to dataset:** dropdown, a text field containing **testset2**, and an **Apply** button. A **Cancel** button with a red X icon is located at the very bottom center.

Figure 5: Bulk Profile Editor

## 8 Searching for matching profiles

A profile match is performed between two sets of profiles selected from the database: notionally a “Crime set” denoted by C-Set, and a “Reference set” denoted by R-Set.

Typically, the C-Set consists of profiles recovered from one or more crime scenes, and the R-Set consists of the profiles of known individuals. However arbitrary sets of profiles may be matched and it does not matter which name is assigned to which set of profiles.

Each profile in the C-Set is matched against each profile in the R-Set. If there are  $N$  profiles in the C-Set and  $M$  profiles in the R-Set then  $N \times M$  pairs of profiles must be matched. If both sets are large then clearly a very large number of individual matches must be performed, and the calculation may take a long time. In such cases the use of hardware acceleration is advantageous (see Section 3 ).

A set of profiles (nominally the C-Set) may be matched against itself. Each profile is matched against every other profile, requiring  $N(N-1)/2$  matches, i.e. approximately  $N^2/2$  matches for large  $N$ .

### 8.1 Selecting Crime and Reference Datasets

The crime and reference datasets are selected using the C-Set and R-Set tabs in the tabbed pane at the left of the GUI (Figure 1). Each of these tabs holds a Select Profiles panel as described in Section 7 for the Bulk Profile Editor (Figure 5).

Note that all three Select Profiles panels in the application (C-Set, R-Set, Bulk Profile Editor) use the same set of named searches, that may be defined or edited in any one of them.

### 8.2 Match parameters

The *Match parameters* tab (Figure 6) allows selection of the frequency database for use in LR calculations, setting delta factors, choosing relationships to match, and setting the LR threshold.

Figure 6: Match parameters tab

### 8.2.1 Frequency database

Frequency databases are provided as text files in external directories. A number of frequency databases are provided with the GPMDNA program. A frequency database is selected using the Browse button in the top right of the tab.

### 8.2.2 Deltas

The crime and reference delta values provide a crude model of errors and mutations. Delta represents the probability that an allele has changed because of an error in the analysis (eg a 'typo') or because of a mutation. See (1) Section 8 and (2) for more details.

### 8.2.3 Relationships to match

Version 1.0 of the software offers the relationships shown in Table 3.

Name	Relationships
IDENT	Identity
D1	Child, Parent.
SIB	Full sibling.
D2	Grandchild, Grandparent, Uncle, Aunt, Nephew, Niece, half-sibling.
D3	Great-grandchild, Great-grandparent, first cousin, (and others ...)

Table 3: GPMDNA relationships

NB: Unless a familial relationship can be excluded, at least the closest familial relationships (D1 and SIB) should always be run alongside IDENT. See (2) for further explanation.

### 8.2.4 LR threshold

Matches are reported in order of descending LR. The LR threshold provides a bound below which

matches are not reported. This is useful when matching familial relationships with large databases, since a large number of low-LR matches will be generated by chance.

**WARNING! When using hardware acceleration, there is a limit on the number of matches than can be reported back to the CPU. If the LR threshold is set too low, this limit may be exceeded, and further matches (even at higher LRs) will be missed. Look out for warnings of this condition reported in the message window, and increase the LR threshold accordingly.**

### 8.3 Processors

On platforms with NVIDIA CUDA-capable GPUs, the choice of processors to use is controlled from the Processors tab (Figure 7).

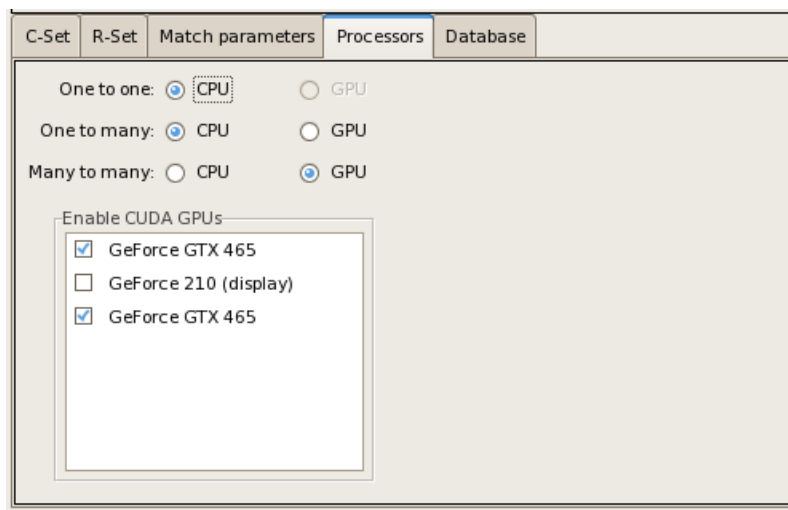


Figure 7: Processors tab

The default behavior is to use the CPU for one-to-one and one-to-many matches, and to use GPUs for many-to-many matches. By default all available GPUs are used except any that are used to drive a display, unless there is only one CUDA-capable GPU, when this is used in any case. However this is not always appropriate and the user may determine which processors to use.

Note that if a GPU is used for a display, any other process (such as GPMDNA) that occupies the GPU for more then five seconds may be terminated by the system. This limits the use of hardware acceleration on single-GPU machines.

### 8.4 Database

By default the GUI attempts to connect to a local MySQL database. However it is possible to connect to a database over the network, provided appropriate users and permissions have been set up on the server. (Figure 8)



C-Set	R-Set	Match parameters	Processors	Database
Server		<input type="text" value="fandserver"/>		<input type="button" value="Change..."/>
User name		<input type="text" value="fanduser"/>		
Password		<input type="text" value="fandpass"/>		
Status:		Connected to fandserver		

*Figure 8: Database tab*

## 9 Results

During the LR calculation the GUI will “freeze”. When the calculation is complete the results are displayed in the lower right pane of the GUI as shown in Figure 9.

Matches are displayed in descending order of LR in the left of the pane. On selecting a result this is displayed in the right of the pane, showing the LR for each relationship tested for these two profiles. The two profiles are displayed in the input/viewing area for ease of comparison.

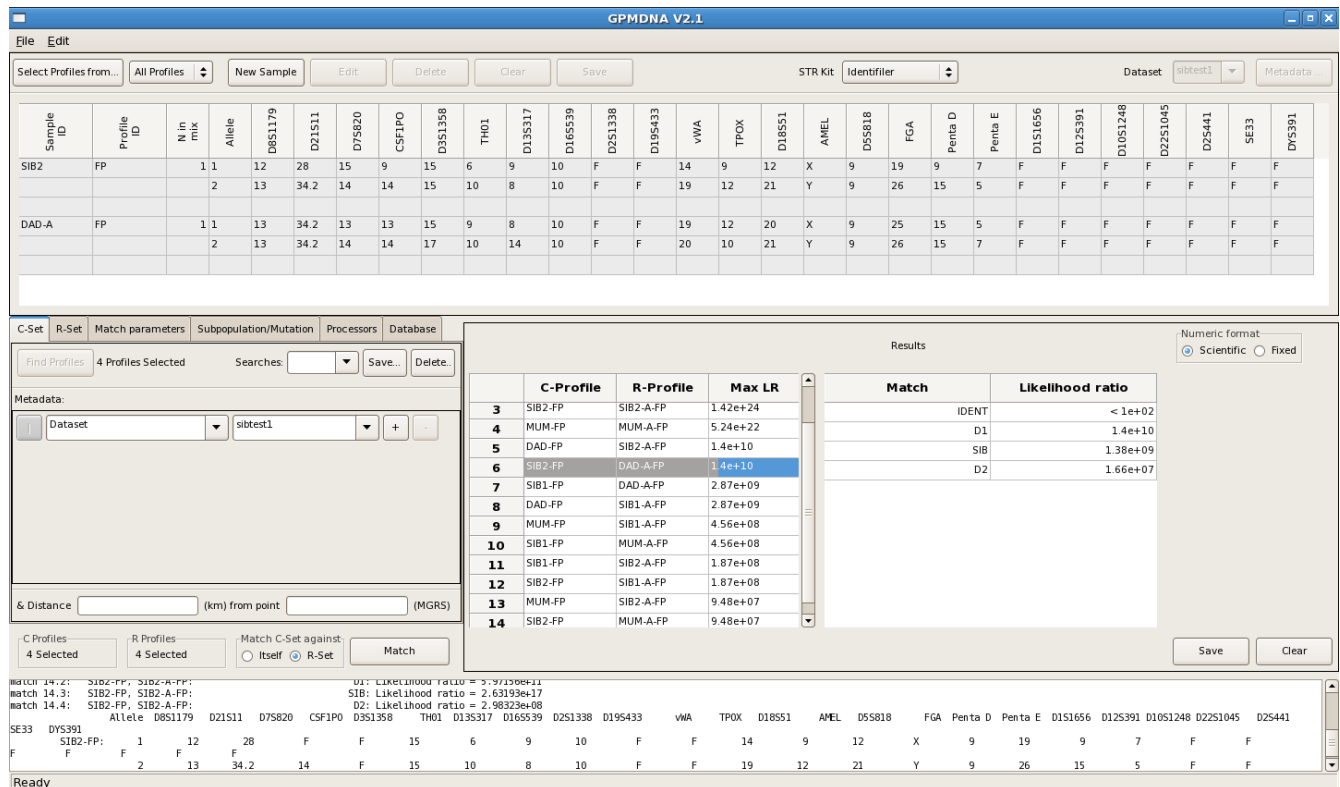


Figure 9: Example of GPMDNA results

### 9.1 Save Results

The Save button at the bottom right of the results pane allows a results set to be exported in CSV format.

### 9.2 Clear Results

The Clear button at the bottom right of the results pane clears the results pane.



## **10      Message window**

Messages including errors and warnings are displayed in the messages window at the bottom of the GUI.