Version 2.1 (Mar, 2008)

User Manual for Software MPDA

The Developing Group of Software MPDA

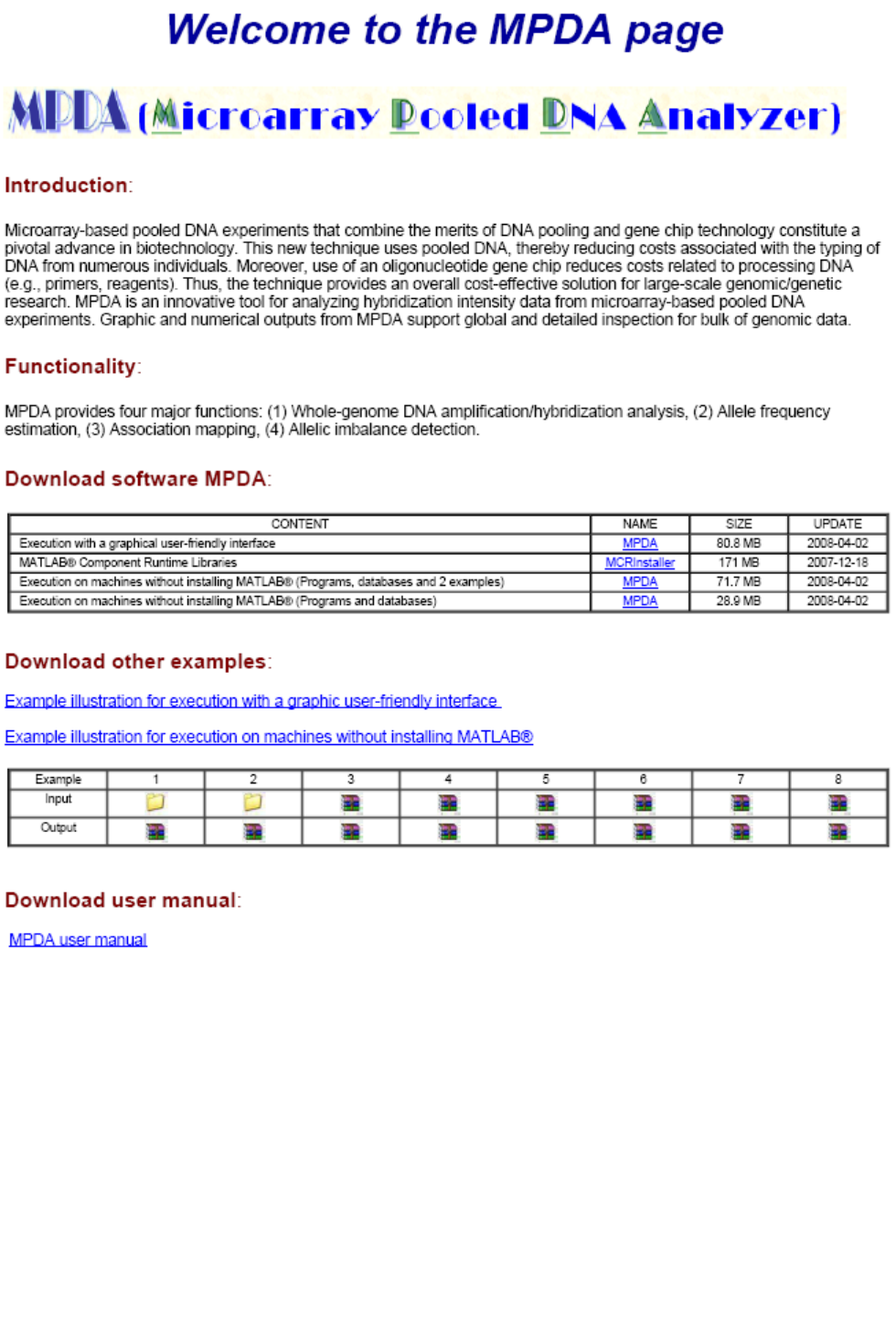
Table of Contents:

1. INTRODUCTION

Software *MPDA* (Microarray Pooled DNA Analyzer) is a powerful tool for pooled DNA data analysis. It can be downloaded at the *MPDA* website (**Figure 1-1**). The main functions of *MPDA* consist of (1) estimating the CPA (coefficient of preferential amplification) and its standard error (s.e.); (2) estimating the AF (allele frequency) and its s.e.; (3) association analysis including single-point pooled DNA association test and multipoint pooled DNA association test; and (4) allelic imbalance analysis including single-point allelic imbalance detection and multipoint allelic imbalance detection.

MPDA was developed under the software platform, MATLAB®, and provided user-friendly interfaces adapted to Windows systems (Windows 98/2000/XP). For users without installing software MATLAB®, we have also developed stand-alone executables generated via the MATLAB® compiler (The newly added function in *MPDA* Version 2.0). In this manual, we outline the downloading and setup procedures in Section 2. We describe the working directories of *MPDA* in Section 3. We introduce the interfaces, analytic functions and operating procedures of *MPDA* in Section 4. Data input formats according to different genotyping platforms and calling algorithms are explained in Section 5. Detailed running procedures of two examples included with *MPDA* are provided in Section 6. Troubleshooting for software installation and execution is given in Section 7. The information of *MPDA* version upgrade is shown in Section 8. Appendices A – D introduce statistical methods used in *MPDA*. In addition, *MPDA* also provides eight examples that the data can be downloaded at the *MPDA* website and the operation procedures of each example can be found in documents, which can also be downloaded at the same website.

Figure 1-1. The *MPDA* website



Welcome to the MPDA page

MPDA (Microarray Pooled DNA Analyzer)

Introduction:

Microarray-based pooled DNA experiments that combine the merits of DNA pooling and gene chip technology constitute a pivotal advance in biotechnology. This new technique uses pooled DNA, thereby reducing costs associated with the typing of DNA from numerous individuals. Moreover, use of an oligonucleotide gene chip reduces costs related to processing DNA (e.g., primers, reagents). Thus, the technique provides an overall cost-effective solution for large-scale genomic/genetic research. MPDA is an innovative tool for analyzing hybridization intensity data from microarray-based pooled DNA experiments. Graphic and numerical outputs from MPDA support global and detailed inspection for bulk of genomic data.

Functionality:

MPDA provides four major functions: (1) Whole-genome DNA amplification/hybridization analysis, (2) Allele frequency estimation, (3) Association mapping, (4) Allelic imbalance detection.

Download software MPDA:

CONTENT	NAME	SIZE	UPDATE
Execution with a graphical user-friendly interface	MPDA	80.8 MB	2008-04-02
MATLAB® Component Runtime Libraries	MCRInstaller	171 MB	2007-12-18
Execution on machines without installing MATLAB® (Programs, databases and 2 examples)	MPDA	71.7 MB	2008-04-02
Execution on machines without installing MATLAB® (Programs and databases)	MPDA	28.9 MB	2008-04-02

Download other examples:

Example illustration for execution with a graphic user-friendly interface

Example illustration for execution on machines without installing MATLAB®

Example	1	2	3	4	5	6	7	8
Input								
Output								

Download user manual:

MPDA user manual

2. SOFTWARE INSTALLATION AND INITIALIZATION

MPDA was developed based on software MATLAB® and adapted to MS Windows® 98/ME/NT/2000/XP/2003. *MPDA* can be executed using different versions of MATLAB®. In addition, stand-alone executables of *MPDA* were also developed. The standalone *MPDA* can be run on machines without installing MATLAB®. We illustrate the installation procedures of the two versions below.

2.1 MPDA with a user-friendly interface (MATLAB® is required)

Installation procedures of *MPDA* using MATLAB® software, version R2006a:

- Step 1 – Download a zip file 'MPDA.rar' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm (**Figure 2.1**).
- Step 2 – Unzip the file 'MPDA.rar' to get the main directory 'MPDA' (**Figure 2.2**).
- Step 3 – Move the directory 'MPDA' to the designated directory where MATLAB® was installed, e.g., 'C:\Program Files\MATLAB\R2006a' (**Figure 2.3**).
- Step 4 – Initialize MATLAB®, and enter its user interface (**Figure 2.4**).
- Step 5 – Click the button 'File' in the command bar of 'MATLAB Command Window', and select the button 'Set Path' (**Figure 2.5**).
- Step 6 – Click the button 'Add Folder', and select the working directory (i.e., 'MPDA') in the designated directory to add folder. Click the button 'Save' to save the path (**Figure 2.6**).
- Step 7 – Click the button 'Add Folder' again, and select the subdirectory 'Database' in the directory 'MPDA' to add a new folder. Click the button 'Save' to save the path (**Figure 2.7**).
- Step 8 – Key the command 'MPDA' in the command line in 'MATLAB Command Window' to enter the *MPDA* environment (**Figure 2.8**).

After finishing the eight steps, the welcome page of *MPDA* will be shown (**Figure 2-9**). Two interfaces of *MPDA* are designed for both the association analysis (**Figure 2-10**) and the allelic imbalance analysis (**Figure 2-11**), respectively.

2.2 MPDA with standalone executables (MATLAB® is not required)

Installation procedures of the standalone *MPDA* and MATLAB® Component Runtime as follows:

- Step 1 – Download an executable file 'MCRInstaller.exe' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm (**Figure 2.12**).
- Step 2 – Execute the executable file 'MCRInstaller.exe' to install MATLAB® Component Runtime (**Figure 2.13**). (If MATLAB® Component Runtime cannot be installed smoothly, please refer to Section 7.1 for troubleshooting.)
- Step 3 – Download a zip file 'MPDA_mcr.rar' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm (**Figure 2.14**).
- Step 4 – Decompress the zip file 'MPDA_mcr.rar' to a destination directory (e.g., 'C:\Program Files\MATLAB\R2006a\MPDA') (**Figure 2.15**).
- Step 5 – Click 'Start' and select 'Run' to open a window of 'Run' (**Figure 2.16**).
- Step 6 – Type command 'cmd' to activate the MS-DOS Prompt window (**Figure 2.17**). (If the appearance of the MS-DOS Prompt window is not proper, please refer to Section 7.2 for troubleshooting.)
- Step 7 – Change the path to the main directory 'MPDA' (**Figure 2.18**).
- Step 8 – Type 'MPDA' and press [Enter] key to execute an executable 'MPDA.exe' (**Figure 2.19**).

After finishing the eight steps, the welcome page of standalone *MPDA* will be shown (**Figure 2.20**).

Figure 2-1. Step 1 – Download a zip file 'MPDA.rar' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm.

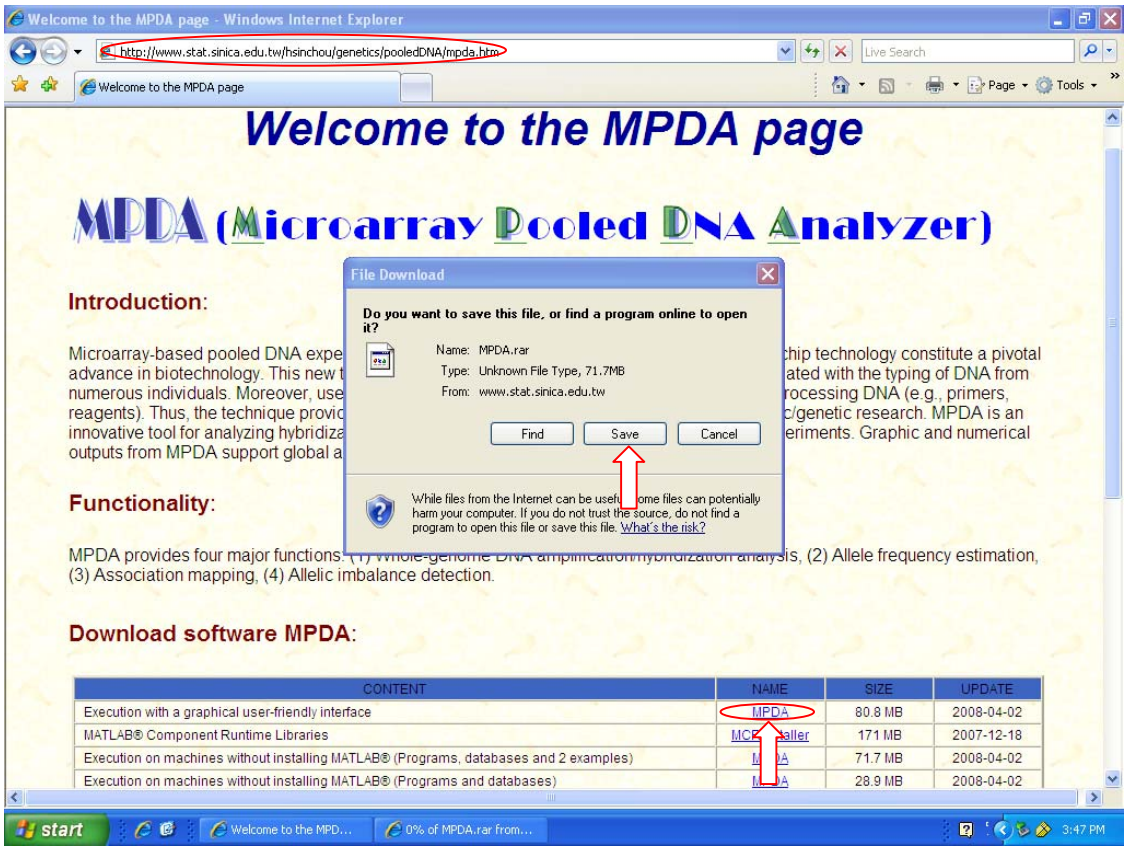


Figure 2-2. Step 2 – Unzip the file ‘MPDA.rar’ to get the main directory ‘MPDA’.

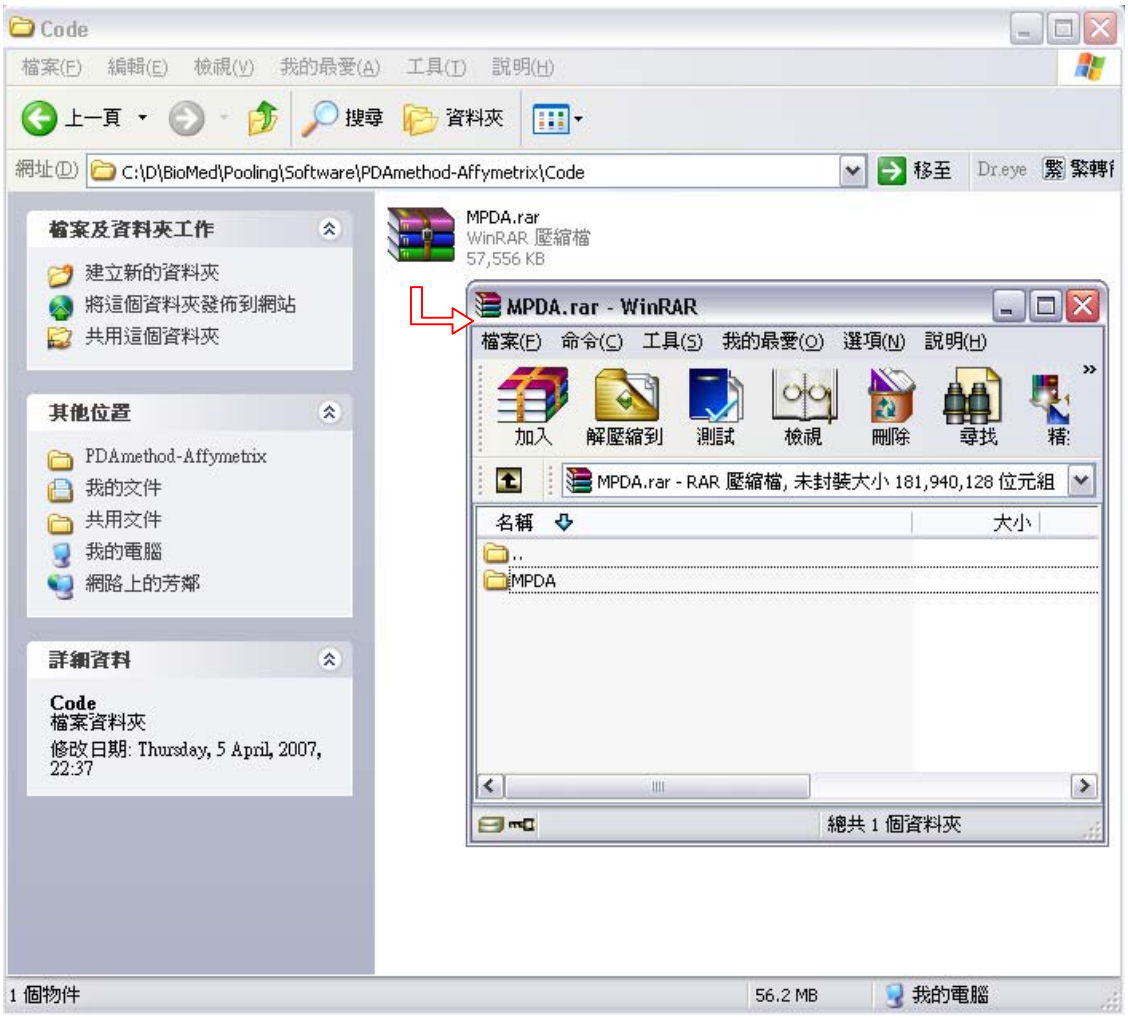


Figure 2-3. Step 3 – Move the directory 'MPDA' to the designated directory where MATLAB® was installed, e.g., 'C:\Program Files\MATLAB\R2006a'.

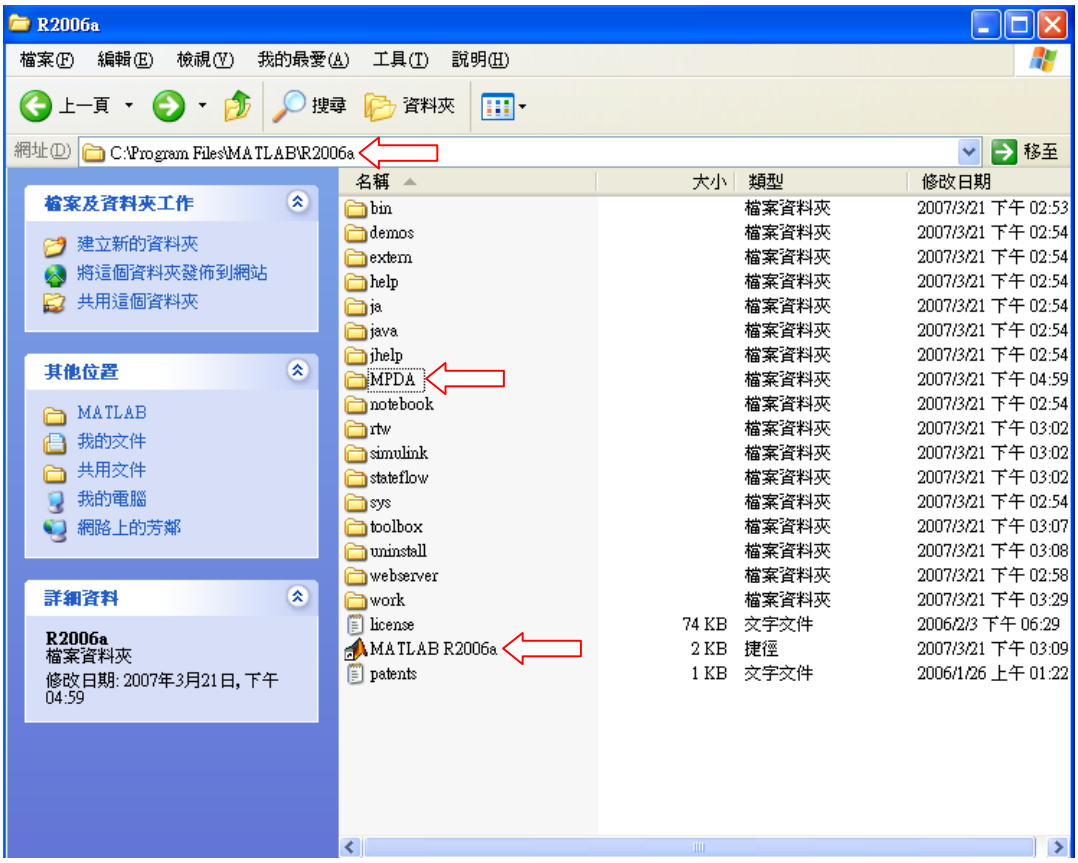


Figure 2-4. Step 4 – Initialize MATLAB®, and enter its user interface.

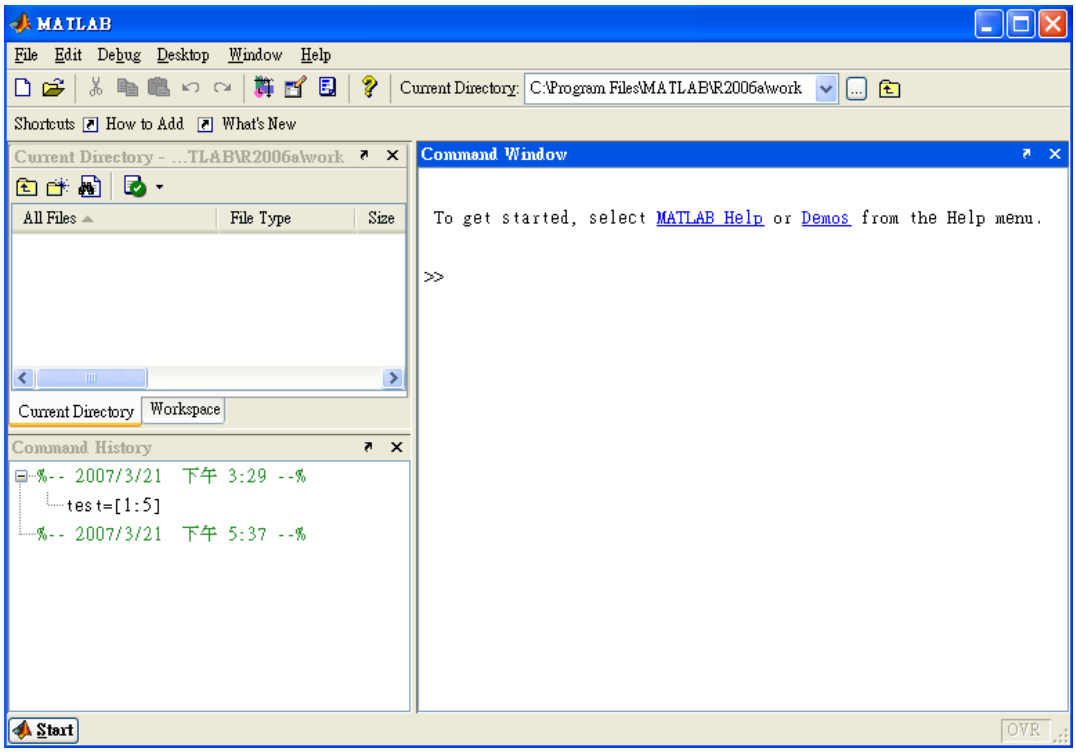


Figure 2-5. Step 5 – Click the button ‘File’ in the command bar of ‘MATLAB Command Window’, and select the button ‘Set Path’.

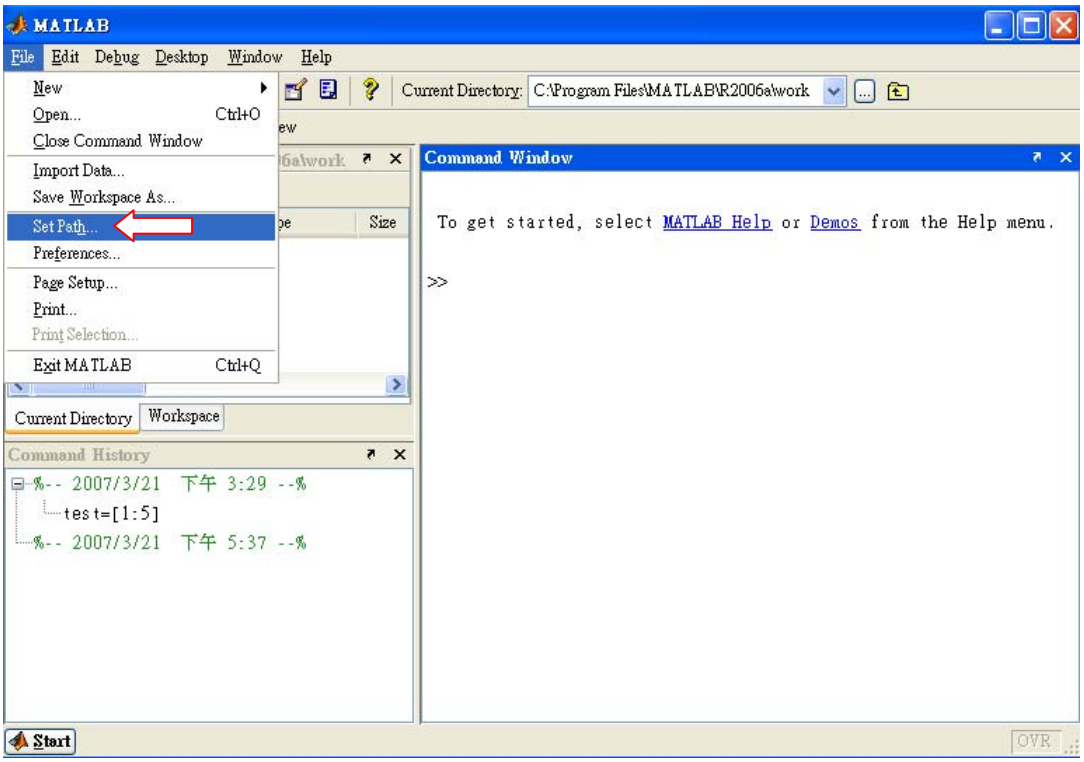


Figure 2-6. Step 6 – Click the button ‘Add path’, and select the working directory (i.e., ‘MPDA’) in the designated directory to add path. Click the button ‘Save’ to save the path.

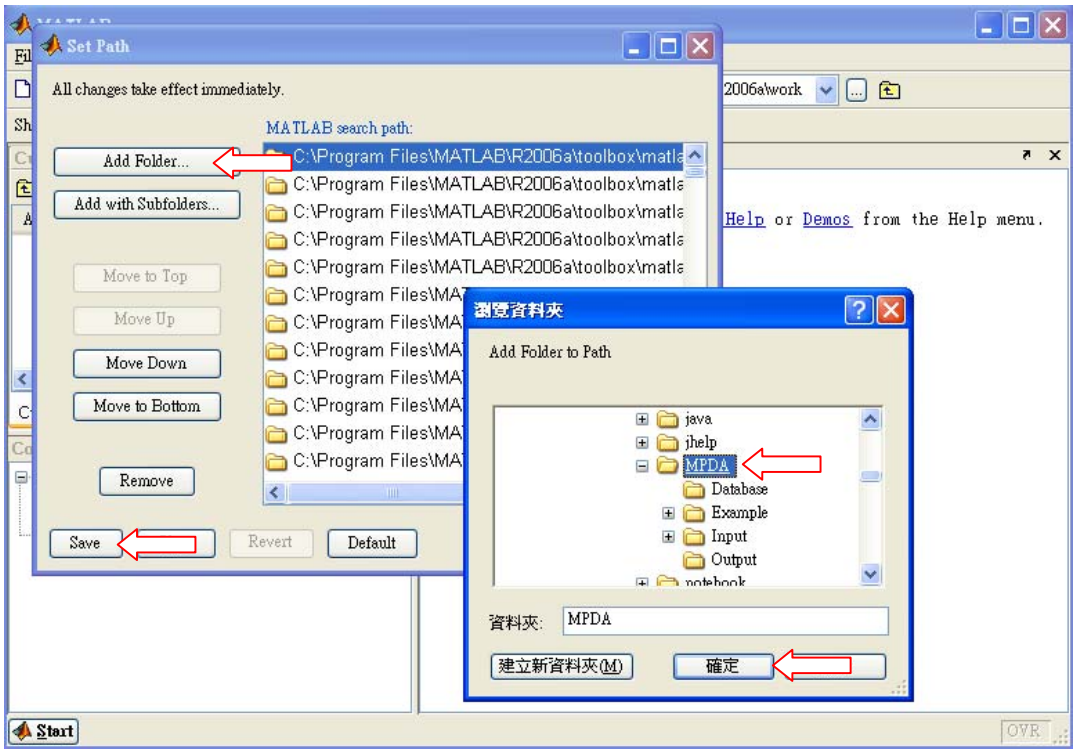


Figure 2-7. Step 7 – Click the button ‘Add path’, and select the subdirectory ‘Database’ to add path. Click the button ‘Save’ to save the path.

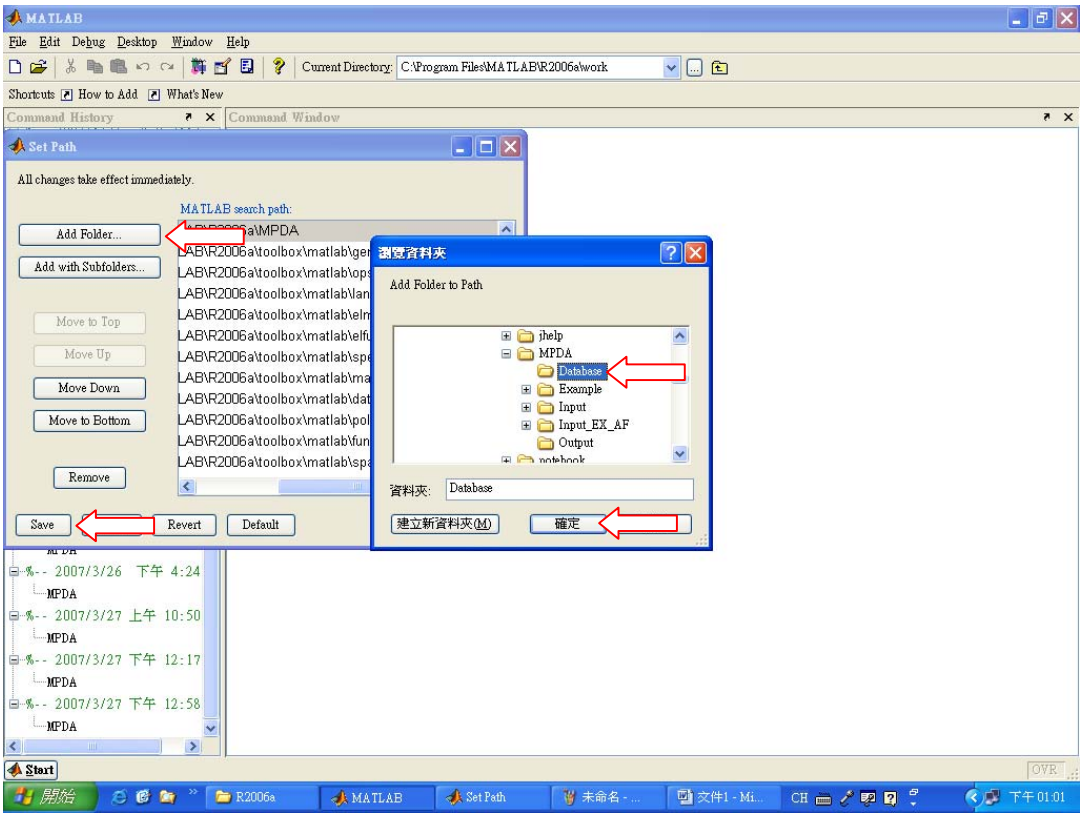


Figure 2-8. Step 8 – Key in the command ‘MPDA’ in the command line in ‘MATLAB Command Window’ to enter the *MPDA* environment.

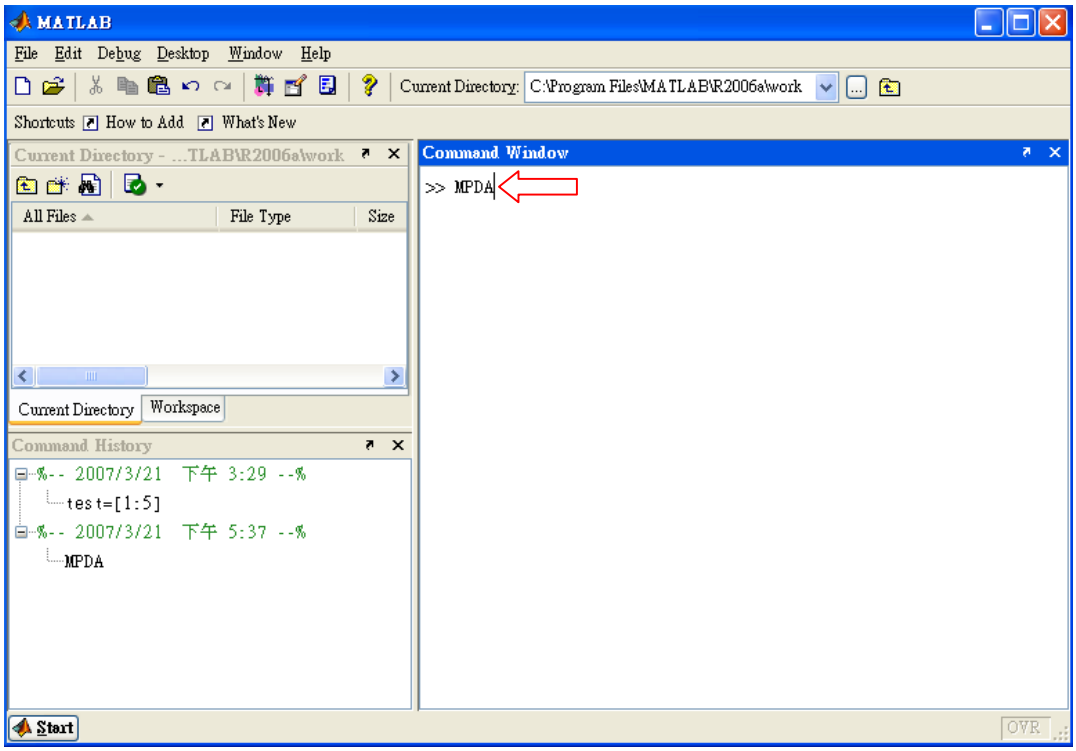


Figure 2-9. Interface 1 of *MPDA* – Main interface

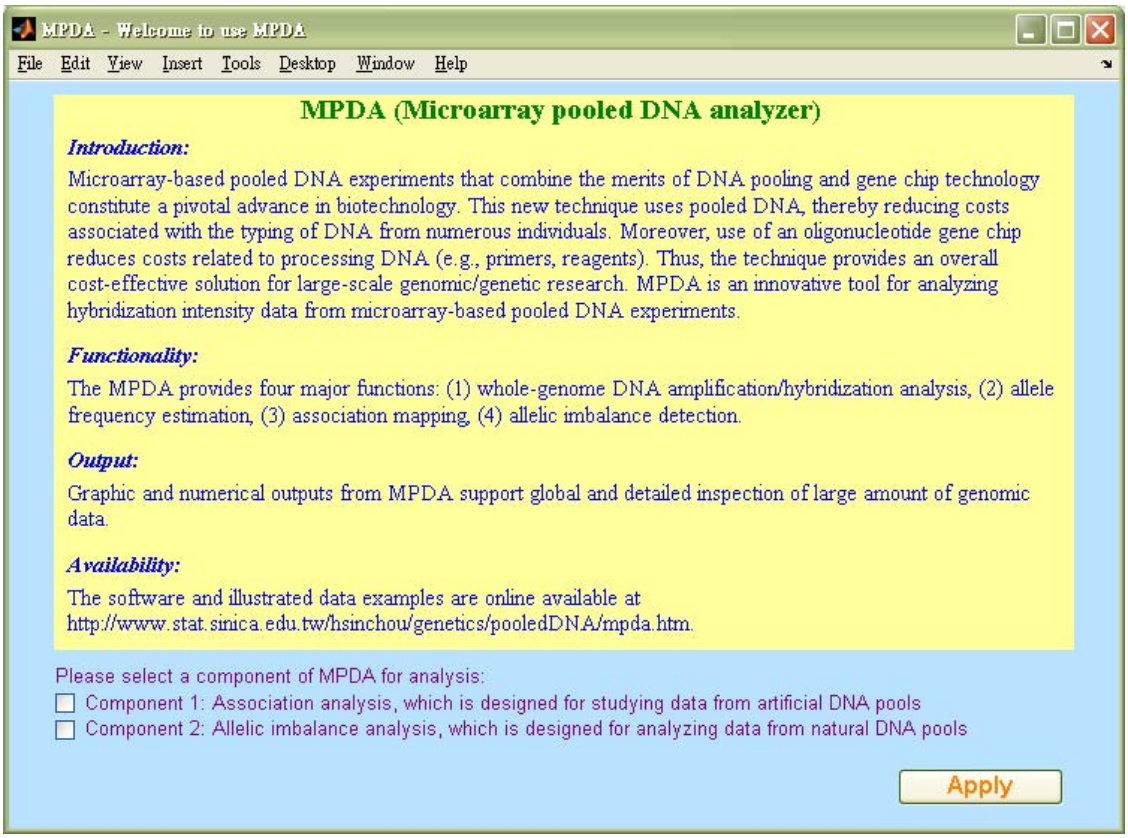


Figure 2-10. Interface 2 of *MPDA* – Association analysis

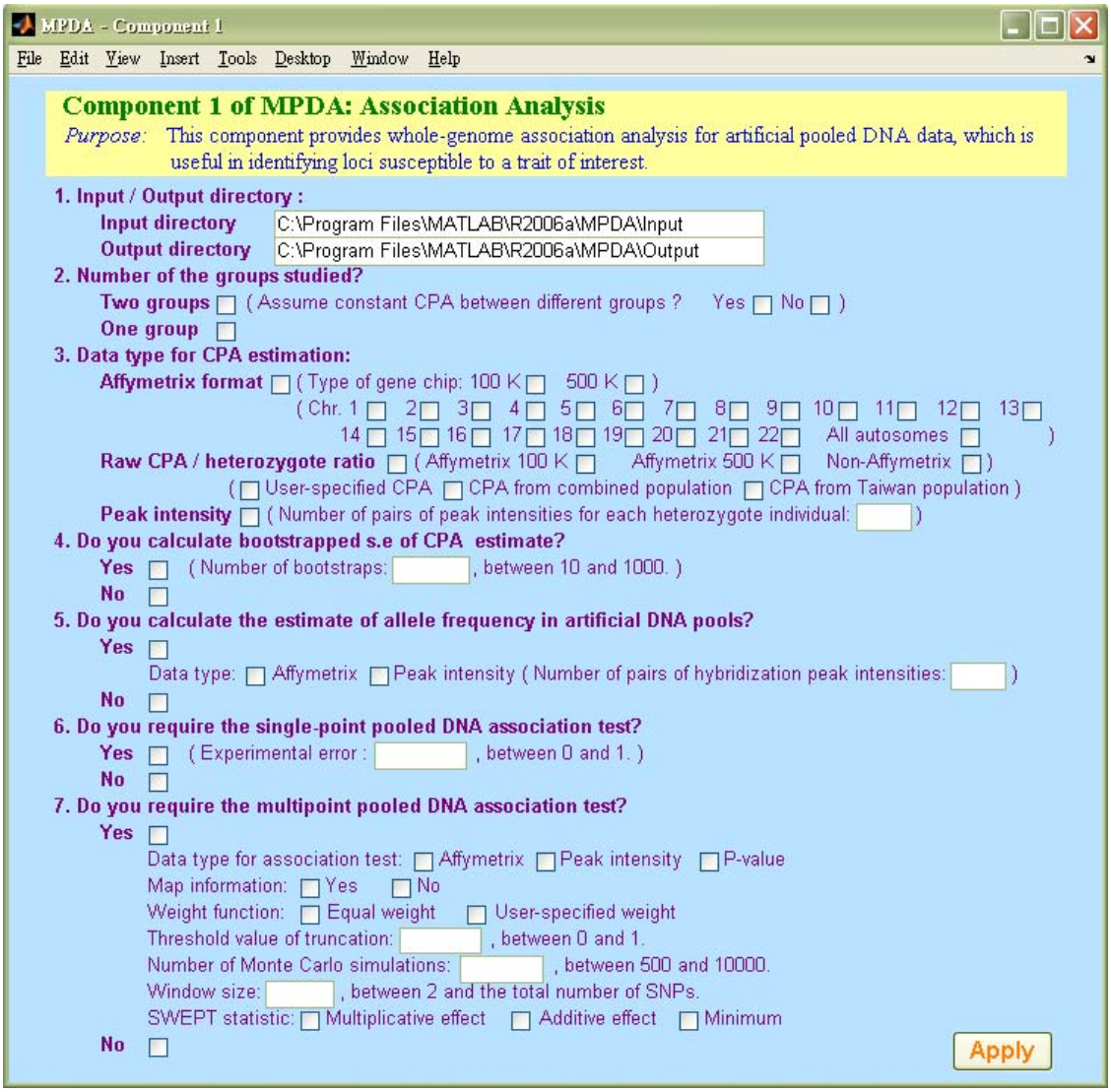


Figure 2-11. Interface 3 of *MPDA* – Allelic imbalance analysis

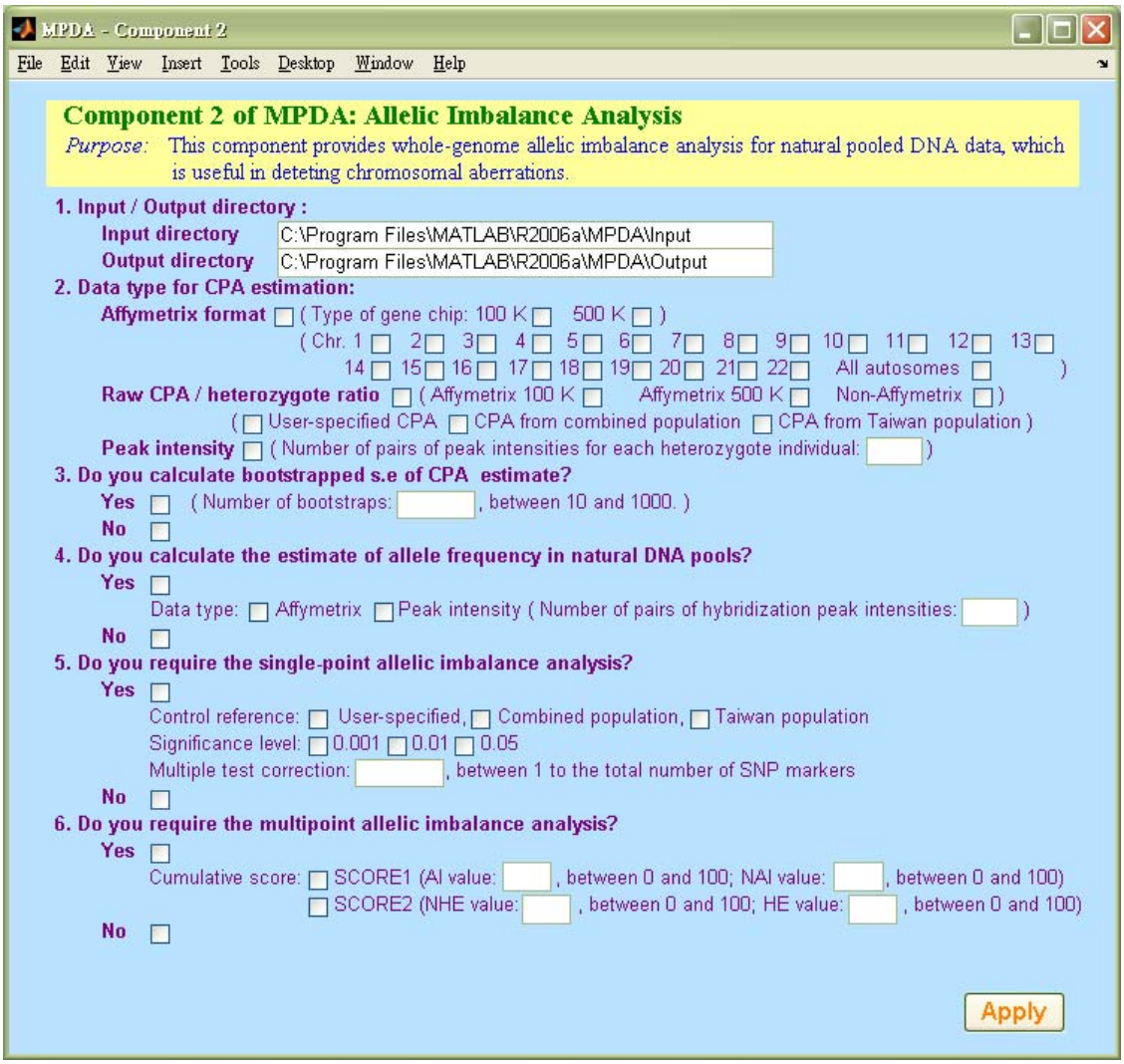


Figure 2-12. Step 1 – Download an executable file 'MCRInstaller.exe' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm

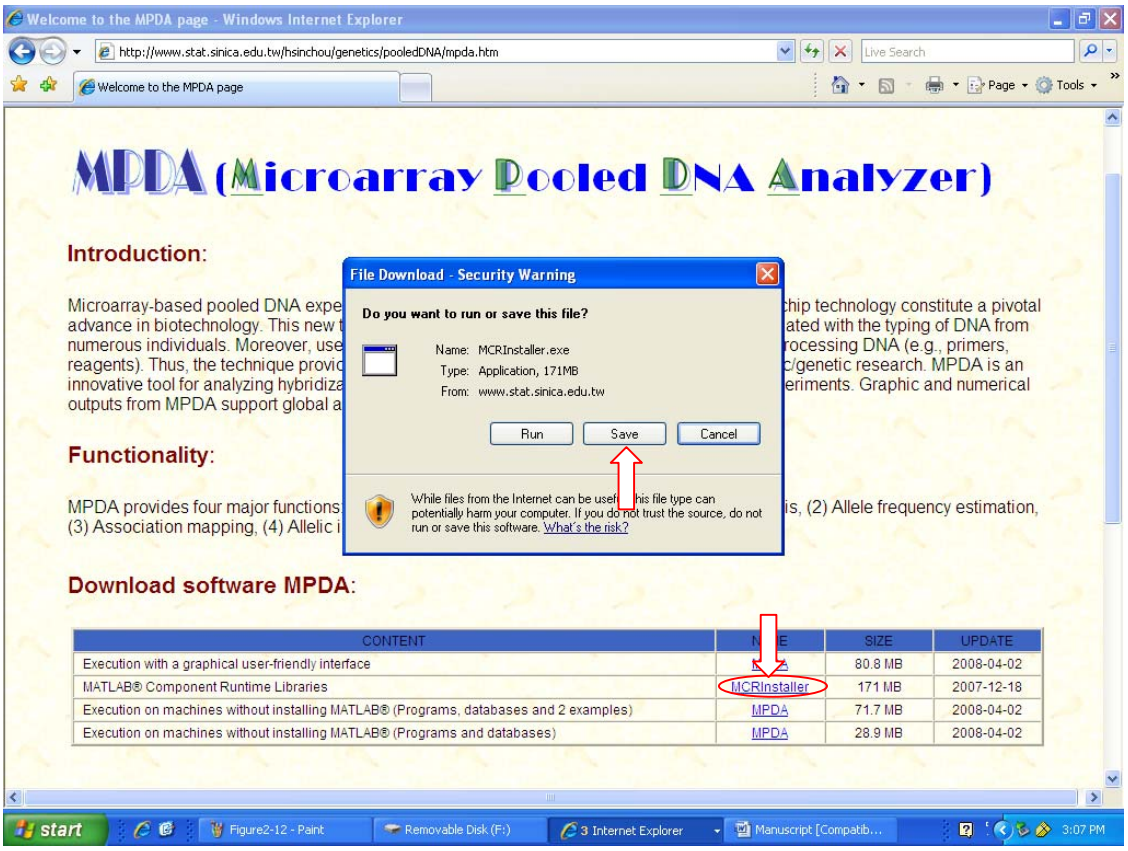


Figure 2-13. Step 2 – Execute the executable file 'MCRInstaller.exe' to install MATLAB® Component Runtime.

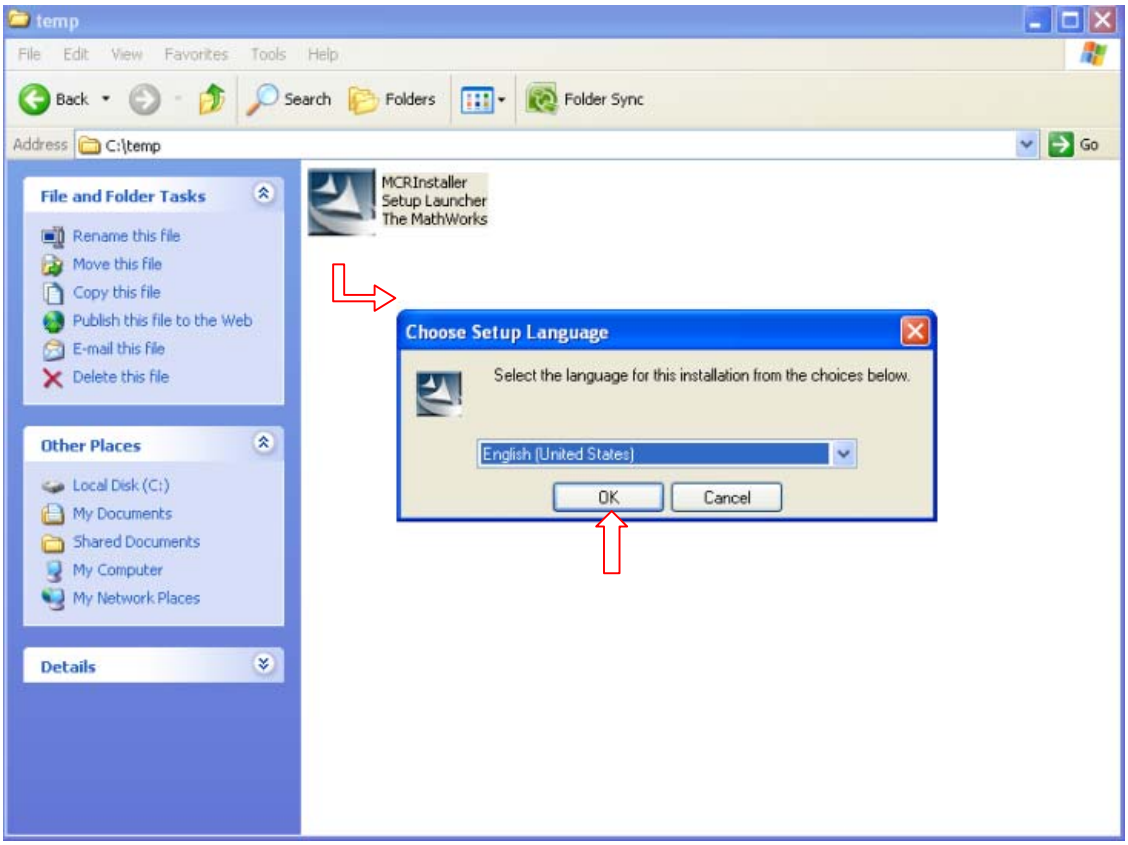


Figure 2-14. Step 3 – Download a zip file 'MPDA_mcr.rar' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm.

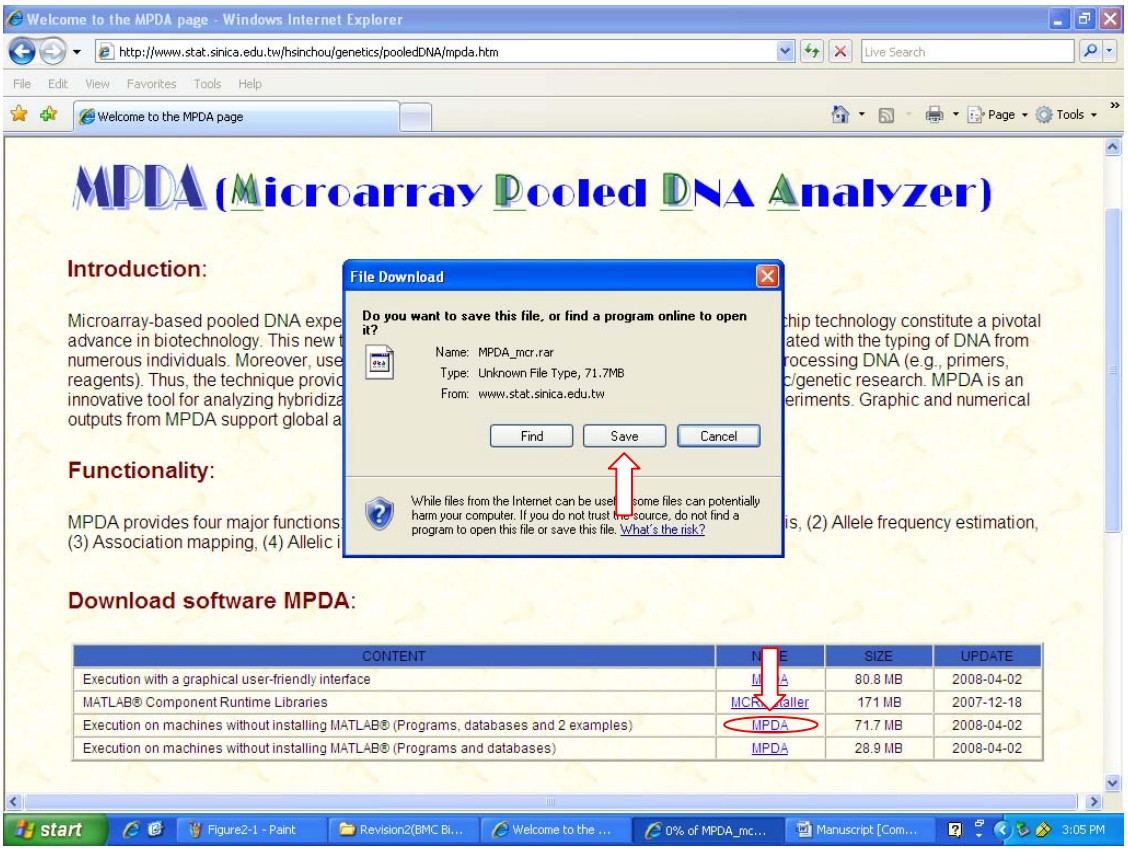


Figure 2-15. Step 4 – Decompress the zip file ‘MPDA_mcr.rar’ to a destination directory ‘C:\Program Files\MATLAB\R2006a\MPDA’.

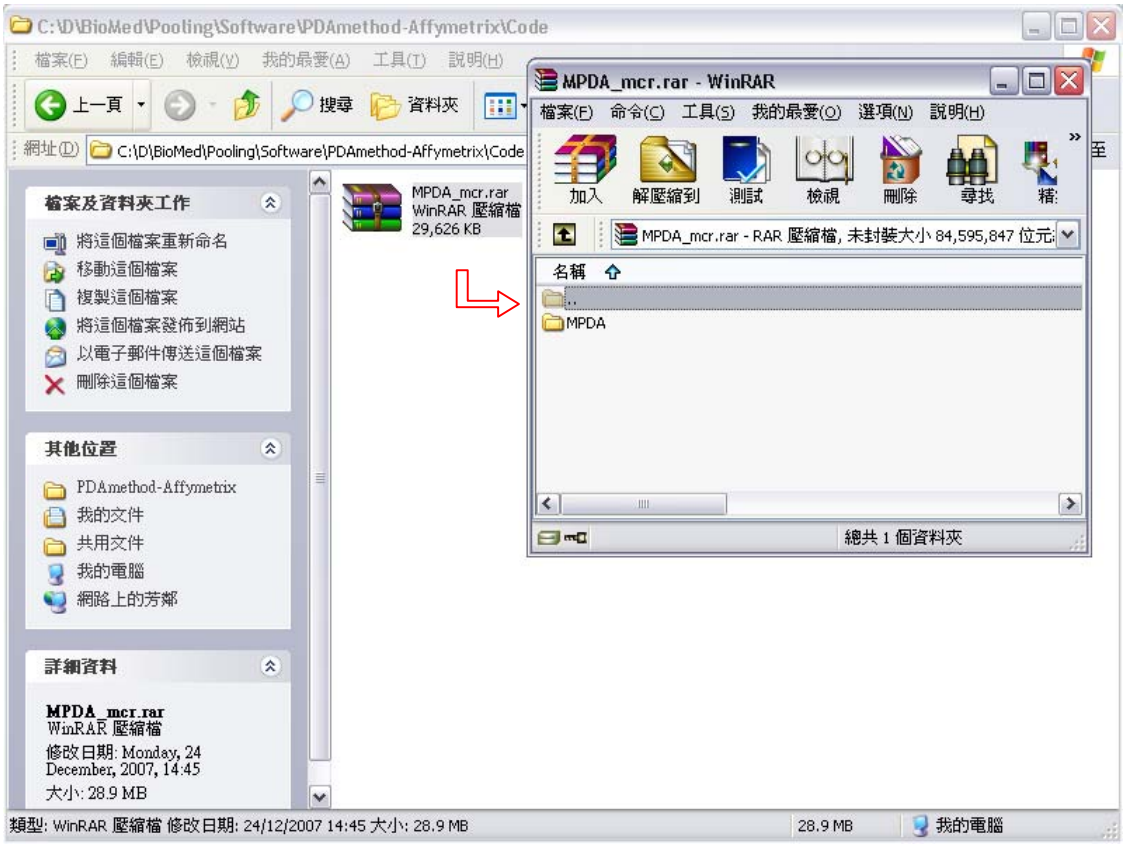


Figure 2-16. Step 5 – Click ‘Start’ and select ‘Run’ to open a window of ‘Run’.

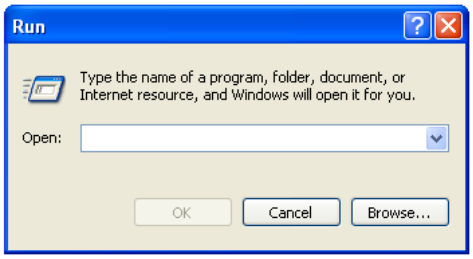


Figure 2-17. Step 6 – Type command ‘cmd’ to activate the MS-DOS Prompt window.

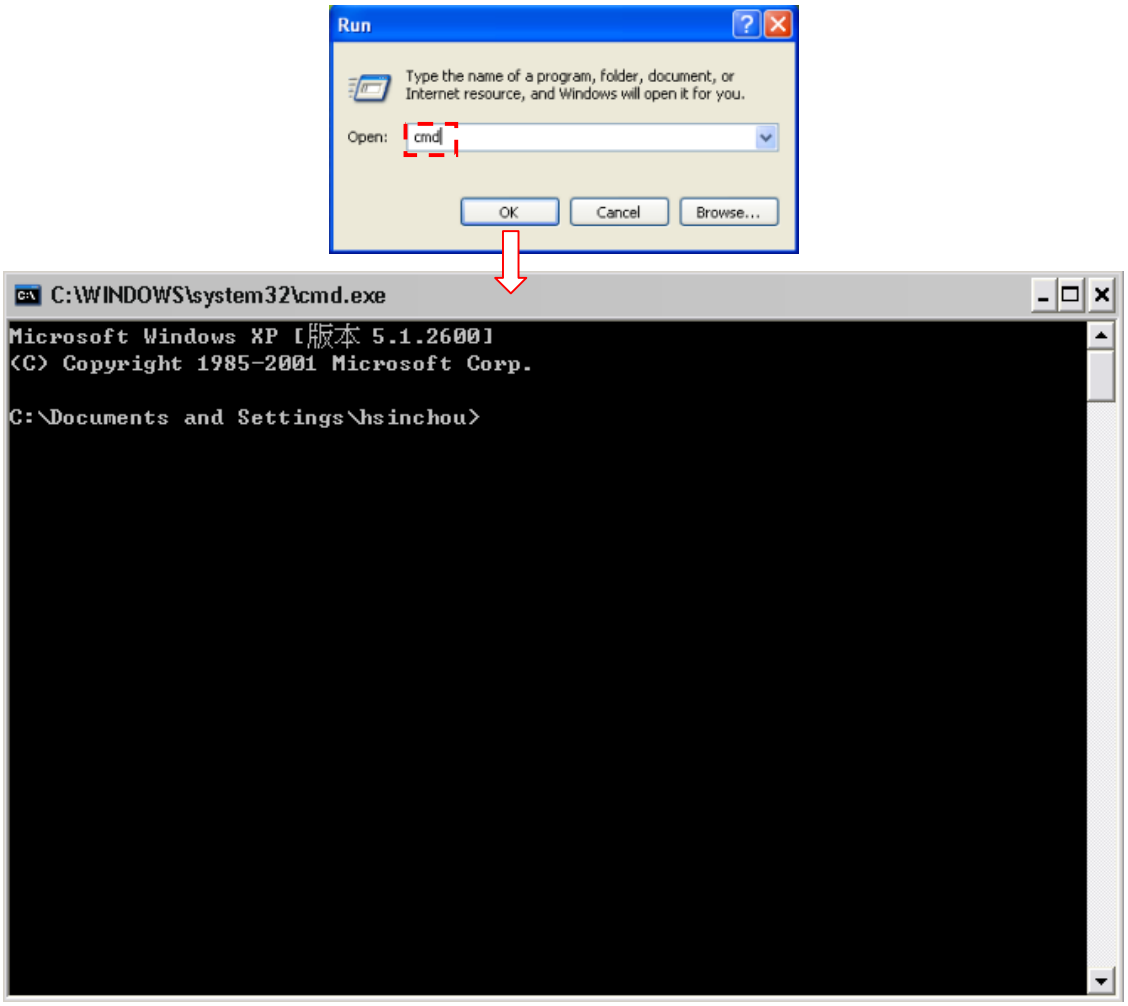


Figure 2-18. Step 7 – Change the path to the main directory ‘MPDA’

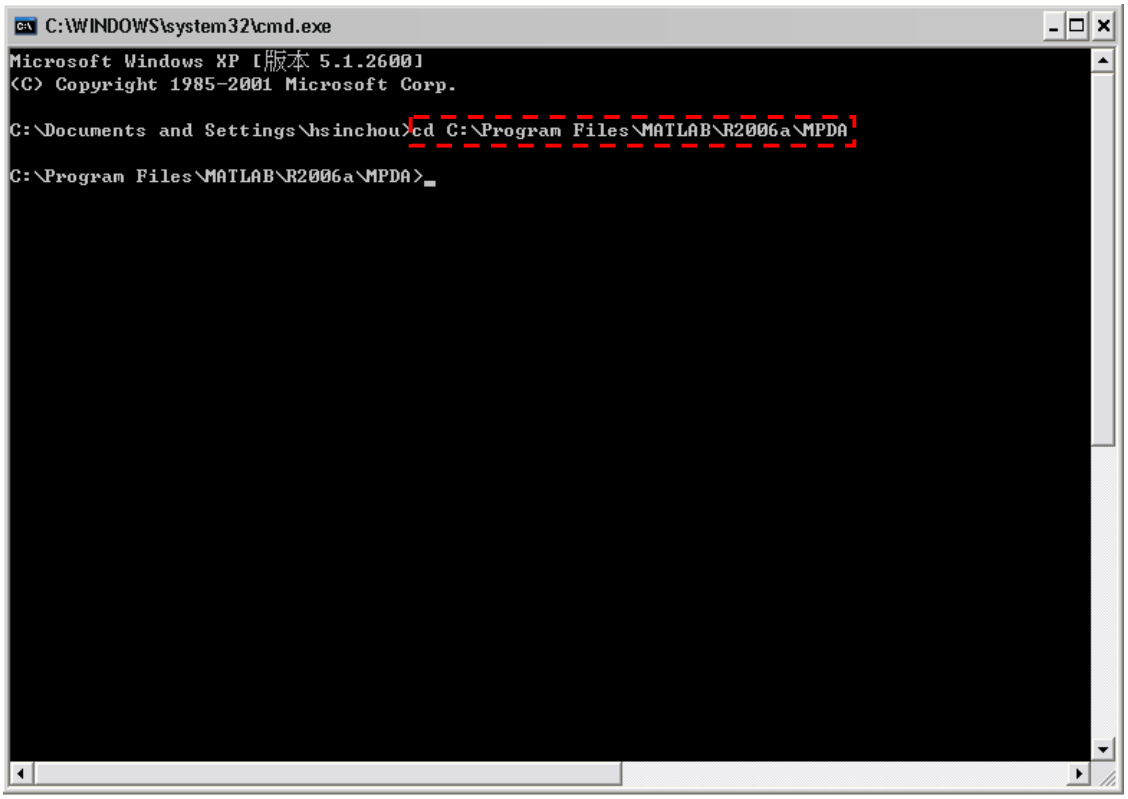


Figure 2-19. Step 8 – Type ‘MPDA’ and press [Enter] key to execute an executable ‘MPDA.exe’

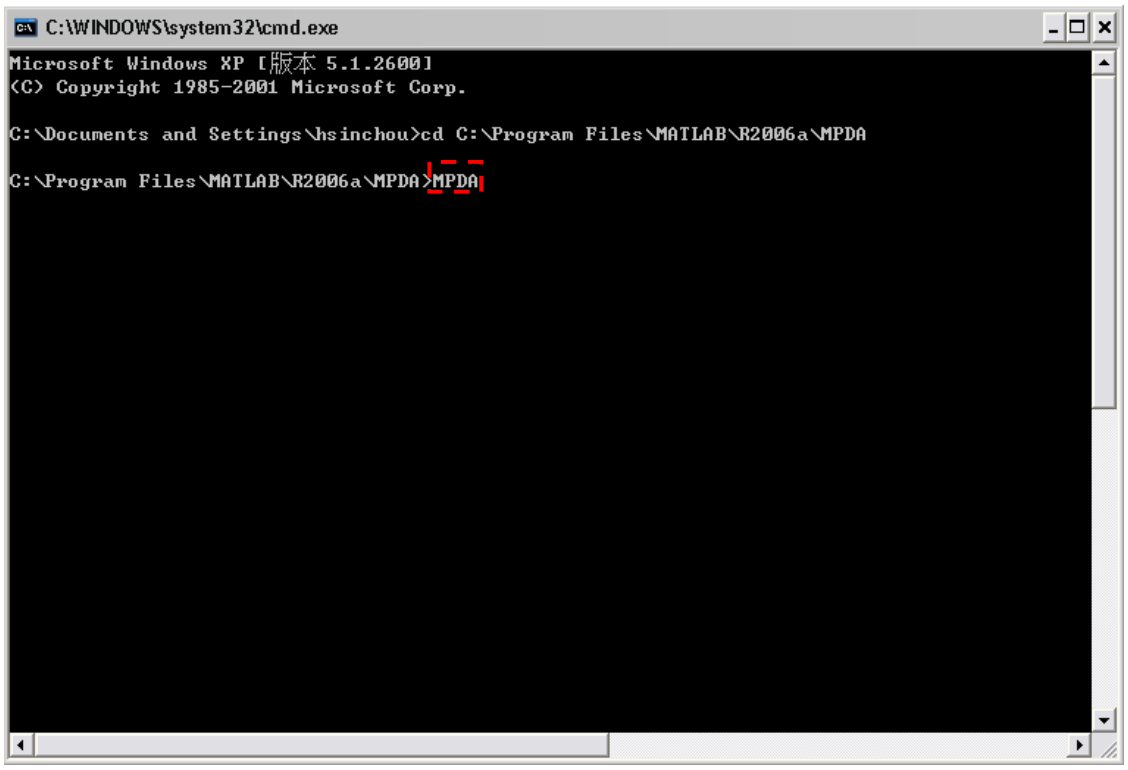
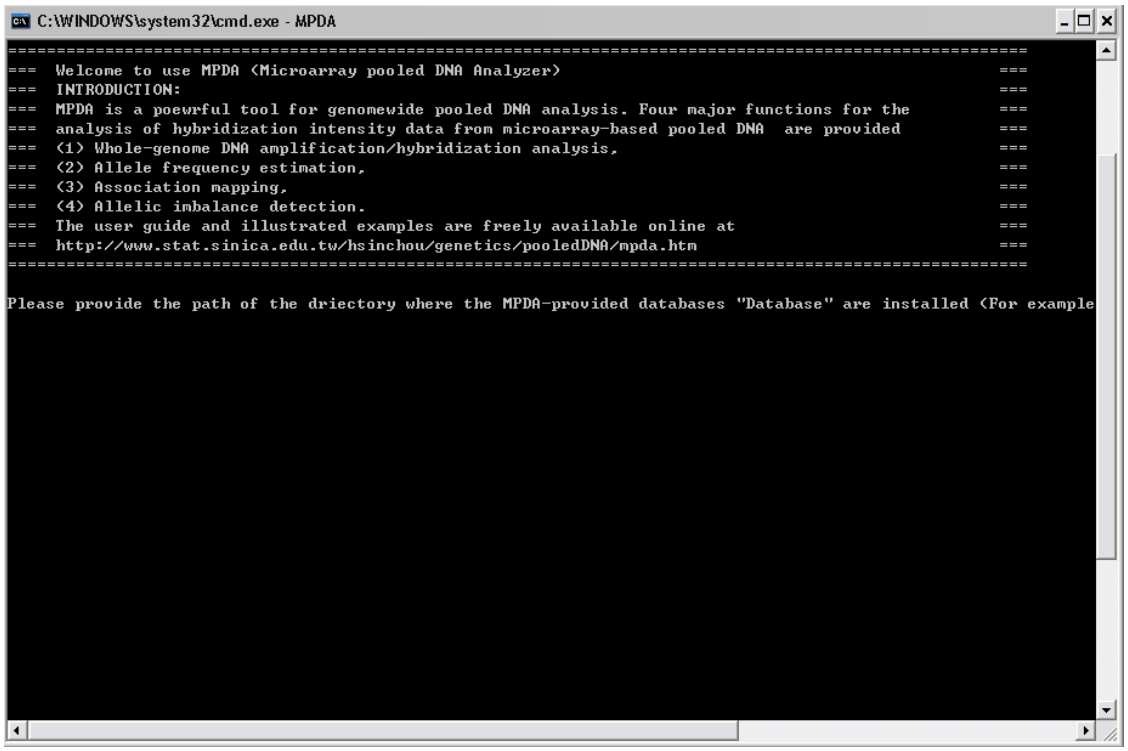


Figure 2-20. A welcome page of standalone *MPDA* will be showed in an MS-DOS Prompt window



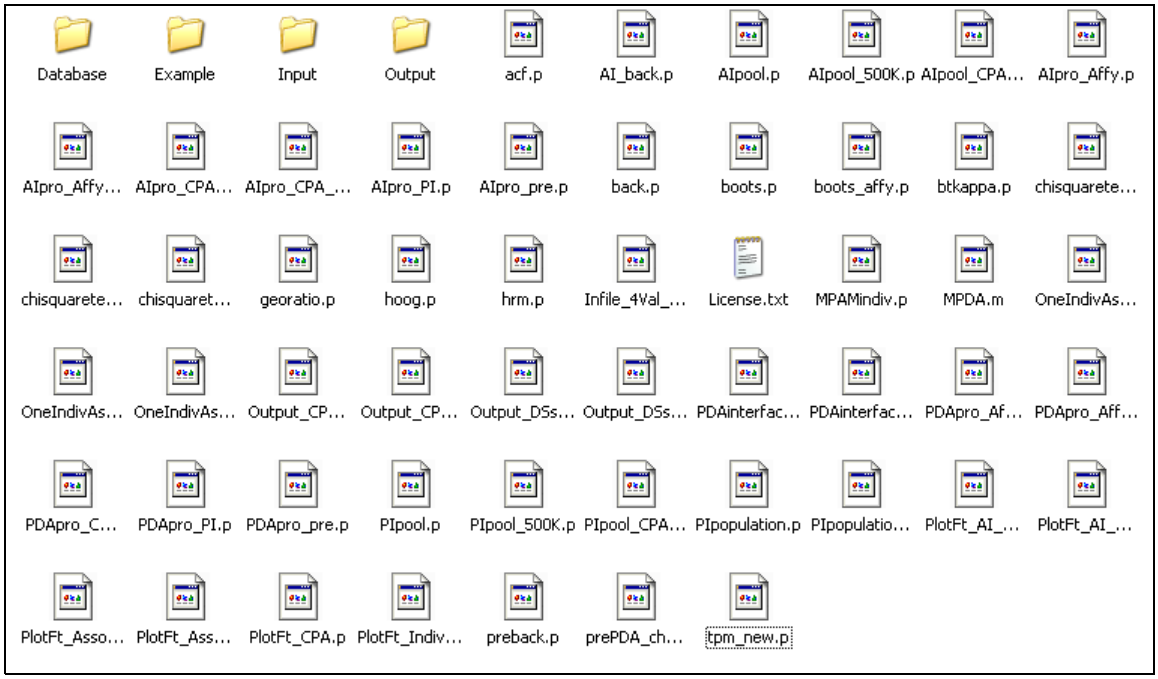
3. DESCRIPTION OF WORKING DIRECTORIES

3.1 MPDA with a user-friendly interface (MATLAB® is required)

The main directory name of *MPDA* is 'MPDA', which consists of four directories, *MPDA* license and some program files (**Figure 3-1**).

- Directory 'Input' – All input data must be saved in this directory.
- Directory 'Output' – All output results will be saved automatically in this directory.
- Directory 'Example' – Two illustrated examples discussed in the *MPDA* paper are provided in this directory.
- Directory 'Database' – Reference databases of CPA and databases of AF mean and prediction error for the analyses of the Affymetrix GeneChip Human Mapping 100K Set and 500K Set data are saved in this directory.
- File 'License.txt' – The *MPDA* license.
- Files *.p – The program source codes of *MPDA*.

Figure 3-1. Working directories and program files of *MPDA*



3.2 MPDA with standalone executables (MATLAB® is not required)

The main directory name of *MPDA* is 'MPDA', which consists of four directories, *MPDA* license and two program files.

- Directory 'Input' – All input data must be saved in this directory.
- Directory 'Output' – All output results will be saved automatically in this directory.
- Directory 'Example' – Two illustrated examples discussed in the *MPDA* paper are provided in this directory.
- Directory 'Database' – Reference databases of CPA and databases of AF mean and prediction error for the analyses of the Affymetrix GeneChip Human Mapping 100K Set and 500K Set data are saved in this directory.
- File 'License.txt' – The *MPDA* license.
- File 'MPDA.exe' – The executable file of standalone *MPDA*.
- File 'MPDA.ctf' – MATLAB® functions and data file that define *MPDA*.

4. MPDA INTERFACE, FUNCTIONS AND OPERATING PROCEDURES

4.1 MPDA with a user-friendly interface (MATLAB® is required)

Three interfaces of *MPDA* contain a welcome page and two main components. The welcome page, i.e., main interface, gives a short introduction to software *MPDA* (**Figure 2-9**). The first component was developed for association analysis (**Figure 2-10**). The second component was developed for allelic imbalance analysis (**Figure 2-11**).

4.1.1 Welcome page

This component provides a short introduction to software *MPDA* and its functionalities. There is one item on this page. Users can check the box 'Component 1' to carry out association analysis or the box 'Component 2' to carry out allelic imbalance analysis.

4.1.2 Component 1 – Association analysis

The purpose of this component is to provide a whole-genome association analysis for the artificially pooled DNA data. The analysis is useful in identifying loci associated with a particular trait of interest. There are seven main items in this component (**Figure 2-10**) as follows.

- Item 1 – Input/Output directory

  Users should provide directories for data input files (e.g., 'C:\Program Files\MATLAB\R2006a\MPDA\Input') and result output files (e.g., 'C:\Program Files\MATLAB\R2006a\MPDA\Output'). *MPDA* will automatically read data from the specified input directory and save outputs in the specified output directory.

- Item 2 – Number of groups studied

  Component 1 in *MPDA* provides a one-group analysis (CPA and AF) or a two-group analysis (CPA, AF and association analysis) for artificially pooled DNA data. Users can check the box 'One group' to estimate CPA and determine whether to calculate adjusted AF. Or, users can check the box 'Two groups' to estimate CPA and determine whether to calculate adjusted AF and carry out association tests.

Moreover, users should specify which CPA calibration methods are applied to the subsequent association analysis. Users can check the box 'Yes' to apply constant CPAs or 'No' to use unequal CPAs between two groups.

- Item 3 – Data type for CPA estimation

  MPDA permits three input data types for CPA estimation as follows:

  (1) *Affymetrix format:* Data files of hybridization intensities, which are obtained from the software GDAS, GCOS, CNAT and BAT (Affymetrix, CA, USA), should be provided. First, users should select which type of Affymetrix gene chips is used, i.e., 100K or 500K. Second, users can check 'All autosomes' to carry out whole-genome analysis or check only some specific chromosomes of interest.

  (2) *Raw CPA/heterozygote ratio:* CPA data and/or the corresponding s.e. should be provided. First, users should select which type of Affymetrix gene chips is used, i.e., 100K or 500K. Second, users can provide their own CPA reference or directly use the *MPDA*-provided CPA reference datasets (CPA from a combined population and CPA from the Taiwanese population) provided by *MPDA*.

  (3) *Peak intensity:* For non-Affymetrix users, data can be reformatted into pairs of peak intensities for each heterozygote individual to calculate CPA. Three CPA estimators (arithmetic mean, geometric mean, and bias-correction CPA) are provided in the analysis of *MPDA*. The statistical formulae are shown in **Appendix A**.

- Item 4 – Calculation of the bootstrapped s.e. of the CPA estimate

  MPDA calculates s.e. of CPA estimate based on a parametric bootstrapping resampling procedure, where composite relative allele signals (CRAS) are modeled by a beta distribution. Users can check the box 'Yes' and then key in the number of bootstrap replications between 10 and 1,000 to calculate s.e. Or, users can check the box 'No' to omit the calculation. The detailed procedure can be found in **Appendix A**.

- Item 5 – Estimation of adjusted AF

  Users can check the box 'Yes' and then select the Affymetrix-format hybridization intensities or pairs of peak intensities to calculate adjusted AFs. Or, users can check the box 'No' to omit the calculation. The estimation procedure can be found in **Appendix B**.

- Item 6 – Single-point pooled DNA association test

  Users can check the box 'Yes' and then key in the experimental s.e. to carry out a single-point pooled DNA association test. Or, users can check the box 'No' to omit the procedure. The detailed testing procedure can be found in **Appendix C**.

- Item 7 – Multipoint pooled DNA association test

  Users can check the box 'Yes' and check seven options to conduct a multipoint pooled DNA association test. The seven options are listed as follows.

  (1) *Data type for the association test:* Users can select the Affymetrix-format hybridization intensities, pairs of peak intensities, or raw p-values obtained from previous single-point association tests for a multipoint association test.

  (2) *Map information:* Users can check the box 'Yes' to input marker positions for the latter graph demonstration of multipoint p-values. If users check the box 'No' to omit the inter-marker distances, then the graph will be shown with an equal intermarker distance.

  (3) *Weight function:* Users can check 'Equal weight' to assign equal weights to all marker loci. Or, users can check 'User-specified weight' and then provide a set of weights.

  (4) *Threshold value of truncation:* Users can specify a truncation threshold. The threshold must be between 0 and 1 to exclude markers whose p-values from the previous single-point association test are greater than the threshold.

  (5) *Number of Monte Carlo simulations:* Users should enter the number of Monte Carlo simulations. The number of simulations must be between 500 and 10000.

  (6) *Window size:* Users should provide a value for the number of markers in a window. The window size must be between 2 and the total number of SNPs in the study.

  (7) *SWEPT statistics:* Users can select the multiplicative p-value test statistic, additive p-value test statistic or minimum p-value statistic for a multipoint pooled DNA association test. The detailed testing procedure can be found in **Appendix C**.

Remark 1: Users are recommended to fill in the calculation conditions from the top to the bottom in the *MPDA* interfaces. Moreover, please re-initialize the *MPDA* interface and

re-input the calculation conditions when running a new analysis.

Remark 2: After options have been selected and parameters for the previous seven items have been specified, users can click [Apply] to submit the job. And, users can click [Ctrl + Break] to terminate the submitted job if necessary.

4.1.3 Component 2 – Allelic imbalance analysis

The purpose of this component is to provide whole-genome allelic imbalance detection for naturally pooled DNA data. The analysis is useful to identify chromosomal aberrations typically associated with cancers. There are six main items in this component (**Figure 2-11**) as follows:

- Item 1 – Input/Output directory

  Users should specify the directories of input and output files before they carry out analyses. *MPDA* will automatically read data and save outputs in the destination input and output directories, respectively.

- Item 2 – Data type for CPA estimation

  MPDA allows for three input data types for CPA estimation as follows:

  (1) *Affymetrix format:* Data files of hybridization intensities, which are obtained from the software GDAS, GCOS, CNAT and BAT (Affymetrix, CA, USA), should be provided. First, users should select which type of Affymetrix gene chips is used, i.e., 100K or 500K. Second, users can check 'All autosomes' to carry out whole-genome analysis or check only some specific chromosomes of interest.

  (2) *Raw CPA/heterozygote ratio:* CPA data and/or the corresponding s.e. should be provided. First, users should select which type of Affymetrix gene chips is used, i.e., 100K or 500K. Second, users can provide their own CPA reference or use two *MPDA*-provided CPA reference datasets (CPA from a combined population and CPA from Taiwanese population).

  (3) *Peak intensity:* For non-Affymetrix users, they can reformat the data into pairs of peak intensities for each heterozygote individual to calculate CPA. Three CPA estimators (arithmetic mean, geometric mean, and bias-correction CPA) are provided in the analysis of *MPDA*. The statistical formulae can be found in **Appendix A**.

- Item 3 – Calculation of the bootstrapped s.e. of the CPA estimate

MPDA calculates the s.e. of CPA estimate based on a parametric bootstrapping resampling procedure, where relative allele intensities are modeled by a beta distribution. Users can check the box 'Yes' and key in the number of bootstrap replications between 10 and 1,000 to calculate s.e. Or, users can check the box 'No' to omit the calculation. The detailed procedures can be found in **Appendix A**.

- Item 4 – Estimation of adjusted AF

  Users can check the box 'Yes' to calculate adjusted AFs or the box 'No' to omit the calculation. The estimation procedures can be found in **Appendix B**.

- Item 5 – Single-point allelic imbalance analysis

  Users can check the box 'Yes' and then answer three questions to conduct a single-point pooled DNA allelic imbalance analysis. The three questions are listed as follows.

  (1) *Control reference:* Users can provide a user-specified reference or utilize the two MPDA-provided databases, i.e., the combined population reference or Taiwanese population reference, according to their study populations.

  (2) *Significance level:* Users can choose a significance level from 0.001, 0.01 and 0.05 for a single-point allelic imbalance analysis.

  (3) *Multiple test correction:* Users should provide a value for the correction number for multiple tests. The correction number must be between 1 and the total number of SNPs in the study. The pre-specified significance level selected in Question 2 will be divided by the correction number to yield an adjusted significance level. The detailed testing procedures can be found in **Appendix D**.

- Item 6 – Multipoint allelic imbalance analysis

  Users can check the box 'Yes' to perform a multipoint allelic imbalance analysis. *MPDA* provides two cumulative scores for a multipoint allelic imbalance analysis. Users can choose one or both scores at the same time in the analysis.

  (1) *SCORE1:* Users should enter a value for an allelic imbalance point. The score is between 0 and 100. This positive value will be assigned to an allelic imbalance SNP. And, users should also enter a value for a non-allelic imbalance point. The value is between 0 and 100. A minus sign for the value is assigned to a non-allelic imbalance SNP.

(2) *SCORE2:* Users should also enter a value for a non-heterozygote point between 0 and 100. This positive value will be assigned to a non-heterozygote SNP. And, users should also enter a value for a heterozygote point between 0 and 100. A minus sign for the value is assigned to a heterozygote SNP. The detailed testing procedures can be found in **Appendix D**.

Remark 1: Users are recommended to fill in the calculation conditions from the top to the bottom in the *MPDA* interfaces. Moreover, please re-initialize the *MPDA* interface and re-input the calculation conditions when running a new analysis.

Remark 2: After options have been selected and parameters for the previous seven items have been specified, users can click [Apply] to submit the job. And, users can click [Ctrl + Break] or [Ctrl + C] to terminate the submitted job if necessary.

4.2 MPDA with standalone executables (MATLAB® is not required)

There is only one interface for the analyses of the standalone *MPDA* software. The interface works interactively in the MS DOS Prompt window (**Figure 2-17**). Users enter their replies to the *MPDA* point-by-point queries, which have been explained in Section 4.1. Computational jobs are submitted once all the *MPDA* queries are answered. And, users can click [Ctrl + C] to terminate the submitted job if necessary.

5. DATA INPUT FORMAT

This section illustrates the data structure/format requested in Component 1 (**Figure 5.1**) and Component 2 (**Figure 5.2**) of *MPDA*. Note that the data input format for the execution of *MPDA* with a user-friendly interface and *MPDA* with standalone executables is identical.

5.1 Component 1 – Association analysis

According to the item functions selected in the interface of Component 1 (**Figure 2.10**), users should prepare their data following the data structure shown in **Figure 5.1**. All the directories and files of the input data must be saved in the working directory, 'C:\Program Files\MATLAB\R2006a\Input'. There are three types of data formats: the Affymetrix format data, intensity pair data, and p-value data. In this section, we introduce these three types of data in order and illustrate all required and provisional directories/files for each data type. Part of the illustrations is adopted from the supplements of PDA (Yang et al., *BMC Bioinformatics*, 2006).

- Affymetrix-format data directories/files:
  - ♣ Directory 'Population 1': [Required]

    The directory "Population 1" must be established under the directory "C:\Program Files\MATLAB\R2006a\Input" regardless of whatever an analysis of 'One group' or 'Two groups' (i.e., the number of the study groups queried in Item 2) is selected. Under the directory "Population 1", several directories should be prepared. Users can find examples for the related directories/files introduced below in Examples 4 and 5 at the MPDA website.

    (1) Directory 'IndGeno': [Optional]

    If users check the box 'Affymetrix format' for the data type for CPA estimation (Item 3), then they should save their Affymetrix genotype data in this directory. If the genotype data are derived from the Affymetrix GeneChip Human Mapping 100K and 500K Sets, the first two data files will be read as genotype data of the first person, the third and fourth data files will be read as genotype

data of the second person, and so on. If the genotype data are derived from the Affymetrix Genome-Wide Human SNP Array 5.0, the genotype data of each person should have been saved in a file.

(1a) Genotype data files for different individuals from individual genotyping:

The files provide the genotype calls and SNP-related information. Each type of genotype file contains five to eight columns, where the variables are arranged according to the particular genotype calling software.

(1a-1) 100K gene chip – *Format for software CNAT:*

Column 1: The index of SNP

Column 2: The name of SNP probe set

Column 3: The index of chromosome

Column 4: Physical Position

Column 5: Genotype call

(1a-2) 100K gene chip – *Format for software GDAS:*

Column 1: The index of SNP

Column 2: The name of SNP probe set

Column 3: The dbSNP RS ID

Column 4: The index of chromosome

Column 5: Physical Position

Column 6: The TSC ID

Column 7: Genotype call

Column 8: Confidence value

(1a-3) 100K gene chip – *Format for software GTYPE:*

Column 1: The index of SNP

Column 2: The name of SNP probe set

Column 3: The index of chromosome

Column 4: Physical Position

Column 5: The dbSNP RS ID

Column 6: The TSC ID

Column 7: Genotype call

Column 8: Confidence value

(1a-4) 500K gene chip – *Format for GTYPE500K:*

Column 1: The index of SNP

Column 2: The name of SNP probe set

Column 3: The index of chromosome

Column 4: Physical Position

Column 5: The dbSNP RS ID

Column 6: Genotype call

Column 7: Confidence value

(1a-5) Array 5.0 gene chip – *Format for Software Bat 2.0:* [Experimental]

Column 1: The index of chromosome

Column 2: Physical Position

Column 3: The name of SNP probe set

Column 4: Genotype call

Column 5: Confidence value

Column 6: Contrast

Column 7: Strength

Column 8: Forced call

(1b) File '!4Var_ColSet.txt': File formats for genotype data differ according to the particular genotyping calling software. The file '!4Var_ColSet.txt' is used to provide *MPDA* which file format is used. This file contains seven rows and two columns. The first row denotes the name of genotyping calling software which was used to format the genotype data. The second to the seventh rows correlated with the corresponding column indices for a starting field, SNP ID, chromosome, physical position, SNP call and an ending field in the provided genotype files, respectively. And, some data in the provided genotype files will not be used in *MPDA*. For example, in the format of GDAS, SNP call is arranged in the 7th column, the ending field. And, the 6th column in the genotype file will not be used. Users can select a format from (1b-1), (1b-2), (1b-3), (1b-4) and (1b-5) according to their genotype data. (Note: The input in the second column is case-sensitive. Capital letters should be used.)

(1b-1) 100K gene chip – *Software CNAT:*

Software	CNAT
SnpID	2
Chromosome	3
PhyPosition	4
Call	5
Starting_field	1
Ending_field	7

(1b-2) 100K gene chip – *Software GDAS:*

Software	GDAS
SnpID	2
Chromosome	4
PhyPosition	5
Call	7
Starting_field	1
Ending_field	7

(1b-3) 100K gene chip – *Software GTYPE:*

Software	GTYPE
SnpID	2
Chromosome	3
PhyPosition	4
Call	7
Starting_field	1
Ending_field	7

(1b-4) 500K gene chip – *GTYPE500K:*

Software	GTYPE500K
SnpID	2
Chromosome	3
PhyPosition	4
Call	6
Starting_field	1

Ending_field 7

(1b-5) Array 5.0 gene chip – *Software Bat 2.0:* [Experimental]

Software ARRAY5

SnpID 3

Chromosome 1

PhyPosition 2

Call 4

Starting_field 1

Ending_field 7

(2) Directory 'IndPI': [Optional]

If users check the box 'Affymetrix format' for the data type for CPA estimation (Item 3), then they should save their Affymetrix intensity data in this directory. *MPDA* will read all intensity data sequentially. If the intensity data are derived from the Affymetrix GeneChip Human Mapping 100K and 500K Sets, the first two data files in the directory will be read as intensity data of the first person, the third and fourth data files will be read as intensity data of the second person, and so on. If the intensity data are derived from the Affymetrix Genome-Wide Human SNP Array 5.0, the intensity data of each person should have been saved in a file.

(2a) For the data from the Affymetrix GeneChip Human Mapping 100K and 500K Sets, the files provide the hybridization intensity data and SNP-related information. Each intensity file contains 58 columns.

Column 1: The index of SNP

Column 2: The name of SNP probe set

Column 3 – the final column: The remaining 56 columns are the hybridization intensity data arranged as follows [Q1-Q7 denote the quartet, PM and MM denote perfect match and mismatch, SS and AS denote sense strand and antisense strand, A and B denote allele A and allele B]:

'(Q1,PM,SS,*A*)', '(Q1,PM,SS,*B*)', '(Q1,MM,SS,*A*)', '(Q1,MM,SS,*B*)', '(Q1,PM,AS,*A*)','(Q1,PM,AS,*B*)', '(Q1,MM,AS,*A*)', '(Q1,MM,AS,*B*)',

……

‘(Q7,PM,SS,*A*)’, ‘(Q7,PM,SS,*B*)’, ‘(Q7,MM,SS,*A*)’, ‘(Q7,MM,SS,*B*)’, ‘(Q7,PM,AS,*A*)’, ‘(Q7,PM,AS,*B*)’, ‘(Q7,MM,AS,*A*)’,’(Q7,MM,AS,*B*)’.

For data obtained from the Affymetrix Genome-Wide Human SNP Array 5.0, the files provide the hybridization intensity data and SNP-related information. Each intensity file contains four columns.

Column 1: The index of chromosome

Column 2: Physical Position

Column 3: The name of SNP probe set

Column 4: The normalized hybridization intensity, where one line per allele of each SNP (Note that: The intensity data can be found in the SNP Intensity Summary File provided by Affymetrix Genome-Wide Human SNP Array 5.0. This function is still experimental.)

(3) Directory ‘PoolAF’: [Optional]

If users check the box ‘Yes’ for the calculation of estimating AF in artificial DNA pools and check the box ‘Affymetrix’ for the data type (Item 5), then this directory should be prepared.

(3a) Subdirectory ‘PoolGeno’:

(3a-1) Genotype data files for an artificial DNA pool must be saved in this directory. The data formats for different genotyping software are introduced in (1a).

(3a-2) File ‘!4Var_ColSet.txt’: This file contains seven rows and two columns as introduced in (1b).

(3b) Subdirectory ‘PoolPI’:

Hybridization intensity data file for an artificial DNA pool must be saved in this directory. The data formats are the same as that introduced in (2a).

(3c) File ‘!poolsize.txt’:

This file records the pool size of the study DNA pool.

♣ Directory ‘Population 2’: [Optional]

This directory must be established if users select ‘Two groups’ for the number of the study groups (Item 2). The data structure is exactly the same as the directory ‘Population 1’.

♣ File 'IndPI.txt': [Optional]

If users select the option 'Raw CPA/heterozygote ratio', check the box 'Affymetrix 100 K' or 'Affymetrix 500 K' and check the box 'User-specified CPA' (Item 3), then this file must be provided. Note that, at the time, they do not need to establish the two directories 'IndGeno' and 'IndPI'. The file 'IndPI.txt' contains the following columns.

(1) Column 1: The index of populations. Code 1 signifies the first population, and code 2, the second population. If only one population is included in the analysis, then this column is a column vector containing only 1s.

(2) Column 2: The SNP probe set name

(3) Column 3: The user-specified CPA

(4) Column 4: The s.e. of the user-specified CPA [Optional]

♣ File 'Weight.txt': [Optional]

If users check 'Yes' for requiring the multipoint pooled DNA association test and check the box 'User-specified weight' for the weight function (Item 7), then this directory should be provided. The file 'Weight.txt' contains two columns.

(1) Column 1: The SNP probe set name

(2) Column 2: The user-specified weight. It must be non-negative real values.

- Non-Affymetrix-format data (Intensity pair data):

Some required and provisional files for the non-Affymetrix format data are introduced. And, users can find an example for the related files introduced below in Example 3 at the *MPDA* website.

♣ File 'SnpName.txt': [Required]

This file contains two columns.

(1) Column 1: The index of SNPs. The SNP names must be arranged in a column (i.e., one row per SNP) and must be Arabic numerals.

(2) Column 2: Physical position [Optional]

♣ File 'IndPI.txt': [Optional]

(1) When users check the box 'Peak intensity' for the data type for CPA estimation (Item 3), they should save their intensity pair data in this file. This file contains

the following columns:

(1a) Column 1: The index of populations. Code 1 signifies the first population, and code 2, the second population. If only one population is included in the analysis, then this column is a column vector containing only 1s.

(1b) Column 2: The index of SNPs. The order of SNPs must match the order in 'SnpName.txt'.

(1c) Column 3 – the final column: Pairs of peak intensities. The third and fourth columns are the peak intensities of the first and second alleles of heterozygous individuals, respectively. If there are more than one pairs of intensities, they should be listed followed by the fourth column.

(2) When users check the box 'Raw CPA/heterozygote ratio' for the data type for CPA estimation, check the box 'Non-Affymetrix' and check the box 'User-specified CPA' (Item 3) for the raw CPA/heterozygote ratio, then this file should be prepared. The file contains the following columns:

(2a) Column 1: The index of populations. Code 1 signifies the first population, and code 2, the second population. If only one population is included in the analysis, then this column is a column vector containing only 1s.

(2b) Column 2: The index of SNPs. The order of SNPs must match the order in 'SnpName.txt'.

(2c) Column 3: The user-specified CPA.

(2d) Column 4: The s.e. of the user-specified CPA [Optional]

♣ File 'PoolAF.txt': [Optional]

If users check the box 'Yes' for the calculation of estimating AF in artificial DNA pools and check the box 'Peak intensity' for the data type (Item 5), then this file should be prepared. This file contains the following columns:

(1) Column 1: The index of populations. Code 1 signifies the first population, and code 2, the second population. If only one population is included in the analysis, then this column is a column vector containing only 1s.

(2) Column 2: The index of SNPs. The order of SNPs must match the order in 'SnpName.txt'.

(3) Column 3: Pool size

(4) Column 4 – the final column: Pairs of peak intensities. The third and fourth columns are the peak intensities of the first and second alleles of heterozygous individuals, respectively. If there are more than one pairs of intensities, they should be listed followed by the fourth column.

♣ File 'Weight.txt': [Optional]

If users check 'Yes' for requiring the multipoint pooled DNA association test and check the box 'User-specified weight' for weight function (Item 7), then this file should be provided. The file 'Weight.txt' contains two columns.

(1) Column 1: The index of SNPs. The index of SNPs. The order of SNPs must match the order in 'SnpName.txt'.

(2) Column 2: User-specified weight. It must be non-negative real value(s).

- P-value data:

Users can perform SWEPT based on the user-specified p-value data. Users can check 'Yes' for requiring the multipoint pooled DNA association test and check the box 'P-value' for the data type for association test. Note that if users want to use this utility, they must first check the box 'Peak intensity' in Item 3. The check will be automatically removed after they check 'P-value'. Users can find an example for the related files introduced below in Example 6 at the *MPDA* website.

♣ File 'SnpName.txt': [Required]

The data format is the same as the previous illustration for the file 'SnpName.txt' in 'Non-Affymetrix-format data'.

♣ File 'Weight.txt': [Optional]

The data format is the same as the previous illustration for the file 'Weight.txt' in 'Non-Affymetrix-format data'.

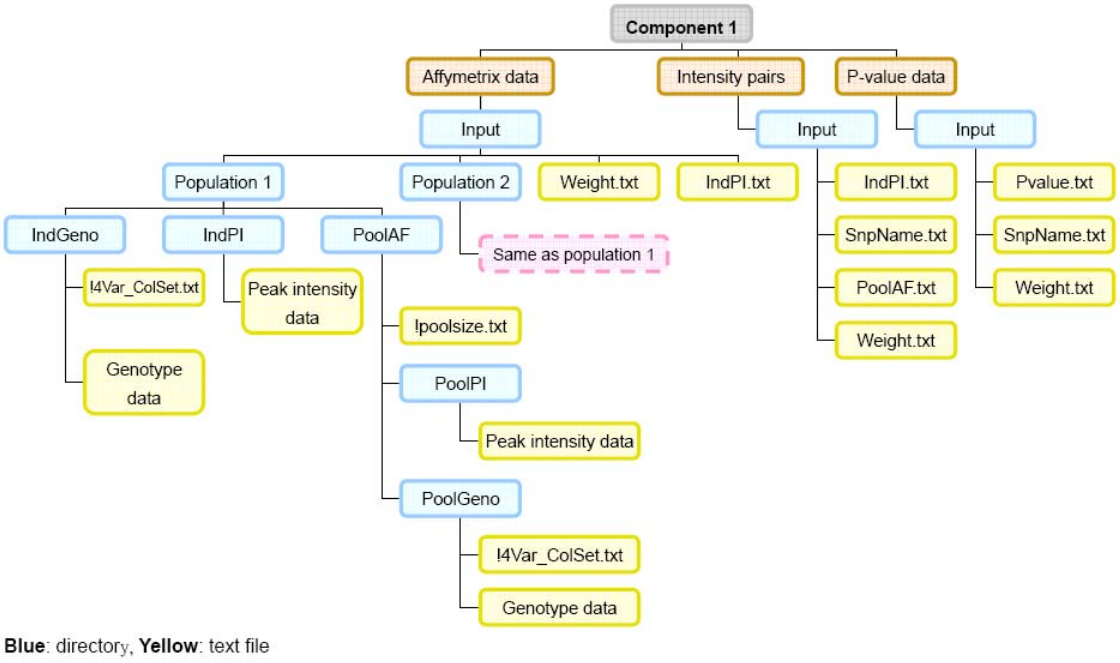
♣ File 'Pvalue.txt': [Optional]

This file contains two columns.

(1) Column 1: The index of SNPs. It must be numerical. The order of SNPs must match the order in 'SnpName.txt'.

(2) Column 2: P-values. Users provide their own p-value for each study SNP. The p-values must be between 0 and 1.

Figure 5-1. The structure of data directory/file for Component 1 analysis



5.2 Component 2 – Allelic imbalance analysis

According to the data types selected in the interface of Component 2 (**Figure 2.11**), users should prepare their data following the data structure shown in **Figure 5.2**. All the directories and files must be saved in the working directory, 'C:\Program Files\MATLAB\R2006a\Input'. There are two types of data formats: the Affymetrix format data and intensity pair data. In this section, we introduce the two types of data in order and illustrate all required and optional directories/files for each data type. Part of the illustration is adopted from the supplements of PDA (Yang et al., *Bioinformatics*, 2006).

- Affymetrix-format data directories/files:

  Users can find examples for the related directories/files introduced below in Examples 7 and 8 at the *MPDA* website.

  ♣ Directory 'IndGeno': [Optional]

  If users check the box 'Affymetrix format' for the data type for CPA estimation

(Item 2), then they should save their Affymetrix genotype data in this directory. This directory contains genotype data files for different individuals from individual genotyping and the file '!4Var_ColSet.txt'. The formats of the files are exactly the same as the illustration for the directory 'IndGeno' in Component 1.

♣ Directory 'IndPI': [Optional]

If users check the box 'Affymetrix format' for the data type for CPA estimation (Item 2), then they should save their Affymetrix intensity data in this directory. This directory contains hybridization intensity data files for different individuals from individual genotyping. The formats of the files are exactly the same as the illustration for the directory 'IndPI' in Component 1.

♣ Directory 'PoolAF': [Optional]

If users check the box 'Yes' for the calculation of estimating AF in artificial DNA pools and check the box 'Affymetrix' for the data type (Item 4), then this directory should be prepared. This directory contains two subdirectories 'PoolPI' and 'PoolGeno'. 'PoolPI' contains data files of hybridization intensity from a natural DNA pool. The formats of the files are exactly the same as that illustration for the directory 'PoolPI' in Component 1. 'PoolGeno' contains genotype data files from a natural DNA pool and a file '!4Var_ColSet.txt'. The formats of the files are exactly the same as that illustration for the directory 'PoolGeno' in Component 1. Because an allelic imbalance analysis focuses on a natural DNA pool instead of an artificial DNA pool, the pool size requested in Component 1 is not necessary in Component 2.

♣ File 'IndPI.txt': [Optional]

This file must be provided when users select the option 'Raw CPA/heterozygote ratio', check the box 'Affymetrix 100 K' or 'Affymetrix 500 K' and check the box 'User-specified CPA' (Item 2) for the data type for CPA estimation. At the time, they do not need to establish the previous two directories 'IndGeno' and 'IndPI'. The file 'IndPI.txt' contains the following columns.

(1) Column 1: The name of SNP probe set

(2) Column 2: The user-specified CPA

(3) Column 3: The s.e. of the user-specified CPA [Optional].

♣ File '!Control_samplesize.txt': [Required]

This file records the number of controls used to establish the normal reference.

♣ File 'ControlReference.txt': [Optional]

This file must be provided when users select the option 'Yes' for single-point allelic imbalance analysis and check the box 'User-specified' for control reference. The file contains the following columns:

(1) Column 1: The name of SNP probe set

(2) Column 2: The sample mean of CRAS for genotype *AA*

(3) Column 3: The sample s.e. of CRAS for genotype *AA*

(4) Column 4: The sample mean of CRAS for genotype *AB*

(5) Column 5: The sample s.e. of CRAS for genotype *AB*

(6) Column 6: The sample mean of CRAS for genotype *BB*

(7) Column 7: The sample s.e. of CRAS for genotype *BB*

- Non-Affymetrix-format data (Intensity pair data):

♣ File 'SnpName.txt': [Required]

This file contains two columns.

(1) Column 1: The index of SNPs. The SNP names must be arranged in a column (i.e., one row per SNP) and must be Arabic numerals.

(2) Column 2: Physical position [Optional]

♣ File 'IndPI.txt': [Optional]

(1) When users check the box 'Peak intensity' for the data type for CPA estimation (Item 2), their intensity pair data should be prepared in this file. This file contains the following columns:

(1a) Column 1: The index of SNPs. The order of SNPs must match the order in 'SnpName.txt'.

(1b) Column 2 – the final column: Pairs of peak intensities. The third and fourth columns are the peak intensities of the first and second alleles of heterozygous individuals, respectively. If there are more than one pairs of intensities, they should be listed followed by the fourth column.

(2) When users check the box 'Raw CPA/heterozygote ratio', check the box

'Non-Affymetrix' and check 'User-specified CPA' for the data type for CPA estimation (Item 2), this file should be prepared. The file contains the following columns:

(2a) Column 1: The index of SNPs. The order of SNPs must match the order in 'SnpName.txt'.

(2b) Column 2: The user-specified CPA

(2c) Column 3: The s.e. of the user-specified CPA [Optional]

♣ File 'PoolAF.txt': [Optional]

This file contains the following columns:

(1) Column 1: The index of individuals (natural pools). It must be Arabic numerals.

(2) Column 2: The index of SNPs. The order of SNPs must match the order in 'SnpName.txt'.

(3) Column 3 – the final column: Pairs of peak intensities. The third and fourth columns are the peak intensities of the first and second alleles of natural DNA pools, respectively. If there are more than one pairs of intensities, they should be listed followed by the fourth column.

♣ File '!Control_samplesize.txt': [Required]

This file records the number of controls used for establishing a normal reference.

♣ File 'ControlReference.txt': [Required]

This file must be provided when users select the option 'Yes' for single-point allelic imbalance analysis and check the box 'User-specified' for control reference. The file contains the following columns:

(1) Column 1: The SNP probe set name.

(2) Column 2: The sample mean of CRAS for genotype *AA*

(3) Column 3: The sample s.e. of CRAS for genotype *AA*

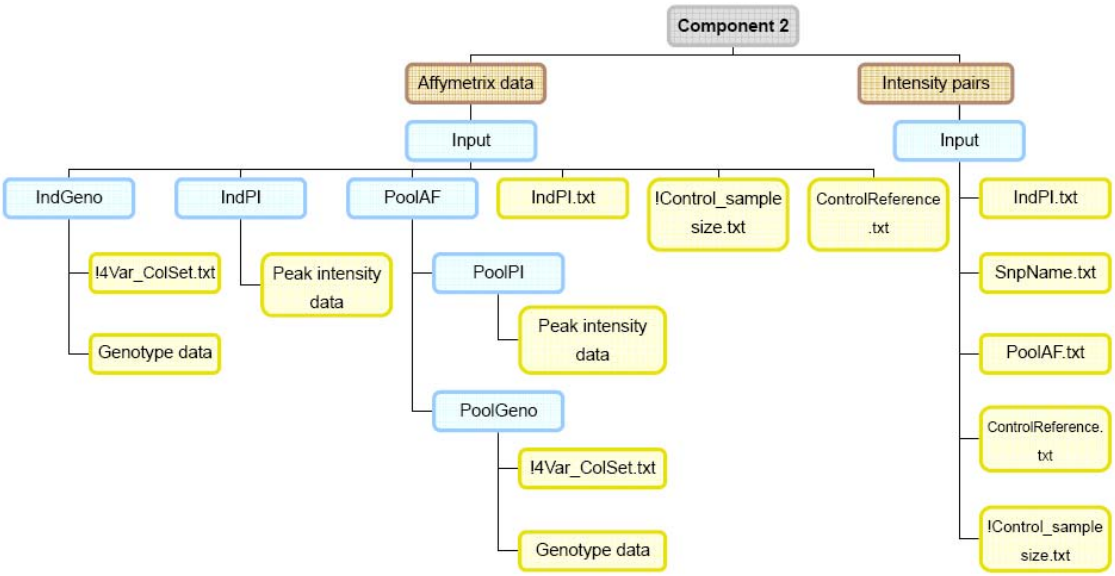
(4) Column 4: The sample mean of CRAS for genotype *AB*

(5) Column 5: The sample s.e. of CRAS for genotype *AB*

(6) Column 6: The sample mean of CRAS for genotype *BB*

(7) Column 7: The sample s.e. of CRAS for genotype *BB*

Figure 5-2. – Structure of input data directories in component 2



Blue: directory, **Yellow**: text file

6. Examples

This section provides detailed running procedures for the two examples included with *MPDA*. The first example (Example_AF) illustrates the calculation of AF and its s.e. based on a CPA reference database. The second example (Example_Asso) illustrates the calculation of AFs and conduction of association mapping.

6.1 Example 1: Calculation of AF and its s.e. based on a CPA reference database

This example calculates AF and its s.e. in an artificial DNA pool with a pool size of 240. We calculate unadjusted AF and adjusted AF based on the Taiwan-specific CPA reference database attached in *MPDA*. The data are provided in the directory 'C:\Program Files\MATLAB\R2006a\MPDA\Example\ Example_AF' included with *MPDA*.

- Step 1 – Download an executable file 'MCRInstaller.exe' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm.
- Step 2 – Execute the executable file 'MCRInstaller.exe' to install MATLAB® Component Runtime.
- Step 3 – Download a zip file 'MPDA_mcr.rar' from the *MPDA* website http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm.
- Step 4 – Decompress the zip file 'MPDA_mcr.rar' to a destination directory 'C:\Program Files\MATLAB\R2006a\MPDA'.
- Step 5 – Click 'Start' and select 'Run' to open a window of 'Run'.
- Step 6 – Type command 'cmd' to activate the MS-DOS Prompt window. (Note: If the appearance of the MS-DOS Prompt window is not proper, you can choose proper parameters of 'Options', 'Fonts', 'Layout' and 'Color' for configuration according to the setting of monitor resolution. For example, if the monitor resolution is set to be 1024×768, then you can choose size 8×16 in the 'Font' tab; width 125 and height 45 for screen buffer size, width 125 and height 45 for window size in the 'Layout' tab. Click [OK] to confirm the parameters and choose 'Save properties for future windows with same title' to save the setting. The MS-DOS Prompt window must be restarted to validate the new settings.)
- Step 7 – Type 'cd C:\Program Files\MATLAB\R2006a\MPDA' to change a working

directory to the *MPDA* directory in the MS-DOS Prompt window.

- Step 8 – Copy a directory, 'C:\Program Files\MATLAB\R2006a\MPDA\Example\Example_AF\Input\population1', and all files in the directory to the input directory 'C:\Program Files\MATLAB\R2006a\MPDA\Input'. If there have been files in the input directory, then they should be removed before data copy.
- Step 9 – Type 'MPDA' in the MS-DOS Prompt window and press [Enter] key to execute an executable 'MPDA.exe'. The welcome page of standalone *MPDA* will be shown after a few seconds.
- Step 10 – Type 'C:\Program Files\MATLAB\R2006a\MPDA' and press [Enter] key to provide the path for the directory where the *MPDA*-provided databases are installed.
- Step 11 – Type '1' to perform an analysis of Component 1.
- Step 12 – Type 'C:\Program Files\MATLAB\R2006a\MPDA\Input' and press [Enter] key to provide the path for the input directory.
- Step 13 – Type 'C:\Program Files\MATLAB\R2006a\MPDA\Output' and press [Enter] key to provide the path for the output directory.
- Step 14 – Type '1' to perform a one-group analysis.
- Step 15 – Type '2' to use CPA databases provided by software *MPDA* directly.
- Step 16 – Type '3' to select the Taiwanese CPA database in this example.
- Step 17 – Type '1' to select a CPA database of Affymetrix 100K.
- Step 18 – Type '1' to calculate the estimate of allele frequency.
- Step 19 – Type '1' to illustrate the input data are from Affymetrix gene chips.
- Step 20 – Please wait for a while. The analysis will be performed and the results will be automatically saved in directory 'C:\Program Files\MATLAB\R2006a\MPDA\Output'.

6.2 Example 2: Calculation of AFs and conduction of association mapping

This example calculates AFs and carries out single-point and multipoint association mapping. We constructed two artificial DNA pools. The first pool has a pool size of 10 and the second pool has a pool size of 30. Both samples are drawn from the same

Taiwanese population. Based on the Taiwan-specific CPA reference database attached in *MPDA*, we estimate adjusted AFs in two groups. Moreover, we carry out association tests. First, we carry out single-point association test, where the experimental s.e. is 0.02. Second, we carry out multipoint association test, where the SWEPT statistic with equal weight, a p-value truncation threshold of 0.05, window size of 5 and a multiplicative effect. The empirical p-values are calculated based on 10,000 Monte Carlos. The data are provided in the directory 'C:\Program Files\MATLAB\R2006a\MPDA\Example\Example_Asso' included with *MPDA*.

- Step 1 – Step 6 shown in Example 1 can be skipped if MATLAB® Component Runtime and *MPDA* have been installed.
- Step 7 – Type 'cd C:\Program Files\MATLAB\R2006a\MPDA' to change a working directory to the *MPDA* directory in the MS-DOS Prompt window.
- Step 8 – Copy the two directories, 'C:\Program Files\MATLAB\R2006a\MPDA\Example\Example_Asso\Input\population1' and 'C:\Program Files\MATLAB\R2006a\ MPDA\Example\Example_Asso\Input\population2' with all files in the two directories to the input directory 'C:\Program Files\MATLAB\R2006a\MPDA\Input'. If there have been files in the input directory, then they should be removed before data copy.
- Step 9 – Type 'mpda' in the MS-DOS Prompt window and press [Enter] key to execute an executable 'MPDA.exe'. The welcome page of standalone *MPDA* will be shown after a few seconds.
- Step 10 – Type 'C:\Program Files\MATLAB\R2006a\MPDA' and press [Enter] key to provide the path for the directory where *MPDA* is installed.
- Step 11 – Type '1' to perform an analysis of Component 1.
- Step 12 – Type 'C:\Program Files\MATLAB\R2006a\MPDA\Input' and press [Enter] key to provide the path for the input directory.
- Step 13 – Type 'C:\Program Files\MATLAB\R2006a\MPDA\Output' and press [Enter] key to provide the path for the output directory.
- Step 14 – Type '2' to perform a two-group analysis.
- Step 15 – Type '2' to use CPA databases provided by software *MPDA* directly.
- Step 16 – Type '3' to select the Taiwanese CPA database in this example.

- Step 17 – Type ‘1’ to select a CPA database of Affymetrix 100K.
- Step 18 – Type ‘1’ to calculate the estimate of allele frequency.
- Step 19 – Type ‘1’ to illustrate the input data are from Affymetrix gene chips.
- Step 20 – Type ‘1’ to perform a single-point pooled DNA association test.
- Step 21 – Type ‘0.02’ to specify the experimental error.
- Step 22 – Type ‘1’ to perform a multipoint pooled DNA association test.
- Step 23 – Type ‘1’ to confirm that map information is provided.
- Step 24 – Type ‘1’ to use an equal-weight statistic.
- Step 25 – Type ‘0.05 to specify the threshold value of p-value truncation.
- Step 26 – Type ‘10000’ to specify the number Monte Carlo simulations.
- Step 27 – Type ‘5’ to specify the window size.
- Step 28 – Type ‘1’ to select a multiplicative SWEPT statistic.
- Step 29 – Please wait for a while. The analysis will be performed and the results will be automatically saved in directory ‘C:\Program Files\MATLAB\R2006a\MPDA\Output’.

7. Troubleshooting

7.1 MATLAB® Component Runtime cannot be installed smoothly?

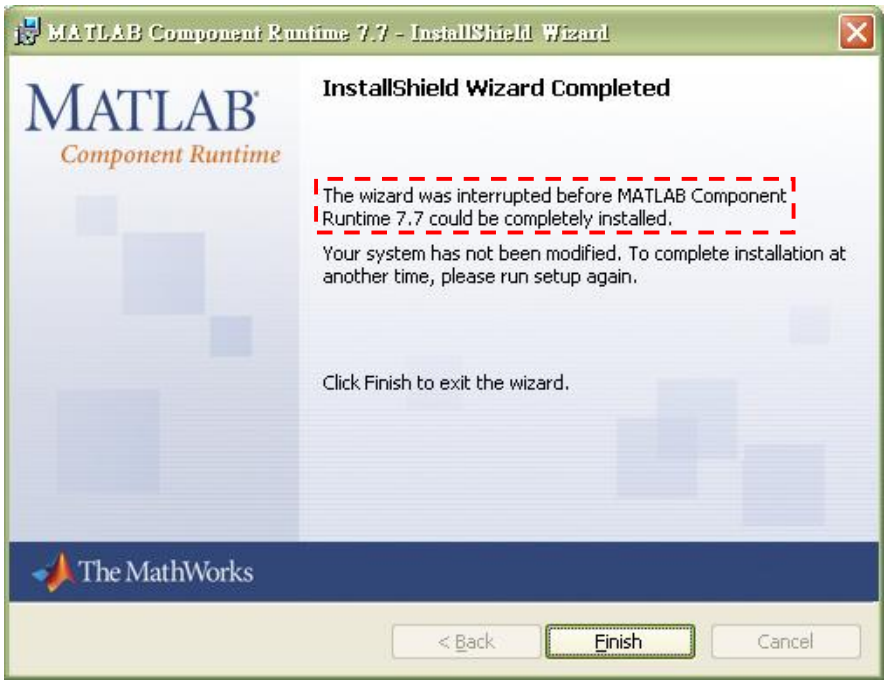
If users encounter an error shown in the InstallShield Wizard of MATLAB® Component Runtime (Figure 7.1) while installing the MATLAB® Component Runtime, the problem can be solved by installing the Microsoft Visual C++ Redistributable Package.

Users can install the Microsoft Visual C++ 2005 Redistributable Package (x86 or x64). The package for a PC with a 32-bit system can be downloaded at http://www.microsoft.com/downloads/details.aspx?FamilyId=32BC1BEE-A3F9-4C13-9C99-220B62A191EE&displaylang=en; the package for a 64-bit PC can be downloaded at http://www.microsoft.com/downloads/details.aspx?familyid=90548130-4468-4bbc-9673-d6acabd5d13b&displaylang=en. Or users can install the Microsoft Visual C++ 2005 SP1 Redistributable Package. The package can be downloaded at http://www.microsoft.com/downloads/details.aspx?displaylang=zh-tw&FamilyID=200b2fd9-ae1a-4a14-984d-389c36f85647. After installing the Microsoft Visual C++ Redistributable Package, the MATLAB® Component Runtime should be re-installed successfully.

7.2 The appearance of the MS-DOS Prompt window is not proper?

Users can choose proper parameters of 'Options', 'Fonts', 'Layout' and 'Color' for configuration according to the setting of monitor resolution. For example, if the monitor resolution is set to be 1024×768, then you can choose size 8×16 in the 'Font' tab; width 125 and height 45 for screen buffer size, width 125 and height 45 for window size in the 'Layout' tab. Press [OK] to confirm the parameters and choose "Save properties for future windows with same title" to save the setting. The MS-DOS Prompt window must be restarted to validate the new settings.

Figure 7-1. An error message is shown in the InstallShield Wizard while installing MATLAB® Component Runtime.



8. MPDA version upgrade

Versions:

MPDA Version 1.0: Jan/2007

MPDA Version 2.0: Dec/2007; *MPDA* Version 2.1: Mar/2008

What are the new features in *MPDA* Version 2.0 and Version 2.1 compared to Version 1.0?

(1) Users can run *MPDA* on machines without installing MATLAB®. A new standalone *MPDA* has been developed in Version 2.0.

(2) New database of CPA and database of AF mean and prediction error for the analysis of Affymetrix GeneChip Human Mapping 500K Set are provided in Version 2.0.

(3) More functions and error messages of input data checking and troubleshooting for the installation of the MATLAB® Component Runtime are added in Version 2.1.

APPENDIX A – CPA ESTIMATE

Suppose that a SNP marker has two alleles *A* and *a*. Let n_{h} denote the number of samples who are heterozygous with respect to this SNP. Let $\{(h_{A,j}, h_{a,j}), j = 1, \cdots, n_{\mathrm{h}}\}$ denote the pairs of CRAS (or standardized intensity) of the n_{h} samples, where $h_{A,j} + h_{a,j} = 1$. Let $(\bar{h}_A, \bar{h}_a)$ denote the arithmetic sample means of the pairs of CRAS, i.e.,

$$\bar{h}_A = \left(\sum_{j=1}^{n_{\mathrm{h}}} h_{Aj}\right) \Big/ n_{\mathrm{h}} \quad \text{and} \quad \bar{h}_a = \left(\sum_{j=1}^{n_{\mathrm{h}}} h_{aj}\right) \Big/ n_{\mathrm{h}} .$$

Three CPA estimators are defined as follows:

- **The arithmetic mean adjustment (arithmetic mean of intensity ratios):**

$$\hat{\kappa}_{\mathrm{H}} = \frac{1}{n_{\mathrm{h}}} \sum_{j=1}^{n_{\mathrm{h}}} \frac{h_{A,j}}{h_{a,j}} .$$

- **The unbiased adjustment:**

$$\hat{\kappa}_{\mathrm{U}} = \hat{\kappa}_{\mathrm{H}} + \frac{n_{\mathrm{h}}}{n_{\mathrm{h}} - 1} \left(\frac{\bar{h}_A}{\bar{h}_a} - \hat{\kappa}_{\mathrm{H}} \right).$$

- **The geometric mean adjustment (geometric mean of intensity ratios):**

$$\hat{\kappa}_{\mathrm{G}} = \sqrt[n_{\mathrm{h}}]{\left(\prod_{j=1}^{n_{\mathrm{h}}} \frac{h_{A,j}}{h_{a,j}} \right)} .$$

- **The s.e. of the CPA estimates:**

  A bootstrapping procedure is applied to calculate s.e. of the estimated CPA estimates. Assume that the observed CRAS of allele *A* follows a beta distribution, i.e.,

$$f(h_{A,j}) \sim [B(\alpha, \beta)]^{-1} \cdot (h_{A,j})^{\alpha - 1} (1 - h_{A,j})^{\beta - 1}, \alpha > 0, \beta > 0, h_{A,j} > 0,$$

  where $B(\alpha, \beta) = [\Gamma(\alpha) \cdot \Gamma(\beta)] \cdot [\Gamma(\alpha + \beta)]^{-1}$. The moment estimators of the two parameters based on data $\{(h_{A,j}, h_{a,j}), j = 1, \cdots, n_{\mathrm{h}}\}$ are calculated as follows:

$$\hat{\alpha} = [(\bar{h}_A)^2 \cdot (1 - \bar{h}_A)] \cdot S_{h_A}^{-1} - \bar{h}_A$$

$$\text{and} \quad \hat{\beta} = [\bar{h}_A \cdot (1 - \bar{h}_A)] \cdot S_{h_A}^{-1} - (\hat{\alpha} + 1),$$

  where S_{h_A} is the standard deviation of CRAS of allele *A*. The yielded empirical beta distribution is used to generate *B* bootstrapping samples, and then calculate new

CPAs, $\{\hat{\kappa}_b, b = 1, \cdots, B\}$. The sample s.e. of the B bootstrapping CPA estimates is used to estimate the s.e. of the estimated CPA, i.e.,

$$S(\hat{\kappa}) = \sqrt{[\sum_{b=1}^{B} (\hat{\kappa}_b - \bar{\hat{\kappa}})^2] / (B-1)},$$

where $\bar{\hat{\kappa}}$ is the sample mean of the bootstrapping CPA estimates.

APPENDIX B – AF ESTIMATE

Let $(h_A^{\text{Pool}}, h_a^{\text{Pool}})$ be a pair of CRAS (or standardized intensities) of alleles *A* and *a* in a DNA pool. Let $S(\hat{\kappa})$ denote the estimated s.e. of estimated CPA and S_{E} denote the experimental s.e. The unadjusted and adjusted AF estimates are written as follows:

- **The unadjusted AFs of alleles *A* and *a* are:**

$$(\tilde{p}_A, \tilde{p}_a) = \left(\frac{h_A^{\text{Pool}}}{h_A^{\text{Pool}} + h_a^{\text{Pool}}}, \frac{h_a^{\text{Pool}}}{h_A^{\text{Pool}} + h_a^{\text{Pool}}} \right), \ \tilde{p}_A + \tilde{p}_a = 1.$$

- **The adjusted AFs of alleles *A* and *a* are:**

$$(\hat{p}_A, \hat{p}_a) = \left(\frac{h_A^{\text{Pool}}}{h_A^{\text{Pool}} + \hat{\kappa} \cdot h_a^{\text{Pool}}}, \frac{\hat{\kappa} \cdot h_a^{\text{Pool}}}{h_A^{\text{Pool}} + \hat{\kappa} \cdot h_a^{\text{Pool}}} \right), \ \hat{p}_A + \hat{p}_a = 1.$$

- **The s.e. of the adjusted AFs in a pooled DNA with a pool size of *n*:**

$$\left(\hat{V}(\hat{p}_A)\right)^{1/2} = \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{2n} + \left[\hat{p}_A(1-\hat{p}_A) \frac{S(\hat{\kappa})}{\hat{\kappa}} \right]^2 + S_{\text{E}}^2}.$$

APPENDIX C – ASSOCIATION ANALYSIS

Consider a case-control study with n_{case} patients and n_{control} normal controls. Let $\hat{p}_A^{\text{case}}$ and $\hat{p}_A^{\text{control}}$ denote the adjusted AFs of allele A in case and control groups. Let $S(\hat{\kappa})$ denote the bootstrapped variance of estimated CPA. Two single-point association test statistics are written as follows:

- **Single-point association test with a constant CPA:**

$$X^2 = \frac{(\hat{p}_A^{\text{case}} - \hat{p}_A^{\text{control}})^2}{S_1^2 + S_2^2 + S_3^2} \xrightarrow{D} \chi_1^2,$$

where $S_1^2 = \frac{\hat{p}_A^{\text{case}}\hat{p}_a^{\text{case}}}{2n_{\text{case}}} + \frac{\hat{p}_A^{\text{control}}\hat{p}_a^{\text{control}}}{2n_{\text{control}}}$, $S_2^2 = \frac{S^2(\hat{\kappa})}{\hat{\kappa}^2}(\hat{p}_A^{\text{case}}\hat{p}_a^{\text{case}} - \hat{p}_A^{\text{control}}\hat{p}_a^{\text{control}})^2$ and $S_3^2 = 2S_{\text{E}}^2$.

- **Single-point association test with unequal CPAs:**

$$X^2 = \frac{(\hat{p}_A^{\text{case}} - \hat{p}_A^{\text{control}})^2}{S_1^2 + \widehat{S}_2^2 + S_3^2} \xrightarrow{D} \chi_1^2,$$

where $\widehat{S}_2^2 = \left(S(\hat{\kappa}^{\text{case}}) \cdot \frac{\hat{p}_A^{\text{case}}\hat{p}_a^{\text{case}}}{\hat{\kappa}^{\text{case}}} - S(\hat{\kappa}^{\text{control}}) \cdot \frac{\hat{p}_A^{\text{control}}\hat{p}_a^{\text{control}}}{\hat{\kappa}^{\text{control}}} \right)^2$.

Next, we introduce the multipoint association test. Let $\{\theta_1, \cdots, \theta_N\}$ denote the vectors of p-values of N SNPs from the previous single-point association tests. Let s denote the number of SNPs in a window. Let μ denote the threshold of p-value truncation. Let $I[E]$ denote the indicator that takes the value of 1 if event E is true. Let w_{ij} denote a weighted value of the p-value θ_j in the ith window and satisfies that the sum of weights in the window is 1. Three sliding window empirical p-value test (SWEPT) statistics are expressed as follows:

- **Additive SWEPT:** $Z_A(i,s) = \sum_{j=i}^{i+s-1} w_{ij} \times \theta_j \times I[\theta_j < \mu], i = 1, \cdots, N+1-s$
- **Multiplicative SWEPT:** $Z_M(i,s) = \prod_{j=i}^{i+s-1} \theta_j^{w_{ij} \times I[\theta_j < \mu]}, i = 1, \cdots, N+1-s$
- **Minimum SWEPT:** $Z_{Min}(i,s) = \min_{j=i,\cdots,i+s-1} \{\theta_j\}, i = 1, \cdots, N+1-s$

The empirical distribution of the previous test statistics can be obtained based on a Monte Carlo procedure. Based on the empirical distribution, an empirical p-value is calculated to test allelic association.

APPENDIX D – ALLELIC IMBALANCE ANALYSIS

Consider a diallelic SNP with alleles *A* and *a*. Let $h_{A,j}^{\text{Control}}$ and $h_{a,j}^{\text{Control}}$ denote the CRAS of alleles *A* and *a* of the *j*th normal control. Let $h_{A,k}^{\text{Case}}$ and $h_{a,k}^{\text{Case}}$ denote the CRAS of alleles *A* and *a* of the *k*th patient. AFs for the *j*th normal control are as follows:

$$(\hat{\lambda}_{A,j}^{\text{Control}}, \hat{\lambda}_{a,j}^{\text{Control}}) = \left(\frac{h_{A,j}^{\text{Control}}}{h_{A,j}^{\text{Control}} + \hat{\kappa} \cdot h_{a,j}^{\text{Control}}}, \frac{\hat{\kappa} \cdot h_{a,j}^{\text{Control}}}{h_{A,j}^{\text{Control}} + \hat{\kappa} \cdot h_{a,j}^{\text{Control}}} \right).$$

AFs for the *k*th patient are as follows:

$$(\hat{\lambda}_{A,k}^{\text{Patient}}, \hat{\lambda}_{a,k}^{\text{Patient}}) = \left(\frac{h_{A,j}^{\text{Patient}}}{h_{A,k}^{\text{Patient}} + \hat{\kappa} \cdot h_{a,k}^{\text{Patient}}}, \frac{\hat{\kappa} \cdot h_{a,j}^{\text{Patient}}}{h_{A,k}^{\text{Patient}} + \hat{\kappa} \cdot h_{a,k}^{\text{Patient}}} \right).$$

- **Single-point allelic imbalance analysis:**

  For different genotypes $uv \in (AA, Aa, aa)$, the α - level prediction bands are defined as $(\hat{\mu}_{uv} - \eta_\alpha \cdot \hat{\sigma}_{uv}, \hat{\mu}_{uv} + \eta_\alpha \cdot \hat{\sigma}_{uv})$, where $(\hat{\mu}_{uv}, \hat{\sigma}_{uv})$ denotes the mean and prediction error for genotype *uv* and η_α is a critical value. Multiple test correction can be incorporated by dividing α by a correction factor. AF on each locus for each patient, $(\hat{\lambda}_{A,k}^{\text{Patient}}, \hat{\lambda}_{a,k}^{\text{Patient}})$, is compared with the prediction bands established based on a large samples of normal controls to test whether the SNP is in allelic imbalance.

- **Multipoint allelic imbalance analysis:**

  Next we introduce two statistics for a multipoint allelic imbalance analysis. Let $\{\hat{p}_i, i = 1, \cdots, N\}$ denote the AFs of *N* SNPs for a patient. Let AI and HE denote the sets of allelic imbalance and the set of heterozygote respectively.

  (1) *SCORE 1:*

  A SNP in which the AF of allele *A* is outside of the three prediction bands is identified as an allelic imbalance point; otherwise, it is classified as a non-allelic imbalance point. A positive score (*s*) is assigned to an allelic imbalance SNP and negative score (*t*) is assigned to a non-allelic imbalance SNP from the previous single-point allelic imbalance tests. The test statistic is a cumulative sum of the scores as follows:

$$\sum_{i=1}^{j}\{s\cdot I[\hat{\lambda}_{A,i}^{Patient}\in \mathrm{AI}]+t\cdot I[\hat{\lambda}_{A,i}^{Patient}\notin \mathrm{AI}]\},\ j=1,\cdots,N.$$

(2) *SCORE 2:*

All SNPs were divided into heterozygous and non-heterozygous SNPs for each patient. If the estimated AF of a SNP was located within the prediction band $\hat{\mu}_{MN}\pm\eta_{\alpha}\cdot\hat{\sigma}_{MN}$, then the SNP was classified as a heterozygous SNP; otherwise, it was classified as a non-heterozygous SNP. A positive score (x) is assigned to non-heterozygous SNPs and a negative score (y) is assigned to heterozygous SNPs. Scores for SNP points were accumulated from the starting SNP on each chromosome. The test statistic is written as follows:

$$\sum_{i=1}^{j}\{x\cdot I[\hat{p}_i\notin \mathrm{HE}]+y\cdot I[\hat{p}_i\in \mathrm{HE}]\},\ j=1,\cdots,N.$$

The Developing Group of Software *MPDA*

Hsin-Chou Yang, Dr
Assistant Research Fellow
Institute of Statistical Science Academia Sinica
Taiwan

Mei-Chu Huang, Miss
Research Assistant
Institute of Statistical Science Academia Sinica
Taiwan

Lin-Hui Li, Dr
Assistant Research Scientist
Institute of Biomedical Sciences, Academia Sinica,
Taiwan

Cathy S. J. Fann, Dr
Research Fellow
Institute of Biomedical Sciences, Academia Sinica,
Taiwan