GMDR User Manual

Version 1.0

Department of Biostatistics
University of Arkansas for Medical Science
Little Rock, Arkansas, U.S.A.

Institute of Bioinformatics
Zhejiang University,
Hangzhou, Zhejiang, P.R. China

November, 2018

1 Introduction to GMDR Software

1.1 GMDR Approach

*G*eneralized *M*ultifactor *D*imensionality *R*eduction (GMDR) is a nonparametric data mining approach that is alternative to the traditional linear or logistic regression methods for detecting and characterizing nonlinear interactions among discrete genetic and environmental attributes. GMDR is a systematic extension of the multifactor dimensionality reduction (MDR) [1, 2]. Compared with the original MDR method [1, 2], the GMDR method [3-5] permits adjustment for discrete and quantitative covariates and is applicable to a breadth of phenotypes such as continuous, count, dichotomous, polytomous nominal, ordinal, time-to-event, and multivariate, in various study designs. GMDR will increase the prediction accuracy and help users draw a more meaningful conclusion.

1.2 GMDR Software

The GMDR is a free, open-source interaction analysis tool to perform detection of gene-gene and/or gene-environment interactions with the generalized multifactor dimensionality methods. The latest version can be downloaded at the website http://ibi.zju.edu.cn/software/.

1.3 System Requirements

The GMDR software can run on a variety of platforms: Microsoft Windows, Linux, and Mac OSX, on which a Java Runtime Environment is installed. The system requirements include:

Operating System: Linux (Kernel v2.2 or higher), Mac OS X ((v10.0 or higher), or Windows (XP Professional, 7.0, or superior)

Java: Java™ Runtime Environment (build 1.8.0 or later)

RAM: ≥ 512 M

2 Formats of GMDR files

The acceptable file format for inputting data used in the GMDR software is the same as the PLINK's. The data can be loaded in either the text or the binary format. The input data are required to contain the fields corresponding to three pieces of information: pedigree structure, attributes, and phenotypes and covariates. (A singleton is treated as a special family that is composed of only one

member whose ancestors and descendants are missing.) The former two are organized into one or several pedigree file(s) in the text format where the extension name of ".ped" is suggested to use, while they are organized into a family structure file and one or several attribute files, respectively, in the binary format where the extension names of ".fam" and ".bed" are suggested to use, respectively. Several pedigree files or attribute files can be merged into a single file for the subsequent analysis or outputting. The phenotypes and covariates are organized into a phenotype file where an extension name of ".phe" is usually used. Additionally, it is also required to provide a map file (".map" in the text format and ".bim" in the binary format) that contains the information of genetic markers/genes such as name, chromosome number, physical and genetic positions. The requirements for data are described as follows.

2.1 PED file

The PED file contains the information of the individuals, such as, the individual's ID, sex and the genotypes. The PED file is a white-space (space or tab) delimited file: the first six columns as follows are mandatory:

Family ID
Individual ID
Paternal ID
Maternal ID
Sex (1=male; 2=female; 0=unknown)
Affection status

And the genotypes of each SNPs is coded as follows:

1 1	Homozygote "1"/"1"
1 2/2 1	Heterozygote
2 2	Homozygote "2"/"2"
0 0	Missing genotype

A part of a sample PED file, example.ped, is shown as follows:

```
1 10000 0       0       1 0 2 2 2 1 1 1 2 1 2 1 2 1 2 2 1 1 2 1 1 1 1 1 2 2 2 1 2 1 2 2
1 10001 0       0       2 0 2 2 1 1 2 1 1 1 2 1 2 1 2 1 2 1 1 1 1 1 1 1 2 1 1 1 2 2 2 2
1 10002 10000 10001 1 0 2 2 2 1 1 1 2 1 2 1 2 1 2 1 1 1 1 1 1 1 1 2 2 2 1 2 2 2 2 2
1 10003 10000 10001 1 1 2 2 1 1 2 1 1 1 2 1 2 1 2 1 1 1 1 1 1 1 1 2 2 1 1 2 2 2 2 2
2 20000 0       0       1 0 2 2 2 1 1 1 1 1 2 2 2 1 2 1 2 1 1 1 2 2 1 1 2 2 1 1 2 1 2 2
2 20001 0       0       2 0 2 1 2 1 1 1 1 2 2 2 1 1 1 1 1 2 1 1 1 2 1 1 1 2 2 1 1 2 2 2 1
2 20002 20000 20001 1 1 2 2 2 2 1 1 1 2 1 2 2 2 1 1 1 1 1 1 1 2 2 1 1 2 2 1 1 2 1 2 2
```

2.2 MAP file

The MAP file consists of information about the markers. By default, each line of the MAP file describes a single marker and must contain exactly 4 columns:

Chromosome (1-22, X, Y or 0 if unplaced)
SNP identifier
Genetic distance (Morgans)
Base-pair position (bp units)

The MAP file must contain as many markers as are in the PED file. A typical description for a MAP file, example.map, looks like:

```
1    snp0    0    27374308
1    snp1    0    27374342
2    snp2    0    27375209
2    snp3    0    27375722
8    snp4    0    27375822
8    snp5    0    27379777
8    snp6    0    27380739
8    snp7    0    27380761
10   snp8    0    27381390
10   snp9    0    27382044
10   snp10   0    27383938
18   snp11   0    27384428
18   snp12   0    27385587
18   snp13   0    27386013
18   snp14   0    27386203
```

To save storage space and processing time, users can input data in the binary files. The information in the PED file(s) will be stored in separate files: the information on pedigree structure, affection status and other non-binary attributes stored in a FAM file (*.fam) and that on SNPs stored in a binary PED file (*.bed) while creating an extended MAP file (*.bim) (which contains, in addition to that in the MAP file, information about the allele names, which would otherwise be lost in the BED file). The .fam and .bim files are still plain text files: these can be viewed with a standard text editor.

2.3 BED file

As same as the file format for the PLINK, the part of information on SNPs in the PED file(s) can be converted and saved in the binary format——BED file. Each biallelic genotype is each biallelic genotype is compressed in 2 bits only, instead of 2 bytes (2*8 = 16 bits).

2.4 FAM file

Going with the procedure of compression, the FAM file is built. The FAM file has the data in the PED file except the genotypes. And that mean the data of the FAM file is the first six columns in PED file. For example, the file example.fam appears below:

```
1    10000    0        0        1    0
1    10001    0        0        2    0
1    10002    10000    10001    1    0
1    10003    10000    10001    1    1
2    20000    0        0        1    0
2    20001    0        0        2    0
2    20002    20000    20001    1    1
2    20003    20000    20001    1    0
3    30000    0        0        1    0
3    30001    0        0        2    0
```

3	30002	30000	30001	1	1
3	30003	30000	30001	1	0
4	40000	0	0	1	0
4	40001	0	0	2	0
4	40002	40000	40001	1	1
4	40003	40000	40001	1	0
5	50000	0	0	1	1
5	50001	0	0	2	1
5	50002	50000	50001	1	0
6	60001	0	0	2	1

2.5 BIM file

The procedure of compressing will generate the BED file, the FAM file and the BIM file.The BIM file is an extended MAP file, which also includes the names of the alleles: (chromosome, SNP, cM, base-position, allele 1, allele 2):

1	snp0	0	27374308	1	2
1	snp1	0	27374342	2	1
2	snp2	0	27375209	2	1
2	snp3	0	27375722	1	2
8	snp4	0	27375822	1	2
8	snp5	0	27379777	1	2
8	snp6	0	27380739	1	2
8	snp7	0	27380761	2	1
10	snp8	0	27381390	2	1
10	snp9	0	27382044	2	1
10	snp10	0	27383938	2	1
18	snp11	0	27384428	1	2
18	snp12	0	27385587	2	1
18	snp13	0	27386013	1	2
18	snp14	0	27386203	1	2

2.6 PHE file

The PHE file contains the data of phenotypes for analysis. The PHE file have a header row, in which case you can use variable names to specify the response variable(s) or covariate(s). The first two columns are the family id and the individual id of each individual to link to the pedigree/attributes file(s) and the next column is the phenotypes/covariates of each individual. For example:

FID	ID	phe0	phe1	phe2
1	10000	0	0.8875	0.1125
1	10001	0	0.875	0.125
1	10002	0	0.87	0.13
1	10003	1	0.88	0.12
2	20000	0	0.6325	0.3675
2	20001	0	0.69	0.31
2	20002	1	0.675	0.
2	20003	0	0.6625	0.3375
3	30000	0	0.865	0.135
3	30001	0	0.8025	0.1975
3	30002	1	0.835	0.165
3	30003	0	0.8425	0.1575
4	40000	0	0.7725	0.2275
4	40001	0	0.76	0.24

3 Basic usage

The GMDR is a java program, and can be started by clicking the icon or executing the java command at the terminal like “java –jar GMAR.jar”. The software has two kinds of user-friendly interfaces: Graphical User Interface (GUI) and Command Line Interface (CLI). GUI can run in majority of desktop systems, and CLI can run in all the popular shell systems. We will use the GUI to exemplify the usage of GMDR with the example files.

3.1 Build a project file

Figure 1 shows the main interface of the GMDR as soon as users start GUI. Clicking the main menu “Project”, users can select desirable items such as ”New” and “Open” to create a new and open an old project file, respectively. The default directory is the same as that in which gmdr software is located and users can change as needed.

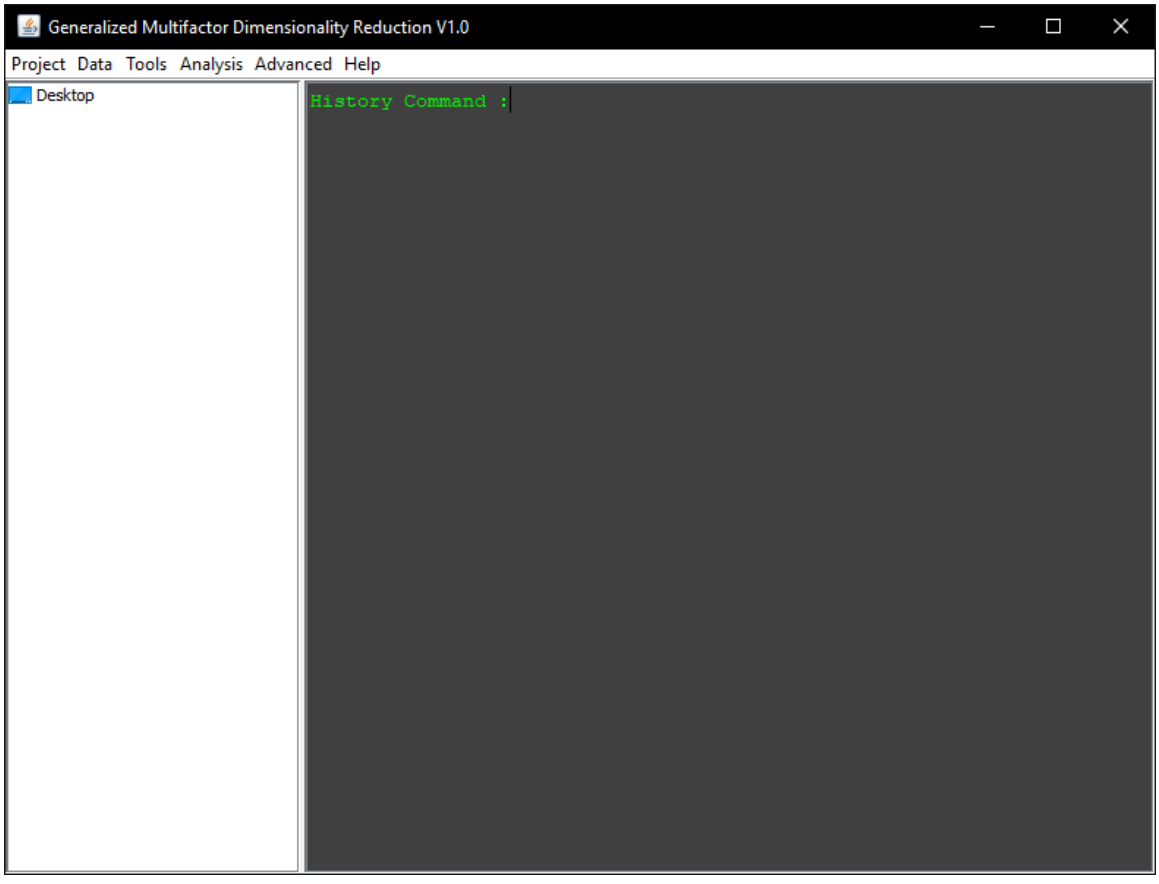

Figure 1. The GUI main interface of GMDR

3.2 Input the files

After having a project file, users need load the data for analysis. As mentioned above, there are two kinds of data input formats, one is the general text files and the other is the binary files. Buttoning the Data->Load Data, a dialog box will appear like Figure 2. The input dialog box offers various choices to load the data. According to the different data files, users choose an appropriate tab to input the data.

In addition to clicking the “Browse” button to load data files individually, users also can use the function of “QuickFileset”. When the “QuickFileset” is chosen, the combobox will become active. All the available files in the directory of the project file will appear.Selecting a file, the program will load the files that have the same stem name in the working directory . For example, if choosing the example.bed in the combobox of the binary tab, the program will automatically load the example.bed, the example.bim and the example.fam if they exist in the same folder. Consistent with the process of loading binary files, we can load the general text files, the PED file and the MAP file.

As for the “Alternate Phenotype” tab, it is used to load the phenotype file which will be read in analysis of model.

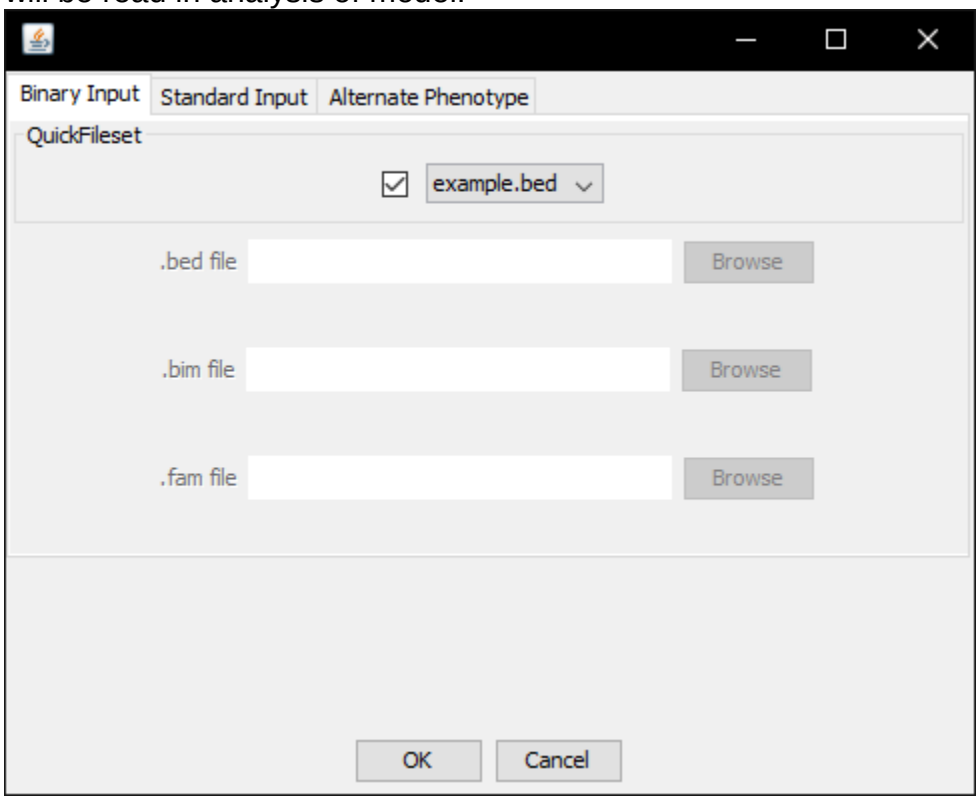

Figure 2. The input dialog box.

3.3 Analysis

After the data files are loaded, users can start to perform the analysis. Clicking the Analysis button on the main window, a tabbed panel will pop out as shown in Figure 3. Users can calculate the residuals, set the configurations to compute membership coefficients, and then conduct the MDR analysis for dimensionality reduction based on the residuals.

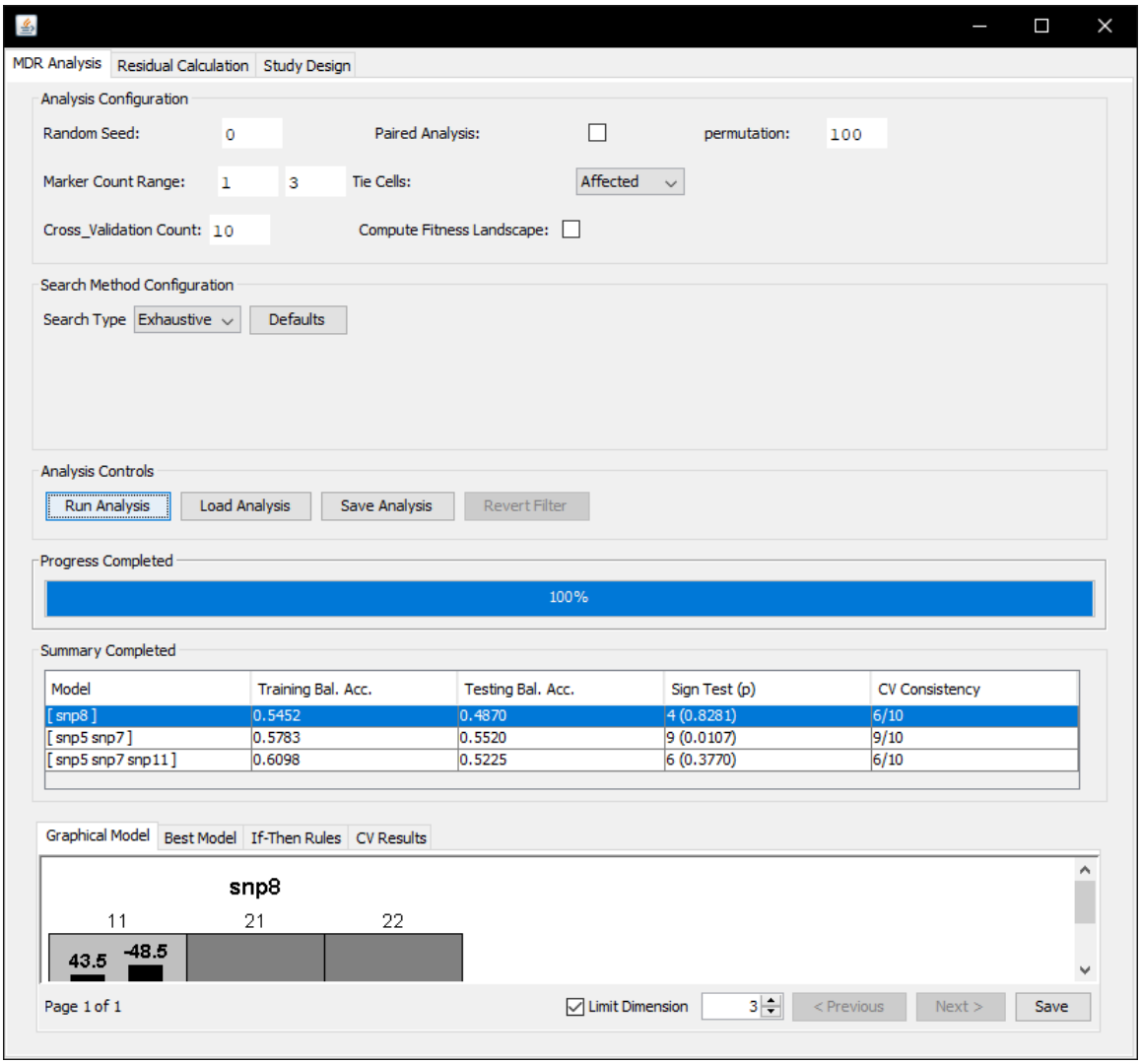

Figure 3. The interface of the analysis

3.3.1 Residual calculation

After clicking the residual calculation tab located on the analysis window the residual calculation tab will appear as shown in Figure 4. If the phenotype file is

not yet loaded, users may use the “Load Phenotype” here. After loading the phenotype file, a table of the data will be available for choose the response variable(s) and covariate(s). By clicking the button of “▼” to add the selected column(s) into the textbox of the “response” and “predictor” as the dependent variable(s) and covariate(s). The link function and estimating method are used to specify the statistical model and estimation procedure. And then click run to compute the residuals under the null hypothesis.. After completing calculation, a list of residuals will be displayed in the textbox of the “Residual”.

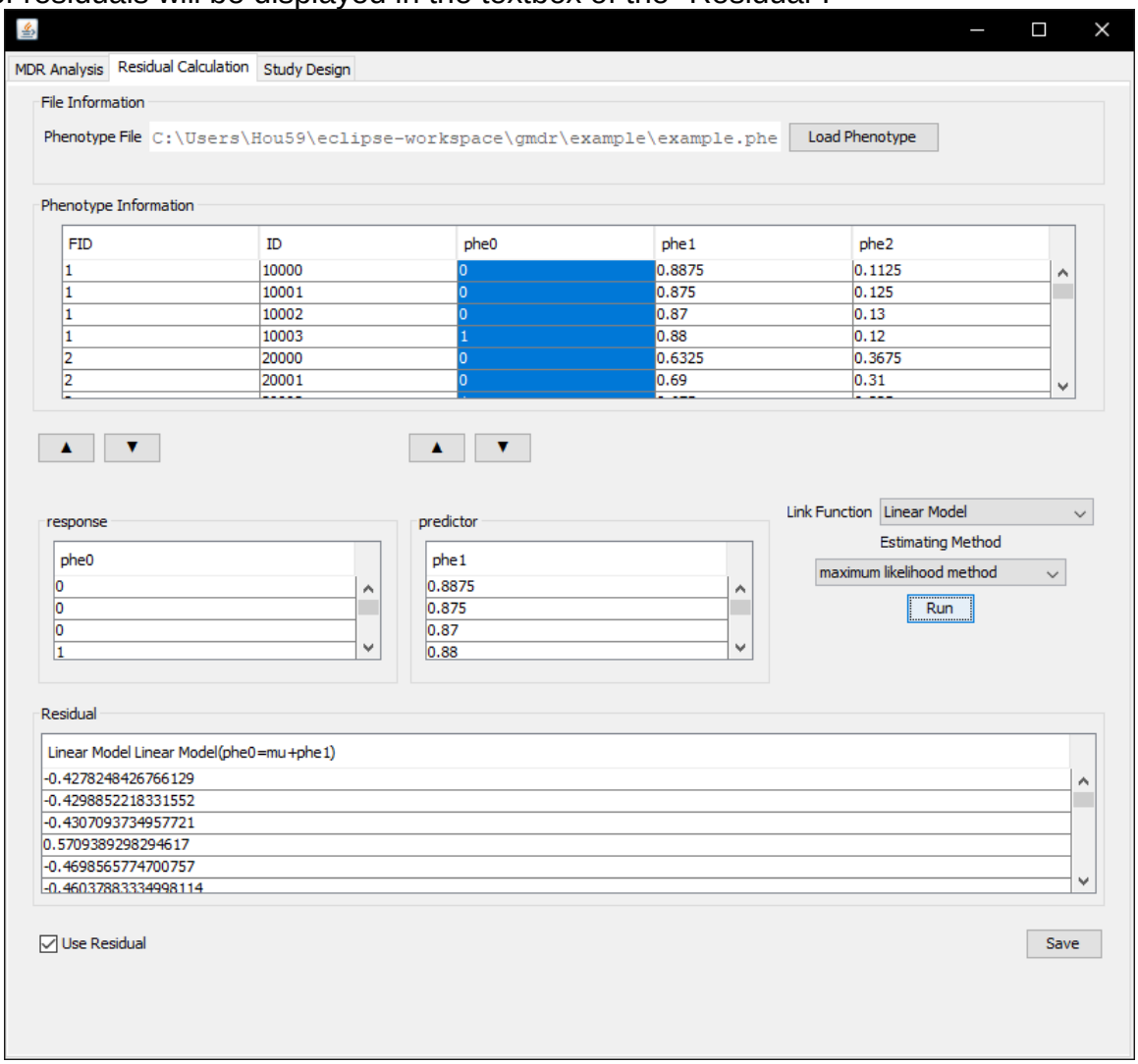

Figure 4. The residual calculation tab

3.3.2 Study Design

Study design table is used to specify the configurations for computation of membership coefficients. If the Population Stratification box is checked, a file on principal components can be loaded for adjust the potential population structure in the unrelated individuals. The membership coefficients will be calculated for

unrelated samples (singletons and founders in families) and nonfounders in families according to the pedigree structure. Either unrelated sample only, nonfounders only, or pooled samples of both unrelated samples and nonfounders can be specified for analysis.

3.3.3 MDR Analysis

After calculating the residuals and membership coefficients, the statistic values, defined as the product of the residual and membership coefficient, are loaded as the input of the MDR analysis and users can perform MDR analysis for dimensionality reduction by the “MDR analysis” tab. As shown in Figure 5, there are many options to change the running arguments that will be mentioned below. After completing analysis, the result of the analysis will be shown in the area of the table named “summary and can export the output data in several popular formats as users need.

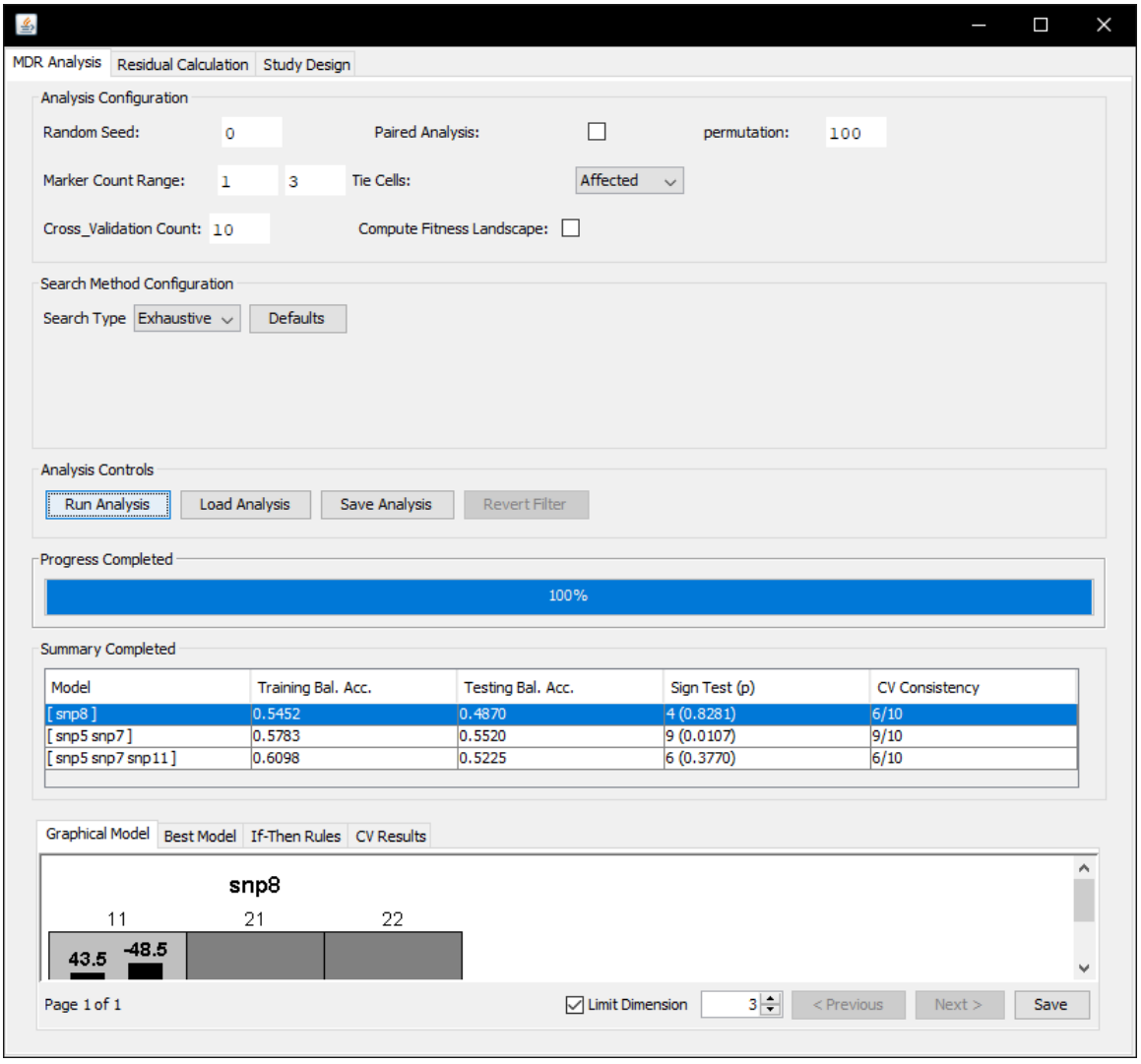

Figure 5. The panel of MDR analysis

4. The GMDR Configuration

As mentioned above, users can change the default running parameters on the panel of MDR analysis, which will be introduced one by one.

4.1 Random Seed

This option sets the seed for the random number generator for randomly dividing the data into cross-validation partitions. Reuse of the same seed will repeat same analysis. The primary function of this random seed is to determine the order of the data when it is shuffled for cross validation, but it is also used for others such as sample selection for ReliefF.

4.2 Attribute Count Range

This field contains the range for the number of attribute combinations to be considered. The default configuration value of 1 to 4 tells GMDR to conduct analysis at all combinations of 1, 2, 3 and 4 attributes.

4.3 Cross-Validation Count

The value in this field is the level of cross validation to be performed during the GMDR analysis. For example, a value of 10 means performing 10-fold cross validation. If the files are previously loaded, the system will check to ensure that the level of cross validation is possible for the available instances. A value of 1 placed in this field indicates that no cross validation is to be performed. If an *even value* greater than or equal to 10 is used in the “CV Count” field, the Sign Test (p) column in the *”Summary Table”* will display values for each combination.

4.4 Paired Analysis

This field indicates that the marker file loaded is using a paired analysis format Alternating rows of cases and controls that are matched in the marker file. For example, this option should be used for matched case-control studies in epidemiology and assumes the matching is 1:1. As a result of such a setting, the

matched pairs are kept together during cross-validation.

NOTE: This option is unselected by default. This option should not be used when importing a non-paired marker file; otherwise, incorrect results will be produced in the GMDR analysis.

4.5 Tie Cells

This field allows the user to indicate how patterns with an equal number of affected and unaffected data within a cross-validation sample should be classified.

NOTE: In a case of continuous statistic values, it is rarely to encounter a tie cell.

4.6 Compute Fitness Landscape

When this option is enabled, the GMDR software will keep track of the training accuracy value for each combination of attributes examined. This information is displayed at the end of the analysis via several different views. As this feature consumes a great deal of memory, it may be advisable to disable it for datasets with too many attributes. This option is disabled by default.

4.7 Search Method Configuration

Broadly speaking, a **search algorithm** is a procedure that takes a problem as input and returns a solution to the problem, usually after evaluation of many possible solutions.

The set of all possible solutions to a problem is called the search space. Brute-force search or "naïve"/uninformed search algorithms use the simplest, most intuitive searching method through the search space, whereas informed search algorithms use heuristics to apply knowledge about the structure of the search space to try to reduce the amount of searching time required.

This section of the configuration allows the user to determine what type of search algorithm will be applied during the GMDR analysis. The present search algorithms available are as follows:

4.7.1 Exhaustive

The “brute force algorithm”, this method attempts to find the solution within the whole search space by evaluating every possible solutions. This is the default

for GMDR and requires no additional parameters.

4.7.2 Forced

This searching strategy allows the user to indicate specific attributes to be used as opposed to consideration of all attributes within the dataset. Values are separated by comma. For example, the list of value 1, 2, 3, 4, 5, 9 would indicate only evaluating the combination of attribute 1, 2, 3, 4, 5, and 9 in the analysis. Also the actual header values may be entered for the forced attributes (such as X0, X1, X3 considering these three attributes only). These values are case sensitive.

4.7.3 Random Search

This searching strategy randomly navigates the available search space and generates random attribute combinations for evaluation. This implementation allows replacement (selection of the same solution may occur more than once for evaluation). This method allows the user to determine the length of the evaluation by the measurement of time (Runtime option) or in terms of evaluations performed (Evaluations option).

The “Evaluations” option allows the user to enter an integer value for the max number of evaluations to be performed per attribute count by the searching algorithm before end of the analysis.

The “Runtime” option allows the user to specifically allot an amount of time for the search to continue. The time may be entered as a real number and allows the user to select the unit of measurement between seconds, minutes, hours and days.

1. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer**. *American Journal of Human Genetics* 2001, **69**:138-147.

2. Moore JH: **Computational analysis of gene-gene interactions using multifactor dimensionality reduction**. *Expert Review of Molecular Diagnostics* 2004, **4**(6):795-803.

3. Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD: **A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence**. *Am J Hum Genet* 2007, **80**(6):1125-1137.

4. Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD: **A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies**. *Am J Hum Genet* 2008, **83**(4):457-467.

5. Lou X-Y: **UGMDR: a unified conceptual framework for detection of multifactor interactions underlying complex traits**. *Heredity* 2015, **114**(3):255-261.