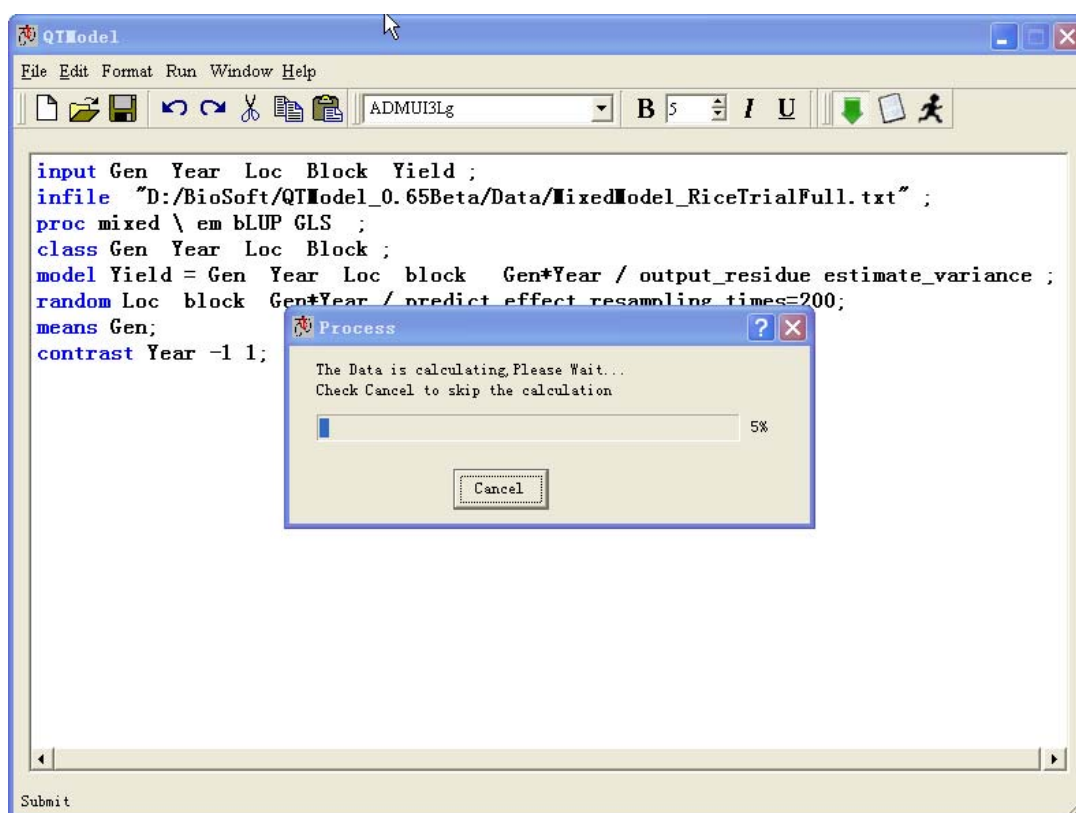


# QTModel User Manual

Software for Quantitative Trait and Statistical Analysis of Experimental Data

(Microarray Data Analysis, Diallele Cross Analysis, Mixed Linear Model Analysis)

Jian Yang, Junqi Cheng, Yangyun Zou, Zhen Xia and Jun Zhu



Copyright 2007 ©

Zhejiang University, China

This software can be freely redistributed with non-commercial purpose and under the condition of no changes to the software and its related documents. No warranty of any sort is provided for use of the software.

## 1. Introduction to QTModel

QTModel is user-friendly computer software which packaged with modules for microarray data analysis, diallele design analysis and mixed model analysis.

### 1.1 Mixed model analysis (*mixed* module)

The *mixed* module is developed for analyzing data from experimental designs with random factors. It is now available for commonly used randomized block design, randomized complete block design, latin square design, factorial design, multi-factor factorial design, nested design, and cross nested design etc. For fixed factors, pair-wised comparisons are done for all possible pairs of fixed effects of one factor. For random factors, some mixed linear model approaches, such as MINQUE, MIVQUE, REML and EM, will be used to estimate the variances of these random factors, and also unbiased prediction methods, such as BLUP, LUP and AUP, are used to predict the random effects of the random factors.

### 1.2 Microarray data analysis (*array* module)

The module of microarray data analysis is based on the HAB-Array method (Henderson III and Bayesian method for microarray analysis), which is proposed for detecting DEGs (Differentially Expressed Genes) under specific treatments. Currently, it has the capacity of analyzing data with one or two treatment factors. If the treatment is treated as random factor, the effect of each treatment level can be predicted by MCMC (Markov Chain Monte Carlo) algorithm or conventional mixed linear model approaches such as BLUP (Best Linear Unbiased Prediction) and LUP (Linear Unbiased Prediction). While if the treatment is treated as fixed factor, pair-wise comparison will be done between all possible pair of treatment levels. The selected DEGs will be exported for cluster analysis by the software ClusterProject.

### 1.3 Diallele cross analysis (*diallel* module)

The *diallel* module is developed for analyzing data from diallel cross designs. It is commonly used to analyze classical quantitative traits, including agronomic trait, seed trait, endosperm trait, animal trait, as well as major-gene trait. For random factors including additive effect, dominance effect, additive by additive epistatic effect, cytoplasmic effect, maternal effects, paternal effect, sex linkage effect, endosperm effects, environment effect etc, MINQUE method will be used to estimate the variance components for single trait and covariance


components for multiple traits of these random factors, and also unbiased prediction method LUP is used to predict the random effects of these random factors.

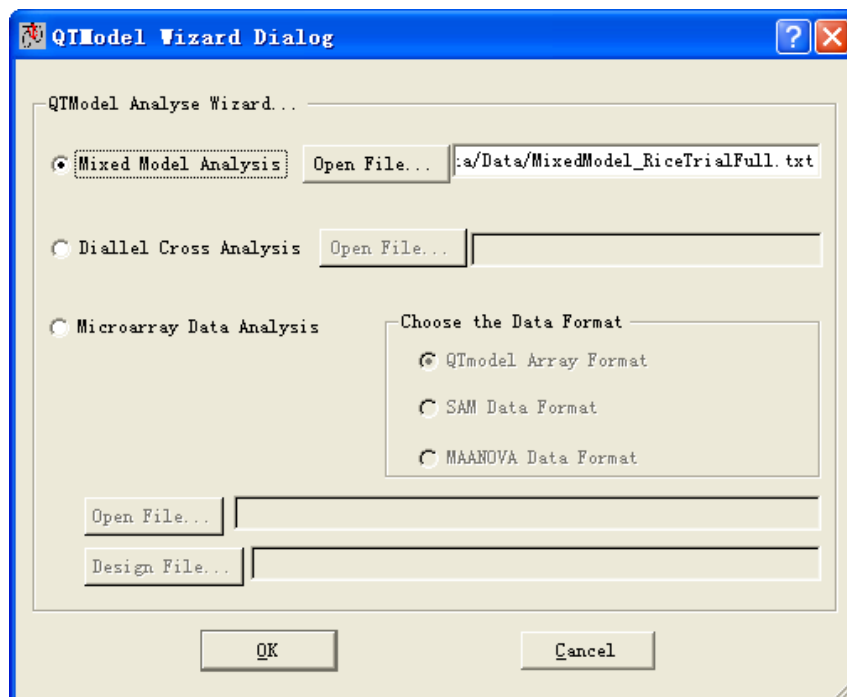
### 1.4 Graphic user interface

The software is programmed by C/C++ programming language. The GUI (Graphic User Interface) of QTModel is developed by QT (a GUI software toolkit). Since QT is a cross-platform C++ class library, any program developed based on QT can be compiled and run any operating system with QT installed. Now QTModel have binary executive programs available for Windows XP, RedHat Linux, and IBM AIX. We will describe the application of QTModel in the following section.

## 2 Starting with QTModel

The software is freely available from the URL <http://ibi.zju.edu.cn/software/qtmodel/>. Download the QTModel setup package QTModel-0.65Beta-Setup.exe, and double click it to install. Click “Start” → “Programs” → “QTModel” → “QTModel.exe” or click the shortcut in desktop to run QTModel software.

Click “File” → “New” menu or the first  button on the Toolbar. A wizard dialog will be popped up for selecting the modules.



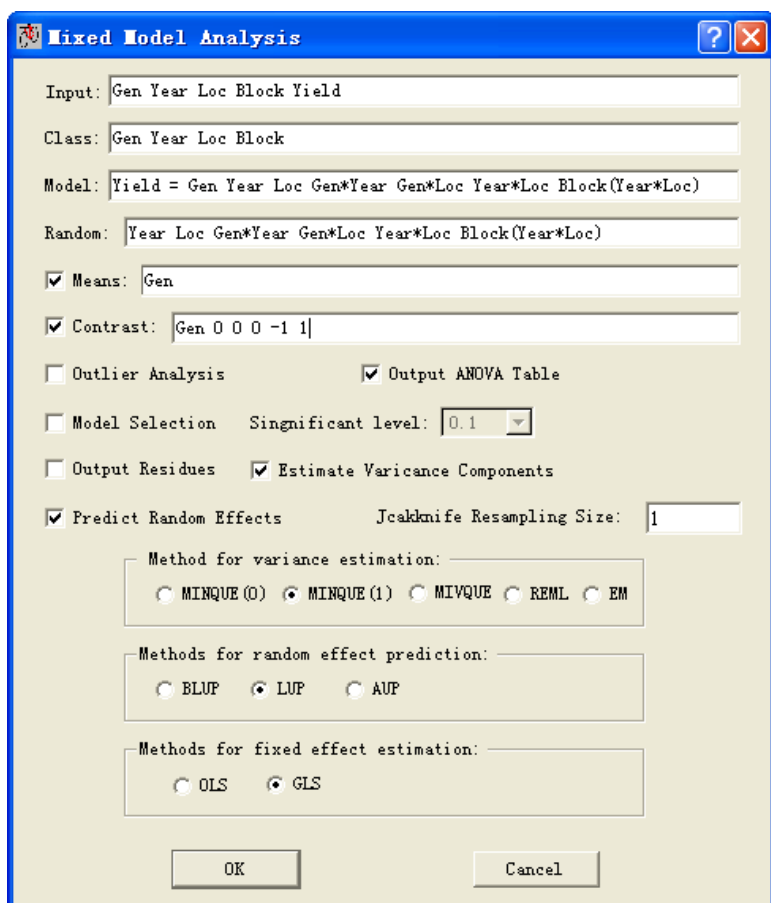
## 2.1 Mixed Linear Model Analysis

Click the “Mixed Model Analysis” check box on the wizard dialog to use the *mixed* module of QTModel. Load the source data file by clicking the button “Open File...”. The data file should be in text format with data points separate by tabs or white-spaces. Missing data points should be represented by dot “.”. Here is an example of the data file.

**Example 1** Data format for *mixed* module of QTModel.

Gen	Year	Loc	Block	Yield
1	1	1	1	19.7
1	1	1	2	31.4
1	1	1	3	29.6
1	1	2	1	40.8
1	1	2	2	.
1	1	2	3	30.2
1	1	3	1	34.7
1	1	3	2	29.1
1	1	3	3	35.1
⋮	⋮	⋮	⋮	⋮
5	2	4	1	32.6
5	2	4	2	40
5	2	4	3	34.2

Click “OK” to enter the dialog to specify the model and options.



**Input:** the name of each column in the data file.

**Class:** the classification variables to be used in the model. These are called class variables (They can also be called categorical, qualitative, discrete, or nominal variables.) Class

variables can be either numeric or character. The values of a class variable are called levels. For example, the class variable Sex has the levels "male" and "female." An independent variable that is not declared in this statement is assumed to be continuous. Continuous variables must be numeric. For example, the heights and weights of subjects are continuous variables.

**Model:** specify the dependent variables and independent effects in the model. Each term in a model, called an effect, is a variable or combination of variables. Effects are specified with a special notation using variable names and operators. There are two primary operators: crossing and nesting. For example, suppose there are three independent effects  $A$ ,  $B$  and  $C$ , we could have the combinations  $A*B$ ,  $A*C$ ,  $C*B$  and  $A*B*C$  for crossing, and  $C(A)$  and  $C(A*B)$  etc. for nesting.

**Random:** specify some independent effects that are assumed to be sampled from a normal population of effects as random effects. These random effects should appear after the “*Model*” statement.

**Means:** do pairwise comparisons among different levels within the effect specified in this statement. The effects specified in this statement must be fixed classification effects and must appear after the “*Model*”.

**Contrast:** enable you to perform custom hypothesis test by specifying an vector  $C$  for testing the hypothesis  $Cb = 0$ , where  $b$  is an vector of the specified effect. The effects specified in this statement must be fixed classification effects and must appear after the “*Model*”. For example, suppose a fixed classification effect  $A$  with three levels, the statement

for testing  $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} [A_1 A_2 A_3] = 0$  can be written as “ $A -1 0 1$ ”.

**Outlier Analysis:** do outlier analysis. In the result, all the data points will be listed by the order of their  $P$ -values. Any data point with  $P$ -values less than the critical  $P$ -value is indicated as an outlier data point. The value “ $CD(b)$ ” represents the influence of this data point to fixed effects. And the value “ $CD(u)$ ” represents the influence of this data point to random effects.

**Output ANOVA Table:** Output a table for the result of ANOVA analysis.

**Model Selection:** select independent effects in the model by model selection technique. Currently, stepwise model selection strategy is used. The “*Significant Level*” specifies the significant level for entry into or staying in the model used in the stepwise selection method. The default value is 0.1.

**Output Residues:** output the estimated residues by the given model.

**Estimate Variance Components:** estimate the variance components of the random effects, for fixed model, only estimate the residual variance. If you choose the option “*Predict Random Effects*”, Jackknife resampling technique will be used to do significance tests for all the variance components.

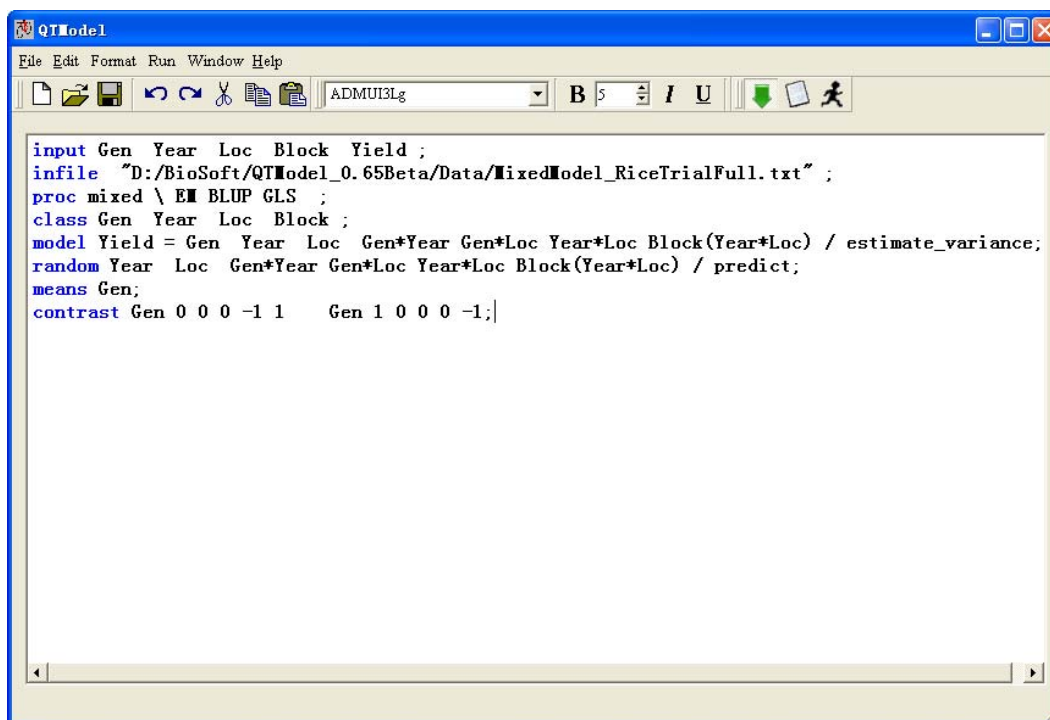
**Predict Random Effects:** predict the random effects as well as do significance tests by jackknife resampling technique. The “*Jackknife Resampling Size*” specify how many data points will be removed for each resampling iterate.


**Method for variance estimation:** choose the method for estimating variance components. MINQUE (Minimum Norm Quadratic Unbiased Estimation), MIVQUE (Minimum Variance Quadratic Unbiased Estimation), REML (Restricted Maximum Likelihood Estimation), and EM (Expectation-Maximization) are provided.

**Method for random effect prediction:** choose the method for predicting random effects. BLUP (Best Linear Unbiased Prediction), LUP (Linear Unbiased Prediction), and AUP (Adjusted Unbiased Prediction) are provided.

**Method for fixed effect estimation:** choose the method for estimating fixed effects. GLS (Generalized Least Square) and OLS (Ordinary Least Square) are provided.

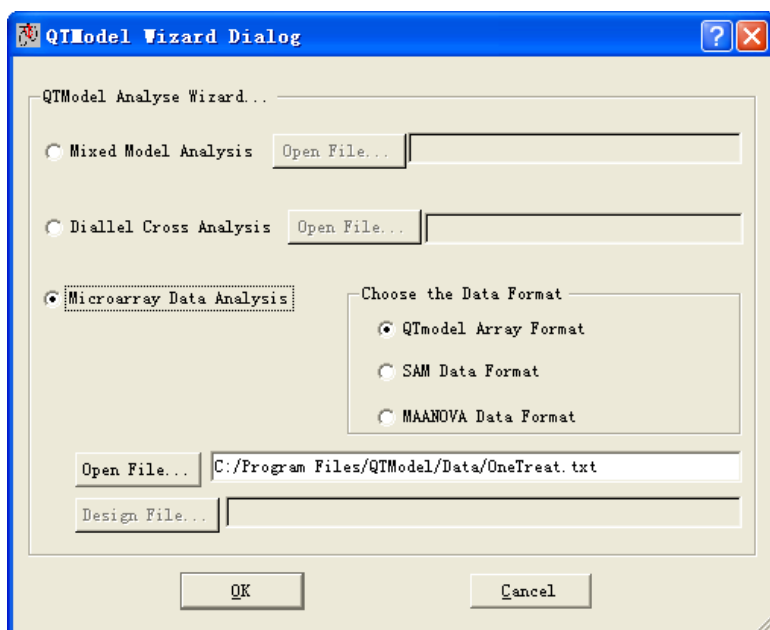
After finish specifying the model and options, click “OK” to close the dialog. All the information in the dialog will be saved in a command file with “qcf” as its extension name and displayed in the “Edit” window of QTModel. You can change the model and options by editing this command file.



Click “Run” → “Submit” menu or the last button  on the toolbar to analyze the data by the specified commands. When computation finishes, the result will be saved in a text file with “qtm” as its extension name and displayed in the “Output” window of QTModel.

## 2.2 Microarray Data Analysis

Click the “Microarray Data Analyse” check box on the wizard dialog to use the *array* module of QTModel. Load the source data file by clicking the button “Open File...”.



### 2.2.1 QTModel Array Format

QTModel can accept three kinds of data formats including QTModel array format, SAM



format and MAANOVA format.

*QTModel array format:* the first column is gene ID. Data for the same gene should be listed together. The last column is response values (gene expression intensities or ratios). Missing expression measurements must be denoted as 99999 or dot. The remaining columns are technical or biological factors, such as Array, Dye for technical factors and treatments for biological factors, which can be denoted as either integers or characters. Here are two examples in QTModel array format for one treatment and two treatments, respectively.

**Example 1** QTModel array format for one treatment cDNA microarray data

GeneID	Array	Dye	Treat	Rep	Intensity(Log)
YHR007C	1	1	1	1	9.89504
YHR007C	1	2	2	1	4.88072
YHR007C	2	1	1	2	99999
YHR007C	2	2	2	2	3.85794
⋮	⋮	⋮	⋮	⋮	⋮
YAL056W	6	1	2	3	12.5749
YAL056W	6	2	3	3	12.1003
YAL056W	7	1	3	1	12.5645
YAL056W	7	2	4	1	11.3168
⋮	⋮	⋮	⋮	⋮	⋮

**Example 2** QTModel array format for two-treatment Affymetrix microarray data

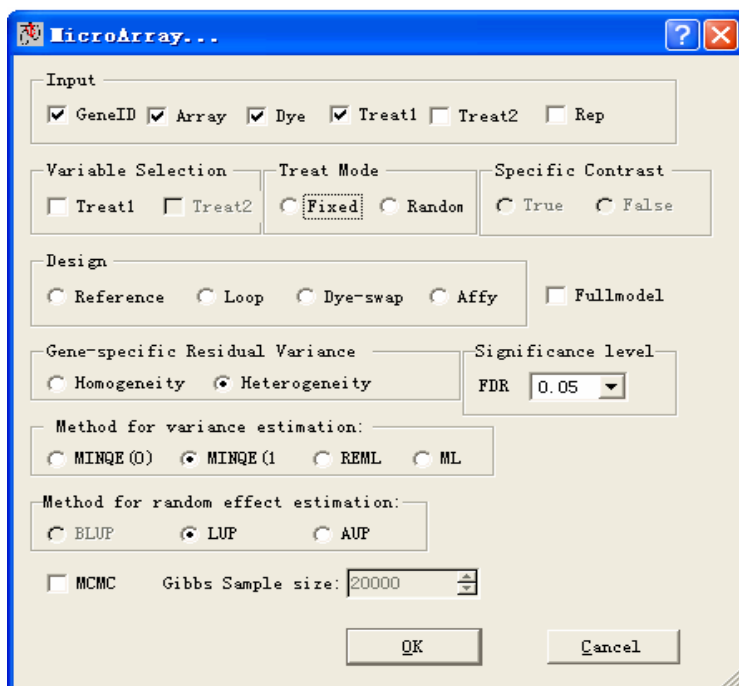
GeneID	Treat_1	Treat_2	Rep	Intensity(Log)
aa114725	1	1	1	7.96
aa114725	1	1	2	8.13955
aa114725	1	2	1	8.30834
aa114725	1	2	2	8.10852
⋮	⋮	⋮	⋮	⋮
aa002704	2	5	1	6.78136
aa002704	2	5	2	5.9542
aa002704	2	6	1	6.89482
aa002704	2	6	2	6.10852
⋮	⋮	⋮	⋮	⋮

QTModel can also handle the data files in SAM (Significant Analysis of Microarray) and MAANOVA (Micro Array ANalysis Of VAriance) formats. Detailed information about these two data formats could be found in the user manual of these two software package. When data file in SAM or MAANOVA format is imported in QTModel, QTModel will automatically

save another file in QTModel array format with the data file name added by “\_QTM” after the original data file name.

### 2.2.2 Run Microarray data by QTModel

The below is MicroArray Analyse box, which will pop out automatically when you clicked OK menu in the above dialog.



**Input:** Experimental variables, which should be correspond to the name of the first line in the data file. The expression value is not included here.

**Variable Selection:** The term is necessary, you should show which factor you are to use for the identification of differentially expressed genes.

**Treat Mode:** This term indicates whether your treat of interest is fixed or random. If “Fixed ” option is chose, all biological factors are set as fixed, and do multiple comparisons for the factor you choose in the term of “Variable Selection”, while “Random” option means all biological factors are random, and it provides the gene by treat interaction for the analysis of gene expression profile.

**Specific Contrast:** A logical value to indicate whether to display the significant genes in the specific level of the other factor in two or multiple –factor experiment. This command is just available in the two factors experiment.

**Design:** The term defines the experimental design of your microarray data which is very

important for statistical model construction, so it is required. Available options includes: reference, loop, dye-swap, affy.

**Full Model:** This term indicates whether to do multiple-gene model fitting which is available for the estimate of the magnitude of the biological variation and other technical variation as well. Because of the large size of microarray data, it is very time-consuming.

**Gene-specific Residual Variance:** Assumption used when fit gene-specific model. Two options are available, “Homogeneity” and “Heterogeneity”. “Homogeneity” is for the assumption of the data with genes in the experiment sharing the equal or common variance, while “Heterogeneity” is for the assumption of gene-specific variance. By default, “Heterogeneity” assumption will be used for data analysis.

**Significance Level:** The term provides the statistical threshold for multiple significance tests. Any value is ok, like 0.1, 0.05, 0.01, ..., depending on your experimental goal. Default FDR is 0.05.

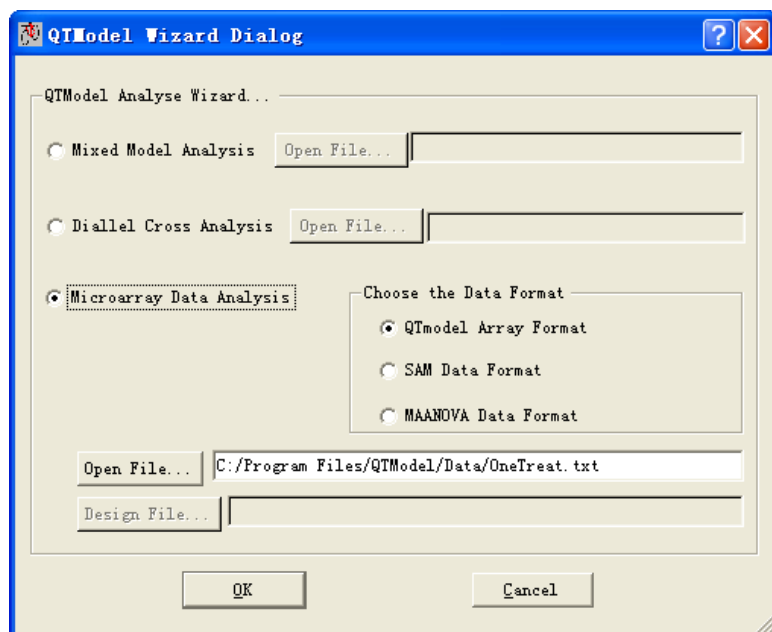
**Method for variance estimation:** There are four methods for you to choose for the estimate of variance components of the random variable in the model. This term is active when the “Treat Mode” option is “Random”.

**Method for random effect estimation:** This command used for the random effect estimation, and works just when “Treat Mode” option is “Random”.

**MCMC:** Markov Chain Monte Carlo (MCMC) is alternative method for the estimation of variance components of random variable and random effects. This method is also time-consuming.

## 2.3 Diallele Cross Analysis

Click the “Diallel Cross Analysis” check box on the wizard dialog to use the *diallel* module. Load the source data file by clicking the button “Open File...”.



### 2.3.1 QTModel Diallel Format

The data file should be in text format with data points separate by tabs or white-spaces. Missing data points should be represented by dot “.”. *Diallel* module can accept the following three kinds of data formats for agronomic trait, seed trait, endosperm trait, animal trait and major-gene trait:

**Example 1** Data format for agronomic trait, seed trait and endosperm trait *diallel* module

Env	Female	Male	Cross	Block	Yield
1	1	1	0	1	54.4
1	1	1	0	2	29.7
1	1	1	0	3	54.3
1	1	4	1	1	38.5
1	2	1	1	2	62.5
1	2	1	1	3	.
⋮	⋮	⋮	⋮	⋮	⋮
2	3	3	0	1	27.8
2	3	3	0	2	19.1
2	3	4	1	1	37.9
2	3	4	1	2	34.2
2	4	4	0	1	49.5
2	4	4	0	2	57.1

**Example 2** Data format for animal trait *diallel* module

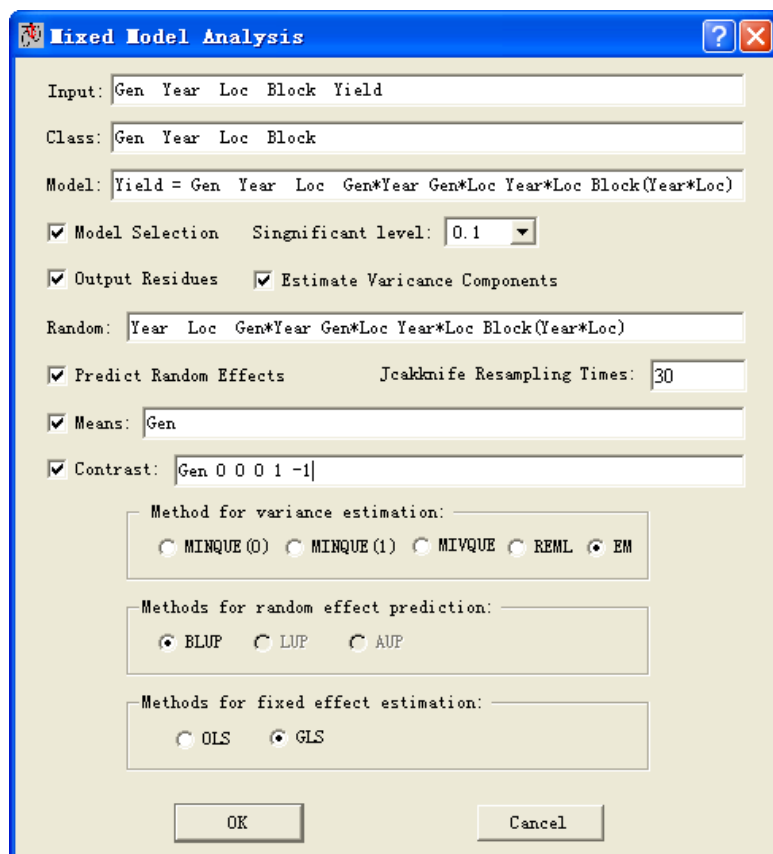
Env	Female	Male	Cross	Block	Sex	TL
1	1	1	0	1	1	93.1
1	1	1	0	1	2	92.6

1	1	2	1	1	1	97.4
1	1	2	1	1	2	94.9
1	1	2	2	1	1	93.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	7	5	1	1	1	98.4
1	7	5	1	1	2	95.2
1	7	5	2	1	1	103.5
1	7	5	2	1	2	95.6
1	7	6	1	1	1	99.1
1	7	6	1	1	2	94.8
1	7	6	2	1	1	99
1	7	7	0	1	1	102
1	7	7	0	1	2	.

**Example 3** Data format for major-gene traits *diallel* module

Env	major_female	major_male	female	male	cross	block	Kernel
1	1	1	1	1	0	1	4.91
1	2	2	2	2	0	1	6.37
1	1	1	3	3	0	1	5.88
1	2	2	4	4	0	1	5.12
1	1	1	5	5	0	1	5.36
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	1	2	7	4	1	2	5.29
2	1	2	6	4	1	2	5.93
2	1	1	7	3	1	2	4.60
2	1	1	8	3	1	2	5.02
2	1	1	8	5	1	2	4.43

Click “OK” to enter the dialog to specify the model and options.



**Input:** choose the name of each column in the data file.

**Trait:** the name of the quantitative traits.

**Diallel Type:** choose the diallel cross type of the data, including “Plant”, “Seed”, “Endosperm”, “Animal”, “MajorGene” for agronomic trait, seed trait, endosperm trait, animal trait and major-gene trait separately. “MajorGene” item should choose the subtype, including “Plant” and “Seed” for agronomic and seed separately.

**Model Analysis:** the software will list the general model and random effects for the chosen diallel type. According to the dataset and analysis purpose, user can add or subtract the effects in “model” and “random”, and choose to estimate the variance components of these effects, covariance components among quantitative traits, heterosis, or condition analysis. The items of “covariance” and “condition” could not be estimated for single trait dataset.

Effects abbreviation:

Env: environment effect

A: additive effect

D: dominance effect

AA: additive by additive effect

M: maternal effect

P: paternal effect

Am: maternal additive effect

Dm: maternal dominance effect

AmAm: maternal additive by maternal additive effect

Aen: endosperm additive effect

Den: endosperm dominance effect

