Manual of QTXnetwork

Content

1 Introduction on QTX

QTXNetwork (Figure 1.1) is GPU-based computation software for linkage and association analyses of epistasis and environment interaction of complex traits. It contains four functional modules (Figure 1.2): quantitative trait locus (QTL) for linkage analysis, quantitative trait SNP (QTS) for genome-wide association analysis, quantitative trait transcript/protein/metabolite (QTT/P/M) for transcriptome/proteome/metabolome association analysis, and GMDR for data filtering which will be further used in Genome-Wide Association Studies (GWAS). By using the massive parallel nature of multi-GPUs, this mapping tool can perform association analyses on large-scale omics data for complex traits.

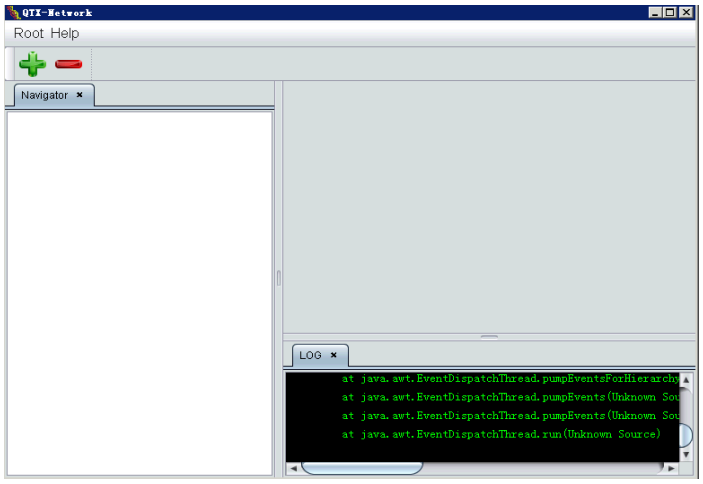


Fig 1.1: The main window of the QTXNetwork

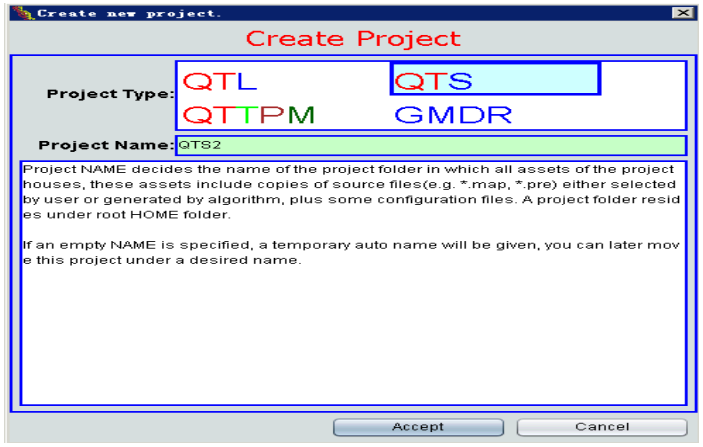


Fig 1.2: The popup window for creating new project for QTL, QTS, QTT/P/M or GMDR

2 User Manual for QTL

2.1 Introduction on QTL

QTL is one module of the software QTXNetwork for mapping quantitative trait loci (QTL) with epistasis effects and QTL by environment (QE) interaction effects in DH, RI, BC1, BC_2, F_2, IF_2 and BxFy populations, and for graphical presentation of QTL mapping results. The software is developed based on the MCIM (Mixed-model based Composite Interval Mapping) method, and programmed by C++ programming language.

2.2 Running QTL

2.2.1 Creating a project for QTL

The QTL program, kept in the folder ~/exe/QTL under the QTXNetwork, can be run directly in "command window" or in "explore window". In default, you can run it via the shell program (QTXNetwork). To run the QTL, you first need running the QTXNetwork for creating a new project for the QTL by clicking the button  in the main window of the QTXNetwork (Fig 2.1), a window for creating new project will be popped up (Fig 2.2). After selecting module type "QTL" and specifying project name, click the "Accept" button (Fig 2.2), a QTL project will be listed in the left pane named Navigator (Fig 2.3).

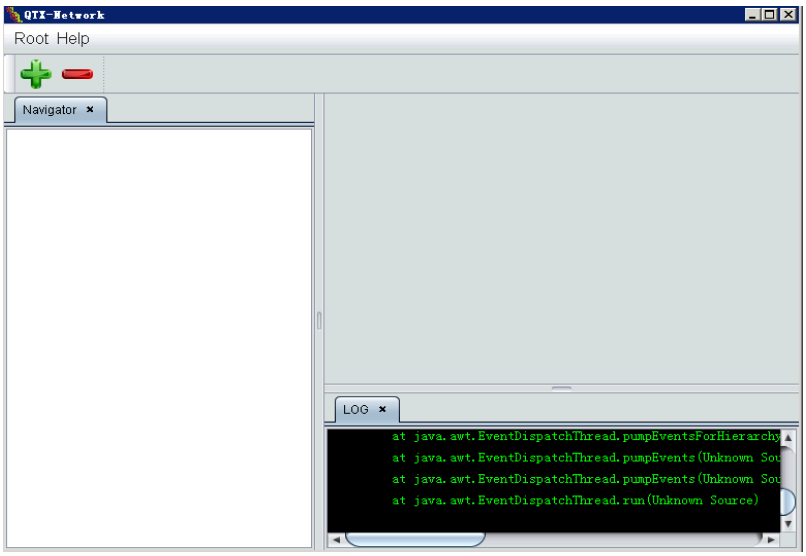


Fig 2.1: The main window of the QTXNetwork

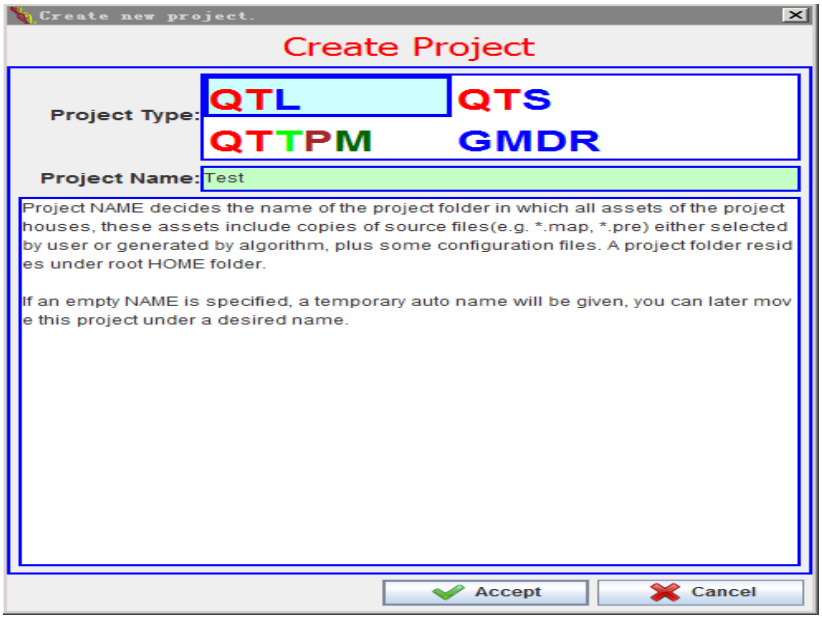


Fig 2.2: The popup window for creating new project for QTL

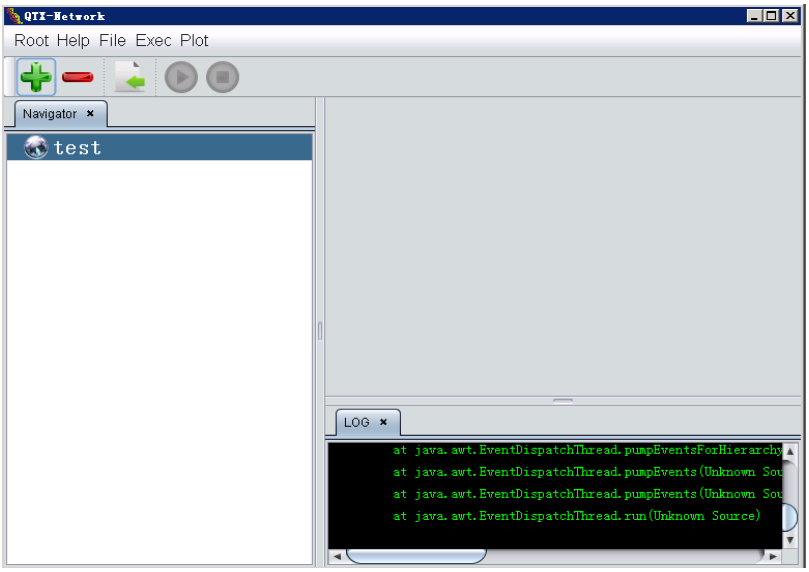


Fig 2.3: The window of QTXNetwork for performing QTL

2.2.2 Deleting QTL project

Select the QTL project to delete in the Navigator pane, and then click the fast button

2.2.3 Specifying data file

In a QTL project window, a button  for “locate resource files” will occur in the toolbar, you can locate and import data files (a marker linkage map file and a data file which contains observations of the markers and the traits under studied for all individuals for QTL mapping in the popup window after clicking the button ). In this panel, there are three sections to set input or output files. The first is the Genetic map section, which contains information about the marker linkage map, such as the number of chromosomes, number and order of markers on each of the chromosomes, flanking marker distances, etc. The second is the Input data section, and you need input data files, which integrated the observations of marker and the trait studied into one file. The third is the “Output report” section, where the output file name can be inputted.

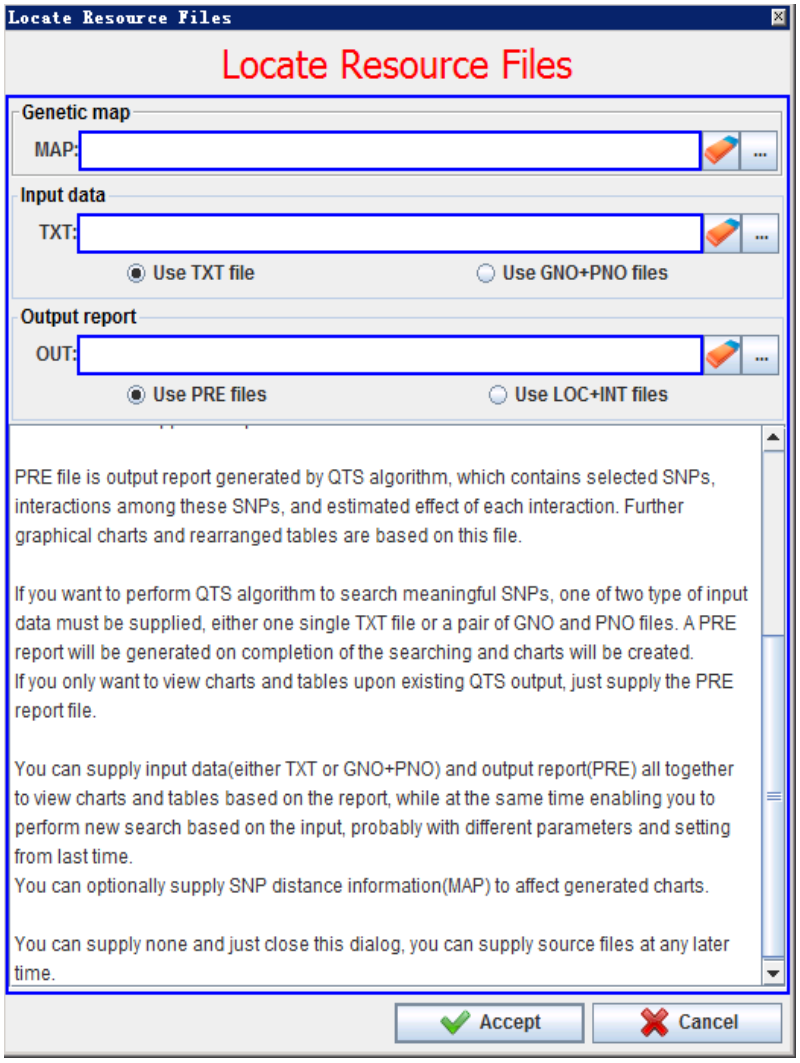


Fig 2.4: The window for inputting data files and output file

2.2.4 Starting QTL

For example, suppose you open a QTL projected named “test”, and also specify the data files for this project. The sample data files are the map file (~\sample\ QTL\SimF2.map), the data file (~\sample\QTL\SimF2.txt), and the report file (~\sample\QTL\simF2.pre), respectively, then, the main window for the QTL project will exhibits as the Fig 2.5.

Finally, you can click the button  (Fig 2.5) to start the QTL analysis if all default algorithm parameters are used. You can adjust QTL algorithm configuration before running, detailed information refer to the chapter 3.3.

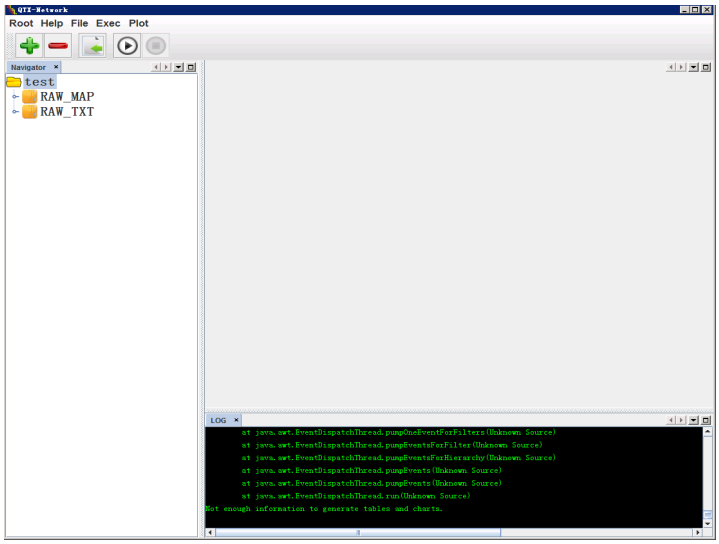


Fig 2.5: The window of QTXNetwork after creating or opening a QTL project

2.3 Data file format and running configuration of QTL

2.3.1 Format of marker linkage map file

This file contains information about the marker linkage map, such as the number of chromosomes, number and order of markers on each of the chromosomes, flanking marker distances, etc. It consists of general description and map body.

General Description: This part is in the front of map file. A typical general description looks like:

_DistanceUnit cM
_Chromosomes 4
_MarkerNumbers 6 4 7 9

There are a total of four possible items for general description. They can be in any order. Each item in general description is a key word followed by certain specification(s). Each key string must be started with an underline “_”, and there should not be any list separator (white space or table) within the key string. The specification(s) must be separated from the key word by at least one list separator, and there must also be at least one list separator between any two neighboring specifications if two or more

specifications are included for the item. A key string and its specification(s) must be placed in the same line. Both key strings and specification (s) (if characters) are not case insensitive.

_DistanceUnit specifies the unit of genetic distances used in the map file. The specification string "cM" stands for centi-Morgan and "M" stands for Morgan.

_Chromosomes is for specifying the total number of chromosomes or linkage groups involved in the map file.

_MarkerNumbers is for specifying the number of markers on each of the chromosomes. The order of the numbers must be consistent with that for genetic-distance columns in the map body.

Map Body: This part starts from key string *MapBegin* and ends at key string *MapEnd*. A typical map body looks like:

MapBegin

Marker#	Ch1	Ch2	Ch3	Ch4
1	0.00	0.00	0.00	0.00
2	9.84	11.26	7.45	9.85
3	10.22	8.69	9.10	10.93
4	8.25	9.87	10.66	10.70
5	9.79		10.16	10.10
6	7.47		8.34	11.30
7			11.21	9.30
8				7.23
9				11.78

MapEnd

The strings (Marker#, Ch1, Ch2, Ch3, Ch4) in the second row show the contents of the columns below them. The Marker# column (first column) is for the order of all markers on each chromosome; the maximum order is equal to the number of markers on the chromosome that has the most markers among all the chromosomes. The Ch1 column (second column) to Ch4 column (last column) each represents a chromosome or linkage group, and contains genetic distances between adjacent markers on the chromosome. Specifically, the genetic distance for the first marker on each chromosome must be set to zero as the start point of the linkage map for the chromosome; the distance for the second marker is between the first and the second markers; the distance for the third marker is between the second and the third markers, and so on. The order of Ch1 column (second column) to Ch4 (last column) must be consistent with that for the numbers following the key string _MarkerNumbers.

2.3.2 Format of data file

The data file contains information on population type, number of genotypes sampled from the population, number of observations, observations for both markers and quantitative traits, etc. It is composed of four parts: general description, marker data body, trait data body, and some comment lines.

General description: This part is for specifying the basic features of the data file, and is usually put in the front of the data file. Like in the map file, each item in general description is a key character string followed by certain specification(s). Each key string must be started with an underline "_", and no white space is allowed within it. There are eight possible items for general description. They can be arranged in any order. A typical description for a data file looks like:

_Population DH
_Genotypes 200
_Observations 400
_Environments Yes
_Replications No
_TraitNumber 1
_TotalMarker 64
_MarkerCode P1=1 P2=2 F1=3 F1P1=4 F1P2=5

_Population specifies the population type used. Some commonly used populations are listed as follows:

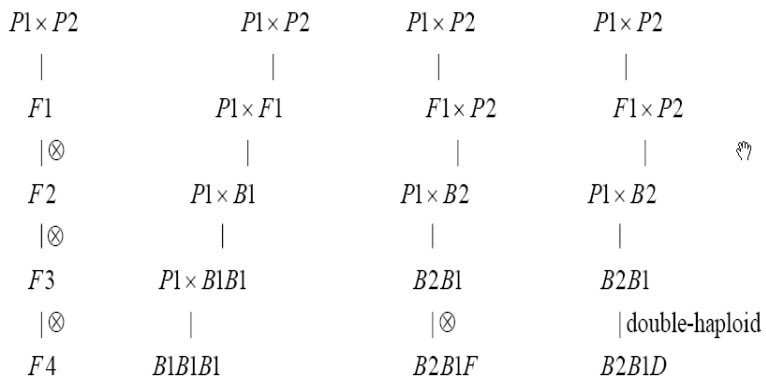
RI population – derived from a cross between two pure-line parents. The specification word for RI population can be RI or RIL.

BC population – derived from crossing F1 with one of the inbred parents. The specification words for BC1 and BC2 populations are B1 and B2, respectively.

F2 population – derived from selfing or sib-mating F1 that is made by crossing two inbred lines.

Immortalized F2 (IF2) population – derived from randomly mating among individuals from DH or RI population (See Ref: Hua JP, Xing YZ, Xu CG, Sun XL, Yu SB and Zhang QF (2002) Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. Genetics 162: 1885–1895). The specification word IF2DH is for IF2 population derived from DH population, and IF2RI for that from RI population.

BxFy Population – derived from F1 backcrossing to one of the inbred parents or selfing for several generations. In each generation, selfing, backcrossing or creating double-haploid is permitted. There are several examples shown as the following figure:



_Genotypes specifies the total number of genotypes sampled from the mapping population.

_Observations specifies the total number of observations for each trait studied.

_Environments specifies the status of experimental design for environments. If the experiment is conducted in multiple environments, write the specification word Yes after the key word _Environments, otherwise write No.

_Replications specifies the status of experimental design for replications or blocks. If the experiment is conducted with replications or blocks, write the specification word Yes after the keyword _Replications, otherwise write No.

_TraitNumber specifies the total number of traits included in the data file.

_TotalMarker specifies the total number of the markers included in the data file. This number must be equal to the summation of the numbers for _MarkerNumbers in the map file.

_MarkerCode defines a marker coding scheme. There are five possible strings for the specifications. Each of the strings looks like an equation, but no white space is allowed within the string. On the left side of the equation symbol is the marker phenotype specification:

P1: Marker phenotype being the same as that of P1;

P2: Marker phenotype being the same as that of P2;

F1: Marker phenotype being the same as that of F1;

F1P1: Marker phenotype that is not P2 type (P1 dominant or undistinguishable between P1 type and F1 type);

F1P2: Marker phenotype that is not P1 type (P2 dominant or undistinguishable between P2 type and F1 type).

On the right side of the equation symbol is the code for the marker type. The marker code should always be a single character (a number or a letter). The symbol dot "." is used to represent missing marker data or trait value. It is not necessary to specify codes for all possible marker types except for F2 population. For example, if your marker data were collected from a DH population, only the specifications for P1 and P2 types are enough.

Marker data body: This part is embraced by two key strings *MarkerBegin* and *MarkerEnd*. The order of the marker data for different marker loci must be consistent with the order of markers on each chromosome determined in the map file. Since electronic table software usually has a limit on the number of columns in spreadsheet, we provide two types of arrangements for marker data.

Type I:

MarkerBegin

#Ind	Mk1	Mk2	Mk3	Mk4	Mk5	Mk6	Mk7	Mk8	Mk9	;
1	1	1	1	2	2	2	2	1	1	;
2	1	1	.	1	1	2	2	2	2	;
3	2	.	2	1	1	1	1	2	2	;
……										;
89	2	2	2	2	.	1	1	.	1	;
90	1	1	2	2	2	2	2	1	1	;

MarkerEnd

Type II:

MarkerBegin

#Mk	1	2	3	4	5	…	48	49	50	…	88	89	90	;
Mk1	1	1	1	2	1		2	2	1		1	2	1	;
Mk2	1	1	1	.	2		1	2	1		1	2	1	;
Mk3	1	.	1	2	2		1	2	2		1	2	2	;
Mk4	2	1	1	1	1		1	2	2		1	2	2	;
Mk5	2	1	.	1	1		1	1	1		2	.	2	;
Mk6	2	2	2	1	1		1	1	1		2	1	2	;
Mk7	2	2	2	1	1		2	1	1		2	1	1	;
Mk8	2	2	2	1	2		2	1	2		1	.	1	;
Mk9	1	2	2	2	2		2	2	2		2	1	1	;

MarkerEnd

Trait data body: This part is between two key strings *TraitBegin* and *TraitEnd*. The data source includes the environment (if available), the replication (if available) and the ID name of subjects, as well as the observations obtained for traits studied. The second row includes the indicator strings and the names of the traits. The number of source strings depends on the experimental design. If both environments and replications are taken, a maximum of three strings must be inputted: the first string for

environment (Env#), the second string for replication (Rep#) and the third string for subject (Geno#). You can use whatever strings to express the sources because they are just used to indicate what the numbers are in the columns below them. If the experiment is conducted without environmental factor or replications, the corresponding column must be removed. And also, a semicolon ";" is required at the end of each observation data row. The following is an example for the trait data body.

```
*TraitBegin*
Env#    Rep#    Geno#   Trait_1 Trait_2 Trait_3 ;
1       1       1       2.44    7.4     10.04   ;
1       1       2       2.4     4.32    8.55    ;
……
1       1       90      3.54    8.19    10.74   ;
1       2       1       3.17    6.91    11.86   ;
1       2       2       1.9     4.31    11.36   ;
……
1       2       90      3.22    10.54   11.48   ;
2       1       1       5.74    12.78   11.27   ;
2       1       2       7.65    7.02    11.96   ;
……
2       1       90      6.58    13.92   9.94    ;
2       2       1       6.01    10.22   9.95    ;
2       2       2       6.22    11.99   7.81    ;
……
2       2       90      7.98    13.21   12.03   ;
*TraitEnd*
```

2.3.3 QTL algorithm and configuration

Before conducting QTL, we need setting the QTL algorithm configuration. Suppose a new QTL project is created or a QTL project is opened in main window of the QTXNetwork, click the QTL project in the Navigator panel (Fig 2.5) to set the algorithm configuration, then activate the command item "Config" under the "Exec" software menu (Fig 2.6), a panel will be popped up for setting configuration (Fig 2.7, Fig 2.8, Fig 2.9, Fig 2.10).

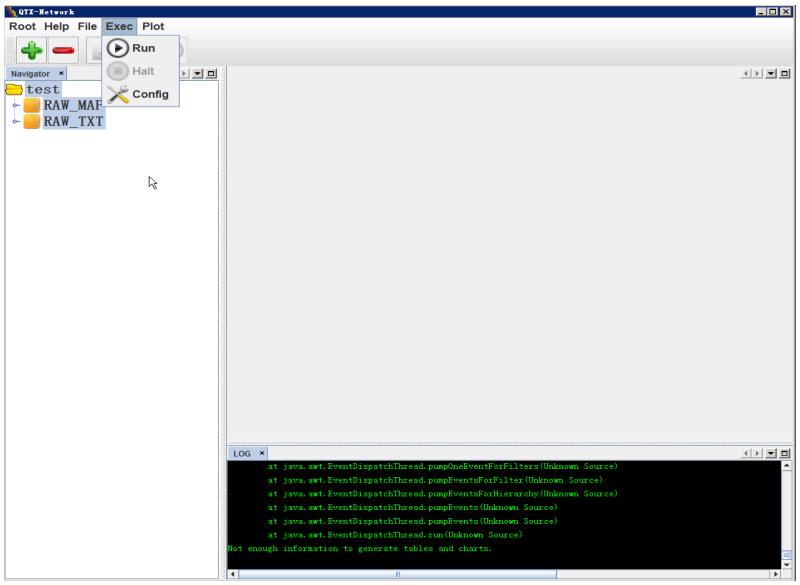


Fig 2.6: Executing QTL under the "Exec" of the menu

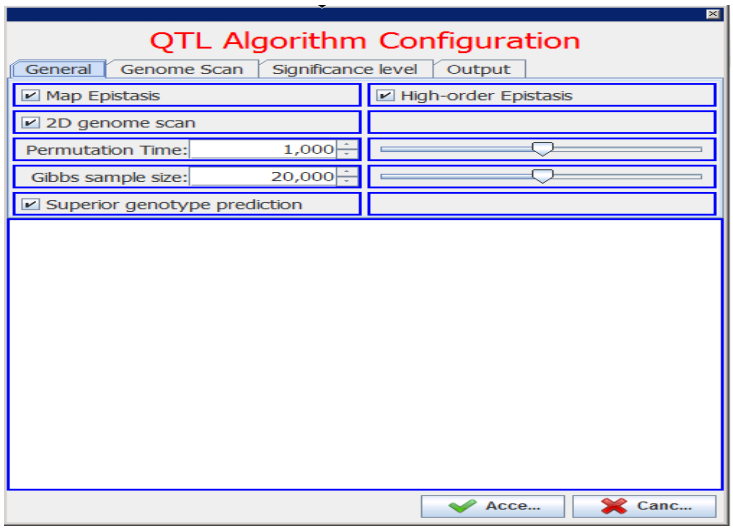


Fig 2.7: General algorithm configuration for QTL

Map Epistasis – Choose this option to detect both single-locus effect QTL and epistasis, otherwise to detect QTL with single-locus effects only.

High Order Epistasis – We define the additive × dominance (AD), dominance × additive (DA) and dominance × dominance (DD) as the high order epistaiss. This option is only available for setting F_2 population. Choose this option to detect QTL with additive (A) and dominance (D) effects, and also epistasis with AA, AD, DA and DD effects, otherwise to detect epistasis with only AA effect.

2D Genome Scan – Choose this option to detect epistasis QTLs with or without single-locus effects. Otherwise the program will only detect the epistasis interaction among QTLs with single-locus effects.

Permutation Time – Set the permutation time to conduct permutation test on significance of QTL effects.

Gibbs sample size – Set this value to predict QTL effects by Monte Carlo Markov Chain method, otherwise by mixed linear model approach.

Superior Genotype Prediction – Choose this option to predict superior genotypes based on QTL effects (Ref. Yang J and Zhu J. (2005). Predicting superior genotypes in multiple environments based on QTL effects. Theoretical and Applied Genetics, 110: 1268-1274.)

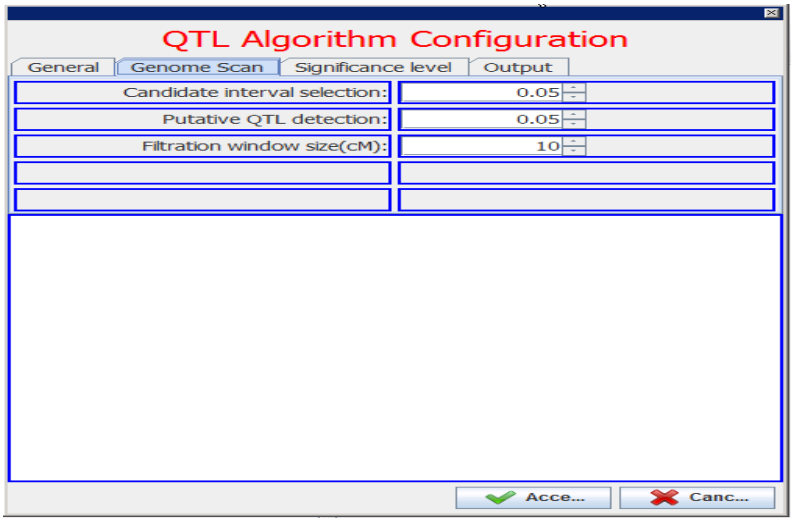


Fig 2.8: Genome scan configuration for QTL

Candidate Interval Selection – The experimental-wise type I error for candidate interval selection.

Putative QTL Detection – The experimental-wise type I error for putative QTL detection.

Filtration window size (cM) – Define a window to distinguish two adjacent test statistic peaks whether they are two QTL or not.

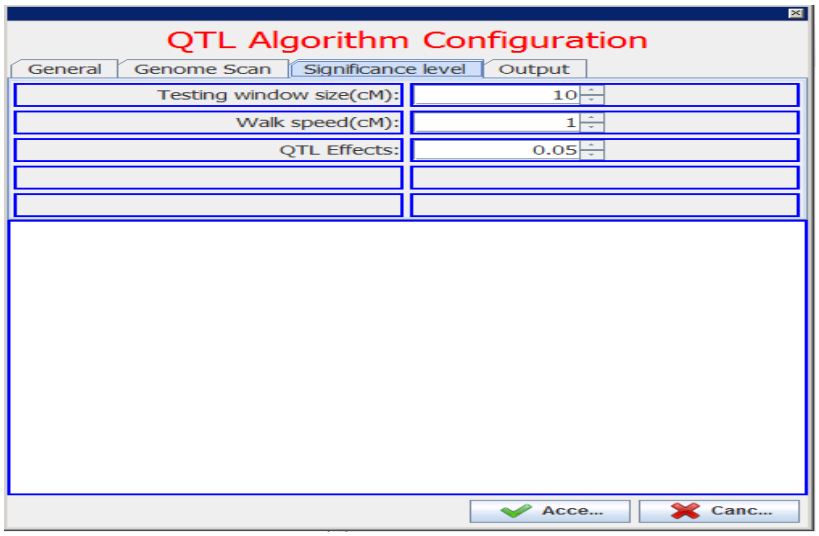


Fig 2.9: Significance level configuration for QTL

Testing window size (cM) – Define a window adjacent to the testing interval (one for each direction). When scan the genome for putative QTL, no marker will be selected as cofactor within the testing window.

Walk speed (cM) – The space between two adjacent test point, when testing for putative QTL along the genome.

QTL effects – The significance level for QTL effects.

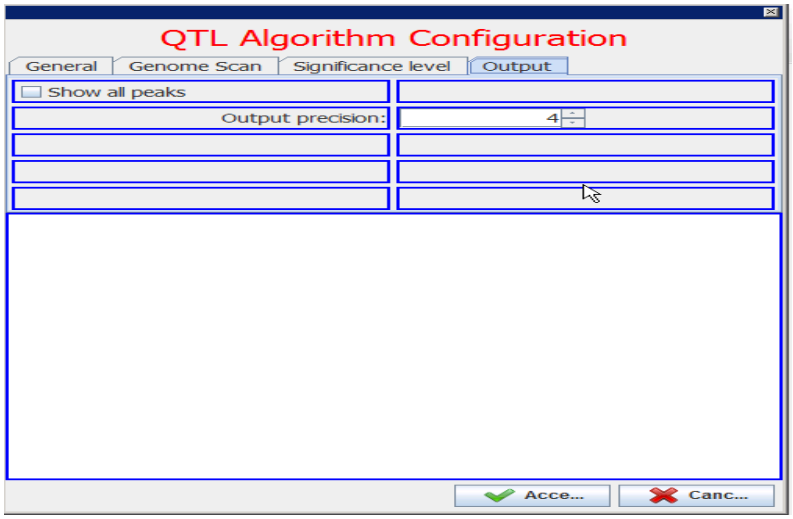


Fig 2.10: Output configuration for QTL

Show all peaks – Choose this option to show all the markers in the linkage group, otherwise only show the markers near QTL.

Output precision – The maximum number of decimal places to display value in the output.

2.4 Saving and understanding pictures and reports

2.4.1 Saving pictures

After finishing analysis of QTL, we can save figures in png or eps format by executing the “Export GG” for GxG plot under the “Plot” menu (Fig 2.11).

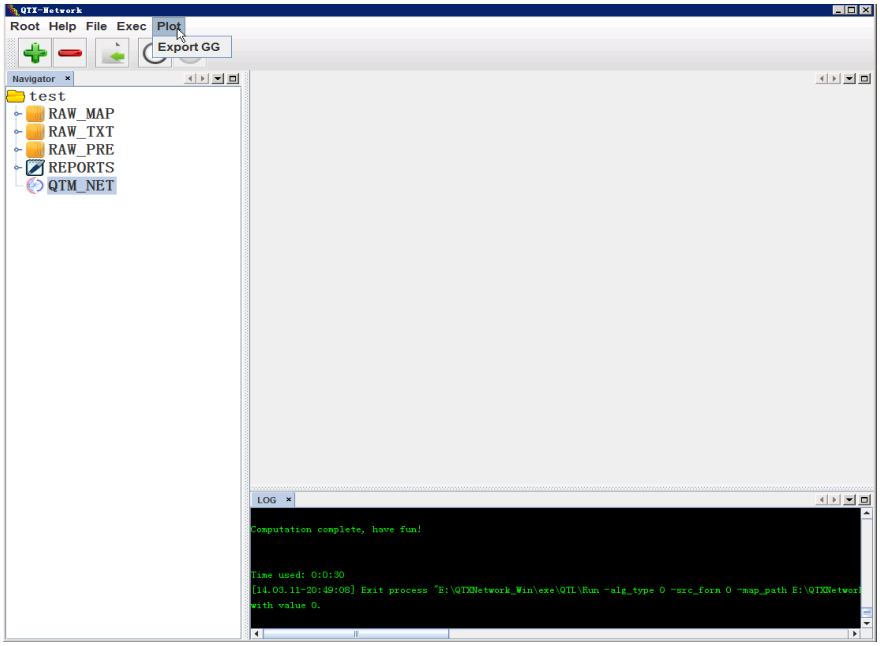


Fig 2.11: Saving picture by using “Plot” menu

2.4.2 Understanding the symbols in figure

For the GxG plot of QTL (Fig 2.12), there are different color circles, squares and lines in plot, which denote different type genetic effects of QTL (Fig 2.13).

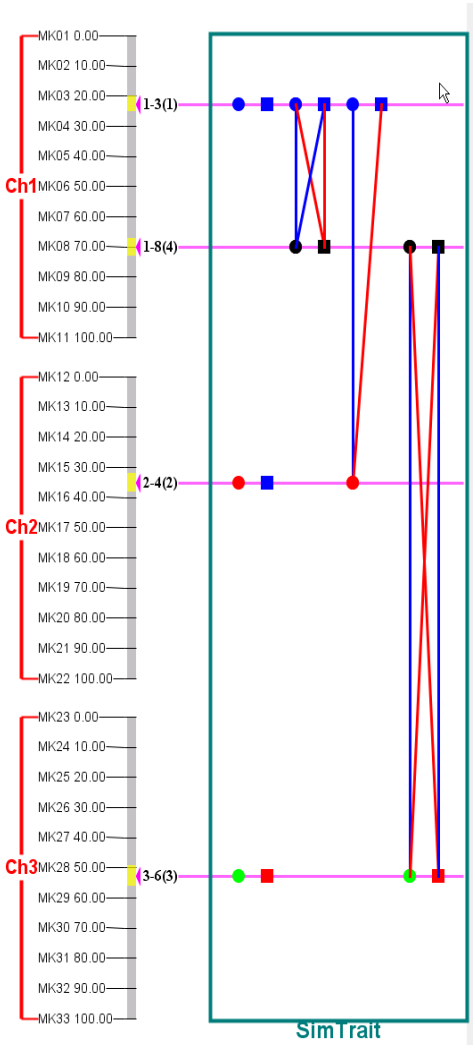


Fig 2.12: The GG plot of QTL

Graphic meta system	Line (Epistasis)	Shape (QTL)	
		Circle	Square
Red	with only epistatic main effect (I)	with only additive effect (A)	with only dominance effct (D)
Green	with only epistasis × environment interaction effect (IE)	with only additive × environment interaction effect (AE)	with only dominance × environment interaction effect (DE)
Blue	with both I and IE	with both A and AE	with both D and DE
Dark	Not available	with no additive related effect	with no dominance related effect

Fig 2.13: The genetic indications for different symbols, lines, colors in GG plot

2.4.3 Understanding text report

In order to understand each item in report file, we firstly define symbols or variables in the report file, and then explain the report file using the Fig 2.14-2.20 coming from the F2 simulation data which can be found in the sample folder (~\Sample\QTL\SimF2.pre) of the QTXNetwork.

A (or a), D (or d), AE (or ae), DE (or de): denote additive effect, dominance effect, additive-environment interaction effect, dominance-environment interaction effect, respectively;

AA (or aa), AD (or ad), DA (or da), DD (or dd), AAE (or aae), ADE (or ade), DAE (or dae), DDE (or dde): denote additive-additive epistasis effect, additive-dominance epistasis effect, dominance-additive epistasis effect, dominance-dominance epistasis effect, and their interaction effects with environments, respectively;

I, IE: denote summation of different types of epistasis effects and environment interaction effects, respectively;

AE*i* (or ae*i*), DE*i* (or de*i*), AAE*i* (or aae*i*), ADE*i* (or ade*i*), DAE*i* (or dae*i*), DDE*i* (dde*i*) denote the i-th environment-specific genetic component effects of QTLs, respectively.

In the Fig 2.14, V(e), V(P) denote the residual variance and the phenotypic variance, respectively, h^2(X) denote the proportion of phenotypic variance contributed to the genetic effects of X, respectively, where X stand for one genetic component effect or summation of some genetic component effects, such as, h^2(A+D) is the proportion of phenotypic variance contributed to additive and dominance effects of detected QTLs.

For the Fig 2.15, the first column is the serial number of each detected significant QTL, the second is the flanking markers of QTL, the third is the distance between QTL and the first marker of the relevant chromosome, the fourth is the support interval of QTL position, the succeeding 3 columns are predicted genetic effects, standard error and significance probability value, respectively. The Fig 2.16 presents the proportions of phenotypic variance explained by the genetic component effects of detected QTLs in the Fig 2.15.

The Fig 2.17 and the Fig 2.18 provide each epistasis effects, epistasis-environment interaction effects of detected paired QTLs (the i-th QTL, j-th QTL), and proportion of phenotypic variance explained by these epistasis effects, respectively.

The Fig 2.19 provides predicted genotypic values based on genetic effects of detected QTLs for several designed genotypes.

QQ: denoting the individual with homozygote alleles for all loci as in P_1;

qq: denoting the individual with homozygote alleles for all detect loci as in P_2;

F1: denoting the individual with heterozygote alleles for all detected loci as in F_1 of P1 × P2;

SuperiorLine(+): denoting predicted pure line with the highest genotypic value;

SuperiorLine(−): denoting predicted pure line with the lowest genotypic value;

SuperiorHybrid(+): denoting predicted highest genotypic value;

SuperiorHybrid(−): denoting predicted lowest genotypic value;

G: designing genotype based on general genetic effects of QTLs;

G+GE*i*: designing genotype based on general and i-th environment specific genetic effects of QTLs;

For the Fig 2.20,

GSL(+/-) denotes designed general superior line with higher/lower trait performance;

SL(+/-)*i* denotes designed superior line with higher/lower trait performance in i-th environment;

GSH(+/-) denotes designed general superior hybrid with higher/lower trait performance;

SH(+/-)*i* denotes designed environment specific hybrid with higher/lower trait performance in i-th environment.

_variance_components (population mean: 100.0234 phenotypic variance: 64.3907 Total Heritability: 0.7534)

h^2(A+D)	h^2(AE+DE)	h^2(I)	h^2(IE)	V(e)/V(P)
0.4414	0.1181	0.1425	0.0313	0.2466

h^2(A)	h^2(D)	h^2(I)	h^2(A+AA)	h^2(AE)	h^2(DE)	h^2(IE)	h^2(AE+AAE)
0.3105	0.1310	0.1425	0.3614	0.0602	0.0579	0.0313	0.0803

Fig 2.14: Variance components and heritability ascribed different genetic components

_1D_effect

QTL	interval	position	range	A	SE	P-Value	D	SE	P-Value	AE1	SE	P-Value	AE2	SE	P-Value	DE1	SE	P-Value	DE2	SE	P-Value
1-3	MK03-MK04	23.0	20.0-25.0	4.6532	0.3100	1.152e-050	4.3235	0.4695	3.646e-020	-1.8818	0.4325	1.360e-005	1.7867	0.4325	3.618e-005	-3.0171	0.6568	4.383e-006	3.2326	0.6568	8.643e-007
2-4	MK15-MK16	35.0	32.0-38.0	-4.2814	0.2879	9.438e-050	-1.1895	0.4747	1.223e-002	--	--	--	--	--	--	-2.2881	0.6578	5.051e-004	2.2414	0.6578	6.567e-004
3-6	MK28-MK29	53.0	49.0-56.0	--	--	--	-3.6915	0.4516	3.135e-016	2.0825	0.4154	5.411e-007	-2.1064	0.4154	4.009e-007	--	--	--	--	--	--

Fig 2.15: Mapping result ascribed to single QTL effects

_1D_heritability

QTL	h^2(a)	h^2(d)	h^2(ae)	h^2(ae1)	h^2(ae2)	h^2(de)	h^2(de1)	h^2(de2)
1-3	0.1681	0.0726	0.0261	0.0137	0.0124	0.0380	0.0177	0.0203
2-4	0.1423	0.0055	---	---	---	0.0199	0.0102	0.0098
3-6	---	0.0529	0.0341	0.0168	0.0172	---	---	---

Fig 2.16: Heritability ascribed to single QTL effects

_2D_effect

QTL_i	interval_i	position_i	range_i	QTL_j	interval_j	position_j	range_j	AA	SE	P-Value	AD	SE	P-Value	DA	SE	P-Value	DD	SE	P-Value	AAE1	SE	P-Value	AAE2	SE	P-Value	ADE1	SE	P-Value	ADE2
1-3	MK03-MK04	23.0	20.0-25.0	1-8	MK08-MK09	70.0	67.0-73.0	2.3253	0.4361	9.820e-008	2.5103	0.6199	5.155e-005	1.4559	0.6407	2.308e-002	4.9151	0.9391	1.676e-007	-1.6547	0.5989	5.731e-003	1.6944	0.5989	5.455e-003	--	--	--	--
1-3	MK03-MK04	23.0	20.0-25.0	2-4	MK15-MK16	35.0	32.0-38.0	1.0728	0.4323	1.310e-002	--	--	--	1.4525	0.6738	3.112e-002	--	--	--	2.4179	0.6015	5.843e-005	-2.2451	0.6015	1.900e-004	--	--	--	--
1-8	MK08-MK09	70.0	67.0-73.0	3-6	MK28-MK29	53.0	49.0-56.0	-2.5607	0.3975	1.202e-010	1.9508	0.6203	1.663e-003	-3.9986	0.5933	1.628e-011	-3.0935	0.9031	6.153e-004	-1.4504	0.5437	7.642e-003	1.4770	0.5437	6.601e-003	--	--	--	--

Fig 2.17: Epistatic effects ascribed to paired QTLs

_2D_heritability

QTL_i	QTL_j	h^2(aa)	h^2(ad)	h^2(da)	h^2(dd)	h^2(aae)	h^2(aae1)	h^2(aae2)	h^2(ade)	h^2(ade1)	h^2(ade2)	h^2(dae)	h^2(dae1)	h^2(dae2)	h^2(dde)	h^2(dde1)	h^2(dde2)
1-3	1-8	0.0210	0.0122	0.0041	0.0234	0.0053	0.0027	0.0027	---	---	---	0.0061	0.0030	0.0031	---	---	---
1-3	2-4	0.0045	---	0.0041	---	0.0106	0.0057	0.0049	---	---	---	---	---	---	---	---	---
1-8	3-6	0.0255	0.0074	0.0310	0.0093	0.0042	0.0020	0.0021	---	---	---	---	---	---	0.0051	0.0024	0.0026

Fig 2.18: Heritability ascribed to epistatic effects of paired QTLs

_genotype_value

Entry	G	G+GE1	G+GE2
P1	0.2579	1.9718	-1.4549
P2	2.8852	6.6775	-0.7330
SuperiorLine(+)	9.2504	22.8634	13.2865
SuperiorLine(-)	-13.3343	-10.3189	-17.4592
F1	0.1767	-1.6882	2.0890
SuperiorHybrid(+)	12.1589	22.8634	13.5513
SuperiorHybrid(-)	-17.4036	-14.4994	-20.5071

Fig 2.19: Genotypic values based on predicted effects of QTL for designed genotypes

_superior_genotype

QTL	GSL(+)	GSL(-)	SL(+)1	SL(+)2	SL(-)1	SL(-)2	GSH(+)	GSH(-)	SH(+)1	SH(+)2	SH(-)1	SH(-)2
1-3	QQ	qq	qq	QQ	qq	qq	Qq	qq	qq	QQ	qq	qq
2-4	qq	QQ	qq	qq	QQ	QQ	qq	QQ	qq	qq	QQ	QQ
3-6	qq	QQ	QQ	qq	qq	QQ	qq	Qq	QQ	Qq	Qq	Qq
1-8	QQ	QQ	qq	QQ	qq	QQ	Qq	Qq	qq	QQ	Qq	Qq

Fig 2.20: The superior and inferior genotypes designed for general environment or specific environment

3 User manual for QTS

3.1 Introduction on QTS

QTS is one module of the software QTXNetwork for mapping SNP loci associated with genetic variation of quantitative traits of analyzed population. This module was developed based on the framework of the QTLNetwork (Bioinformatics, 23: 1527-1536; Bioinformatics, 24(5):721-723). The mixed linear model approach combining with Markov Chain Monte Carlo (MCMC) method were employed to perform association analysis between quantitative traits and SNP loci, predict and test genetic effects ascribed to significant SNP locus. Parallel computing base on Graphics Processing Unit (GPU) has been used in some key time-costive subroutines. The QTS can be run on Window or Mac operation system with or without GPU hardware.

3.2Running QTS

3.2.1 Creating a project for QTS

The QTS program, kept in the folder ./exe/QTS under the QTXNetwork, can be run directly in "command window" or in "explore window". In default, we run it via the shell program (QTXNetwork). To run the QTS, we first need running the QTXNetwork for creating a new project for the QTS by clicking the button  in the main window of the QTXNetwork (Fig 3.1), a window for creating new project will be popped up (Fig 3.2). After selecting module type "QTS" and specifying project name, click the "Accept" button (Fig 3.2), a QTS project will be listed in the left pane named Navigator (Fig 3.3).

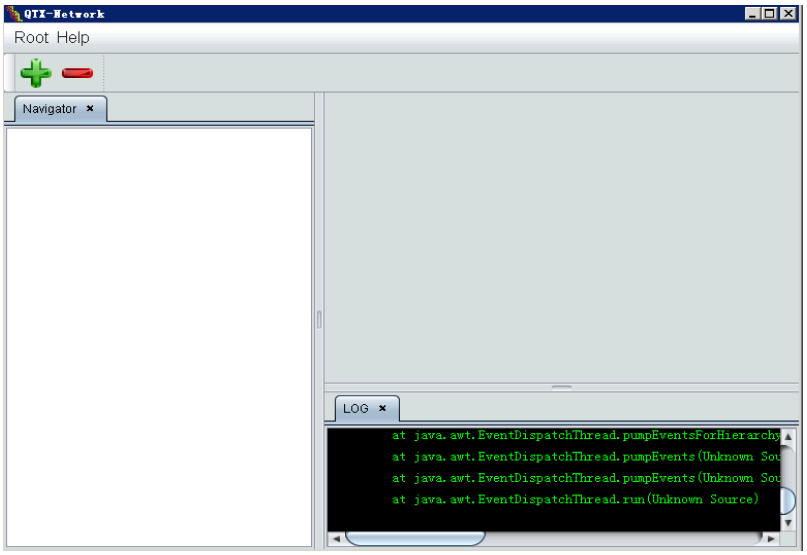


Fig 3.1: The main window of the QTXNetwork

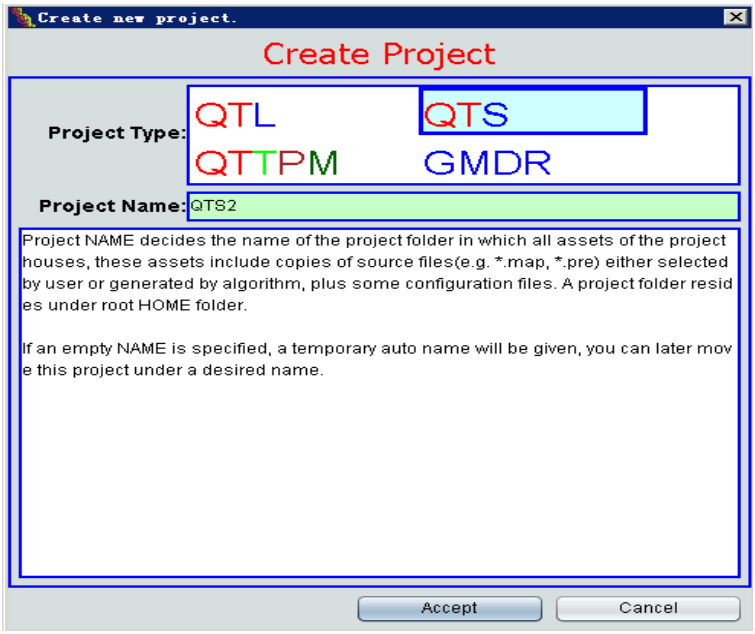


Fig 3.2: The popup window for creating new project for QTS

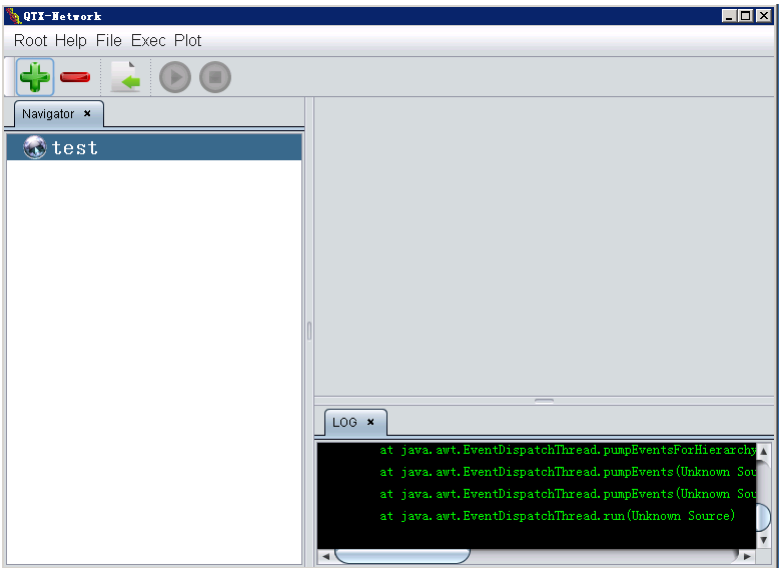


Fig 3.3: The window of QTXNetwork for performing QTS

3.2.2 Deleting QTS project

Select the QTS project to delete in the Navigator pane, and then click the fast button

3.2.3 Specifying data file

In a QTS project window, a button  for "locate and import source file" will occur in the toolbar, we can locate and import data files (genotype data file and phenotype data file for QTS in the popup window after clicking the button  . In this panel, there are three sections to set input or output files (Fig 3.4). The first is the Genetic map section, which will be used when Genetic linkage map is available. Generally, genetic map is not required for association mapping. The second is the Input data section, if the genotype data and phenotype are saved in one file, we need selecting "Use TXT file" option, otherwise, selecting "Use GNO+PNO files" to input genotype data and phenotype files, individually. The third is the "Output report" section, where the output file name can be inputted.

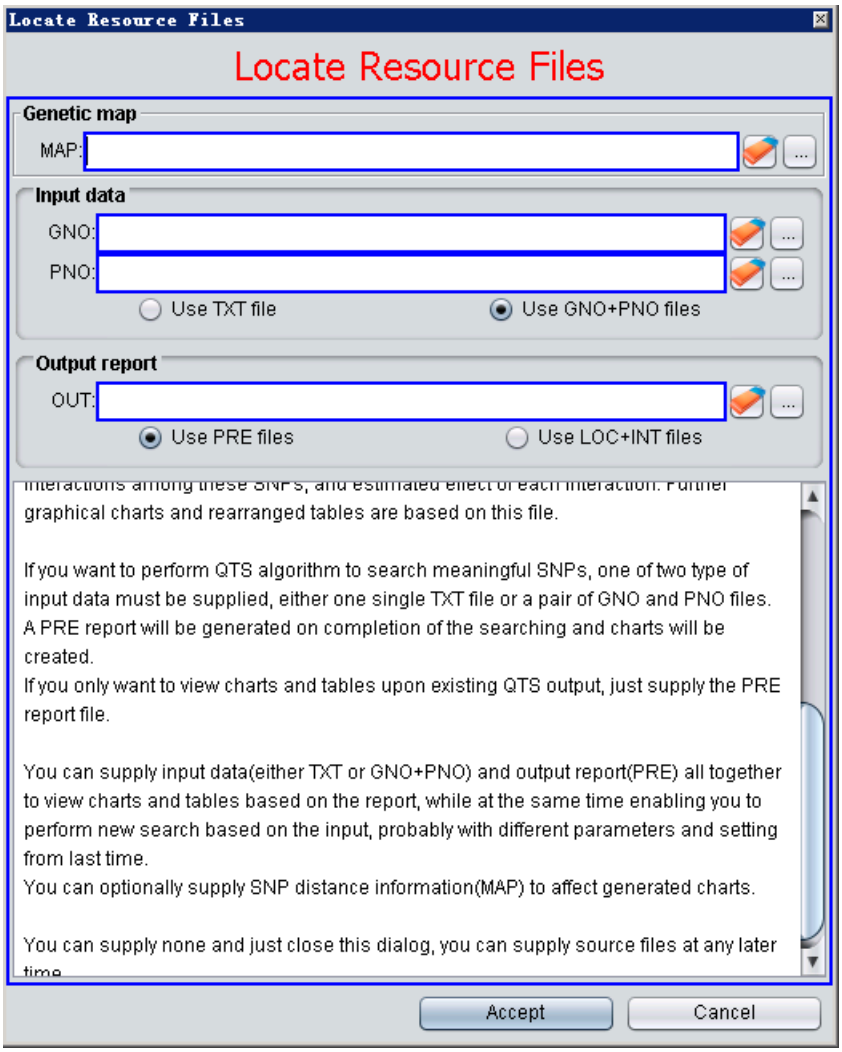


Fig 3.4: The window for inputting data files and output file

3.2.4 Starting QTS

For example, suppose we open a QTS projected named “test”, and also specify the data files for this project. The sample data files are the genotype data file (~\sample\QTS\SimF2_600.gen), the phenotype data file (~\sample\QTS\SimF2_600.phe), and the report file (~\sample\QTS\SimF2_600.pre), respectively, then, the main window for the QTS project will exhibits as the Fig 3.5.

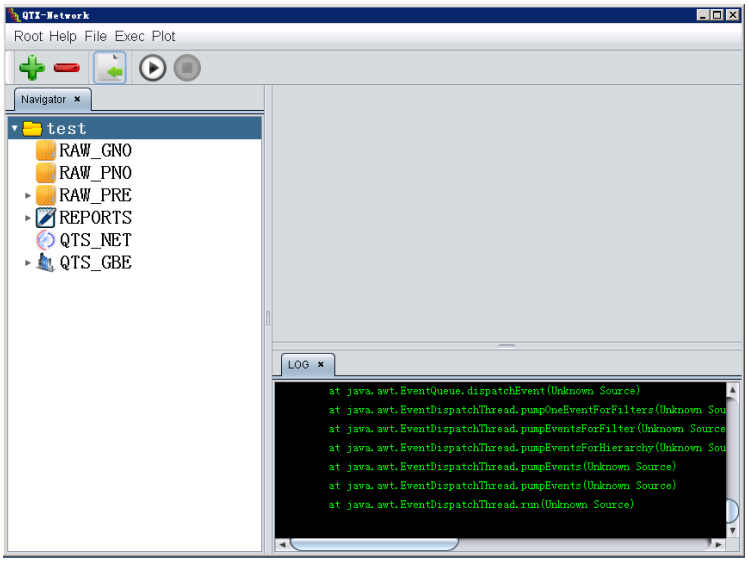


Fig 3.5: The window of QTXNetwork after creating or opening a QTS project

Finally, we can click the button  (Fig 3.5) to start the QTS analysis if all default algorithm parameters are used. We can adjust QTS algorithmn configuration before running, detailed information refer to the chapter 3.3.2.

3.3Data file format and running configuration of QTS

3.3.1 Data format

For performing analyses with QTS, two source data files are required: a genotype data file, or together with a marker linkage map file (not requisite) and a phenotype data file. The linkage map file contains information about the order and genetic distances of all observed markers on the chromosomes or linkage groups. The genotype data file contains marker information for each genotype; the phenotype data file contains observation values of traits under study for all individuals. We provide some sample files for briefly demonstrating the format of source data files for QTS in the sub-directory (~\Sample\QTS) where QTXNetwork has been installed. The genotype data file with extension name of “.Gen” and phenotype data files with extension name of “.Phe” for QTS can be directly used by QTS.

The genotype data can be organized in two different formats which are summarized in Table 3.1 and 3.2. For the first one (Table 3.1), the first line consists of the keyword “#Ind” and molecular marker

names, the first column is for the ID or name of individual subjects. In the second format, the first line consists of the keyword "#Mk" and ID name of individual subjects, the first column is for the ID or name of molecular markers. The other data are the codes of different genotype in each molecular marker locus for all individuals; these codes should be defined in phenotype data file corresponding this genotype data file. It should be noted that the the "I" in "#Ind" and the "M" in "Mk" should be as capital letter, the others are small letters.

Table 3.1: The first type of genotype data format

#Ind	MK1	MK2	MK3	MK4	MK5	…	MKm
1	H	H	A	A	A	…	A
2	H	B	H	B	B	…	H
3	A	A	H	H	H	…	B
…	…	…	…	…	…	...	…
n	A	A	A	A	A	…	H

Table 3.2: The second type of genotype data format

#Mk	1	2	3	4	5	…	n
MK1	H	H	A	A	A	…	B
MK2	H	B	A	A	A	…	B
MK3	A	H	H	A	A	…	B
MK4	A	B	H	A	A	…	B
MK5	A	B	H	A	A	…	H
…	…	…	…	…	…	…	v
MK7	A	H	B	A	H	…	H

General description: This part is for specifying the basic features of the phenotype data file, and is usually put in front of the data file. Like in the map file, each item in general description is a key character string followed by certain specifications. Each key string must be started with an underline "_", and no white space is allowed within it. There are nine possible items for general description. They can be arranged in any order. A typical description for a phenotype data file looks like:

_Population F2
_Genotypes 200
_Observations 600
_Environments Yes
_Replications No

_TraitNumber 1

_Chromosomes 3 Chr1 Chr2 Chr3

_TotalMarker 165 50 50 65

_MarkerCode P1=A P2=B F1=H F1P1=C F1P2=D

_Population specifies the population type used. Some commonly used populations are listed as follows:

RI population – inbreeding lines derived from a cross between two pure-line parents. The specification word for the inbreeding population can be RI or DH.

F2 population – random matting population or selfing population of F_1 hybrid crossing by inbred lines.

_Genotypes specifies the total number of unique subjects used in the mapping population.

_Observations specifies the total number of observations for each trait studied.

_Environments specifies the status of experimental design for environments. If the experiment is conducted in multiple environments, write the specification word "Yes" after the key word _Environments, otherwise write "No".

_Replications specifies the status of experimental design for replications or blocks. If the experiment is conducted with replications or blocks, write the specification word "Yes" after the keyword _Replications, otherwise write "No".

_TraitNumber specifies the total number of traits included in the phenotype data file.

_Chromosomes specifies the chromosome number and the names in marker file. If all molecular markers are regarded as lying in one chromosome, thus, the values after this keyword are "1" and "Chr1".

_TotalMarker specifies the total number of the markers included in the marker data file and the number of markers in each chromosome. The first number must be equal to the summation of all the markers.

_MarkerCode defines a marker coding scheme. There are five possible strings for the specifications. Each of the strings looks like an equation, but no white space is allowed within the string. On the left side of the equation symbol is the marker phenotype specification:

P1: Markers of two alleles being the same as that of P_1 or major allele;

P2: Markers of two alleles being the same as that of P_2 or minor allele;

F1: Markers of two alleles being the same as that of F_1;

F1P1: Markers of two alleles undistinguishable between P1 type and F1 type;

F1P2: Markers of two alleles undistinguishable between P2 type and F1 type.

Trait data body: This part is between two key strings *TraitBegin* and *TraitEnd*. The data source includes the environment (if available), the replication (if available) and the ID name of subjects, as well as the observations obtained for traits studied. The following is an example for the trait data body.

TraitBegin

Env#	Rep#	Geno#	Trait_1	Trait_2	Trait_3	;
1	1	1	2.44	7.4	10.04	;
1	1	2	2.4	4.32	8.55	;

……

```
1       1       90      3.54    8.19    10.74   ;
1       2       1       3.17    6.91    11.86   ;
1       2       2       1.9     4.31    11.36   ;
……
1       2       90      3.22    10.54   11.48   ;
2       1       1       5.74    12.78   11.27   ;
2       1       2       7.65    7.02    11.96   ;
……
2       1       90      6.58    13.92   9.94    ;
2       2       1       6.01    10.22   9.95    ;
2       2       2       6.22    11.99   7.81    ;
……
2       2       90      7.98    13.21   12.03   ;
*TraitEnd*
```

The second row includes the indicator strings and the names of the traits. The number of source strings depends on the experimental design. If both environments and replications are taken, a maximum of three strings must be inputted: the first string for environment (Env#), the second string for replication (Rep#) and the third string for subject (Geno#). You can use whatever strings to express the sources because they are just used to indicate what the numbers are in the columns below them. If the experiment is conducted without environmental factor or replications, the corresponding column must be removed. And also, a semicolon ";" is required at the end of each observation data row.

3.3.2 QTS algorithm and configuration

Before conducting QTS, we first need setting the QTS algorithm configuration. Suppose a new QTS project is created or a QTS project is opened in main window of the QTXNetwork, click the QTS project in the Navigator panel (Fig 3.5) to set the algorithm configuration, then activate the command item "Config" under the "Exec" software menu (Fig 3.6), a panel will be popped up for setting configuration (Fig 3.7, Fig 3.8).

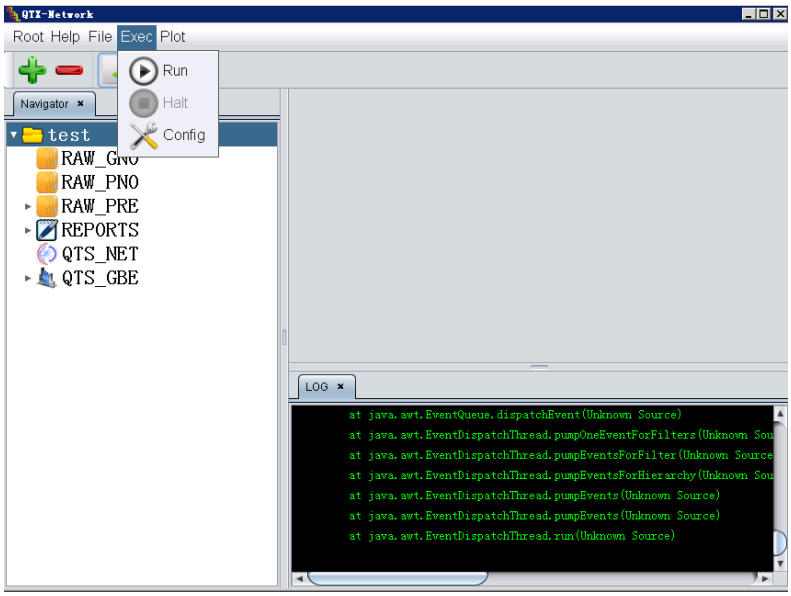


Fig 3.6: Executing QTS under the "Exec" of the menu

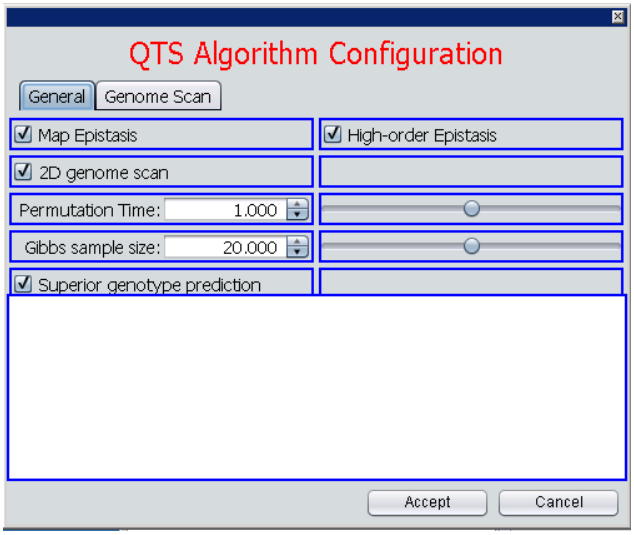


Fig 3.7: General algorithm configuration for QTS

Map Epistasis – Choose this option to detect both single-locus effect QTS and epistasis, otherwise to detect QTS with single-locus effects only.

High Order Epistasis – We define the additive × dominance (AD), dominance × additive (DA) and dominance × dominance (DD) as the high order epistaiss. This option is only available for setting F2 population. Choose this option to detect QTL with additive (A) and dominance (D) effects, and also epistasis with AA, AD, DA and DD effects, otherwise to detect epistasis with only AA effect.

Do 2D Genome Scan – Choose this option to detect epistasis QTSs among markers with single-locus effects and also without single-locus effects. Otherwise the program will only detect the epistasis interaction QTSs among markers with single-locus effects.

Permutation Time – Set the permutation time to conduct permutation test on significance of QTS effects.

Gibs sample size – Set this value to predict QTS effects by Monte Carlo Markov Chain method, otherwise by mixed linear model approach.

Superior Genotype Prediction – Choose this option to predict superior genotypes based on QTS effects (Ref. Yang J and Zhu J. (2005). Predicting Superior Genotypes in Multiple Environments Based on QTL Effects. Theoretical and Applied Genetics, 110: 1268-1274.)

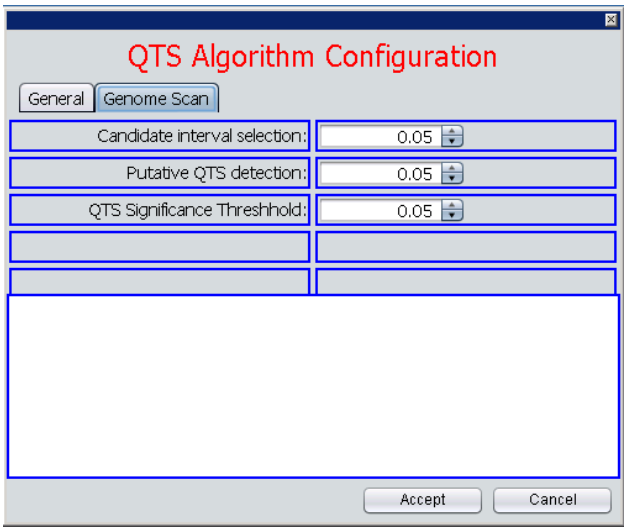


Fig 3.8: Genome scan configuration for QTS

Candidate Interval Selection – The experimental-wise type I error for candidate interval selection.

Putative QTS Detection – The experimental-wise type I error for putative QTS detection.

QTL Effects – The significant level for QTS effects.

3.4 Saving and understanding pictures and report

3.4.1 Saving pictures

After finishing analysis of QTS, we can save figures in png or eps format by executing the "Export GG" for GxG plot or "Export GE" for GxE plot under the "Plot" menu (Fig 3.9).

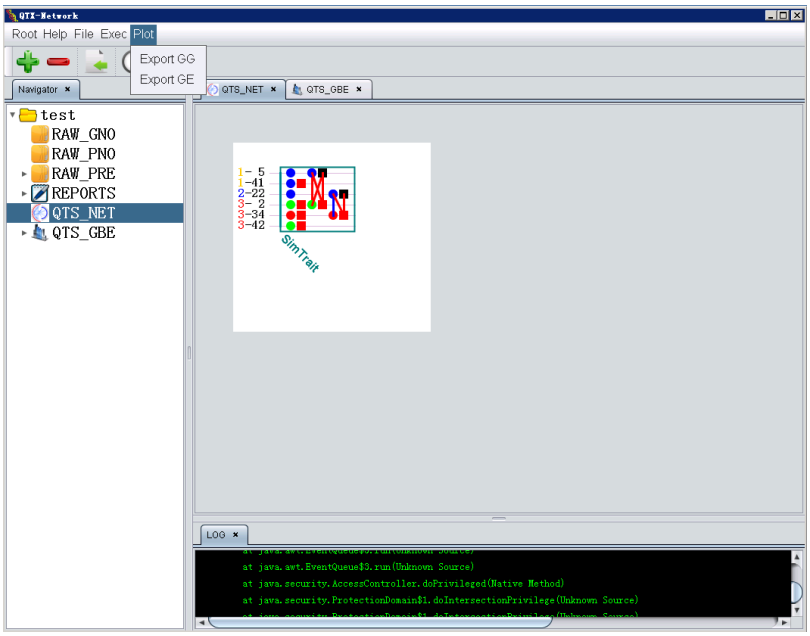


Fig 3.9: Saving picture by using “Plot” menu

3.4.2 Understanding the symbols in figure

For the GxG plot of QTS (Fig 3.10), there are different color circles, squares and lines in plot, which denote different type genetic effects of QTS; detailed definition is presented in the Fig 3.11.

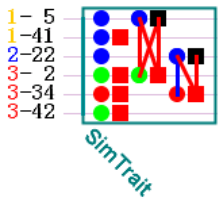


Fig 3.10: The GG plot of QTS

Graphic meta system	Line (Epistasis)	Shape (QTL)	
		Circle	Square
Red	with only epistatic main effect (I)	with only additive effect (A)	with only dominance effct (D)
Green	with only epistasis × environment interaction effect (IE)	with only additive × environment interaction effect (AE)	with only dominance × environment interaction effect (DE)
Blue	with both I and IE	with both A and AE	with both D and DE
Dark	Not available	with no additive related effect	with no dominance related effect

Fig 3.11: The genetic indications for different symbols, lines, colors in GG plot

In the GxE plot of QTS (Fig 3.12), the X-axis are additive (A), dominance (D), and epistasis (AA, AD, DA, DD) effects ascribed to detected QTSs, the Y-axis is the values of effects, the pink rectangle bars stand for genetic main effects, the green line (with one integer n on top) denote the n-th environment-specific effect.

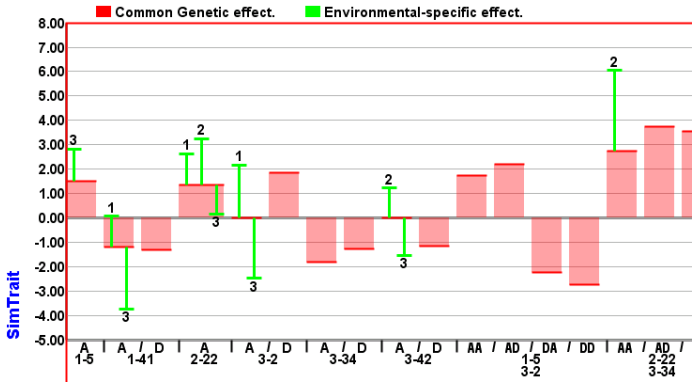


Fig 3.12: The GE plot of QTS

3.4.3 Understanding text report

In order to understand each item in report file, we firstly define symbols or variables in the report file, then explain the report file using the Fig 3.13-3.19 coming from the F2 simulation data which can be found in the sample folder (~\Sample\QTS\SimF2_y600.pre) of the QTXNetwork.

A (or a), D (or d), AE (or ae), DE (or de): denote additive effect, dominance effect, additive-environment interaction effect, dominance-environment interaction effect, respectively;

AA (or aa), AD (or ad), DA (or da), DD (or dd), AAE (or aae), ADE (or ade), DAE (or dae), DDE (or dde): denote additive-additive epistasis effect, additive-dominance epistasis effect, dominance-additive epistasis effect, dominance-dominance epistasis effect, and their interaction effects with environments, respectively;

I, IE: denote summation of different types of epistasis effects and environment interaction effects, respectively;

AE*i* (or ae*i*), DE*i* (or de*i*), AAE*i* (or aae*i*), ADE*i* (or ade*i*), DAE*i* (or dae*i*), DDE*i* (dde*i*) denote the *i*-th environment-specific genetic component effects of SNPs, respectively.

In the Fig 3.13, V(e), V(P) denote the residual variance and the phenotypic variance, respectively, h^2(X) denote the proportion of phenotypic variance contributed to the genetic effects of X, respectively, where X stand for one genetic component effect or summation of some genetic component effects, such as, h^2(A+D) is the proportion of phenotypic variance contributed to additive and dominance effects of detected QTSs.

For the Fig 3.14, the first column is the serial number of each detected significant SNP, the second is the SNP ID, the succeeding 3 columns are predicted genetic effects, standard error and significance probability value, respectively. The Fig 3.15 presents the proportions of phenotypic variance explained by the genetic component effects of detected QTSs in the Fig 3.14.

The Fig 3.16 and the Fig 3.17 provide each epistasis effects, epistasis-environment interaction effects of detected paired QTSs (the i-th SNP, j-th SNP), and proportion of phenotypic variance explained by these epistasis effects, respectively.

The Fig 3.18 provides predicted genotypic values based on genetic effects of detected QTSs for several designed genotypes.

QQ: denoting the individual with homozygote alleles for all loci as in P_1;

qq: denoting the individual with homozygote alleles for all detect loci as in P_2;

F1: denoting the individual with heterozygote alleles for all detected loci as in F_1 of $P_1 \times P_2$;

SuperiorLine (+): denoting predicted pure line with the highest genotypic value;

SuperiorLine (−): denoting predicted pure line with the lowest genotypic value;

SuperiorHybrid (+): denoting predicted F_1 with the highest genotypic value;

SuperiorHybrid (−): denoting predicted F_1 with the lowest genotypic value;

G: designing genotype based on general genetic effects of QTS;

G+GE*i*: designing genotype based on general and *i*-th environment specific genetic effects of QTSs;

For the Fig 3.19,

GSL(+/-) denotes designed general superior line with higher/lower trait performance;

SL(+/-)*i* denotes designed superior line with higher/lower trait performance in *i*-th environment;

GSH(+/-) denotes designed general superior hybrid with higher/lower trait performance;

SH(+/-)*i* denotes designed environment specific hybrid with higher/lower trait performance in *i*-th environment.

_variance_components (population mean: 100.5358 phenotypic variance: 144.3982 Total Heritability: 0.7678)

h^2(A+D)	h^2(AE+DE)	h^2(I)	h^2(IE)	V(e)/V(P)			
0.1046	0.0732	0.4624	0.0638	0.2322			
h^2(A)	h^2(D)	h^2(I)	h^2(A+AA)	h^2(AE)	h^2(DE)	h^2(IE)	h^2(AE+AAE)
0.0556	0.0490	0.4624	0.0625	0.0662	0.0070	0.0638	0.1301

Fig 3.13: Variance components and heritability ascribed different genetic components

_1D_effect

QTL	SNPID	A	SE	P-Value	D	SE	P-Value	AE1	SE	P-Value	AE2	SE	P-Value
1-5	MK5	1.3339	0.3171	2.599e-005	---	---	---	---	---	---	---	---	---
1-41	MK41	-0.9428	0.3171	2.947e-003	-0.9452	0.4465	3.429e-002	1.3273	0.5384	1.370e-002	---	---	---
2-22	MK72	1.6001	0.3291	1.173e-006	---	---	---	1.3725	0.5567	1.369e-002	1.9603	0.5567	4.302e-004
3-2	MK102	---	---	---	1.5494	0.4465	5.215e-004	2.2231	0.5795	1.253e-004	---	---	---
3-34	MK134	-1.6728	0.3501	1.783e-006	-1.3635	0.4465	2.264e-003	---	---	---	---	---	---
3-42	MK142	---	---	---	-1.3860	0.4465	1.912e-003	---	---	---	1.0535	0.5135	4.020e-002

Fig 3.14: Mapping result ascribed to single SNP effects

_1D_heritability

QTL	h^2(a)	h^2(d)	h^2(ae)	h^2(ae1)	h^2(ae2)	h^2(ae3)	h^2(de)	h^2(de1)	h^2(de2)	h^2(de3)
1-5	0.0123	---	0.0042	---	---	0.0126	---	---	---	---
1-41	0.0062	0.0062	0.0159	0.0122	---	0.0354	---	---	---	---
2-22	0.0177	---	0.0132	0.0130	0.0266	---	---	---	---	---
3-2	---	0.0166	0.0245	0.0342	---	0.0391	---	---	---	---
3-34	0.0194	0.0129	---	---	---	---	---	---	---	---
3-42	---	0.0133	0.0085	---	0.0077	0.0177	0.0070	0.0209	---	---

Fig 3.15: Heritability ascribed to single QTS effects

_2D_effect

QTL_i	SNPID_i	QTL_j	SNPID_j	AA	SE	P-Value	AD	SE	P-Value	DA	SE	P-Value	DD	SE	P-Value	AAE1	SE	P-Val
1-5	MK5	3-2	MK102	---	---	---	1.4571	0.6033	1.574e-002	-1.8711	0.6458	3.769e-003	-2.7097	0.9053	2.763e-003	1.6820	0.7558	2.606
1-41	MK41	3-34	MK134	---	---	---	-1.2162	0.6033	4.382e-002	---	---	---	-3.1859	0.9053	4.337e-004	---	---	---
2-22	MK72	3-34	MK134	---	---	---	2.6623	0.6240	1.993e-005	---	---	---	4.2525	0.9053	2.652e-006	---	---	---
3-7	MK107	3-42	MK142	---	---	---	---	---	---	-2.0709	0.5976	5.304e-004	---	---	---	---	---	---
3-34	MK134	3-38	MK138	1.0000	0.4552	2.806e-002	---	---	---	---	---	---	---	---	---	---	---	---
3-34	MK134	3-42	MK142	---	---	---	---	---	---	---	---	---	3.4220	0.9053	1.572e-004	---	---	---

Fig 3.16: Epistatic effects ascribed to paired QTSs

_2D_heritability

QTL_i	QTL_j	h^2(aa)	h^2(ad)	h^2(da)	h^2(dd)	h^2(aae)	h^2(aae1)	h^2(aae2)	h^2(aae3)	h^2(ade)	h^2(ade1)	h^2(ade2)	h^2(ade3)	h^2(dae)
1-5	3-2	---	0.0147	0.0242	0.0509	0.0197	0.0196	---	0.0395	---	---	---	---	---
1-41	3-34	---	0.0102	---	0.0703	0.0041	---	---	0.0124	---	---	---	---	---
2-22	3-34	---	0.0491	---	0.1252	0.0400	---	0.1201	---	---	---	---	---	---
3-7	3-42	---	---	0.0297	---	---	---	---	---	---	---	---	---	---
3-34	3-38	0.0069	---	---	---	---	---	---	---	---	---	---	---	---
3-34	3-42	---	---	---	0.0811	---	---	---	---	---	---	---	---	---

Fig 3.17: Heritability of epistatic effects ascribed to paired QTSs

_genotype_value

Entry	G	G+GE1	G+GE2	G+GE3
QQ	3.3551	10.8276	10.5327	-0.4795
qq	1.6795	-0.6937	2.8295	7.6159
SuperiorLine(+)	7.3701	14.1733	11.2021	8.5127
SuperiorLine(-)	-4.3354	-8.1927	-11.5130	-11.7067
F1	-0.6279	-1.4955	-0.6279	-0.6279
SuperiorHybrid(+)	11.1090	14.1733	14.9409	13.9984
SuperiorHybrid(-)	-10.1362	-13.7103	-13.1500	-16.0966

Fig 3.18: Genotypic values based on predicted effects of QTS for designed genotypes

_superior_genotype

QTL	GSL(+)	GSL(-)	SL(+)1	SL(+)2	SL(+)3	SL(-)1	SL(-)2	SL(-)3	GSH(+)	GSH(-)	SH(+)1	SH(+)2	SH(+)3	SH(-)1	SH(-)2	SH(-)3
1-5	QQ	qq	QQ	QQ	QQ	QQ	qq	qq	QQ	qq	QQ	QQ	QQ	QQ	qq	qq
1-41	qq	QQ	QQ	qq	qq	qq	QQ	QQ	Qq	Qq	QQ	Qq	qq	Qq	Qq	QQ
2-22	QQ	qq	QQ	QQ	QQ	qq	qq	qq	QQ	qq	QQ	QQ	QQ	qq	qq	Qq
3-2	QQ	qq	QQ	QQ	QQ	qq	qq	QQ	Qq	qq	QQ	Qq	Qq	qq	qq	QQ
3-34	qq	QQ	qq	QQ	QQ	QQ	QQ	QQ	qq	Qq	qq	QQ	QQ	Qq	Qq	QQ
3-42	QQ	qq	QQ	QQ	qq	qq	qq	QQ	QQ	qq	QQ	QQ	qq	Qq	qq	Qq
3-7	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	QQ/qq	Qq	ALL	QQ/qq	ALL
3-38	qq	qq	qq	QQ	QQ	qq	qq	qq	qq	Qq/qq	qq	QQ	QQ	Qq/qq	Qq/qq	qq

Fig 3.19: The superior and inferior genotypes designed for general environment or specific environment

4 User manual for QTT/P/M

4.1 Introduction on QTT/P/M

QTT/P/M is one module of the software QTXNetwork for association mapping genetic variation of quantitative trait transcripts (QTT), quantitative trait proteins (QTP), and quantitative trait metabolites (QTM) with phenotypic variation. The mixed linear model approach combining with Markov Chain Monte Carlo (MCMC) method are employed to perform association analysis, and to predict and test genetic effects of significant QTTs, QTPs, and QTMs. Parallel computing base on Graphics Processing Unit (GPU) is used in some key time-costive subroutines. The QTT/P/M can also be run on Windows or Mac operation system with or without GPU hardware.

4.2Running QTT/P/M

4.2.1 Creating a project for QTT/P/M

The QTT/P/M program, kept in the folder ~/exe/QTT under the QTXNetwork, can be run directly in "command window" or in "explore window". In default, we run it via the shell program (QTXNetwork). To run the QTT/P/M, we first need running the QTXNetwork for creating a new project for the QTT/P/M by clicking the button  in the main window of the QTXNetwork (Fig 4.1), a window for creating new project will be popped up (Fig 4.2). After selecting module type "QTTPM" and specifying project name, click the "Accept" button (Fig 4.2), a QTT/P/M project will be listed in the left pane named Navigator (Fig 4.3).

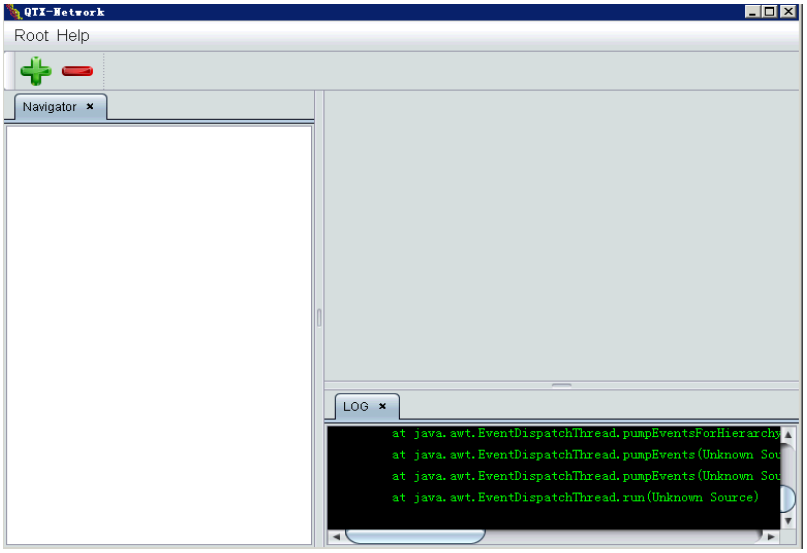


Fig 4.1: The main window of the QTXNetwork

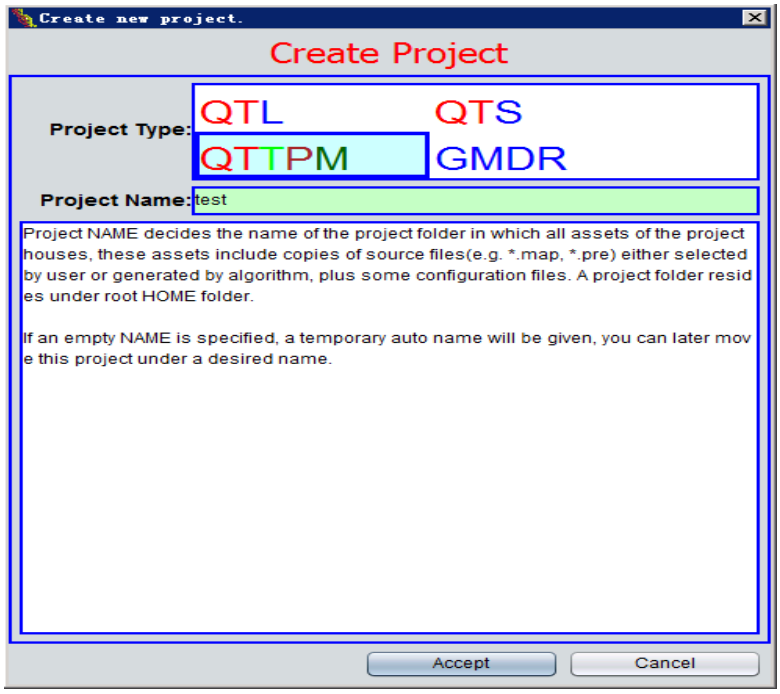


Fig 4.2: The popup window for creating new project for QTT/P/M

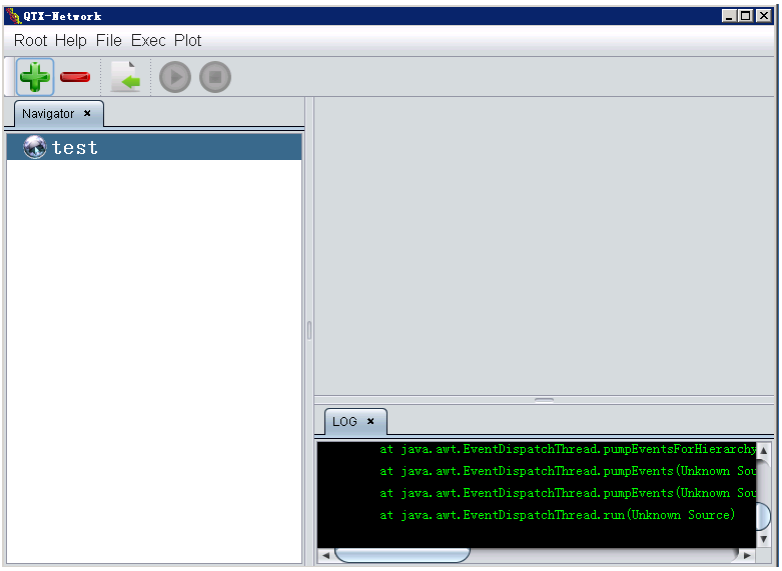


Fig 4.3: The window of QTXNetwork for performing QTT/P/M

4.2.2 Deleting QTT/P/M project

Select the QTT/P/M project to delete in the Navigator pane, and then click the fast button

4.2.3 Specifying data file

In a QTT/P/M project window, a button  for "locate and import source file" will occur in the toolbar, we can locate and import data files (genotype data file and phenotype data file for QTT/P/M in the popup window after clicking the button ) . In this panel (Fig 4.4), there are three sections to set input or output files. Since there is no need for loading genetic map in association mapping for QTT/P/M, only data of genotypes and phenotypes are required for loading. If the genotype data and phenotype are saved in one file, we need selecting "Use TXT file" option, otherwise, selecting "Use GNO+PNO files" to input data files of genotypes and phenotypes, individually. The third is the "Output report" section, where the output file name can be inputted.

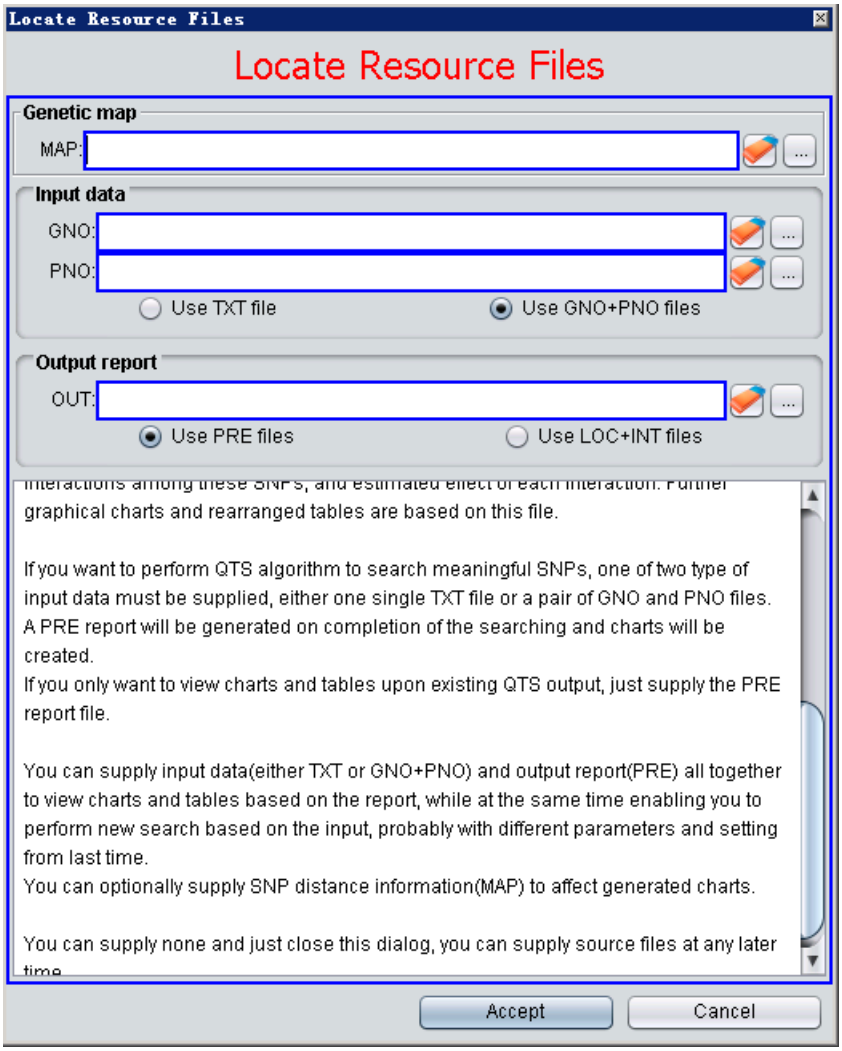


Fig 4.4: The window for inputting data files and output file

4.2.4 Starting QTT/P/M

For example, suppose we open a QTT projected named “test”, and also specify the data files for this project. The sample data files are the genotype data file (~\sample\QTT\SampleQTT.gen), the phenotype data file (~\sample\QTT\SampleQTT.phe), and the report file (~\sample\QTT\SampleQTT.pre), respectively, then, the main window for the QTT project will exhibits as the Fig 4.5.

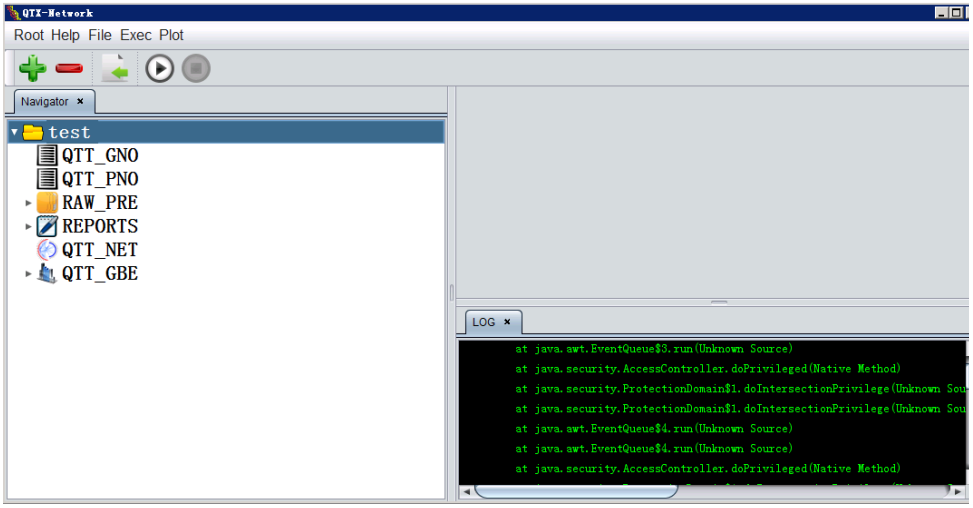


Fig 4.5: The window of QTXNetwork after creating or opening a QTT/P/M project

Finally, we can click the button  (Fig 4.5) to start the QTT analysis if all default algorithm parameters are used. We can adjust QTT algorithmn configuration before running, detailed information refer to the chapter 4.3.2.

4.3 Data file format and running configuration of QTT/P/M

4.3.1 Data format

For performing analyses with QTT/P/M, two source data files are required: a genotype data file and a phenotype data file. The genotype data file contains expression variation for each genotype of all the subjects, and the phenotype data file contains observation values of traits under study for all subjects. Some sample files for briefly demonstrating the format of source data files for QTT/P/M in the sub-directory (~\Sample\QTT) where QTXNetwork has been installed. The genotype data file with extension name of “.Gen” and phenotype data files with extension name of “.Phe” for QTT can be directly used by QTT/P/M.

The genotype data can be organized in two different formats, which are summarized in Table 4.1 and 4.2. For the first one (Table 4.1), the first line consists of the keyword “#Ind” and molecular names, the first column is for the ID or name of individual subjects. In the second format, the first line consists of the

keyword "#Mk" and ID name of individual subjects, the first column is for the ID or name of molecules. The other data are the codes of different genotype in each molecular locus for all subjects. These codes need to be defined in phenotype data file corresponding this genotype data file.

Table 4.1: The first type of genotype data format

#Ind	MK1	MK2	MK3	MK4	MK5	…	MKm
1	0.19	1.37	0	0	0.68	…	0.04
2	0.21	1.5	0	0	0.84	…	0.02
3	0.04	0.71	0	0.64	0	…	0.03
…	…	…	…	…	…	...	…
n	0.05	1.07	0.04	0	1.07	…	0.05

Table 4.2: The second type of genotype data format

#Mk	1	2	3	4	5	…	n
MK1	0.19	0.21	0.04	0.04	0.11	…	0.05
MK2	1.37	1.5	0.71	0.93	1.14	…	1.07
MK3	0	0	0	0	0.02	…	0.04
MK4	0	0	0.64	0.74	0	…	0
MK5	0.68	0.84	0	0	1.86	…	1.07
…	…	…	…	…	…	…	…
MKm	0.04	0.02	0.03	0.03	0.03	…	0.05

General description: This part is for specifying the basic features of the phenotype data file, and is usually put in front of the data file. Each key string must be started with an underline "_", and no white space is allowed within it. There are seven possible items for general description. They can be arranged in any order. A typical description for a phenotype data file looks like:

_Population EXP
_Observations 33
_Chromosomes 1 Chr1
_TotalTranscript 119 119
_TraitNumber 6
_Environments No
_Replications Yes

_Population specifies the population type used. In QTT/P/M mapping, the genotype values are quantitative, so we define the population type as "EXP".

_Observations specifies the total number of observations for each trait studied.

_Environments specifies the status of experimental design for environments. After the key word _Environments, write the specification word "Yes" for having multiple environments or, "No" for only one environment.

_Replications specifies the status of experimental design for replications or blocks. If the experiment is conducted with replications or blocks, write the specification word "Yes" after the keyword _Replications, otherwise write "No".

_TraitNumber specifies the total number of traits included in the phenotype data file.

_Chromosomes specifies the chromosome number and the names in marker file. In QTT/P/M, all molecular markers are regarded as lying in one chromosome, thus, the values after this keyword are "1" and "Chr1".

_TotalTranscript specifies the total number of the transcripts included in the marker data file and the number of markers in each chromosome. The first number must be equal to the summation of all the markers.

Trait data body: This part is between two key strings *TraitBegin* and *TraitEnd*. The data source includes the environment (if available), the replication (if available) and the ID name of subjects, as well as the observations obtained for traits studied. The following is an example for the trait data body.

TraitBegin

Env#	Rep#	Geno#	Trait_1	Trait_2	Trait_3	;
1	1	1	2.44	7.4	10.04	;
1	1	2	2.4	4.32	8.55	;
……						
1	1	90	3.54	8.19	10.74	;
1	2	1	3.17	6.91	11.86	;
1	2	2	1.9	4.31	11.36	;
……						
1	2	90	3.22	10.54	11.48	;
2	1	1	5.74	12.78	11.27	;
2	1	2	7.65	7.02	11.96	;
……						
2	1	90	6.58	13.92	9.94	;
2	2	1	6.01	10.22	9.95	;
2	2	2	6.22	11.99	7.81	;
……						
2	2	90	7.98	13.21	12.03	;

TraitEnd

The second row includes the indicator strings and the names of the traits. The number of source strings depends on the experimental design. If both environments and replications are taken, a maximum of three strings must be included: the first string for environment (Env#), the second string for replication (Rep#)

and the third string for subject (Ind#). You can use whatever strings to express the sources because they are just used to indicate the conents in the columns below them. If the experiment is conducted without environmental factor or replications, the corresponding column must be removed. And also, a semicolon ";" is required at the end of each observation data row.

4.3.2 QTT/P/M algorithm and configuration

Before conducting QTT/P/M, we first need setting the QTT/P/M algorithm configuration. Suppose a new QTT project is created or a QTT project is opened in main window of the QTXNetwork, click the QTT project in the Navigator panel (Fig 4.5) to set the algorithm configuration, then activate the command item "Config" under the "Exec" software menu (Fig 4.6), a panel will be popped up for setting configuration (Fig 4.7, Fig 4.8).

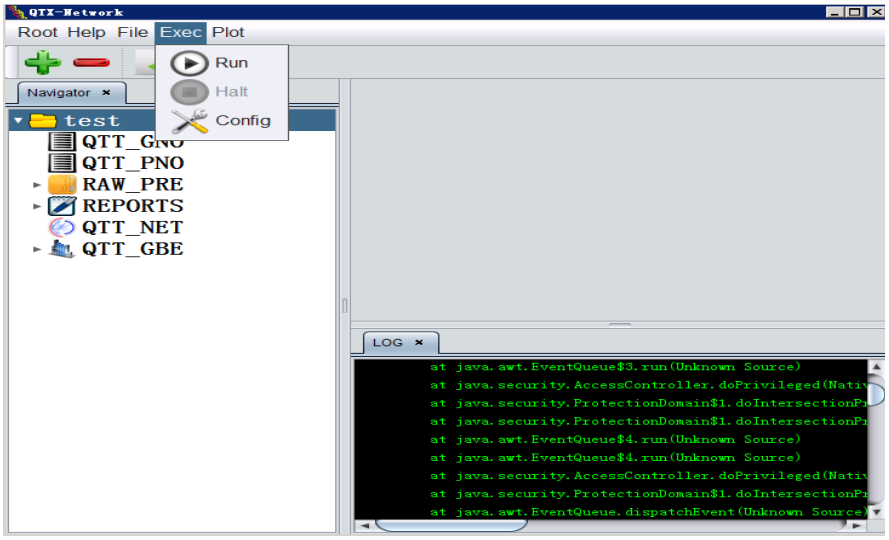


Fig 4.6: Executing QTT/P/M under the "Exec" of the menu

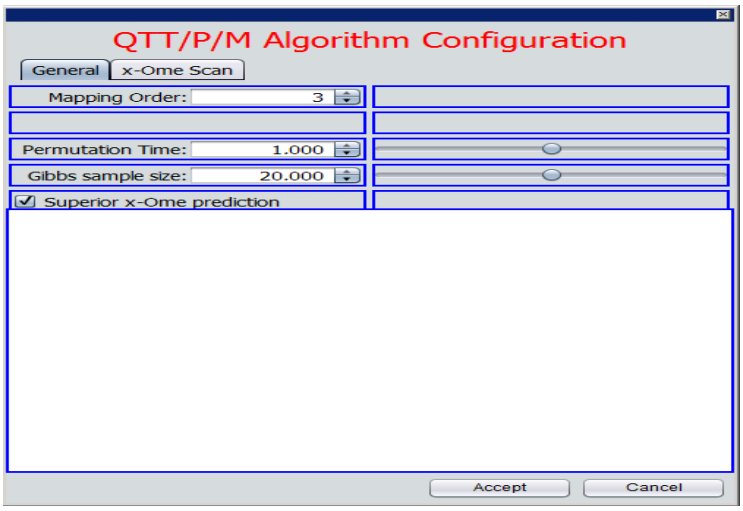


Fig 4.7: General algorithm configuration for QTT/P/M

Mapping Order – Set the mapping order to decide mapping with or without two-locus or three-locus epistasis. "1" denotes that the program only detects single-locus effect QTT/P/M; "2" denotes that the program detects both single-locus effect QTT/P/M and two-locus epistasis; "3" denotes that the program not only detects single-locus and two-locus effect, but also three-locus epistasis.

Permutation Time – Set the permutation time to conduct permutation test on significance of QTT/P/M effects.

Gibbs sample size – Set this value to predict QTT/P/M effects by Monte Carlo Markov Chain method, otherwise by mixed linear model approach.

Superior x-Ome Prediction – Choose this option to predict superior genotypes based on QTT/P/M effects (Ref. Yang J and Zhu J. (2005). Predicting Superior Genotypes in Multiple Environments Based on QTL Effects. Theoretical and Applied Genetics, 110: 1268-1274.)

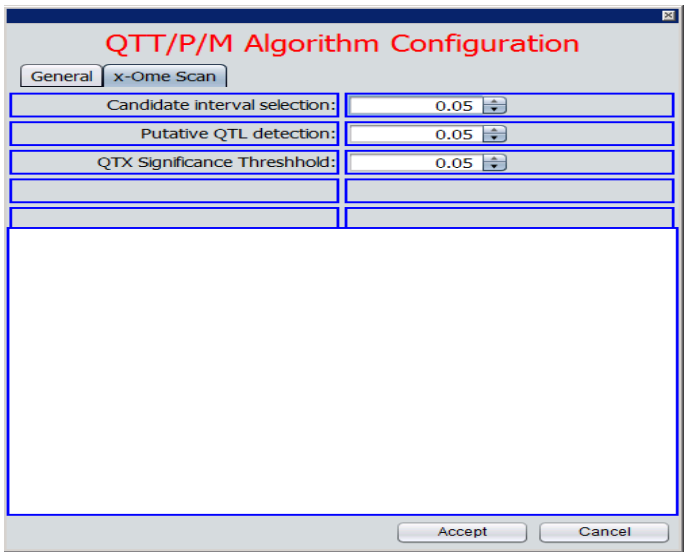


Fig 4.8: Genome scan configuration for QTT/P/M

Candidate interval selection – The experimental-wise type I error for candidate interval selection.

Putative QTL Detection – The experimental-wise type I error for putative QTT/P/M detection.

QTX Significant Threshold – The significant level for QTT/P/M effects.

4.4 Saving and understanding pictures and report

4.4.1 Saving pictures

After finishing analysis of QTT/P/M, we can save figures in png or eps format by executing the "Export GG" for GxG plot or "Export GE" for GxE plot under the "Plot" menu (Fig 4.9).

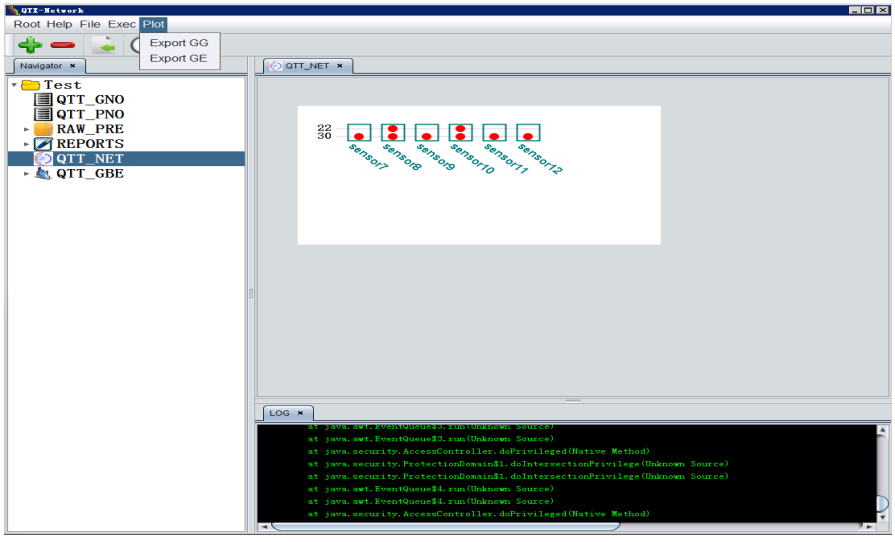


Fig 4.9: Saving picture by using "Plot" menu

4.4.2 Understanding the symbols in figure

For the GxG plot of QTT (Fig 4.10), the red circles denote the main effects of detected significant transcripts.

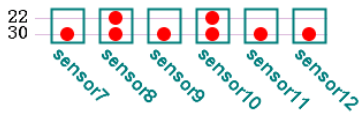


Fig 4.10: The GG plot of QTT

4.4.3 Understanding text report

In order to understand each item in report file, we firstly define symbols or variables in the report file, then explain the report file using the Fig 4.11-4.14 coming from the simulation data which can be found in the sample folder (~\Sample\QTT\SampleQTT.pre) of the QTXNetwork.

Q (or q) and QE (or qe): denote main effect and their interaction effect with environment, respectively;

QQ (or qq) and QQE (or qqe): denote epistasis effect and their interaction effects with environments, respectively;

QE*i* (or qe*i*), QQE*i* (or qqe*i*) denote the *i*-th environment-specific genetic component effects of transcript, protein or metabolism, respectively.

In the Fig 4.11, V(e), V(P) denote the residual variance and the phenotypic variance, respectively, h^2(X) denote the proportion of phenotypic variance contributed to the genetic effects of X, respectively, where X stand for one genetic component effect or summation of some genetic component effects, such as, h^2(Q) is the proportion of phenotypic variance contributed to main effect of detected QTT/P/M.

_variance_components (population mean: 0.7788 phenotypic variance: 8.3179 Total Heritability: 0.9999)

h^2(Q)	V(e)/V(P)
0.9999	0.0001

Fig 4.11: Variance components and heritability ascribed different genetic components

For the Fig 4.12, the first column is the serial number of each detected significant QTT/P/M, the second is the QTT/P/M ID, the succeeding 3 columns are predicted genetic effects, standard error and significance probability value, respectively. The Fig 4.13 presents the proportions of phenotypic variance explained by the genetic component effects of detected QTT/P/M in the Fig 4.12.

QTL	interval	T	SE	P-Value
1-30	ST14	-2.8840	0.4539	2.141e-010

Fig 4.12: Mapping result ascribed to single transcript effects

QTL	h^2(t)
1-30	0.9999

Fig 4.13: Heritability ascribed to single transcript effects

The Fig 4.14, the first column is the individual ID, and the second is the predicted individual genotype values based on detected significant transcript.

5 User manual for GWAS-GMDR

5.1 Introduction on GWAS-GMDR

GWAS-GMDR is a program which implements the GMDR algorithm specifically for analyzing Genome-Wide Association Analysis (GWAS) data. GWAS-GMDR assumes that all the attributes are SNPs, and accelerates the computation speed by the many-core parallel computation technique of Graphics Processing Units (GPU), to overcome the memory and computation burdens of exhaustively searching among millions of SNPs, so that detecting gene-gene interactions become feasible in real GWAS data by GMDR.

5.2 Running GWAS-GMDR

5.2.1 Creating a project for GMDR

The GMDR program, kept in the folder ~/exe/MDR under the QTXNetwork, can be run directly in "command window" or in "explore window". In default, you can run it via the shell program (QTXNetwork). To run the GMDR, you first need running the QTXNetwork for creating a new project for the GMDR by clicking the button  in the main window of the QTXNetwork (Fig 5.1), a window for creating new project will be popped up (Fig 5.2). After selecting module type "GMDR" and specifying project name, click the "Accept" button (Fig 5.2), a GMDR project will be listed in the left pane named Navigator (Fig 5.3).

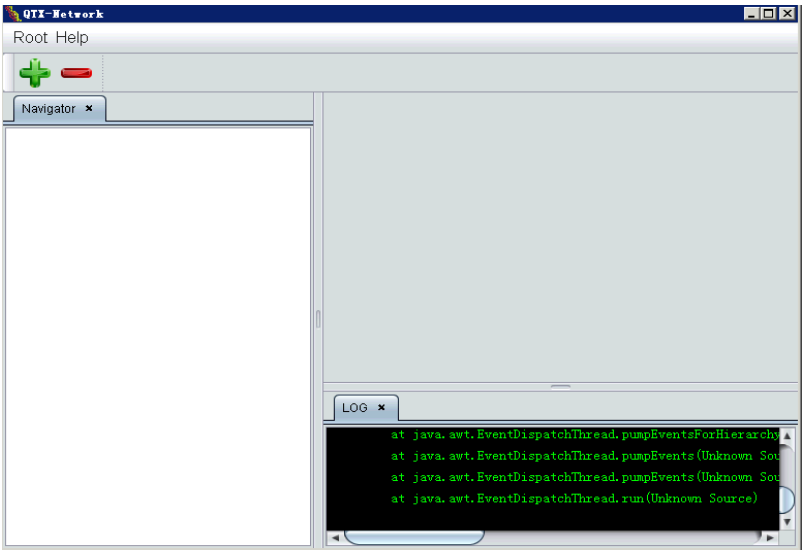


Fig 5.1: The main window of the QTXNetwork

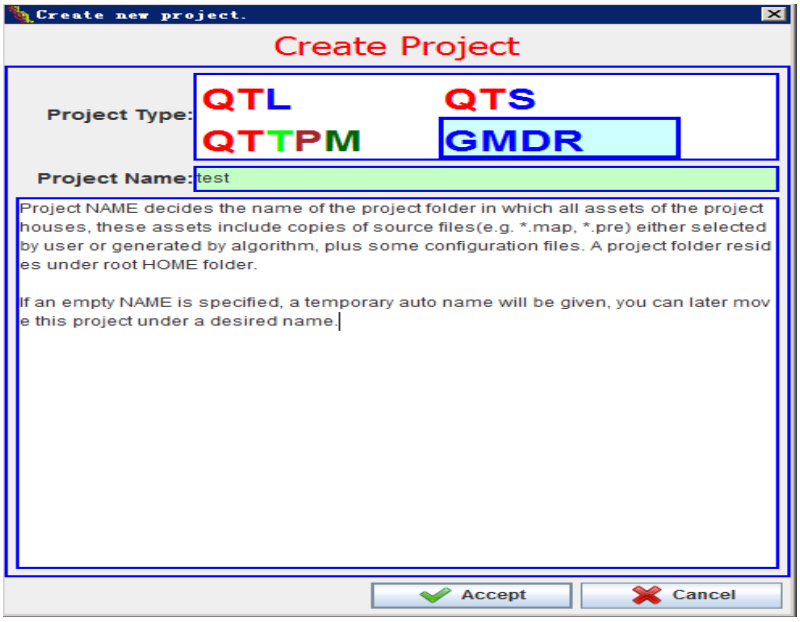


Fig 5.2: The popup window for creating new project for GMDR

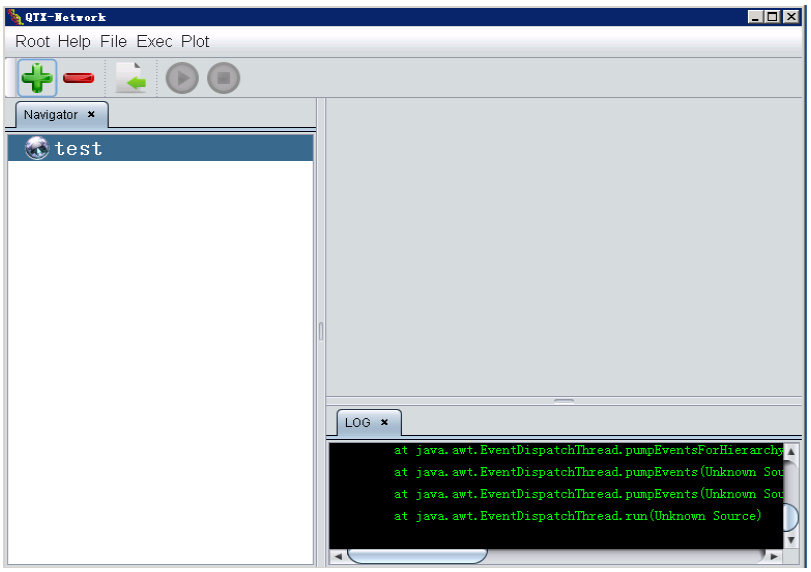


Fig 5.3: The window of QTXNetwork for performing GMDR

5.2.2 Deleting GMDR project

Select the GMDR project to delete in the Navigator pane, and then click the fast button

5.2.3 Specifying data file

In a GMDR project window, a button  for “locate and import source file” will occur in the toolbar, you can locate and import data files (genotype data file and phenotype data file for GMDR in the popup window after clicking the button ). In this panel, there are three sections to set input or output files (Fig 5.4). The first is the Genetic map section, which will be used when Genetic linkage map is available. Generally, genetic map is not required for association mapping. The second is the Input data section, and you need input data files of genotypes and phenotypes, individually. The third is the “Output report” section, where the output file name can be inputted.

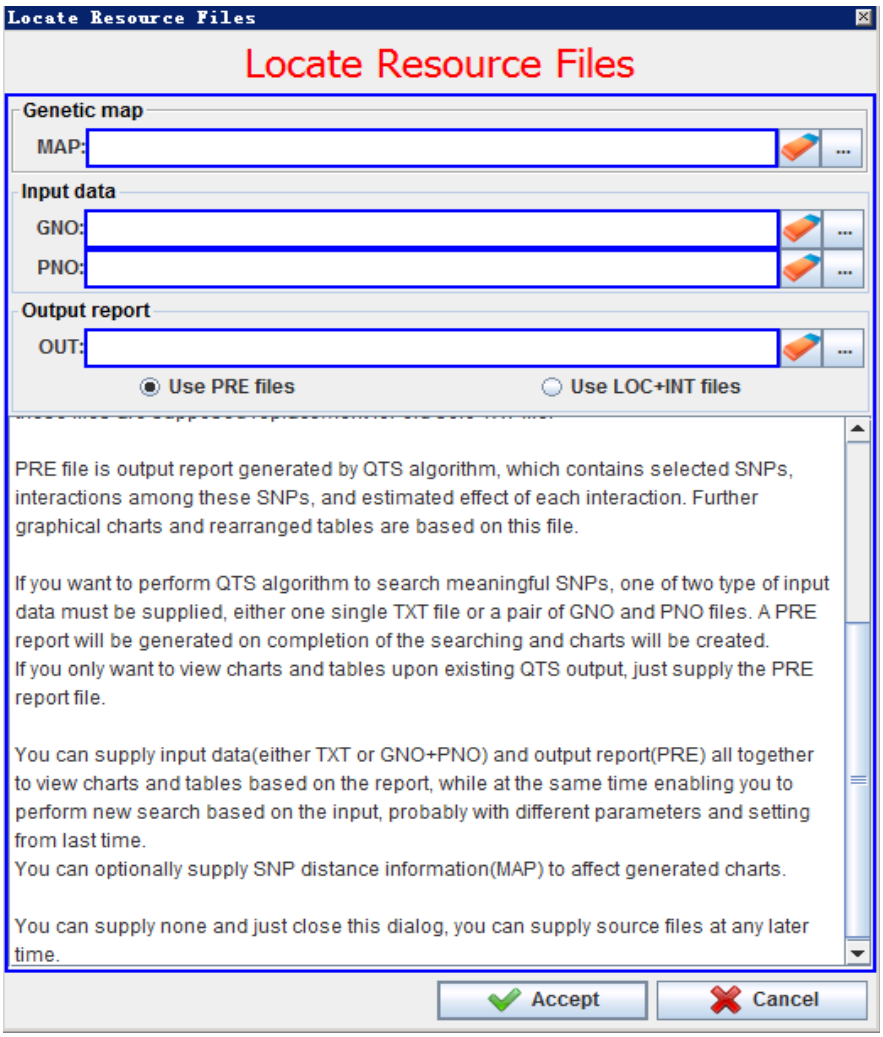


Fig 5.4: The window for inputting data files and output file

5.2.4 Starting GMDR

For example, suppose you open a GMDR projected named “test”, and also specify the data files for this project. The sample data files are the genotype data file (~\sample\ MDR\SampleData.gen), the phenotype data file (~\sample\MDR\SampleData.phe), and the report file (~\sample\MDR\SampleData.pre), respectively, then, the main window for the GMDR project will exhibits as the Fig 2.5.

Finally, you can click the button  (Fig 5.5) to start the GMDR analysis if all default algorithm parameters are used. You can adjust GMDR algorithmn configuration before running, detailed information refer to the chapter 5.3.2.

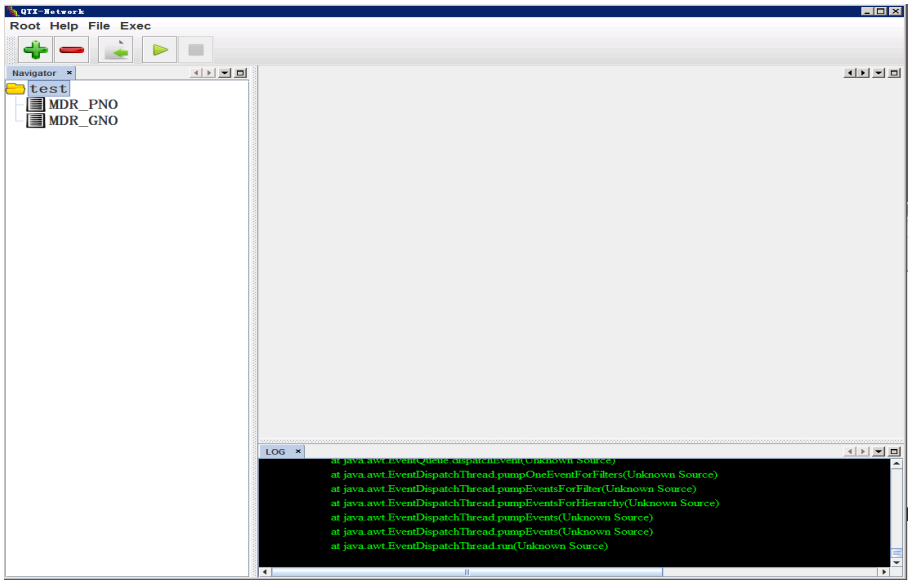


Fig 5.5: The window of QTXNetwork after creating or opening a GMDR project

5.3 Data file format and running configuration of GMDR

5.3.1 Data format

For performing analyses with GMDR, two source data files are required: a genotype data file and a phenotype data file. The genotype data file contains marker information for each genotype; the phenotype data file contains covariate values and observation values of traits under study for all individuals. Some sample files are provided for briefly demonstrating the format of source data files for GMDR in the sub-directory (~\Sample\MDR) where QTXNetwork has been installed. The genotype data file with extension name of “.Gen” and phenotype data files with extension name of “. Phe” for GMDR can be directly used by GMDR.

The genotype data can be organized in two different formats which are summarized in Table 5.1 and 5.2. For the first one (Table 5.1), the first line consists of the keyword “#Ind” and molecular marker

names, the first column is for the ID or name of individual subjects. In the second format, the first line consists of the keyword "#Mk" and ID name of individual subjects, the first column is for the ID or name of molecular markers. The other data are the codes of different genotype in each molecular marker locus for all individuals;

able 5.1: The first type of genotype data format

#Ind	MK1	MK2	MK3	MK4	MK5	…	MKm
1	H	H	A	A	A	…	A
2	H	B	H	B	B	…	H
3	A	A	H	H	H	…	B
…	…	…	…	…	…	...	…
n	A	A	A	A	A	…	H

Note: There are three types of markers: A or 0 for *QQ* genotype, B or 2 for *qq* genotype, and H or 1 for *Qq* genotype.

Table 5.2: The second type of genotype data format

#Mk	1	2	3	4	5	…	n
MK1	H	H	A	A	A	…	B
MK2	H	B	A	A	A	…	B
MK3	A	H	H	A	A	…	B
…	…	…	…	…	…	…	…
MKm	A	H	B	A	H	…	H

General description: This part is for specifying the basic features of the phenotype data file, and is usually put in front of the data file. Each key string must be started with an underline "_", and no white space is allowed within it. There are five possible items for general description. They can be arranged in any order. A typical description for a phenotype data file looks like:

_TotalMarkers 100
_Observations 1800
_Chromosomes 3 Chr1 Chr3 Chr5
_TotalSnps 100 53 10 37
_TraitNumber 2

_TotalMarkers: specifies the total number of markers included in the marker data file.

_Observations: specifies the total number of subjects for each trait studied.

_Chromosomes: specifies the chromosome number and the names in marker file. If all molecular markers are regarded as lying in one chromosome, thus, the values after this keyword are "1" and "Chr1".

_TotalSnps: specifies the total number of the markers included in the marker data file and the number of markers in each chromosome. The first number must be equal to the summation of all the markers.

_TraitNumber: specifies the total number of traits included in the phenotype data file.

Trait data body: This part is betyouen two key strings *TraitBegin* and *TraitEnd*. The data source includes the covariate values and the ID name of subjects, as well as the phenotype values obtained for traits studied. The following is an example for the trait data body.

TraitBegin

Sex	Age	Height	SubjectId	Trait1	Trait2	;
1	72.1	1.74	Subject1	1	37.40640189	;
2	40.5	1.687	Subject2	1	38.286138	;
2	39.4	1.722	Subject3	1	36.91420189	;
2	50.2	1.697	Subject4	1	31.370838	;
1	52.2	1.78	Subject5	1	35.74930189	;
2	39.9	1.687	Subject6	1	35.572038	;
2	72.6	1.713	Subject7	0	22.91250189	;
1	36.2	1.685	Subject8	0	24.341	

……

TraitEnd

There is also another phenotype data file format which does not include above key strings. If the file does not include some covariates, the data source only includes the ID name of subjects and the phenotype values obtained for traits studied. The following is an example for the phenotype data file which seems much easier than the previous one.

SubjectId	Trait1	Trait2	;
Subject1	1	37.40640189	;
Subject2	1	38.286138	;
Subject3	1	36.91420189	;
Subject4	1	31.370838	;
Subject5	1	35.74930189	;
Subject6	1	35.572038	;
Subject7	0	22.91250189	;
Subject8	0	24.341	;

……

5.3.2 GMDR algorithm and configuration

Before conducting GMDR, you first need setting the GMDR algorithm configuration. Suppose a new GMDR project is created or an existed GMDR project is opened in main window of the QTXNetwork, click the GMDR project in the Navigator panel (Fig 5.5) to set the algorithm configuration, then activate the command item “Algorithm Config” under the “Exec” software menu (Fig 5.6), a panel will be popped up for setting configuration (Fig 5.7).

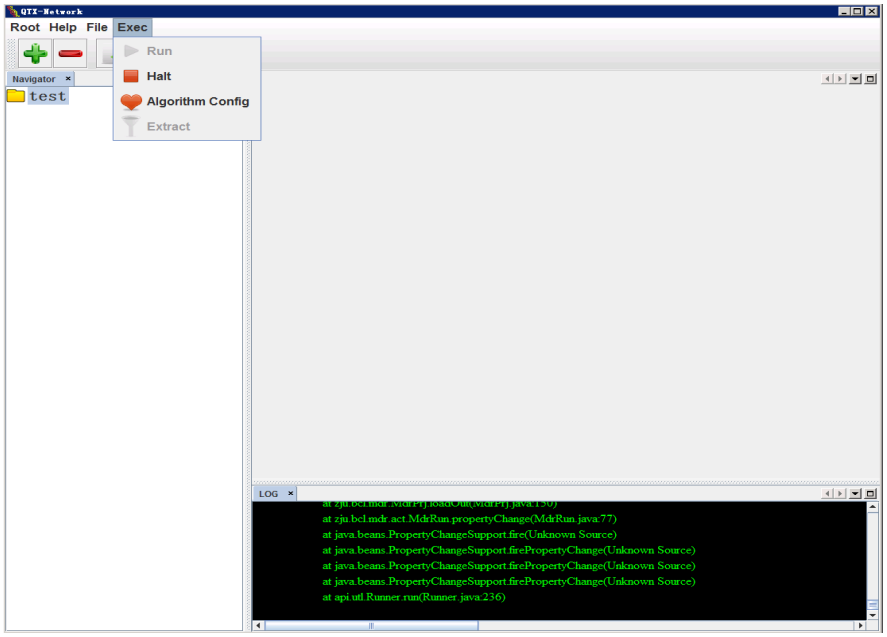


Fig 5.6: Executing GMDR under the “Exec” of the menu

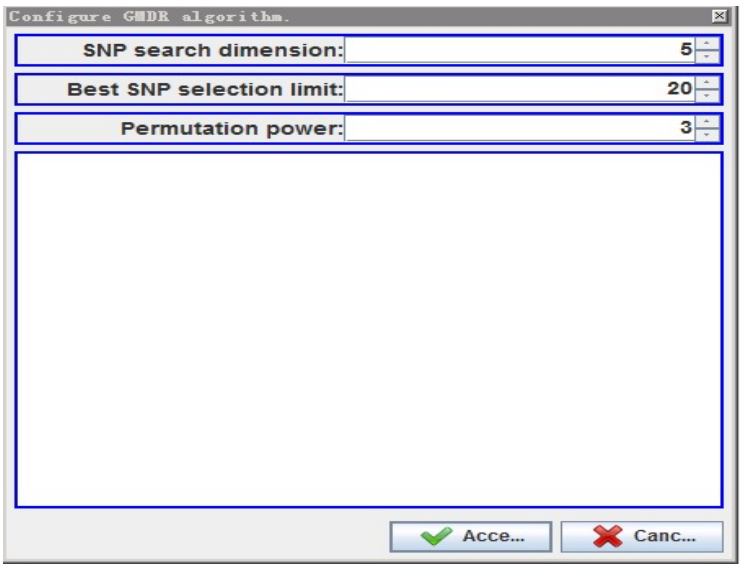


Fig 5.7: General algorithm configuration for GMDR

SNP search dimension – This option sets the search dimension (1~5) and the max dimension is 5.

Best SNP selection limit – This option sets the number of best SNP combinations selected in each training set.

Permutation power – This option set the number of permutations used in testing stage for each selected SNP combination. If this option is set as *n*, then 10^n permutations will be performed for each SNP combination selected in training stage; otherwise the pre-setting 10^3 =1,000 permutations will be performed for each SNP combination selected in training stage.

5.4 Extract results of GMDR for GWAS

After finishing analysis of GMDR, you can extract the result file with GWAS format by executing the "Extract" under the "Exec" menu (Fig 5.6).