KGG: A systematic biological Knowledge-based mining system for Genome-wide Genetic studies (Version 4.1)

User Manual

Miao-Xin Li, Lin Jiang and Jiang Li

Laboratory of Precision Medical Genomics
Zhongshan School of Medicine
Sun Yat-Sen University,
Guangzhou City, Guangdong Province, 510080, PRC

Content

Hints for large GWAS dataset (around or over 2.5 million SNPs)

Set or change large memory for KGG4 say, 2000MB, by *Tools->Set System Memory*.

1. Introduction and References

1.1 Introduction

KGG[1] (Knowledge-based mining system for Genome-wide Genetic studies) is a software tool to perform knowledge-based secondary analysis with summary statistics from genome-wide association studies (GWAS). At present, the version 4 has been equipped with main functions to perform 5 types secondary Knowledge-based analysis by using SNP p-values from GWAS:

- Gene-based association[2,3],
- Conditional gene-based association[2],
- Multivariate gene-based association[4],
- Gene-pair interaction-based association[5],
- Geneset based association [6].

In addition, KGG has provided direct hyperlinks to several useful bioinformatics annotation databases on sequence variants (http://jjwanglab.org/gwasrap), genes (GeneCards, http://www.genecards.org/) and pathways (MsigDB, http://www.broadinstitute.org/gsea/msigdb). A number of functions to model emerging epigenomic regulatory data for prioritizing association signals are still under development.

1.2 References

1. Li MX, Sham PC, Cherny SS, Song YQ. A knowledge-based weighting framework to boost the power of genome-wide association studies. PLoS One. 2010 Dec 31;5(12):e14480.
2. Li MX, Gui HS, Kwan JS, Sham PC. GATES: A rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet. 2011 Mar 11;88(3):283-293.
3. Li et al. A powerful approach isolates independently associated genes of Schizophrenia with summary statistics from large-scale whole genome association meta-analysis (Submitted)
4. Sluis et al. MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. Bioinformatics. 2015 Apr 1;31(7):1007-15.
5. Li MX*, Kwan JS*, Sham PC. HYST: A hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. Am J Hum Genet. 2012 Sep 7;91(3):478-88.
6. Gui et al. Genome-wide gene- and gene-set-based association analyses identify novel patterns of genetic sharing across complex phenotypes(Submitted)
7. Li J, Sham PC, Song Y, Li M. SPS: a simulation tool for calculating power of set-based genetic association tests. Genet Epidemiol. 2015;39(5):395-7

2. Installation

2.1 Install Java Runtime Environment (JRE)

The Java Runtime Environment (JRE) v1.7 (or higher version) is required to run KGG4 on any operating systems (OS). It can be downloaded from http://java.sun.com/javase/downloads/index.jsp for free. Installing the JRE is very easy on Windows OS and Mac OS X.

On Linux, you have more work to do. Details of the installation can be found at http://www.java.com/en/download/help/linux_install.xml. In Ubuntu, if you have an error message like: "Exception in thread "AWT-EventQueue-0" java.awt.HeadlessException …", then please installs the Sun Java Running Environment (JRE) first.

To install the Sun JRE on Ubuntu(10.04), please use the following commands:
sudo add-apt-repository "deb http://archive.canonical.com/ lucid partner"
sudo apt-get update
sudo apt-get install sun-java7-jre sun-java7-plugin sun-java7-fonts

Detailed explanation of above commands can be found at http://www.ubuntugeek.com/how-install-sun-java-runtime-environment-jre-in-ubuntu-10-04-lucid-lynx.html.

Note: After completing Java installation, please make sure that not only the java is executable but also the extracted jre/bin directory is added to the PATH, otherwise KGG4 would not start properly. This is easily achieved by executing the following command on the terminal:

```
echo 'export PATH=/path/to/installed/jre/bin:$PATH' >> ~/.bashrc && source ~/.bashrc
```

Thanks **Attila Pulay** for the suggestion!

2.2 Install and initiate KGG

To simplify the installation, we still keep KGG as a green tool (i.e., no formal installation procedure guided by an installation wizard). After decompressing the kgg4.zip file, you will see a "bin" folder where there are 3 script files to initiate KGG4. On Microsoft Windows, please double click kgg4.exe or kgg464.exe file. On Linux, Mac OS X and Solaris, please type the kgg4 in a Command-line Terminal.

If you have over 4 million variants, you are suggested set a larger memory for KGG. The default setting is 4GB. To do it, please click Tools → Set system memory

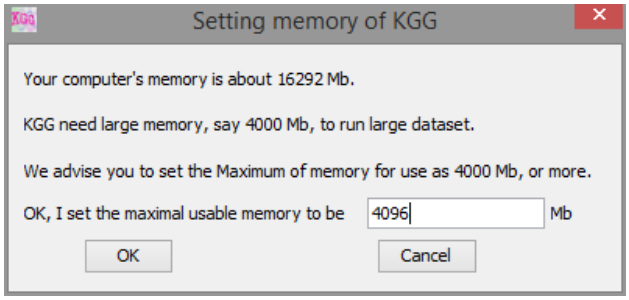


Figure 2.0 Set system memory

3. Interface and main functions

Figure 3.1 shows a typical interface of KGG with an active project.

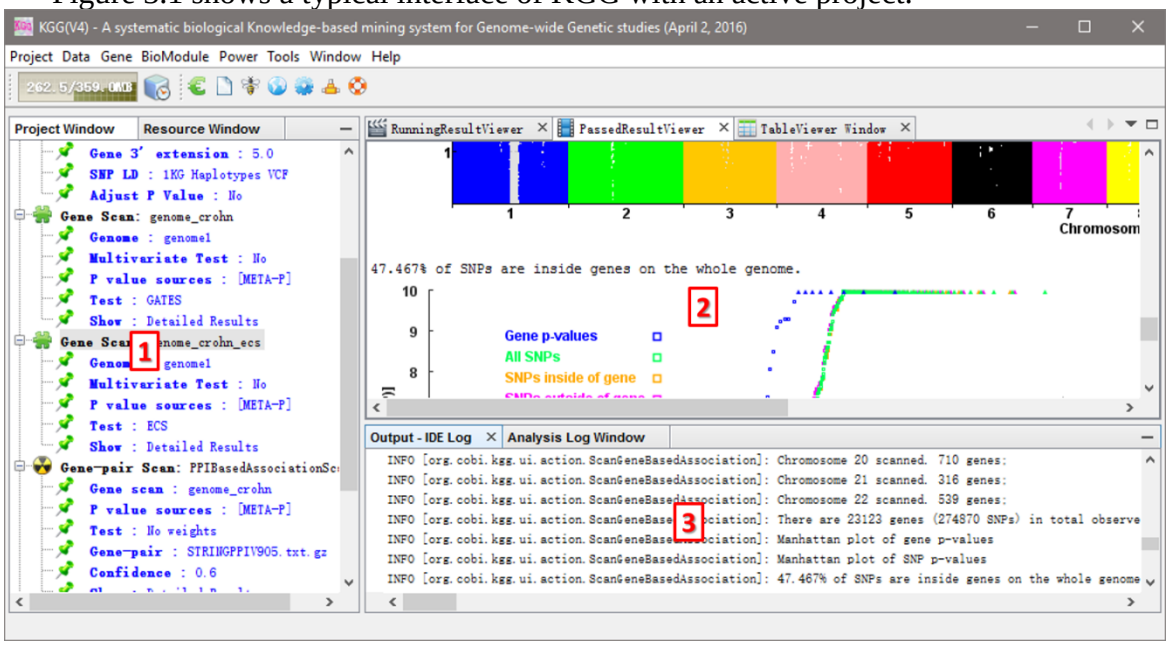


Figure 3.1 A typical KGG interface

Illustration:

Frame 1: tree-structured branches to manage input data and analysis results of a KGG project;

Frame 2: view of input data or output results;

Frame 3: running log of KGG analysis results;

The graphic dialogs of KGG are self-explanatory. Therefore, we will not elaborate the function of each buttons.

3.1 Project

- **Create Project**: create a new KGG project.
- **Open Project:** open an existing KGG project.
- **Close Project:** close the current project.
- **Exit:** exit the KGG application.

3.2 Data

- **Load P value file:** import your association summary results (e.g., the plink output).
- **Define Seed Genes:** tell KGG the known causal genes of the disease you are studying.
- **Build Analysis Genome:** build an analysis genome in which KGG maps all SNPs to their gene features and calculates the r-square or genotypic correlation of SNPs within genes.

3.4 Gene

- **Univariate Association:** conduct a univariate gene-based association scan.
- **Multivariate Association:** conduct a multivariate gene-based association scan.
- **View Genes:** view and export gene-based association results.
- **LD Plot:** view the LD pattern of variants within genes.
- **Conditional Association:** conduct a conditional gene-based association scan.

3.5 BioModule

- **Gene-pair-based Association:** conduct an association scan at gene pairs.
- **View gene-pairs:** view association p-values of gene pairs.
- **Geneset-based Association:** conduct an association scan at genesets.
- **View Geneset:** view p-values of geneset-based association analysis.

3.6 Power

- **Calculator**: SPS-a simulation tool for calculating power of set-based genetic association tests.

3.5 Tools

- **Set System Memory**: set the memory of KGG.

3.6 Window

- **Analysis Log:** show some summary results and log.
- **Project:** show the structure of the working project.
- **Resource:** show the resource that KGG contains.
- **PassedResultViewer:** show the log of a secondary analysis.
- **RunningResultViewer:** show the real-time running log when performing a secondary analysis.
- **TableViewer:** display the content of input p-value file and annotation file.
- **Output:** show the IDE results.

4. Input files

4.1 Input file 1 (GWAS results)

KGG focuses on secondary analysis of GWAS p-values. The major input of KGG is the association p-values (produced by conventional statistical genetic methods, such as PLINK) in a text file. KGG supports a user-customized format for the association p-values. Once chromosome number (or chromosome number and physical position)

and p-values columns are available in a file, you can define the column order by yourselves on KGG. The input files are allowed to include more than one p-value column. The following is an example.

Example input format of KGG:

CHR	SNPID	SNPPOS	P-value1	P-value2	P-value3	…
4	Snp1	100001	0.02301	0.8815	0.007688	…
4	Snp2	110011	0.4384	0.9575	0.006112	…
4	Snp3	120001	0.002688	0.007688	0.4893	…
4	Snp4	130011	0.01115	0.006112	0.119	…
4	Snp5	140001	0.005892	0.4893	0	…
…	…	…	…	…	…	…

Additionally, a p-value column could include values of different models. During building the analysis genome, KGG can recognize this format ?(which format? The follow one?) with the input format "multiple tests per column".

Example a more complex input format of KGG:

CHR	SNP	P-value1	Test-Mode	P-value2	…
4	rs1513559	0.02301	additive	0.007688	…
4	rs1513559	0.4384	recessive	0.006112	…
4	rs1513559	0.002688	dominant	0.4893	…
4	rs1841043	0.01115	additive	0.119	…
4	rs1841043	0.005892	recessive	0	…
…	…	…	…	…	…

4.2 Input file 2 (Candidate Gene list)

Candidate genes could be loaded one by one or imported from a TXT file. The input file has only one column without header, in which one row contains only one gene (symbol or ID).

5. Tutorial of knowledge-based secondary association analysis

We use a real dataset of Crohn's disease (available at http://grass.cgs.hku.hk/limx/kgg/download/KGGSample.zip) as an example to demonstrate how to use KGG for a series of knowledge-based secondary association analysis of conventional p-values from GWAS. This dataset was originally downloaded from a public domain released by (Barrett, et al., 2008) and have SNP ID conversion by SNPTracker (http://grass.cgs.hku.hk/limx/snptracker/) for coordinates of Hg19. It includes 7 columns, as CHR, SNP, POS, RISK, NONRISK, META-Z and META-P. The effective input data in the input summary statistics file are chromosome (CHR), coordinate (POS) and variants' p-values (META-P). The main analysis procedure is illustrated in Figure 5.0.

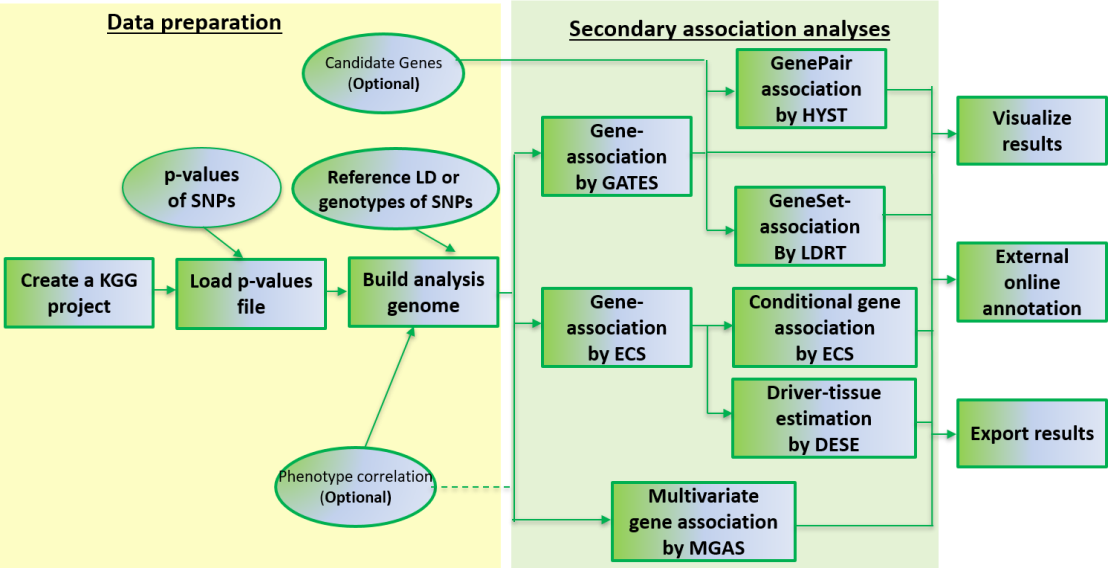


Figure 5.0 Pipeline chart of KGG analysis (version 4)

Notes: Circle nodes stand for data and files (input, output), rectangles denote an analytical procedure, a dashed line stand for virtual relationship between a dataset and an analysis.

5.1: Data preparation

To create a new project, please click the menu Project → Create Project, with a name 'CrohnDisease', and set the project path at C:\KGG (or other path defined by users).

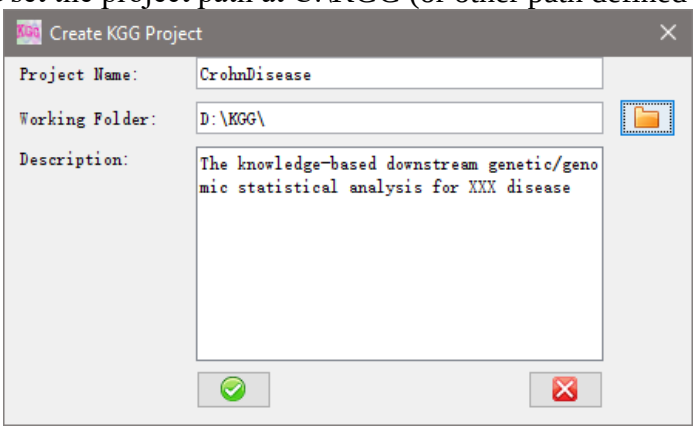
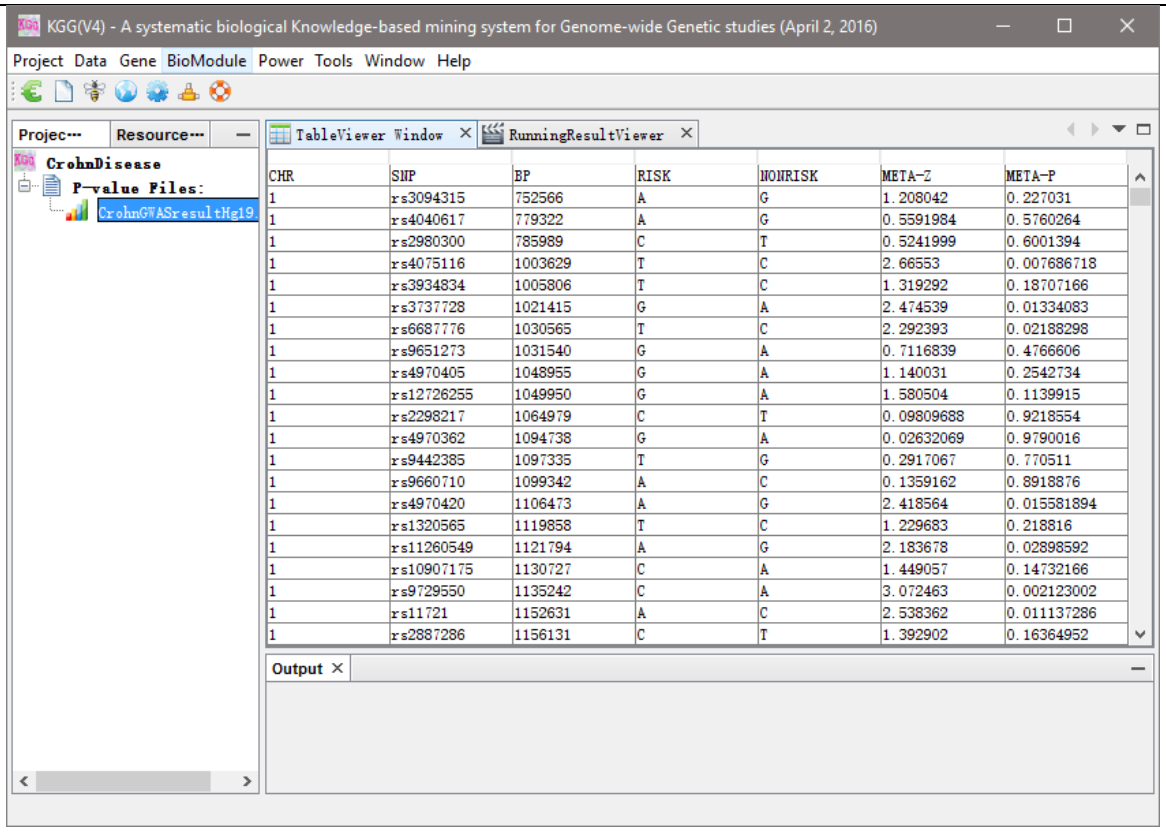


Figure 5.1.1 Dialog of creating project

Load the p-value file into the project. Please select the menu Data →Load P Value File and choose the file 'CrohnGWASresultHg19.txt' containing whole-genome association p-values for Crohn disease in KGGSample folder.



CHR	SNP	BP	RISK	NONRISK	META-Z	META-P
1	rs3094315	752566	A	G	1.208042	0.227031
1	rs4040617	779322	A	G	0.5591984	0.5760264
1	rs2980300	785989	C	T	0.5241999	0.6001394
1	rs4075116	1003629	T	C	2.66553	0.007686718
1	rs3934834	1005806	T	C	1.319292	0.18707166
1	rs3737728	1021415	G	A	2.474539	0.01334083
1	rs6687776	1030565	T	C	2.292393	0.02188298
1	rs9651273	1031540	G	A	0.7116839	0.4766606
1	rs4970405	1048955	G	A	1.140031	0.2542734
1	rs12726255	1049950	G	A	1.580504	0.1139915
1	rs2298217	1064979	C	T	0.09809688	0.9218554
1	rs4970362	1094738	G	A	0.02632069	0.9790016
1	rs9442385	1097335	T	G	0.2917067	0.770511
1	rs9660710	1099342	A	C	0.1359162	0.8918876
1	rs4970420	1106473	A	G	2.418564	0.015581894
1	rs1320565	1119858	T	C	1.229683	0.218816
1	rs11260549	1121794	A	G	2.183678	0.02898592
1	rs10907175	1130727	C	A	1.449057	0.14732166
1	rs9729550	1135242	C	A	3.072463	0.002123002
1	rs11721	1152631	A	C	2.538362	0.011137286
1	rs2887286	1156131	C	T	1.392902	0.16364952

Figure 5.1.2 Input GWAS original result file

Define a number of candidate genes. Click the menu Data →Define Seed Genes to import file 'CrohnCandidateGeneSet.txt' as input of candidate genes. Define all genes as seed genes and save them as candidategeneset_crohn. Note: this step is optional and the seed genes will be only used to highlight gene pairs and gene sets.

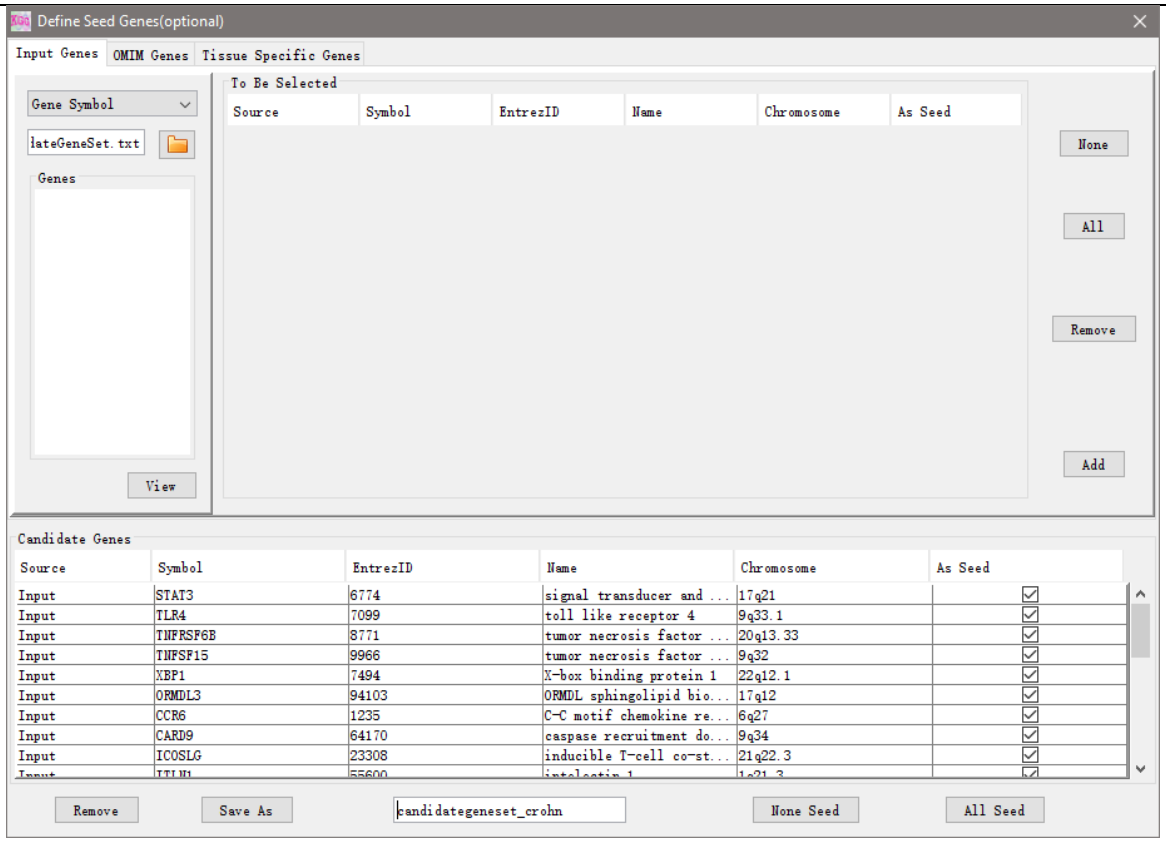


Figure 5.1.3 Input candidate gene set for crohn's disease

Build an analysis genome. Click the Menu Data → Build Analysis Genome.

Download ancestry matched genotypes in 1000 Genomes Project to adjust LD between variants. Click the 'Download' hyperlink on the Dialog to go to a web-page for downloading the genotypes at http://grass.cgs.hku.hk/limx/kgg/phasedgty.html. Choose 'EUR(495 subjects)' and click the hyperlink 1kg.phase3.v5.shapeit2.eur.hg19.tar.gz . Unpack the data into 23 compressed VCF files corresponding to 23 chromosomes. Click to load all the "vcf.gz"files on the "Build Analysis Genome" dialog.

Select META-P for building analysis genome with the default setting and clink the button Build to build analysis genome. It will take a round 15 miniutes on a notebook computer with a CPU with 2.0GHz .

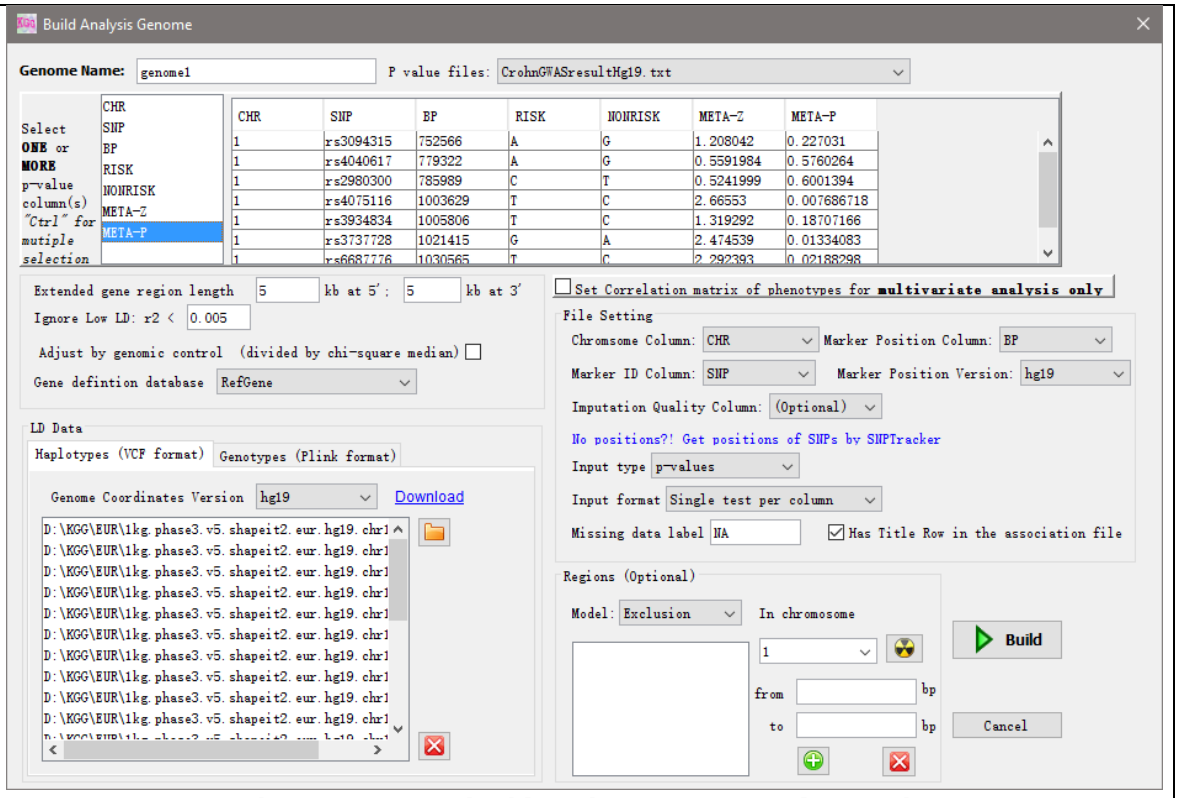


Figure 5.1.4 Select META-P to build analysis genome and name the genome as genome_crohn

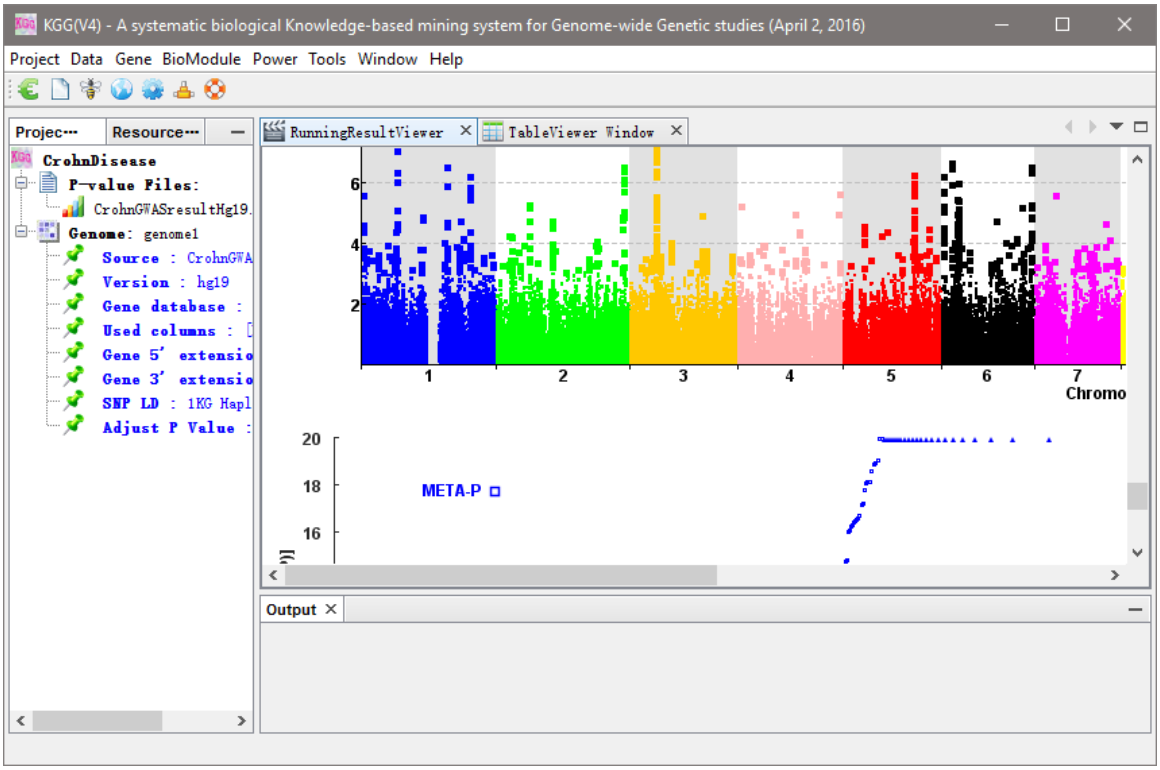


Figure 5.1.5 The display after building analysis genome

5.2 Secondary knowledge-based association analysis

5.2.1 Gene-based association analysis by GATES

Click Gene →Univariate Association to set the parameters as Figure 5.5.1. Set the Scan name as 'genome_crohn', and select SNP p-values to integrate the analysis genome, then choose 'Extended Simes test (GATES, more powerful for a gene with one or a few independent causal variants' method. It should be noted that exported Manhattan plots and QQ plots will be shown in "Running Result Viewer Window" (Figure 5.5.2).

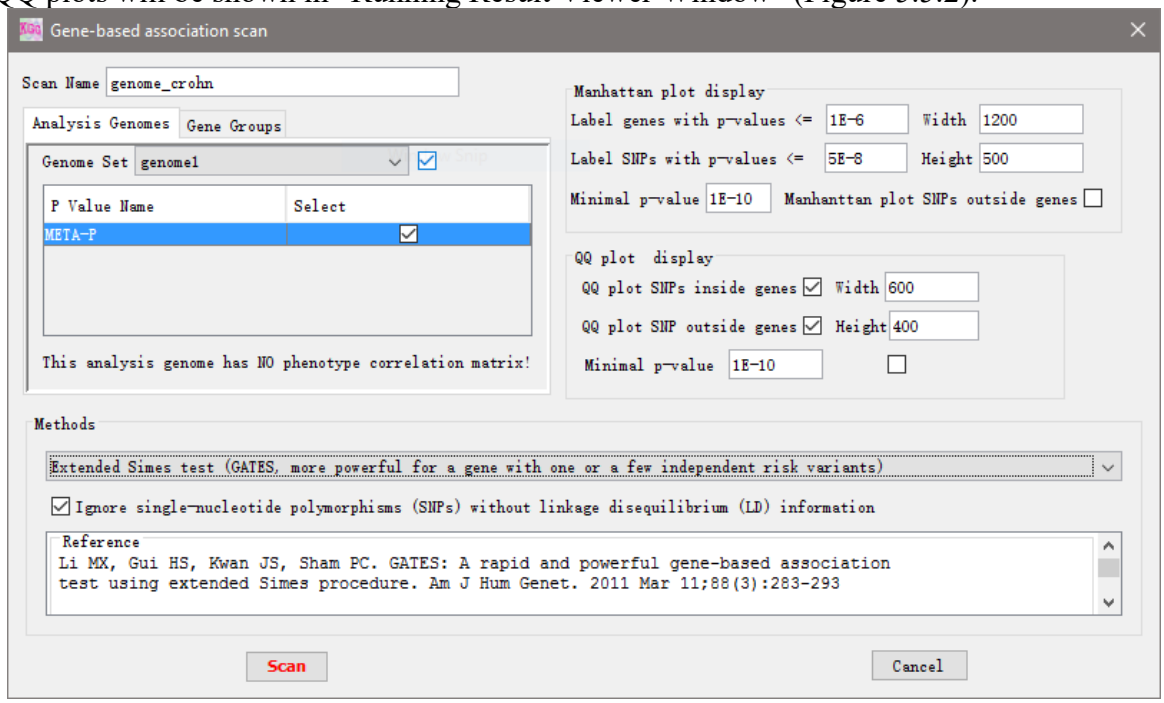


Figure 5.2.1.1 Setting for gene-based scans

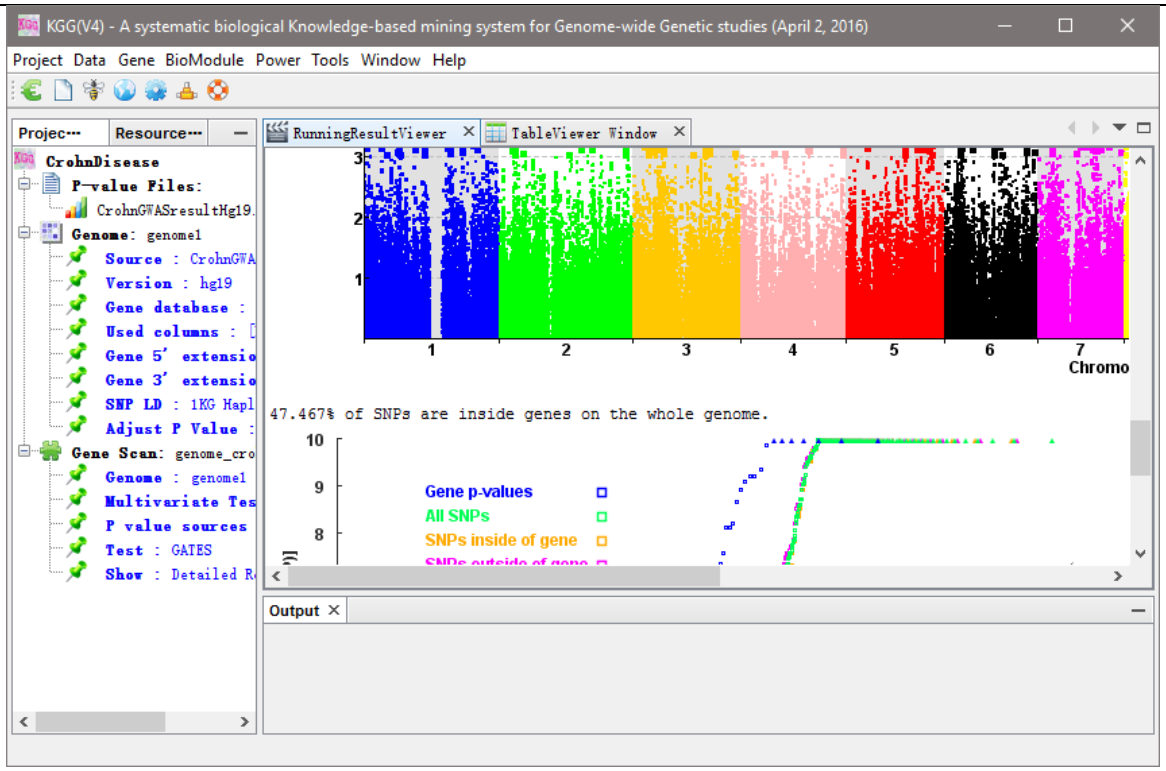


Figure 5.2.1.2 The display after gene-based scan

If you want to view the detail results, please click the "Show: Detailed Results" node under "Genome Scan" in the left frame. The new tab named "ViewGenes" will be created to provide you more information about the result (Figure 5.5.3). You can also export the results you want in this tab.

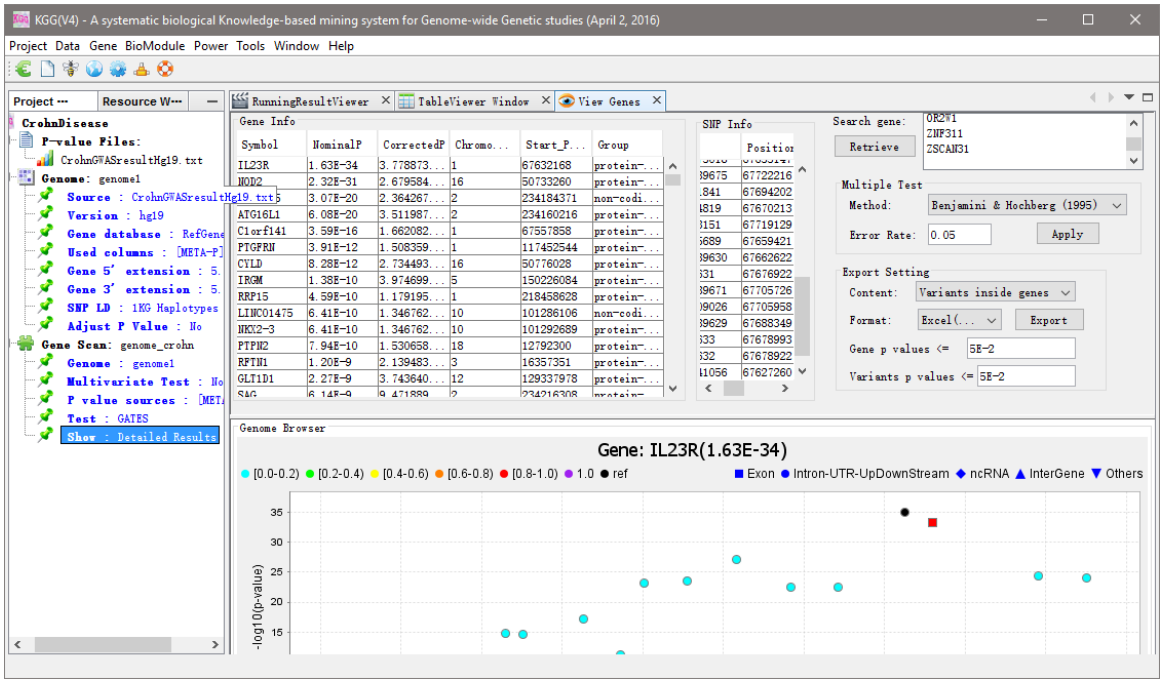


Figure 5.2.1.3 Function of displaying the gene-based association scan result

5.2.2 Gene-based association analysis by ECS

Similarly, you can use another approach, an effective chi-squared test (ECS), to perform the gene-based association analysis. Effective chi-squared test (ECS, more powerful for a gene with multiple dense independent risk variants). The ECS is more powerful than GATES when a gene has multiple dense independent risk variants. ECS also lends itself for a conditional gene-based association analysis.

5.2.3 Conditional associational analysis by ECS on significant or interested genes

Click the menu Gene → Conditional Associational to set the parameters as Figure 5.6.1. Click Load Genes and choose 'Method: Benjamini & Hochberg (1995)' to selected significant genes. By default, genes within 5 MB form a group. After the analysis, the significance of the conditional associational analysis will be shown in the column "ISScore", which is a p-value-like measurement of the significance.

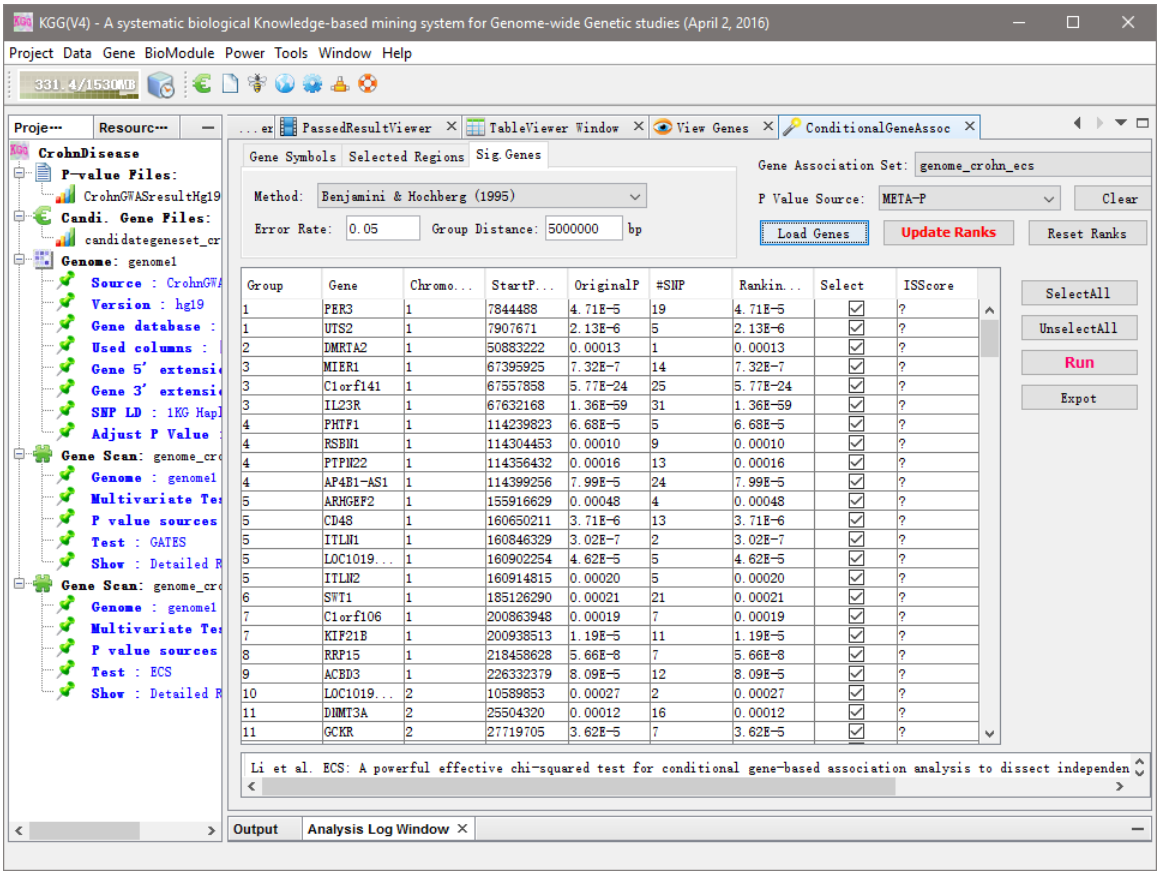


Figure 5.2.3.1 Setting for conditional associational analysis

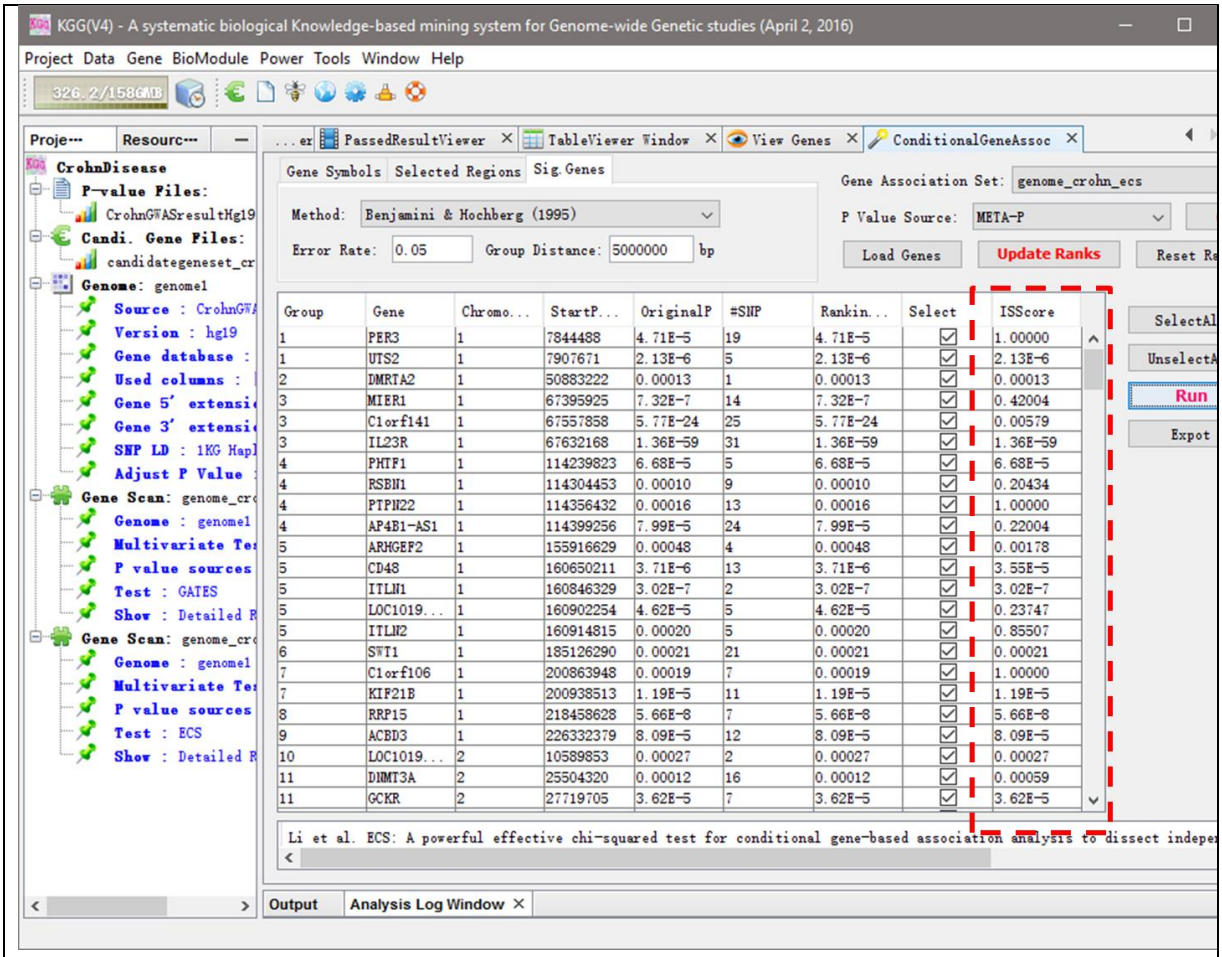


Figure 5.2.3.2 The display after conditional associational analysis

5.2.4 Gene-pair-based association analysis by HYST

Click the menu BioModule → Gene-pair-based Association to set the parameters as Figure 5.7.1. Note: the gene-pair based association analysis should use gene-based association analysis results by GATES as an input. A QQ plot of the gene-pair-based association p-value will be shown at the end of the analysis (Figure 5.7.2).

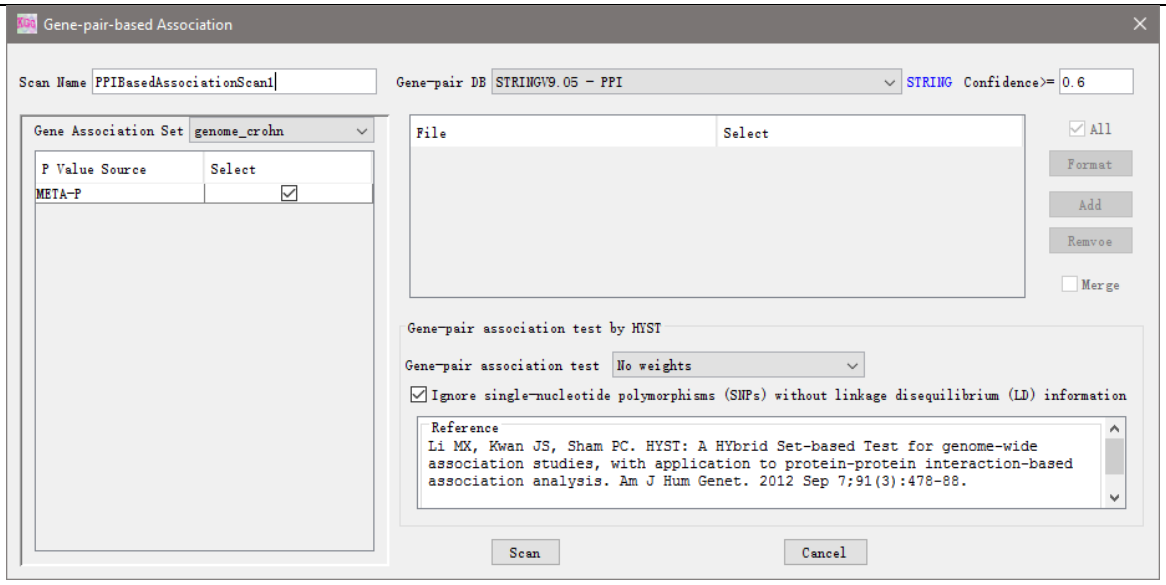


Figure 5.2.4.1 PPI association scan by gene-based p-values

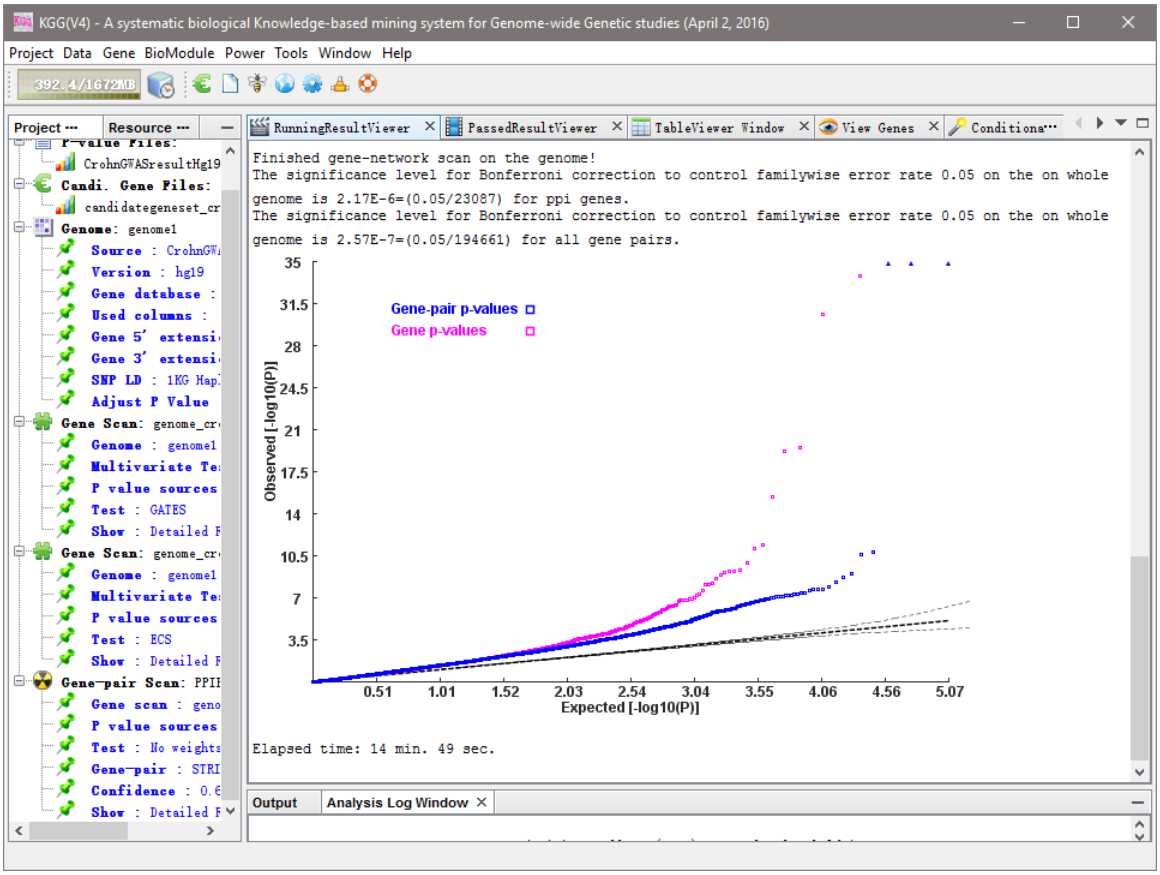


Figure 5.2.4.2 The display after running PPI-based association scan

Click the node "Show: Detailed Results" under 'Gene-pair scan', and you will get the graph of gene pairs. You can also export the results you want in this tab.

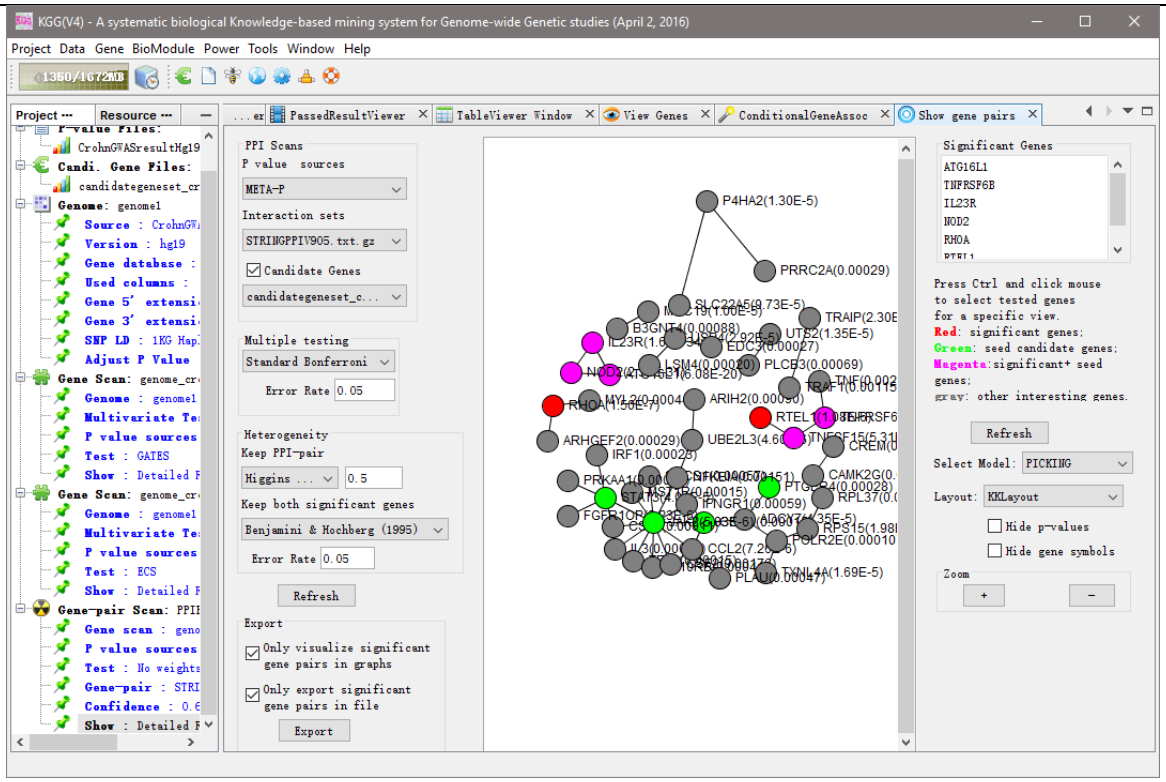


Figure 5.2.4.3 Function of displaying the results of PPI-based association scan

5.2.5 Multivariate gene-based association analysis by MGAS

The multivariate gene-based association analysis is different from the above knowledge-based secondary association analyses that are designed for multivariate analysis. Therefore, in the example dataset (available at http://grass.cgs.hku.hk/limx/kgg/download/KGGSample.zip), we prepared another real example from a published paper [Nat Genet. 2009 Jan;41(1):35-46] to demonstrate the analysis. In the KGGSample\MultiPhenos folder, there are two files, 9MetabolicPhenotypesPhg19.txt and 9MetabolicPhenotypesCorr.txt, which contains p-values and Pearson correlation of 9 quantitative metabolic traits respectively. Similarly, you should load the p-value file in to the KGG project at first.

Compare to the univariate analysis, there is one unique setting for the multivariate analysis when the analysis genome is built. The Pearson correlation should be specified by clicking "Set Correlation matrix of phenotypes for multivariate analysis only" in the "Build Analysis Genome" Dialog (Figure 5.2.5.1).

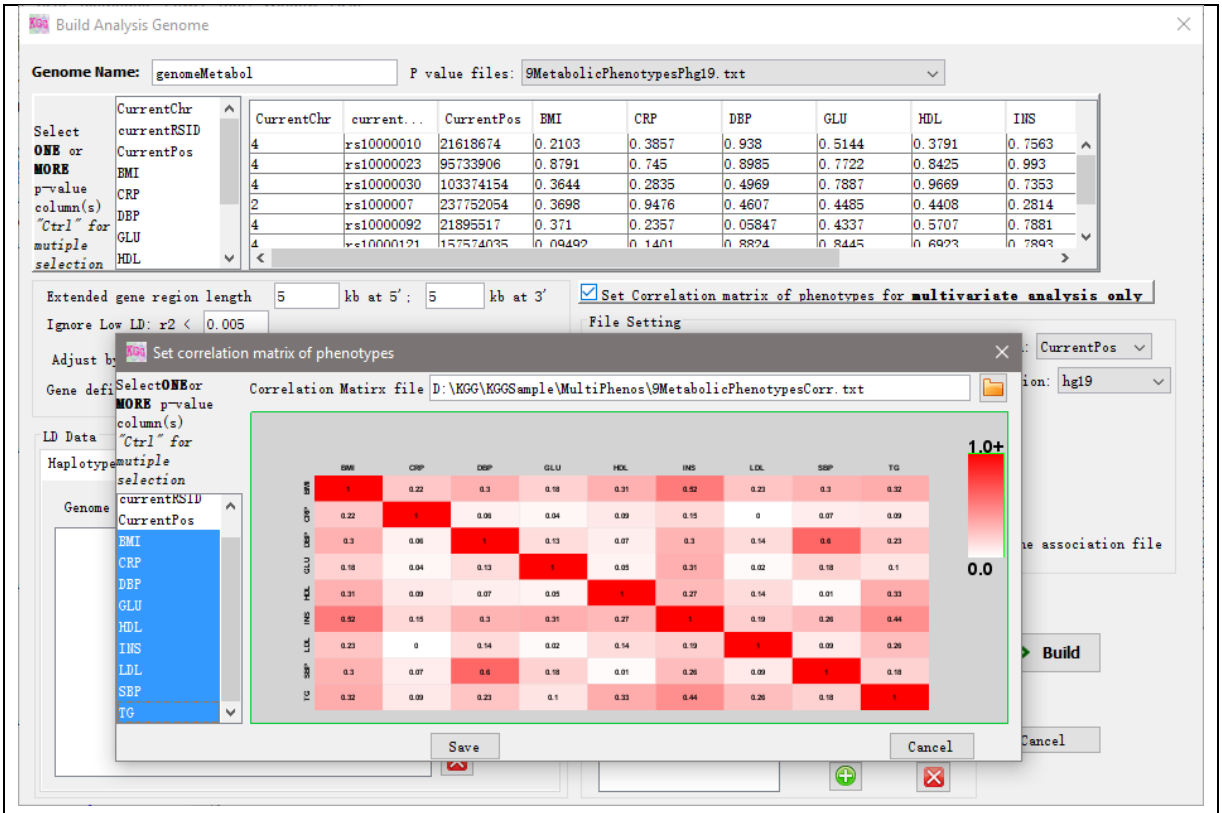


Figure 5.2.5.1 Specify the correlation of phenotypes analysis

Click the menu Gene → Multivariate Association to set the parameters as Figure 5.9.2. The QQ plot will be shown in the RunningResultViewer (Figure 5.9.3). The detailed gene-based p-values can be viewed by clicking the node "Show Detailed Result" (Figure 5.9.4).

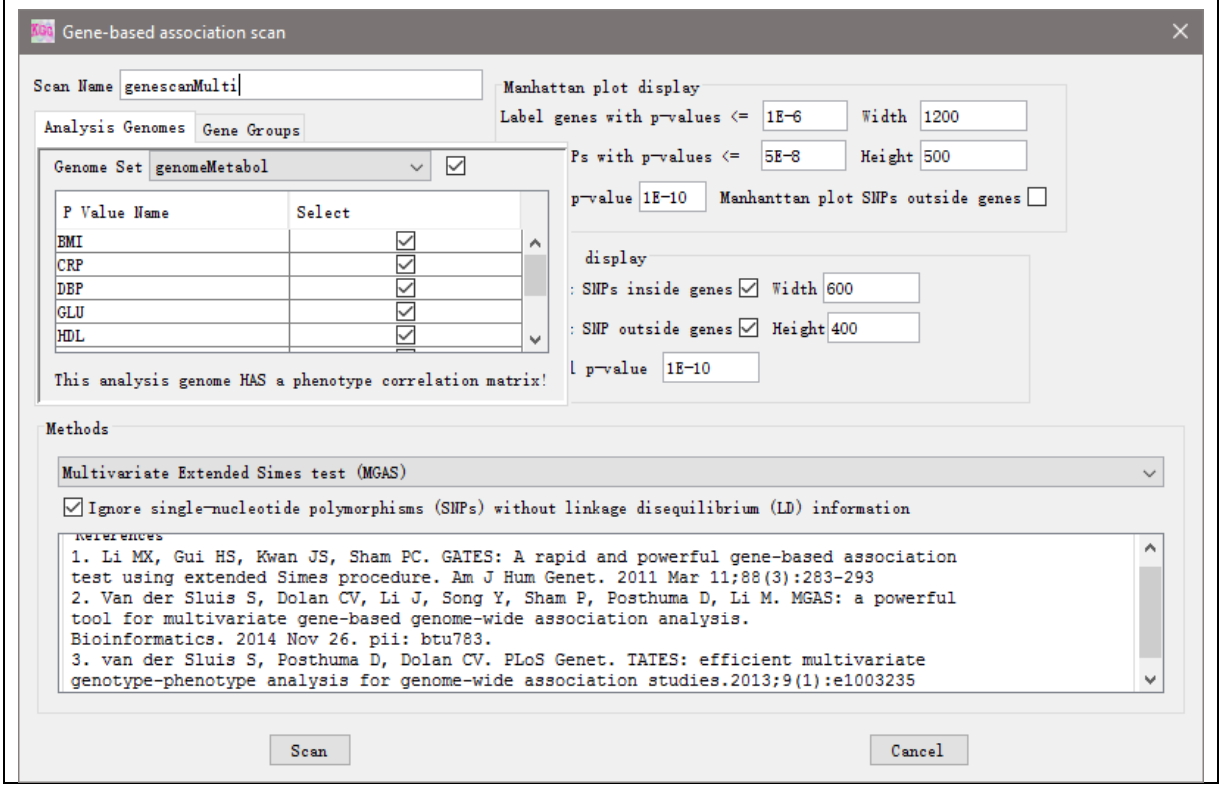


Figure 5.2.5.2 Multivariate gene-based association analysis

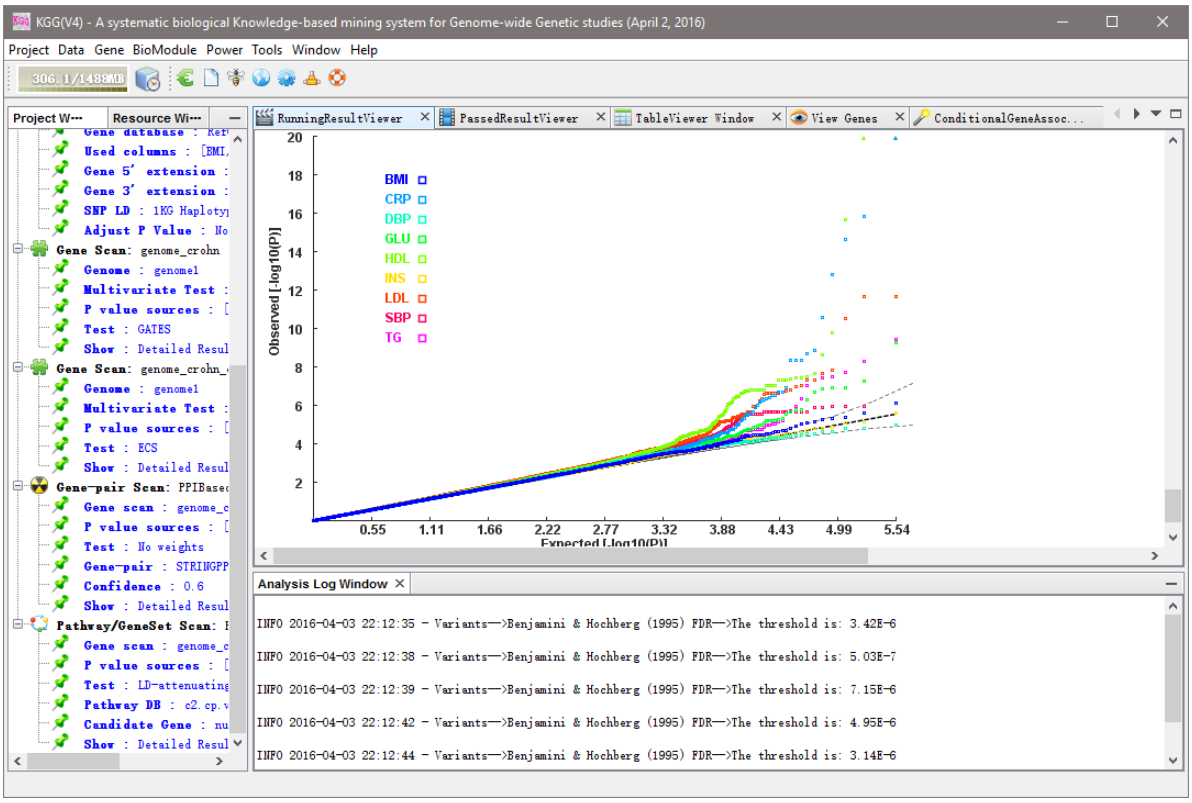


Figure 5.2.5.3 A viewer showing the p-value QQ plot of the multivariate gene-based association analysis

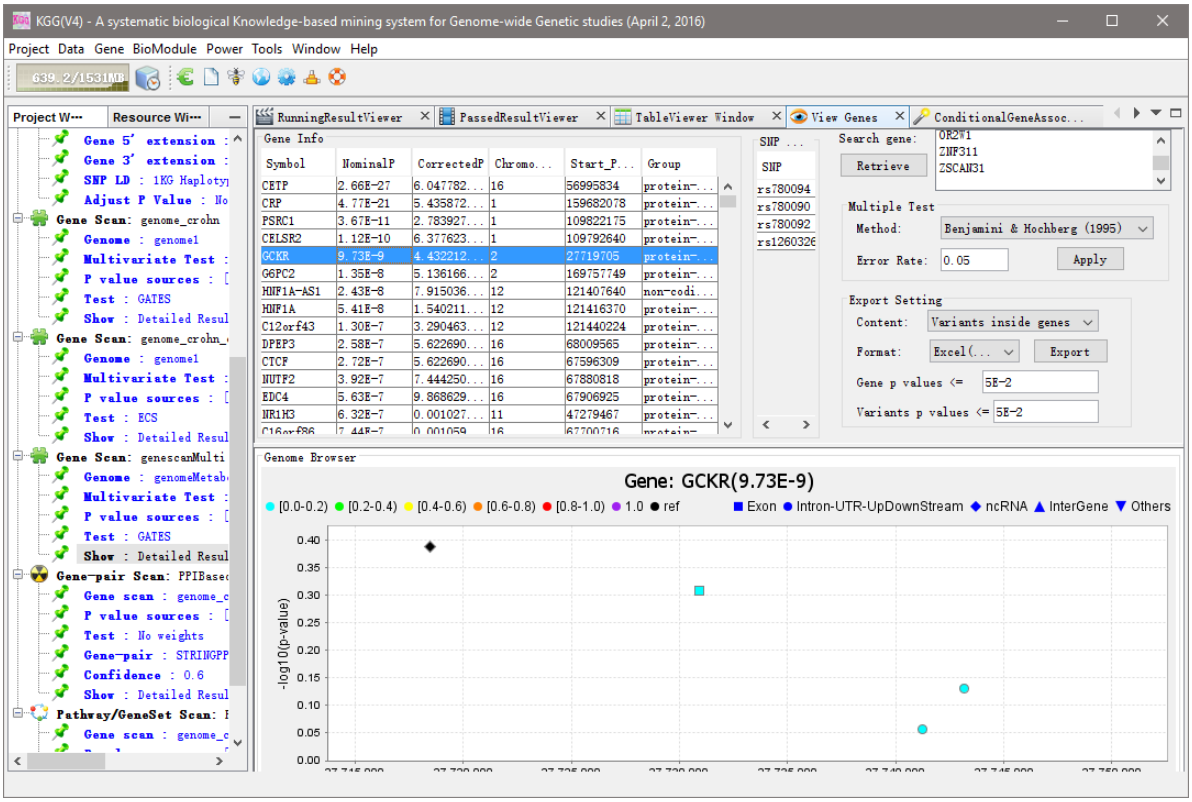


Figure 5.2.5.4 P-values of the multivariate gene-based association analysis

5.2.6 Estimating driver-tissues by selective expression of genes associated with complex diseases or traits

This function is designed to estimate driver tissues by tissue-selective expression of phenotype-associated genes in GWAS. The tissues in which causal or susceptibility genes initiate the phenotypes are called driver or causal tissues.

For this function, KGG requires two types of inputs data, gene expression values of multiple tissues and GWAS summary statistics or association p-values at variants for a tested disease. The expression values at genes and transcripts or even exons can be used for the estimation. The GWAS p-values are used to detect susceptibility genes by a conditional gene-based association test (See 5.2.3).

Click the menu Gene → Driver Tissue (DESE) to set the parameters and input specify an expression file, as Figure 5.2.6.1. The estimating driver-tissues analysis should use gene-based association analysis results by ECS, and input an expression file which contains the expression values of each genes in every tissue. Please download the expression file from the address: http://grass.cgs.hku.hk/limx/rez/. Then, click the button Load Genes to load significant phenotype-associated genes according the threshold of multiple testing (Figure 5.2.6.1). Next, click the button Run to estimate driver-tissues. It will take you two or more hours.

The estimated driver-tissues will be prioritized according to their statistical significance. Four types selective-expression measures (robust-regression z-score, conventional z-score, MAD robust z-score, and ratio of vector-scalar projection) are used in the estimation analysis. A combined prioritization is generated by averaging the -log10(p) based on the four measures (as shown in Figure 5.2.6.2 and Figure 5.2.6.3).

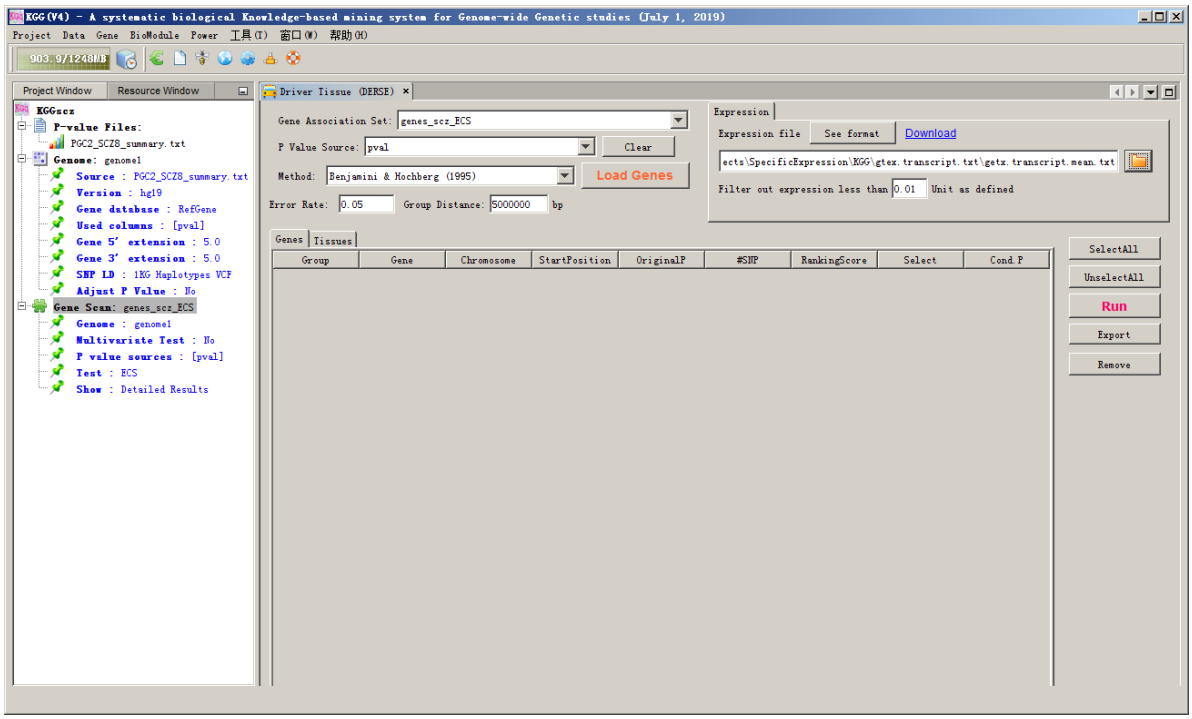


Figure 5.2.6.1 Setting for estimating driver-tissues analysis

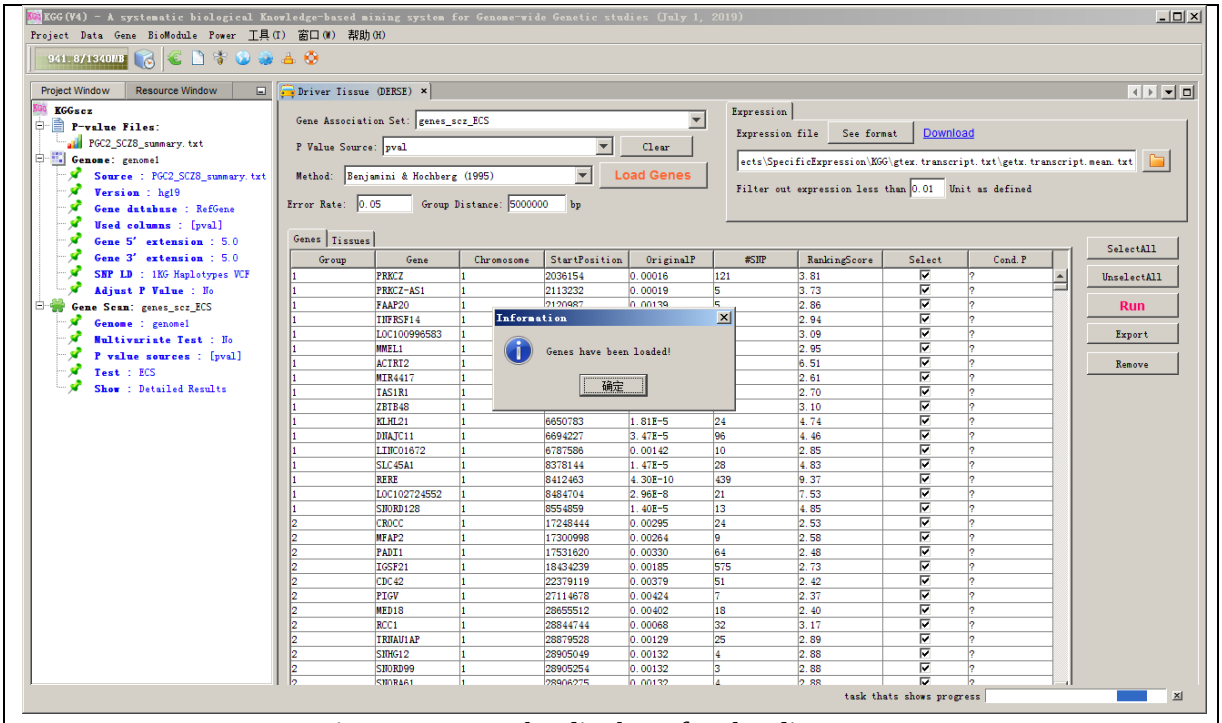


Figure 5.2.6.2 The display after loading genes

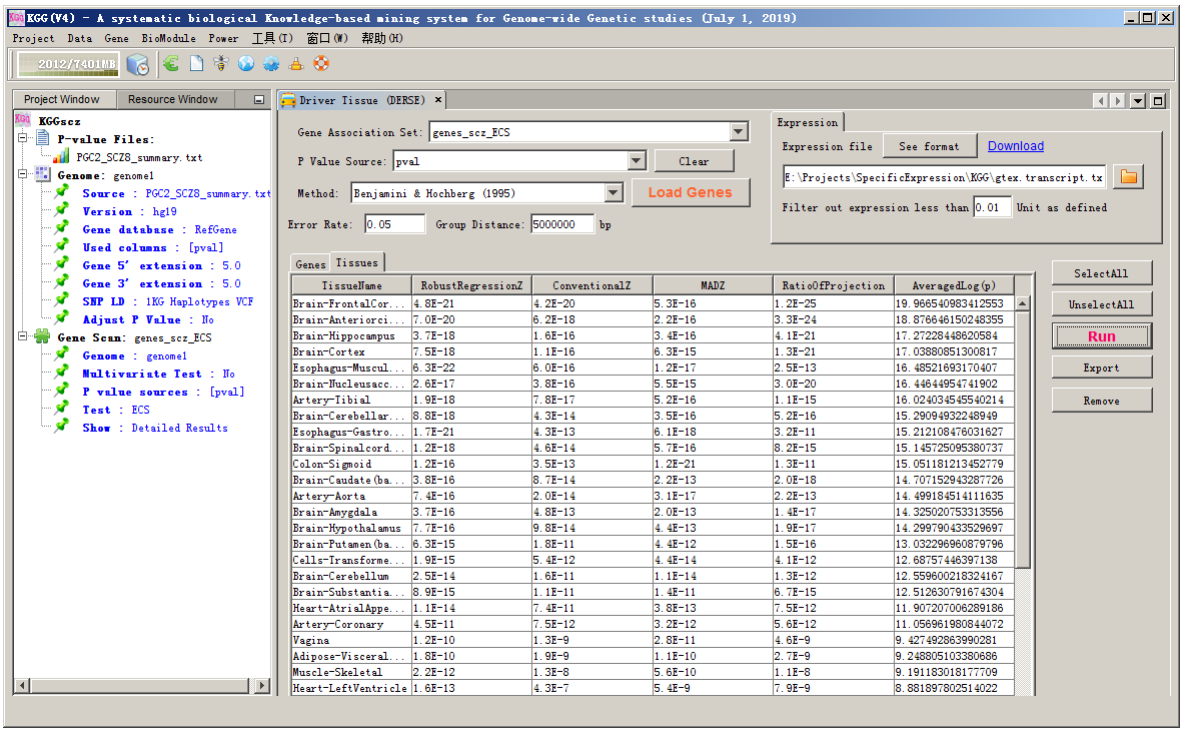


Figure 5.2.6.3 The display after estimating driver-tissues analysis

6. Power estimation of set-based tests by SPS

STEP 1: Power estimation, Power → calculator. The interface is divided into two parts. Set the basic parameters on the left, then you can get the results on the right.

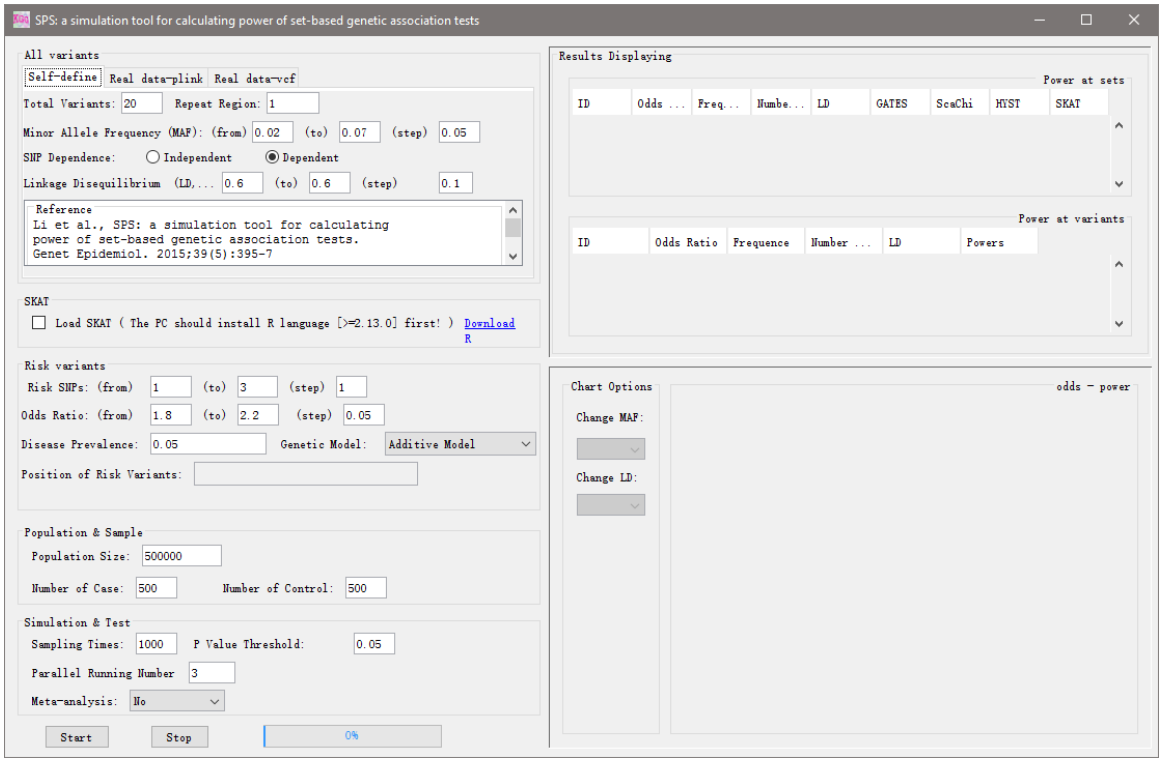


Figure 6.1. The Main interface of SPS.

STEP 2: Set the parameters of all variants, including the number of SNPs, the minor allele frequency (MAF) and LD information. If these SNP markers are divided into several LD blocks, the markers in the same LD block have the same LD with each other. But the LD is set to 0 when the markers belong to different blocks. All of these markers and their LD pattern can be replicated to make up of a larger marker set. Some of these parameters can vary in a certain region, such as MAF and LD, so that the users can investigate how powerful will be affected by changing the critical parameters conveniently. In addition, these parameters can also read from the real data (Plink binary genotype files and vcf file). In this case, the LD information will be calculated from the input genotypes.

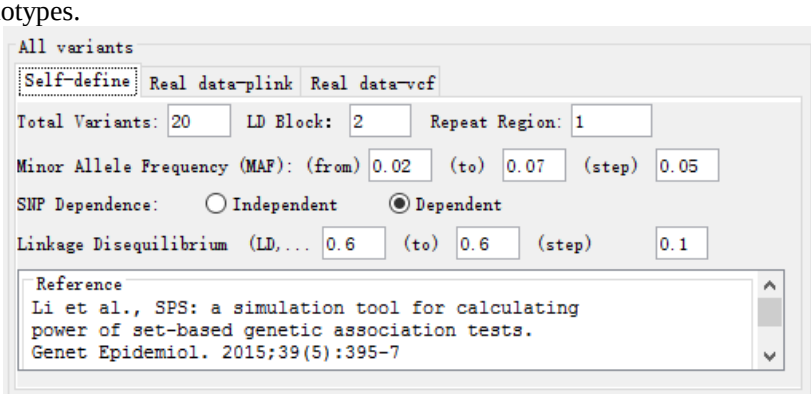


Figure 6.2.1 Set parameters by users.

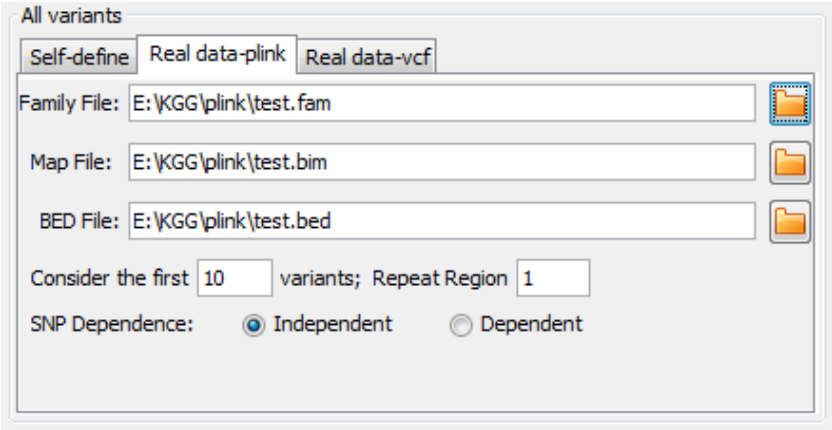


Figure 6.2.2 Set parameters by plink file.

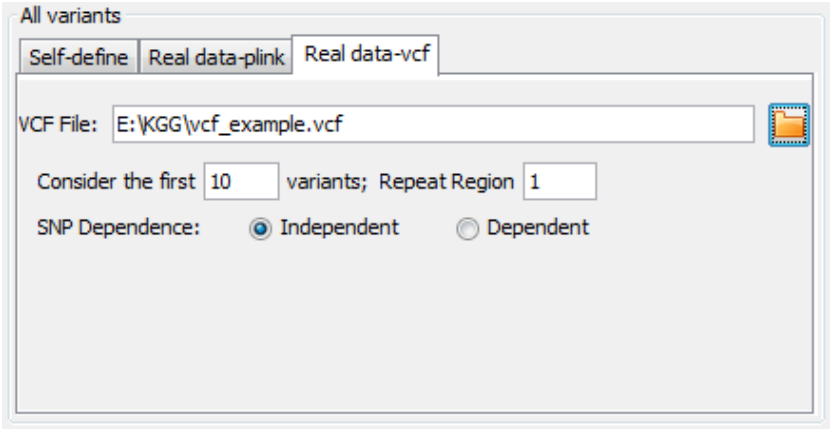


Figure 6.2.3 Set parameters by vcf file.

Table to list parameters:

Parameter	**Description**
Total Variants	The total number of SNPs tested in a set
LD Block	The number of LD blocks. Variants in the same block are in LD and that in different blocks have no LD.
Repeat Region	The number of copies of SNPs. The SNP will be copied for several times to form a larger set and so does the LD pattern of the.
Minor Allele Frequency	The frequency of the least common allele occurs in the population. The MAF can increase from a initial value to a terminal value according to a step value that set from the GUI.
SNP Dependence	The relationship between SNPs. If the SNPs are dependent, the user should set the LD value (r), otherwise 0 is set as default. The LD information can also be read from the real data, where it will be calculated based on the allele frequency.
Linkage Disequilibrium (LD, r)	The r score used to represent LD information. The SNPs in the same block are dependent and keep the same r value, while SNPs in the different blocks are independent with each other and the r value is set as 0. The r value can also increase from an initial value to a final value by a step value.
Family File Map File BED File	The path of the Plink files. The valid file path can be input by the button on the right. If the three files have the same file prefix and are stored in the same directory, the other file paths will be filled automatically when one file is set.
Consider the first several SNPs	The number of SNP that input from the real data. The real data usually include large size of SNPs, which is unnecessary for our simulation. Hence, we just consider the first several SNPs as our

	study objects.
VCF File	The path of a VCF file.

STEP 3: Set parameters of risk variants.

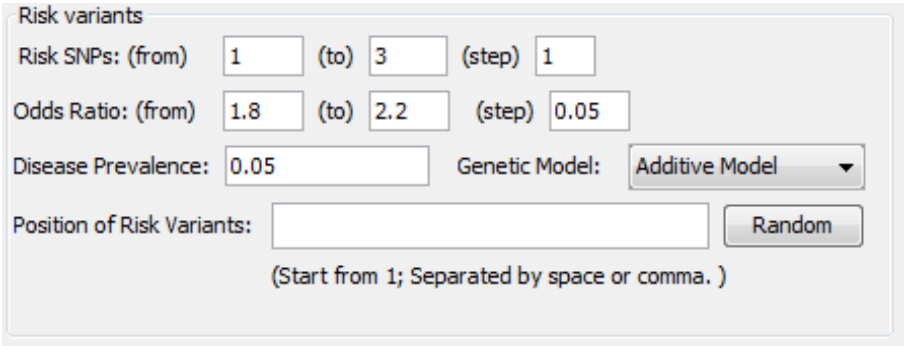


Figure 6.3 Set parameters about risk variants.

Table to list parameters:

Parameter	Description
Risk SNPs	The number of risk SNPs. This parameter can increase from a smaller to a larger value step by step.
Odds Ratio	The value used to quantify the association between risk SNPs and disease. This parameter can increase from a smaller to a larger value step by step.
Disease Prevalence	The proportion of a population found to suffer the disease. This will be used in the genetic model.
Genetic Model	The genetic model of risk loci. The additive model and multiplicative model are candidates in SPS.
Position of Risk Variants	The location information of risk variants within the total variants. The users can click the random button for automatic setting or set by themselves.

STEP 4: Set population and sample. The larger population size and number of case and control are recommended, because they make the result more accurate and stable, but it will take more time correspondingly. So the user should keep balance between them.

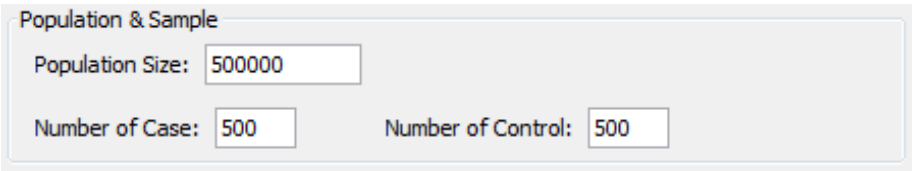


Figure 6.4. Set population and sample.

Table to list parameters:

Parameter	Description
Population Size	The number of individuals in a population generated by simulation according to the certain genotype and phenotype.
Number of Case	The number of individuals that suffer the disease.
Number of Control	The number of individuals that do not suffer the disease.

STEP 5: Set simulation and meta-analysis parameters. A number of case-control samples will be randomly drawn with replacement from the population. And they are subject to calculate the *p* value of the set-based test. The number of *p* values passing the threshold will be counted to calculate the power. In order to speed up the simulation process, you can set several parallel

threads, but more memory resource is needed.

The meta-analysis can be carried out at the variant level or set level. At variant level, the p values of variants in different studies will be combined according Fisher's Combination Test and these meta-p values will be treated by GATES, ScaChi and HYTS. Alternatively, at set level, the p value of variants in a set should be conducted by GATES, ScaChi and HYTS, and then the set-based p values in different studies are aggregated. SPS can also mimic locus heterogeneity by randomizing risk loci of each study in meta-analysis.

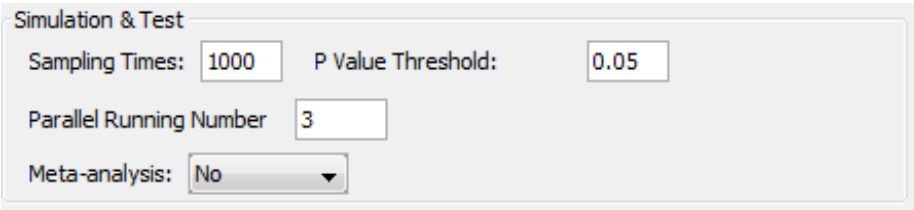


Figure 6.5.1 Set simulation without meta-analysis.

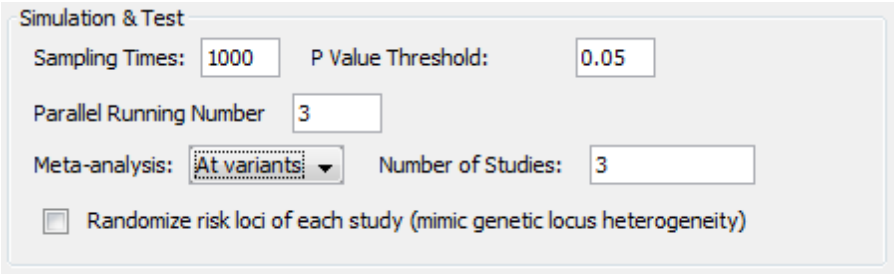


Figure 6.5.2 Set simulation with meta-analysis.

Table to list parameters:

Parameter	**Description**
Sampling Times	The number samples randomly drawn from the case and control group. For each time, a case-control study is achieved.
P Value Threshold	The threshold of type I error that used in the case-control study. For SNP-based test, the bonferroni correction is conducted as default.
Parallel Running Number	The number of threads that running concurrently. The multiple threads mechanism is used here to speed up the running of program. However, this may cost a large volume of memory.
Meta-analysis	Whether to perform meta-analysis. If performed, the users should choose the meta-analysis at variants level or at set level.
Number of Studies	The number of studies considered in the meta-analysis.
Randomize risk loci of each study	Whether to consider the genetic heterogeneity. If considered, the position of risk loci of each study will be set randomly to mimic the heterogeneity.

STEP 6: Run the program. Click the Start button and run the program. The results from tables are shown in the right part immediately. The progress bar provides the real time information of running. If you want to stop the running program, just click the "stop" button.

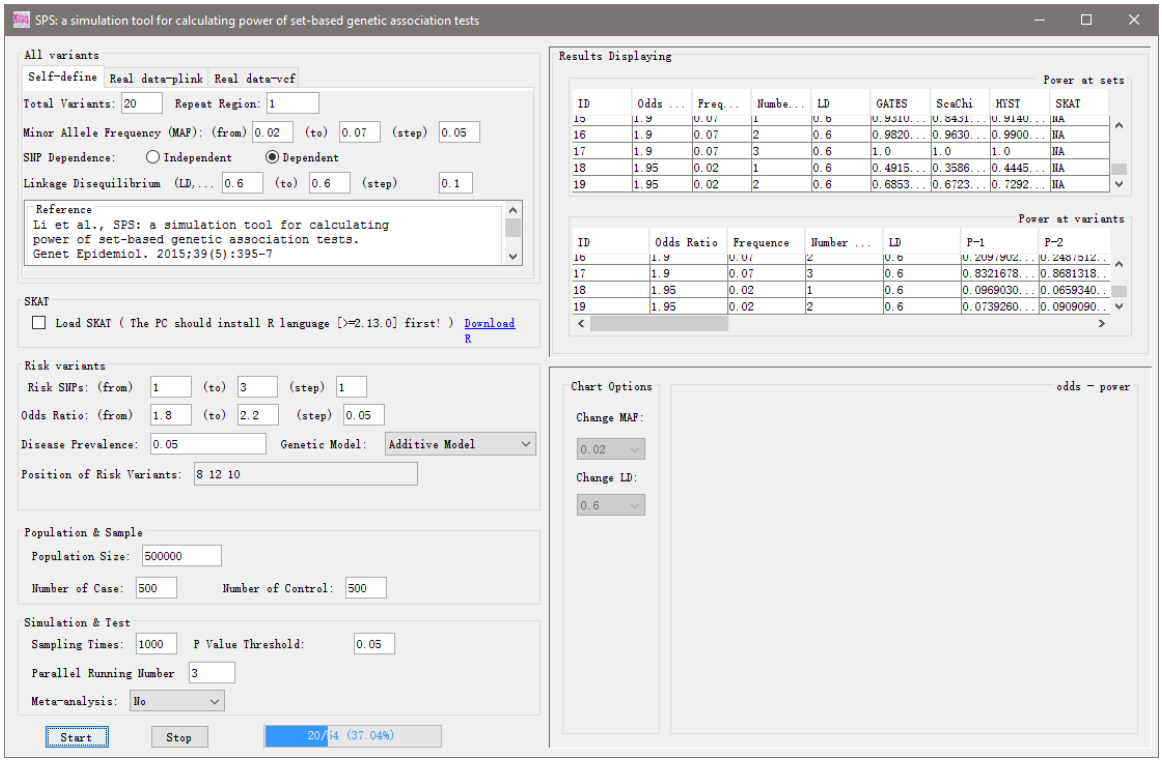


Figure 6.6 Run the program.

STEP 7: Save the result. Users can review the power from tables at the SNP level and set level. A line chart is drawn to show the variation of power within different odd ratios with the given MAF and LD information. You can also change the MAF and LD values to update the chart, and right-click on the tables to save the results as excel files or txt files. The chart can be saved by right-click as well.

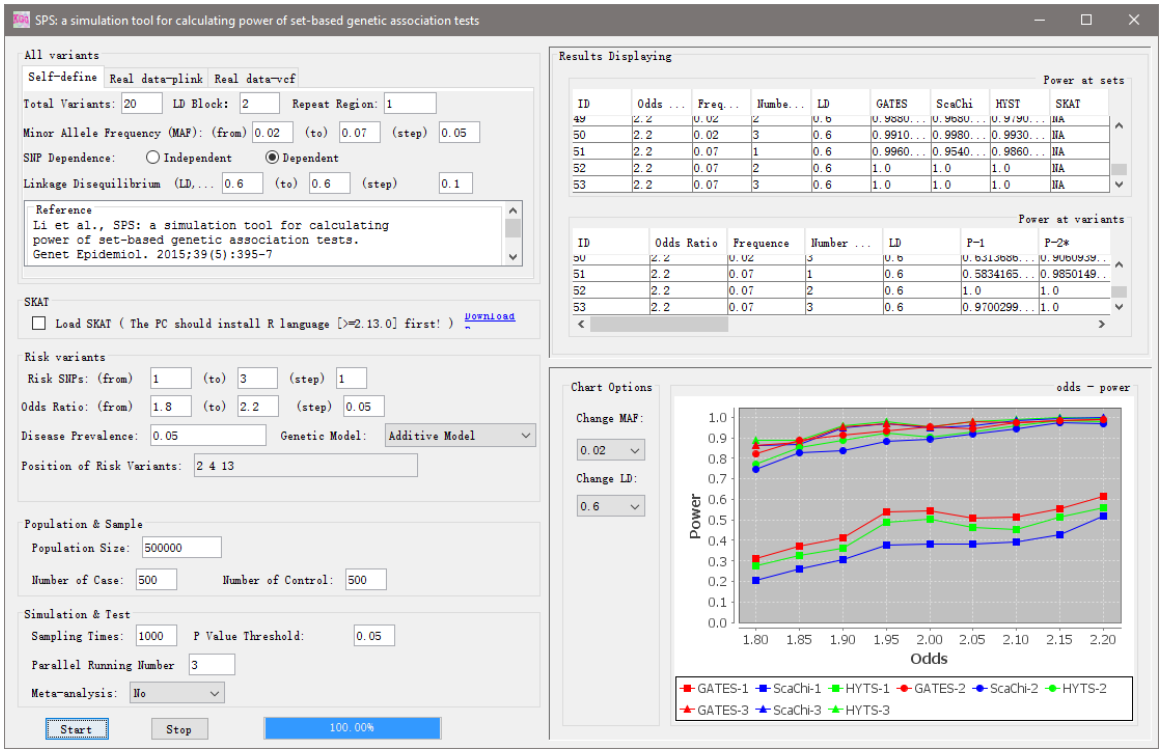


Figure 6.7.1 The output of SPS.

ID	Odds Ratio	Frequence	er of Risk A	LD	GATES	ScaChi	HYST
0	1	0.1	1	0.5	0.055	0.025	0.048
1	1	0.1	1	0.6	0.04	0.016	0.033
2	1	0.1	2	0.5	0.042	0.025	0.041
3	1	0.1	2	0.6	0.047	0.021	0.047
4	1	0.1	3	0.5	0.049	0.017	0.048
5	1	0.1	3	0.6	0.049	0.029	0.044
6	1	0.1	4	0.5	0.044	0.028	0.049
7	1	0.1	4	0.6	0.062	0.027	0.055

Figure 6.7.2 The saved table of set-based power.

ID	Odds Ratio	Frequence	er of Risk A	LD	P-1	P-2	P-3*	P-4	P-5	P-6	P-7*
0	1	0.1	1	0.5	0.002	0.004	0.005	0.001	0.001	0.002	0.001
1	1	0.1	1	0.6	0.003	0.001	0	0.001	0	0.003	0.001
2	1	0.1	2	0.5	0.003	0.003	0.001	0.002	0.001	0.001	0.002
3	1	0.1	2	0.6	0.003	0.006	0.005	0.002	0.002	0.001	0.004
4	1	0.1	3	0.5	0.004	0.002	0.001	0.003	0.003	0	0.001
5	1	0.1	3	0.6	0.003	0.002	0.003	0.001	0.002	0.001	0.004
6	1	0.1	4	0.5	0.003	0.004	0.003	0.004	0.003	0.004	0.004
7	1	0.1	4	0.6	0.001	0.002	0.001	0.003	0.004	0.002	0.001
8	1	0.2	1	0.5	0.004	0.004	0.005	0.002	0.004	0.002	0.005
9	1	0.2	1	0.6	0.008	0.004	0.003	0.001	0.003	0	0.006

Figure 6.7.3 The saved table of variant-based power.

STEP 8: The SKAT tool is integrated into SPS to detect the significant SNPs, especially for rare variants. In order to running SKAT, you should install R software (version >=2.13.0) firstly. Please click the tag 'Download R' to download R, and several R packages are needed too. These packages are "RServe", "SKAT" and "snow". "RServe" is a java-r interface. "Snow" is used to run SKAT in parallel. If you don't know how to install these packages, just paste the prompt message in your R platform. When SKAT is ticked, the power calculated by SKAT will be added to the table and chart. Although SPS can run SKAT in parallel, this is also a time-consuming part and the sampling time should not be set too large.

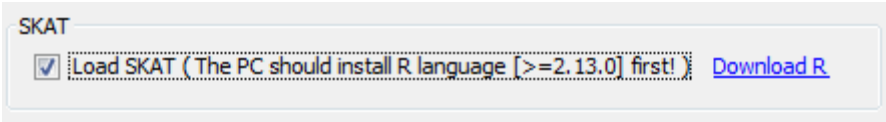


Figure 6.8.1 SKAT option.

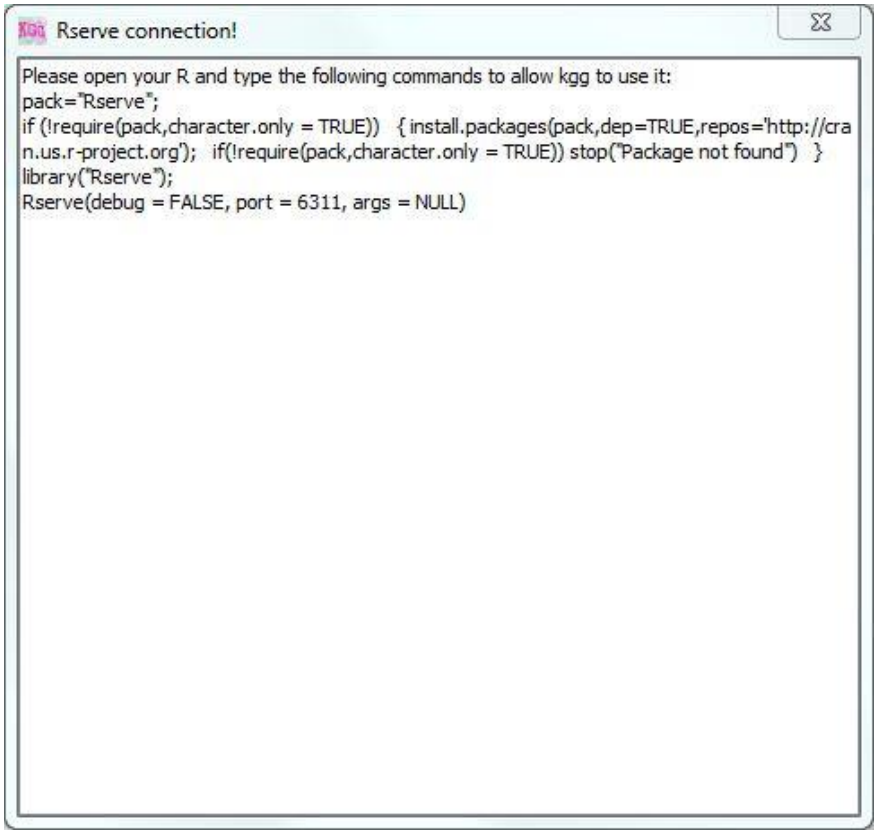


Figure 6.8.2 The prompt message.

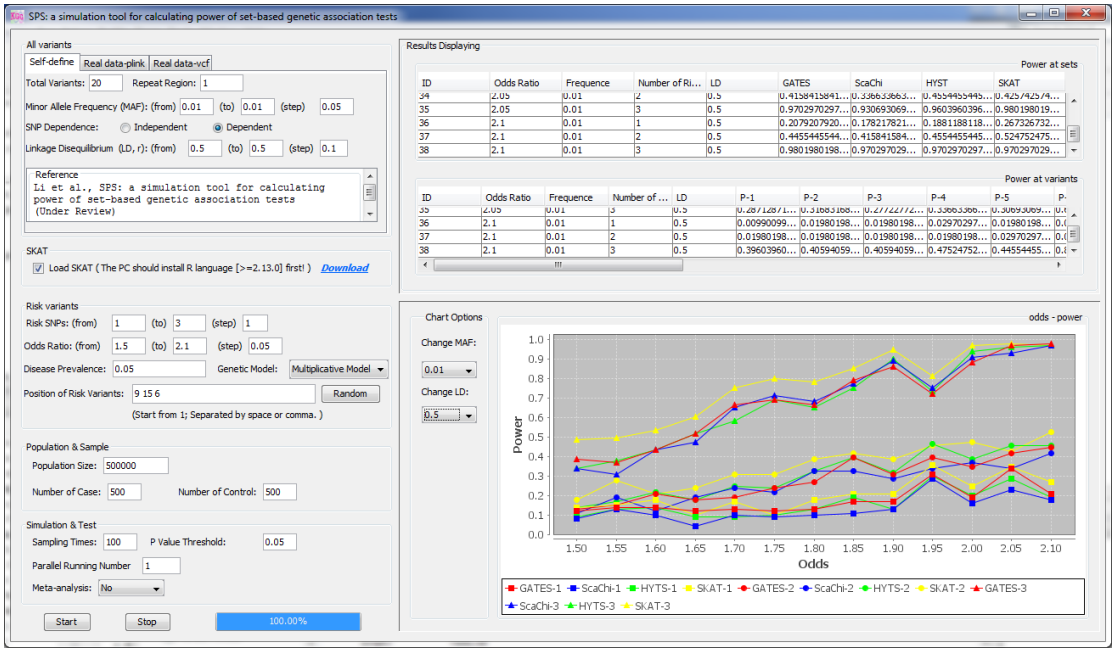
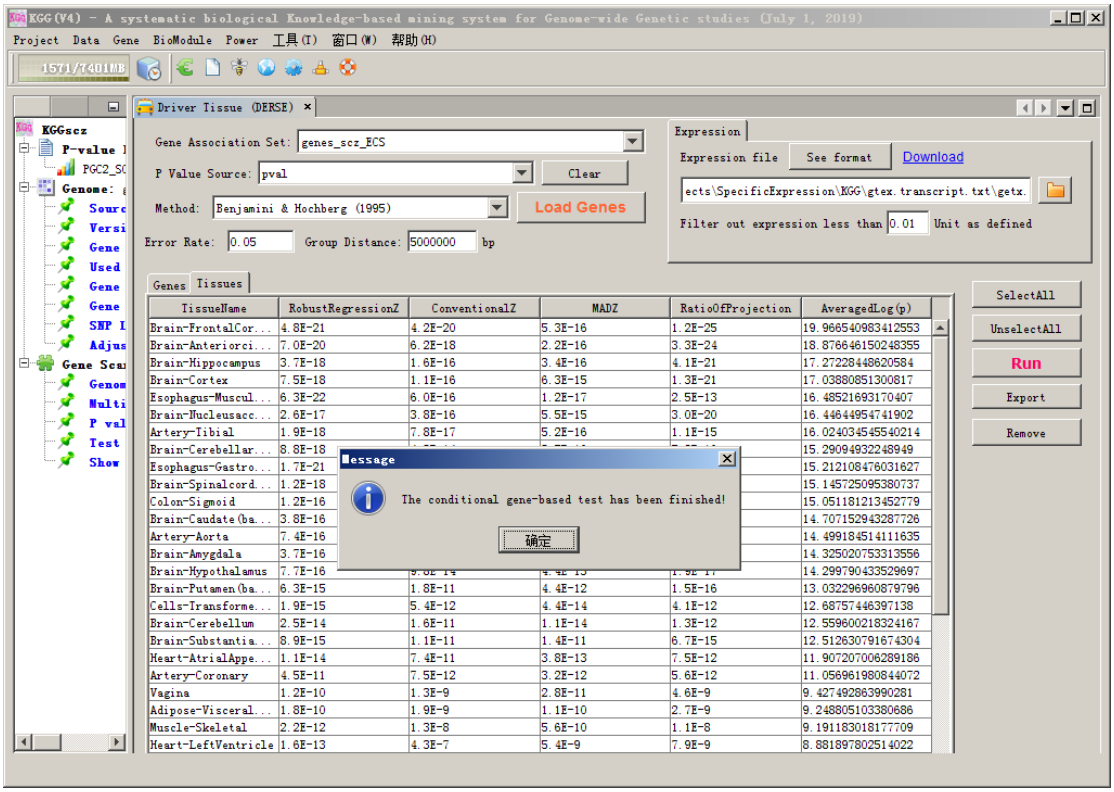


Figure 6.8.3 The output when SKAT added.

7. Updates from KGG3.5 to KGG4.0

Much progress was made from KGG 3.0 to KGG 3.5, mainly including:

1) A more powerful gene-based association analysis by effective Chi-squared test;
2) A powerful conditional gene-based association analysis by effective Chi-squared test;
3) A robust geneset based association analysis by LD-attenuating rank-sum test;



TissueName	RobustRegressionZ	ConventionalZ	MADZ	RatioOfProjection	AveragedLog(p)
Brain-FrontalCor...	4.8E-21	4.2E-20	5.3E-16	1.2E-25	19.966540983412553
Brain-Anteriorci...	7.0E-20	6.2E-18	2.2E-16	3.3E-24	18.876646150248355
Brain-Hippocampus	3.7E-18	1.6E-16	3.4E-16	4.1E-21	17.27228448620584
Brain-Cortex	7.5E-18	1.1E-16	6.3E-15	1.3E-21	17.03880851300817
Esophagus-Muscul...	6.3E-22	6.0E-16	1.2E-17	2.5E-13	16.48521693170407
Brain-Nucleusacc...	2.6E-17	3.8E-16	5.5E-15	3.0E-20	16.44644954741902
Artery-Tibial	1.9E-18	7.8E-17	5.2E-16	1.1E-15	16.024034545540214
Brain-Cerebellar...	8.8E-18				15.29094932248949
Esophagus-Gastro...	1.7E-21				15.212108476031627
Brain-Spinalcord...	1.2E-18				15.145725095380737
Colon-Sigmoid	1.2E-16				15.051181213452779
Brain-Caudate(ba...	3.8E-16				14.707152943287726
Artery-Aorta	7.4E-16				14.499184514111635
Brain-Amygdala	3.7E-16				14.325020753313556
Brain-Hypothalamus	7.7E-16				14.299790433529697
Brain-Putamen(ba...	6.3E-15	1.8E-11	4.4E-12	1.5E-16	13.032296960879796
Cells-Transforme...	1.9E-15	5.4E-12	4.4E-14	4.1E-12	12.68757446397138
Brain-Cerebellum	2.5E-14	1.6E-11	1.1E-14	1.3E-12	12.559600218324167
Brain-Substantia...	8.9E-15	1.1E-11	1.4E-11	6.7E-15	12.512630791674304
Heart-AtrialAppe...	1.1E-14	7.4E-11	3.8E-13	7.5E-12	11.907207006289186
Artery-Coronary	4.5E-11	7.5E-12	3.2E-12	5.6E-12	11.056961980844072
Vagina	1.2E-10	1.3E-9	2.8E-11	4.6E-9	9.427492863990281
Adipose-Visceral...	1.8E-10	1.9E-9	1.1E-10	2.7E-9	9.248805103380686
Muscle-Skeletal	2.2E-12	1.3E-8	5.6E-10	1.1E-8	9.191183018177709
Heart-LeftVentricle	1.6E-13	4.3E-7	5.4E-9	7.9E-9	8.881897802514022