KGG: A systematic biological Knowledge-based mining system for Genome-wide Genetic studies (Version 3.5)

User Manual

Miao-Xin Li, Jiang Li

Department of Psychiatry
Centre for Genomic Sciences
Department of Biochemistry
The University of Hong Kong
Pokfulam, Hong Kong SAR, China

Content

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

Set or change large memory for KGG3 say, 2000MB, by *Tools->Set System Memory*.

1. Introduction and general pipeline

KGG (Knowledge-based mining system for Genome-wide Genetic studies) is a software tool to perform knowledge-based analysis for genome-wide association studies (GWAS). At present, the version 3 has been equipped with main functions to conduct multivariate/univariate gene-based association tests using SNP p-values from GWAS[1,2,3] and to carry out advanced univariate biological module-based association analysis (pathway enrichment and protein-protein interaction (PPI) network association) by a set-based test [2]. 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.

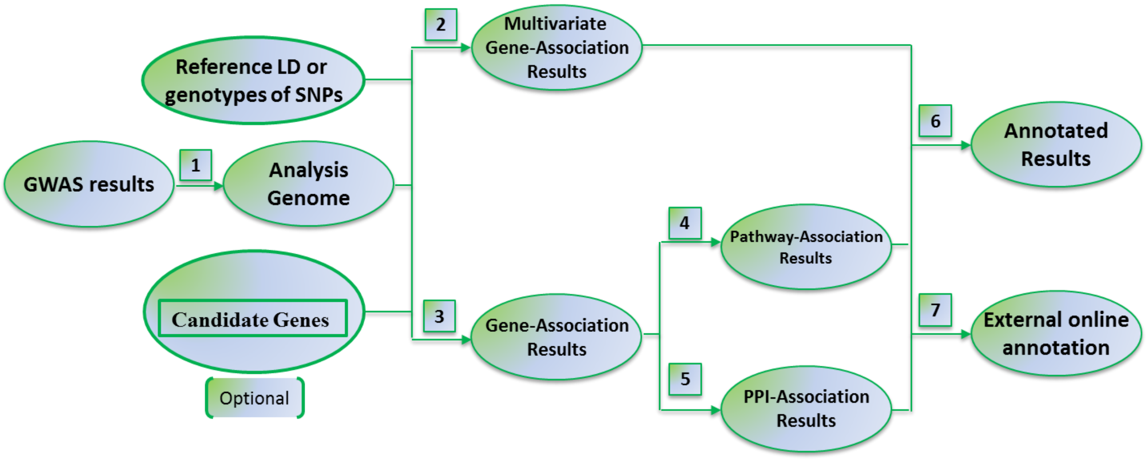

Reference LD or genotypes of SNPs
GWAS results
1
Analysis Genome
Candidate Genes
Optional
2
Multivariate Gene-Association Results
3
Gene-Association Results
4
Pathway-Association Results
5
PPI-Association Results
6
Annotated Results
7
External online annotation

Figure 1.1 Pipeline chart of KGG analysis (version 3)

Notes: Circle nodes stand for data and files (input, output); single directional arrows stand for analytical procedures involved.

Main steps involved:

1) Build an **Analysis Genome**: generate an intermediate dataset which integrates original GWAS p-values, SNP annotation and gene annotation, and LD between SNPs WITIN genes together. It is a unified dataset which will be used for all kinds of analyses on KGG.
2) Conduct a **multivariate gene-based association test**: calculate gene-based p-values of multiple phenotypes by a method [3] .
3) Conduct **gene-based association test**: calculate gene-based p-values of a single phenotype by GATES[1] or HYST[2] .
4) Explore significantly **associated pathways** by HYST [2] and enriched with susceptibility genes by hypergeometric distribution test. One can use either the integrated pathways (gene sets) from MsigDB (http://www.broadinstitute.org/gsea/msigdb) or his or her self-customized pathways on KGG.
5) Explore statistically significant **associated PPI pairs** by HYST [2] which may work together to contribute to the development of the disease or traits. Again, one can use either the integrated PPI pairs from the STRING PPI (http://string-db.org/) or his or her self-customized PPI pairs on KGG.
6) Annotate and export significant SNPs, genes, pathways and PPIs.
7) View **external bioinformatics annotation results** of statistically significant SNPs, genes and pathways.

Other plug-in:

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

References

1. 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.
2. 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.

3. Sluis et al. MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. Bioinformatics (In press)

2. Installation

2.1 Installation of Java Runtime Environment (JRE)

The Java Runtime Environment (JRE) v1.7 (or higher version) is required to run KGG3 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 in Windows OS and Mac OS X.

In 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 the extracted jre/bin directory is added to the PATH, otherwise KGG3 would not start properly. This is easily achievable by executing the following command on the terminal:

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

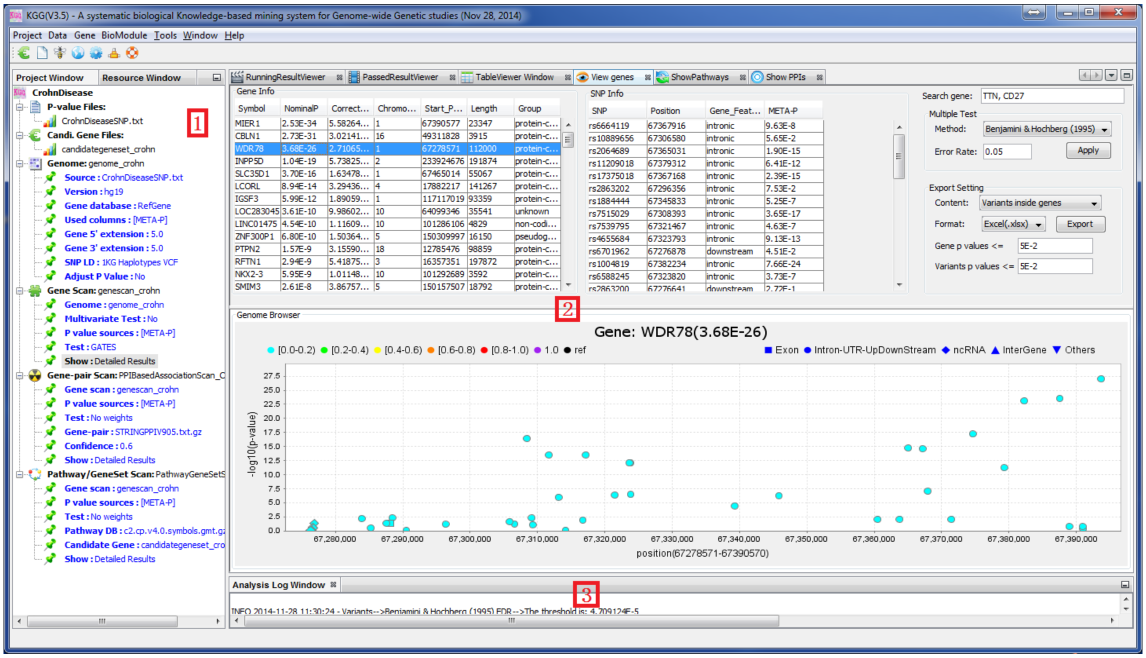
Thanks **Attila Pulay** for the suggestion!

2.2 Installation of 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 kgg3.zip file, you will see a “bin” folder where there are 3 script files to initiate KGG3. On Microsoft Windows, please double click kgg3.exe or kgg364.exe file. On Linux, Mac OS X and Solaris, please type the kgg3 in a Command-line Terminal.

3. Interface and functions

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



CrohnDisease
- **P-value Files:**
  - CrohnDiseaseSNP.txt
- **Candi. Gene Files:**
  - candidategeneset_crohn
- **Genome:** genome_crohn
  - **Source :** CrohnDiseaseSNP.txt
  - **Version :** hg19
  - **Gene database :** RefGene
  - **Used columns :** [META-P]
  - **Gene 5' extension :** 5.0
  - **Gene 3' extension :** 5.0
  - **SNP LD :** 1KG Haplotypes VCF
  - **Adjust P Value :** No
- **Gene Scan:** genescan_crohn
  - **Genome :** genome_crohn
  - **Multivariate Test :** No
  - **P value sources :** [META-P]
  - **Test :** GATES
  - **Show :** Detailed Results
- **Gene-pair Scan:** PPIBasedAssociationScan_C
  - **Gene scan :** genescan_crohn
  - **P value sources :** [META-P]
  - **Test :** No weights
  - **Gene-pair :** STRINGPPIV905.txt.gz
  - **Confidence :** 0.6
  - **Show :** Detailed Results
- **Pathway/GeneSet Scan:** PathwayGeneSetS
  - **Gene scan :** genescan_crohn
  - **P value sources :** [META-P]
  - **Test :** No weights
  - **Pathway DB :** c2.cp.v4.0.symbols.gmt.gz
  - **Candidate Gene :** candidategeneset_cro
  - **Show :** Detailed Results

Symbol	NominalP	Correct...	Chromo...	Start_P...	Length	Group
MIER1	2.53E-34	5.58264...	1	67390577	23347	protein-c...
CBLN1	2.73E-31	3.02141...	16	49311828	3915	protein-c...
WDR78	3.68E-26	2.71065...	1	67278571	112000	protein-c...
INPP5D	1.04E-19	5.73825...	2	233924676	191874	protein-c...
SLC35D1	3.70E-16	1.63478...	1	67465014	55067	protein-c...
LCORL	8.94E-14	3.29436...	4	17882217	141267	protein-c...
IGSF3	5.99E-12	1.89059...	1	117117019	93359	protein-c...
LOC283045	3.61E-10	9.98602...	10	64099346	35541	unknown
LINC01475	4.54E-10	1.11609...	10	101286106	4829	non-codi...
ZNF300P1	6.80E-10	1.50364...	5	150309997	16150	pseudog...
PTPN2	1.57E-9	3.15590...	18	12785476	98859	protein-c...
RFTN1	2.94E-9	5.41875...	3	16357351	197872	protein-c...
NKX2-3	5.95E-9	1.01148...	10	101292689	3592	protein-c...
SMIM3	2.61E-8	3.86757...	5	150157507	18792	protein-c...

SNP	Position	Gene_Feat...	META-P
rs6664119	67367916	intronic	9.63E-8
rs10889656	67306580	intronic	5.65E-2
rs2064689	67365031	intronic	1.90E-15
rs11209018	67379312	intronic	6.41E-12
rs17375018	67367168	intronic	2.39E-15
rs2863202	67296356	intronic	7.53E-2
rs1884444	67345833	intronic	5.25E-7
rs7515029	67308393	intronic	3.65E-17
rs7539795	67321467	intronic	4.63E-7
rs4655684	67323793	intronic	9.13E-13
rs6701962	67276878	downstream	4.51E-2
rs1004819	67382234	intronic	7.66E-24
rs6588245	67323820	intronic	3.73E-7
rs2863200	67276641	downstream	2.72E-1

Search gene: TTN, CD27

Multiple Test
Method: Benjamini & Hochberg (1995)
Error Rate: 0.05
Apply

Export Setting
Content: Variants inside genes
Format: Excel(.xlsx)
Export
Gene p values <= 5E-2
Variants p values <= 5E-2

```
INFO 2014-11-28 11:30:24 - Variants-->Benjamini & Hochberg (1995) FDR-->The threshold is: 4.709124E-5
```

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 KGGs are self-explaining. 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

- **Gene-based association scan:** conduce the gene-based association scans.
- **View genes:** view and export gene-based association results.

3.5 Module

- **PPI-based association scan:** conduct PPI based association scan.
- **View PPIs:** view significant PPI pairs.
- **Pathway-based association scan:** conduct pathway based association scan.
- **View pathways:** view significant pathways.

3.5 Tools

- **Set system memory**: set the memory of KGG.
- **Power calculator**: SPS-a simulation tool for calculating power of set-based genetic association tests.

3.6 Window

- **AnalysisOutput: show the results when performing multiple tests.**
- **Project: depict the structure of the project.**
- **Resource: show the resource that KGG contains.**
- **ResultViewer: give the real-time results when performing the concrete analysis.**
- **RunningResultViewer: record the parameters using in each analysis.**
- **TableViewer: display the content of some files.**
- **Output: show all the IDE output.**

4. Input files

4.1 Input file 1 (GWAS results)

KGG focuses on the downstream analysis of GWA studies, where statistical association p-values (or chi-square values) at SNPs have been generated by conventional statistical genetic methods (such as PLINK). Therefore, the association p-values are the major input of our KGG. KGG flexible supports a user-customized format for the association p-values.

Once three columns of information, chromosome number and SNP IDs (or physical position) and p-values are available in a file, you can define the column order by yourselves on KGG. The input file can include more than one p-value column. The following is an example.

Example input format (with rsID) of KGG:

CHR	SNP	P-value1	P-value2	P-value3	…
4	rs1513559	0.02301	0.8815	0.007688	…
4	rs294755	0.4384	0.9575	0.006112	…
4	rs835316	0.002688	0.007688	0.4893	…
4	rs1841043	0.01115	0.006112	0.119	…
4	rs11726946	0.005892	0.4893	0	…
…	…	…	…	…	…

Example input format (with only position) 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	…
…	…	…	…	…	…	…

Moreover, a p-value column could include values of different models. KGG will recognize this format if you select the input format as “multiple tests per column” when building the analysis genome.

Example a more complex input format of KGG:

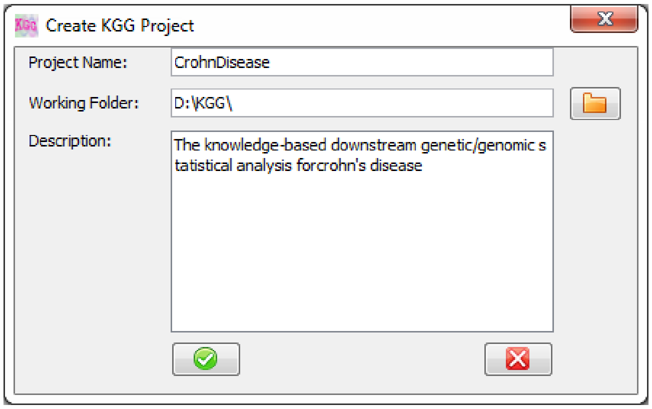
CHR	SNP	P-value1	Test-Mode	P-value2	…
4	rs1513559	0.02301	additive	0.007688	…
4	rs1513559	0.4384	recessive	0.006112	…
4	rs1513559	0.002688	dominant	0.4893	…
4	rs1841043	0.01115	additive	0.119	…
4	rs1841043	0.005892	recessive	0	…
…	…	…	…	…	…

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, while one row contains one gene (symbol or ID).

5. Set-based association analysis tutorial

Step 1: create a new project, named ‘CrohnDisease’, and set the project path at C:\KGG\Tutorial (or other path defined by user).



Create KGG Project

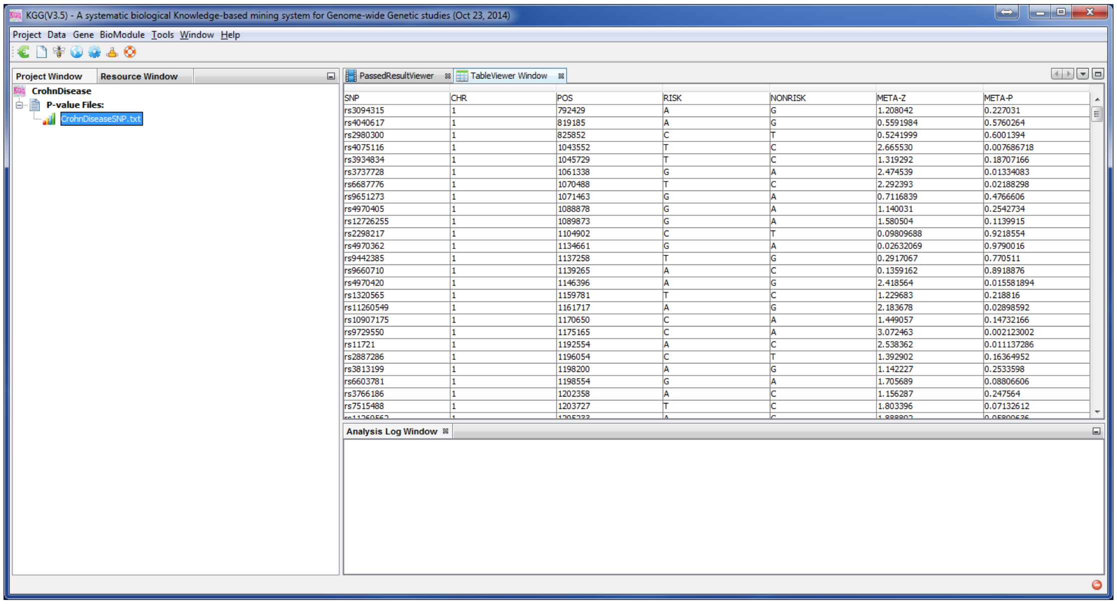
Project Name: CrohnDisease

Working Folder: D:\KGG\

Description: The knowledge-based downstream genetic/genomic statistical analysis forcrohn's disease

Figure 5.1 Create project

Step 2: select the menu Data>Load P Value File and choose ‘CrohnDiseaseSNP.txt’ file which contains the whole-genome association p-values for Crohn disease at SNP-level. This dataset was downloaded from a public domain released by (Barrett, et al., 2008). It includes 7 columns, as SNP, CHR, POS, RISK, NONRISK, META-Z and META-P.

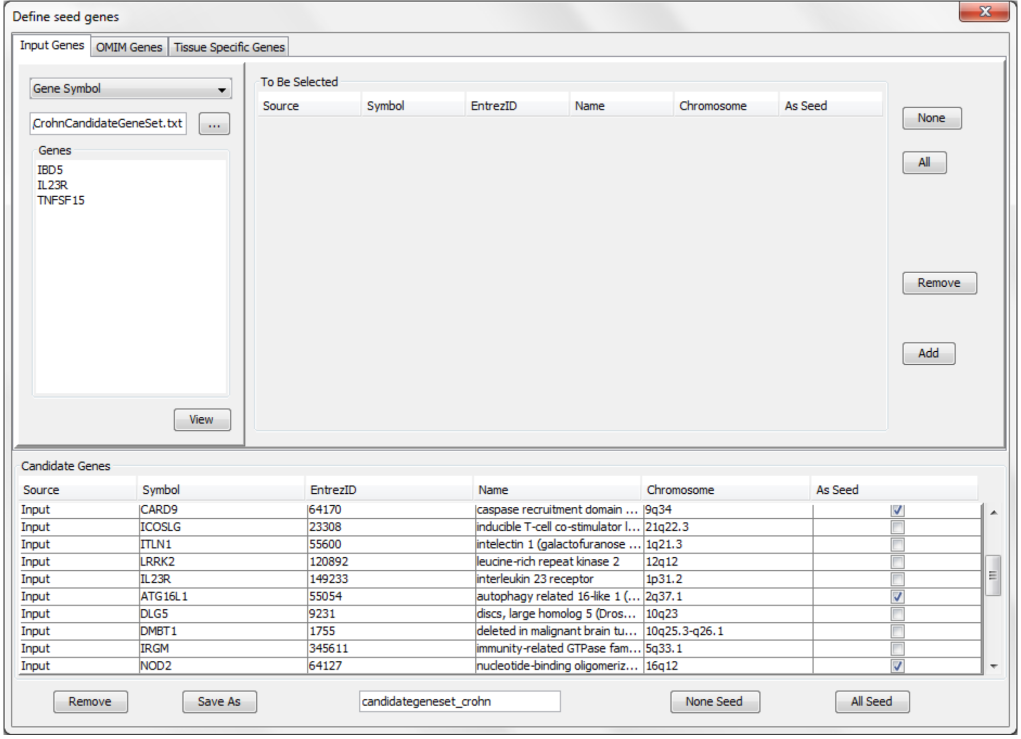

KGG(V3.5) - A systematic biological Knowledge-based mining system for Genome-wide Genetic studies (Oct 23, 2014)
Project Data Gene BioModule Tools Window Help
Project Window
Resource Window
PassedResultViewer
TableViewer Window
Analysis Log Window

CrohnDisease
P-value Files:
CrohnDiseaseSNP.txt

SNP	CHR	POS	RISK	NONRISK	META-Z	META-P
rs3094315	1	792429	A	G	1.208042	0.227031
rs4040617	1	819185	A	G	0.5591984	0.5760264
rs2980300	1	825852	C	T	0.5241999	0.6001394
rs4075116	1	1043552	T	C	2.665530	0.007686718
rs3934834	1	1045729	T	C	1.319292	0.18707166
rs3737728	1	1061338	G	A	2.474539	0.01334083
rs6687776	1	1070488	T	C	2.292393	0.02188298
rs9651273	1	1071463	G	A	0.7116839	0.4766606
rs4970405	1	1088878	G	A	1.140031	0.2542734
rs12726255	1	1089873	G	A	1.580504	0.1139915
rs2298217	1	1104902	C	T	0.09809688	0.9218554
rs4970362	1	1134661	G	A	0.02632069	0.9790016
rs9442385	1	1137258	T	G	0.2917067	0.770511
rs9660710	1	1139265	A	C	0.1359162	0.8918876
rs4970420	1	1146396	A	G	2.418564	0.015581894
rs1320565	1	1159781	T	C	1.229683	0.218816
rs11260549	1	1161717	A	G	2.183678	0.02898592
rs10907175	1	1170650	C	A	1.449057	0.14732166
rs9729550	1	1175165	C	A	3.072463	0.002123002
rs11721	1	1192554	A	C	2.538362	0.011137286
rs2887286	1	1196054	C	T	1.392902	0.16364952
rs3813199	1	1198200	A	G	1.142227	0.2533598
rs6603781	1	1198554	G	A	1.705689	0.08806606
rs3766186	1	1202358	A	C	1.156287	0.247564
rs7515488	1	1203727	T	C	1.803396	0.07132612

Figure 5.2 Input GWAS original result file

Step 3: import file 'CrohnCandidateGeneSet.txt' as input of candidate gene; define ATG16L1, CARD9, IBD5, IL23R, NOD2 and TNFSF15 as seed genes. Then, save it as candidategeneset_crohn.



CrohnCandidateGeneSet.txt

IBD5
IL23R
TNFSF15

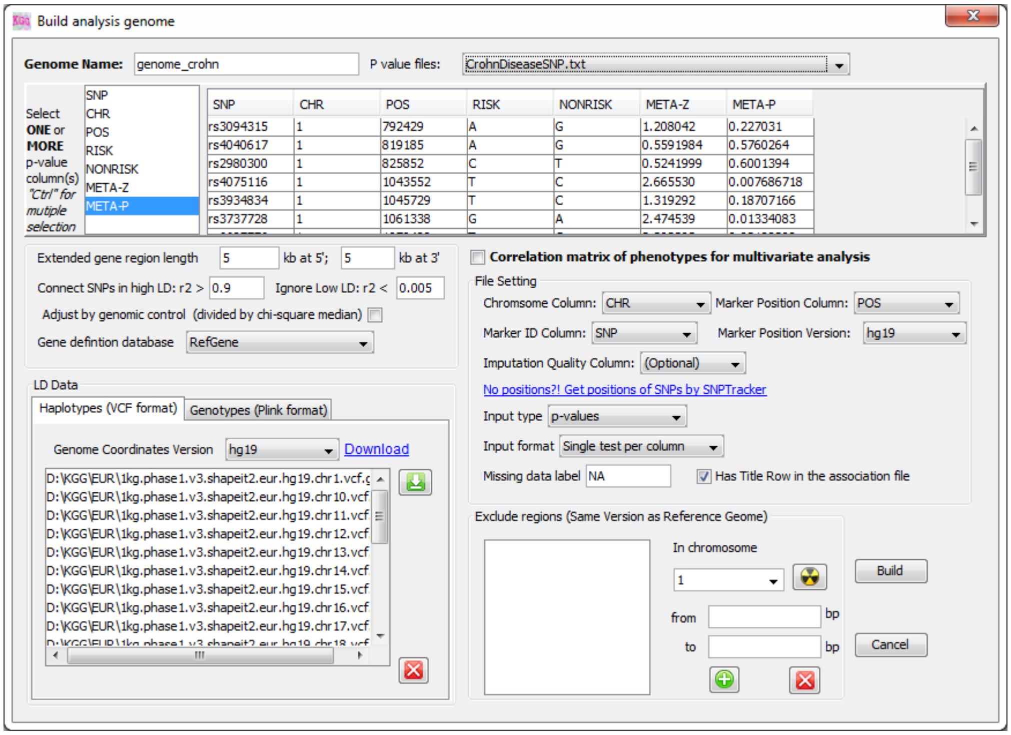
Candidate Genes

Source	Symbol	EntrezID	Name	Chromosome	As Seed
Input	CARD9	64170	caspase recruitment domain ...	9q34	☑
Input	ICOSLG	23308	inducible T-cell co-stimulator l...	21q22.3	☐
Input	ITLN1	55600	intelectin 1 (galactofuranose ...	1q21.3	☐
Input	LRRK2	120892	leucine-rich repeat kinase 2	12q12	☐
Input	IL23R	149233	interleukin 23 receptor	1p31.2	☐
Input	ATG16L1	55054	autophagy related 16-like 1 (...	2q37.1	☑
Input	DLG5	9231	discs, large homolog 5 (Dros...	10q23	☐
Input	DMBT1	1755	deleted in malignant brain tu...	10q25.3-q26.1	☐
Input	IRGM	345611	immunity-related GTPase fam...	5q33.1	☐
Input	NOD2	64127	nucleotide-binding oligomeriz...	16q12	☑

candidategeneset_crohn

Figure 5.3 Input candidate gene set for crohn's disease

Step 4: select META-P for building analysis genome; extend gene region to its flanking 5 kb region in both sides; and use LD SNP coefficients from 1000 Genome Project to adjust LD.



Build analysis genome

Genome Name: genome_crohn   P value files: CrohnDiseaseSNP.txt

Select ONE or MORE p-value column(s) "Ctrl" for mutiple selection: SNP, CHR, POS, RISK, NONRISK, META-Z, META-P

SNP	CHR	POS	RISK	NONRISK	META-Z	META-P
rs3094315	1	792429	A	G	1.208042	0.227031
rs4040617	1	819185	A	G	0.5591984	0.5760264
rs2980300	1	825852	C	T	0.5241999	0.6001394
rs4075116	1	1043552	T	C	2.665530	0.007686718
rs3934834	1	1045729	T	C	1.319292	0.18707166
rs3737728	1	1061338	G	A	2.474539	0.01334083

Extended gene region length 5 kb at 5'; 5 kb at 3'

Connect SNPs in high LD: r2 > 0.9   Ignore Low LD: r2 < 0.005

Adjust by genomic control (divided by chi-square median) □

Gene defintion database RefGene

LD Data

Haplotypes (VCF format) | Genotypes (Plink format)

Genome Coordinates Version hg19 Download

```
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr1.vcf.g
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr10.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr11.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr12.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr13.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr14.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr15.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr16.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr17.vcf
D:\KGG\EUR\1kg.phase1.v3.shapeit2.eur.hg19.chr18.vcf
```

□ Correlation matrix of phenotypes for multivariate analysis

File Setting

Chromsome Column: CHR   Marker Position Column: POS

Marker ID Column: SNP   Marker Position Version: hg19

Imputation Quality Column: (Optional)

No positions?! Get positions of SNPs by SNPTracker

Input type p-values

Input format Single test per column

Missing data label NA   ☑ Has Title Row in the association file

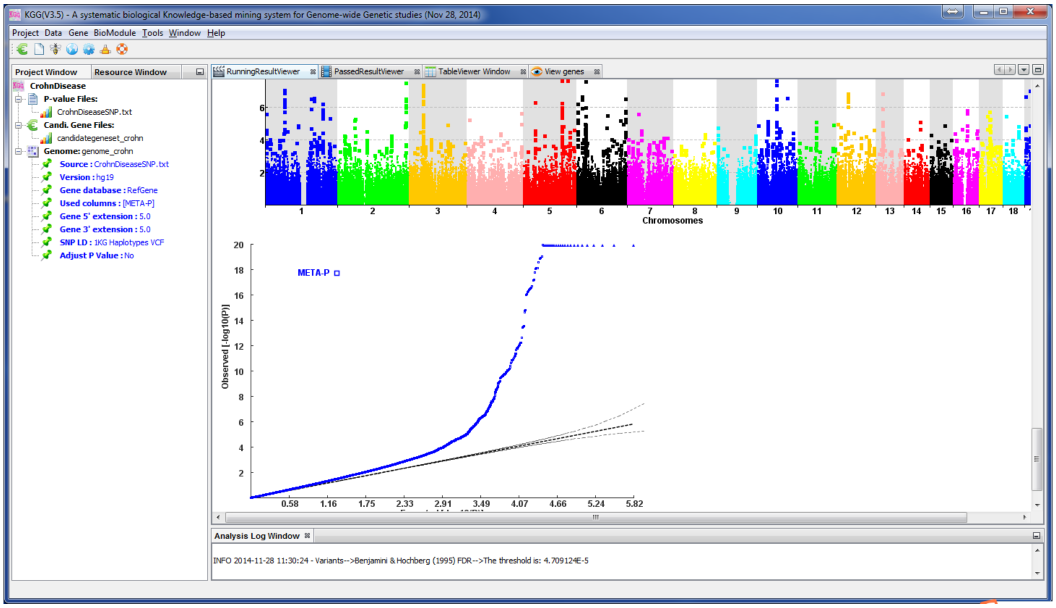
Exclude regions (Same Version as Reference Geome)

In chromosome 1

from bp

to bp

Figure 5.4.1 Select META-P to build analysis genome and name the genome as genome_crohn



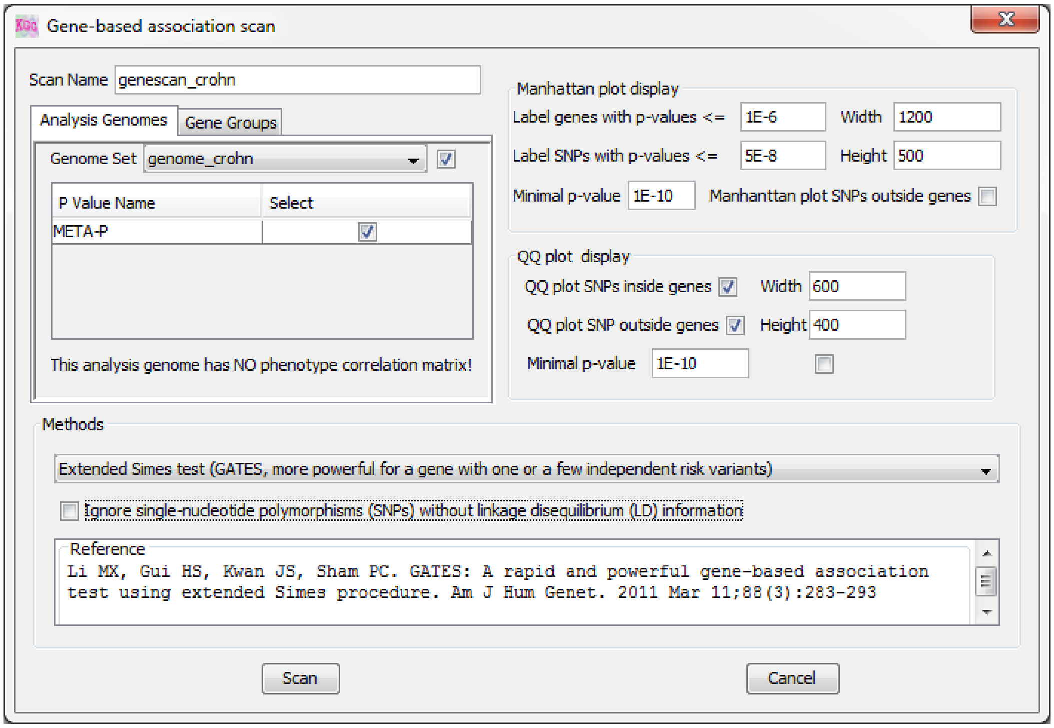
KGG(V3.5) - A systematic biological Knowledge-based mining system for Genome-wide Genetic studies (Nov 28, 2014)

CrohnDisease
- P-value Files:
  - CrohnDiseaseSNP.txt
- Candi. Gene Files:
  - candidategeneset_crohn
- Genome: genome_crohn
  - Source : CrohnDiseaseSNP.txt
  - Version : hg19
  - Gene database : RefGene
  - Used columns : [META-P]
  - Gene 5' extension : 5.0
  - Gene 3' extension : 5.0
  - SNP LD : 1KG Haplotypes VCF
  - Adjust P Value : No

```
INFO 2014-11-28 11:30:24 - Variants-->Benjamini & Hochberg (1995) FDR-->The threshold is: 4.709124E-5
```

Figure 5.4.2 The display after building analysis genome

Step 5: do a gene-based association scan using SNP p-values integrated in the analysis genome named genome_crohn, select 'Extended Simes test(GATES, more powerful for a gene with one or a few independent causal variants' method. Set the parameters as Figure 5.5.1; and name the result as genescan_crohn. Remember that exported Manhattan plots and QQ plots will be shown in "Running Result Viewer Window" (Figure 5.5.2).



Scan Name: genescan_crohn

Genome Set: genome_crohn ☑

P Value Name	Select
META-P	☑

This analysis genome has NO phenotype correlation matrix!

Manhattan plot display

Label genes with p-values <= 1E-6 Width 1200

Label SNPs with p-values <= 5E-8 Height 500

Minimal p-value 1E-10 Manhanttan plot SNPs outside genes □

QQ plot display

QQ plot SNPs inside genes ☑ Width 600

QQ plot SNP outside genes ☑ Height 400

Minimal p-value 1E-10 □

Methods

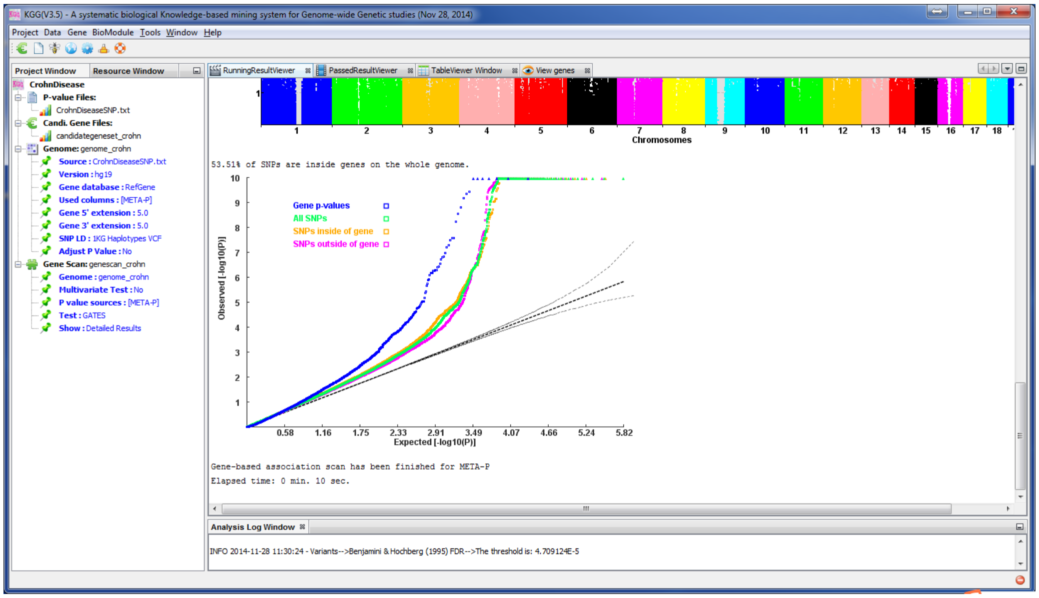
Extended Simes test (GATES, more powerful for a gene with one or a few independent risk variants)

□ Ignore single-nucleotide polymorphisms (SNPs) without linkage disequilibrium (LD) information

Reference

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

Figure 5.5.1 Setting for gene-based scans



CrohnDisease
- P-value Files:
  - CrohnDiseaseSNP.txt
- Candi. Gene Files:
  - candidategeneset_crohn
- Genome: genome_crohn
  - Source : CrohnDiseaseSNP.txt
  - Version : hg19
  - Gene database : RefGene
  - Used columns : [META-P]
  - Gene 5' extension : 5.0
  - Gene 3' extension : 5.0
  - SNP LD : 1KG Haplotypes VCF
  - Adjust P Value : No
- Gene Scan: genescan_crohn
  - Genome : genome_crohn
  - Multivariate Test : No
  - P value sources : [META-P]
  - Test : GATES
  - Show : Detailed Results

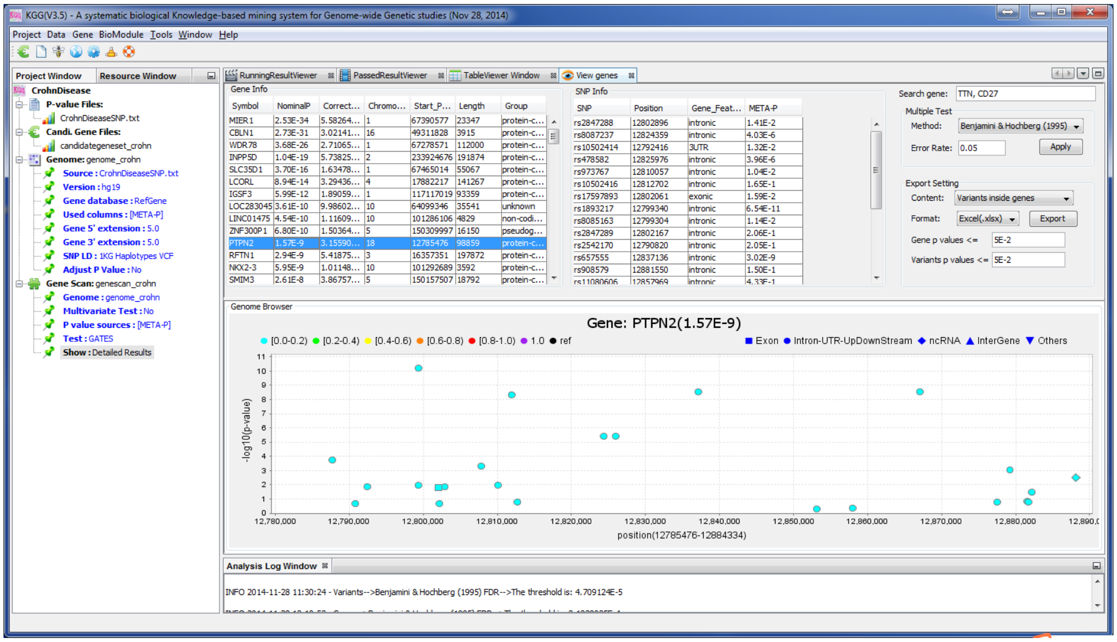
```
53.51% of SNPs are inside genes on the whole genome.
```

```
Gene-based association scan has been finished for META-P
Elapsed time: 0 min. 10 sec.
```

```
INFO 2014-11-28 11:30:24 - Variants-->Benjamini & Hochberg (1995) FDR-->The threshold is: 4.709124E-5
```

Figure 5.5.2 The display after gene-based scan

Step 6: Click the "Show: Detailed Results" node under "Genome Scan" and a new tab "ShowGenes" will be created to provide you more information about the result (Figure 5.6). You can also export the results you want in this tab.



- **CrohnDisease**
  - **P-value Files:**
    - CrohnDiseaseSNP.txt
  - **Candi. Gene Files:**
    - candidategeneset_crohn
  - **Genome:** genome_crohn
    - **Source :** CrohnDiseaseSNP.txt
    - **Version :** hg19
    - **Gene database :** RefGene
    - **Used columns :** [META-P]
    - **Gene 5' extension :** 5.0
    - **Gene 3' extension :** 5.0
    - **SNP LD :** 1KG Haplotypes VCF
    - **Adjust P Value :** No
  - **Gene Scan:** genescan_crohn
    - **Genome :** genome_crohn
    - **Multivariate Test :** No
    - **P value sources :** [META-P]
    - **Test :** GATES
    - **Show :** Detailed Results

Gene Info

Symbol	NominalP	Correct...	Chromo...	Start_P...	Length	Group
MIER1	2.53E-34	5.58264...	1	67390577	23347	protein-c...
CBLN1	2.73E-31	3.02141...	16	49311828	3915	protein-c...
WDR78	3.68E-26	2.71065...	1	67278571	112000	protein-c...
INPP5D	1.04E-19	5.73825...	2	233924676	191874	protein-c...
SLC35D1	3.70E-16	1.63478...	1	67465014	55067	protein-c...
LCORL	8.94E-14	3.29436...	4	17882217	141267	protein-c...
IGSF3	5.99E-12	1.89059...	1	117117019	93359	protein-c...
LOC283045	3.61E-10	9.98602...	10	64099346	35541	unknown
LINC01475	4.54E-10	1.11609...	10	101286106	4829	non-codi...
ZNF300P1	6.80E-10	1.50364...	5	150309997	16150	pseudog...
PTPN2	1.57E-9	3.15590...	18	12785476	98859	protein-c...
RFTN1	2.94E-9	5.41875...	3	16357351	197872	protein-c...
NKX2-3	5.95E-9	1.01148...	10	101292689	3592	protein-c...
SMIM3	2.61E-8	3.86757...	5	150157507	18792	protein-c...

SNP Info

SNP	Position	Gene_Feat...	META-P
rs2847288	12802896	intronic	1.41E-2
rs8087237	12824359	intronic	4.03E-6
rs10502414	12792416	3UTR	1.32E-2
rs478582	12825976	intronic	3.96E-6
rs973767	12810057	intronic	1.04E-2
rs10502416	12812702	intronic	1.65E-1
rs17597893	12802061	exonic	1.59E-2
rs1893217	12799340	intronic	6.54E-11
rs8085163	12799304	intronic	1.14E-2
rs2847289	12802167	intronic	2.06E-1
rs2542170	12790820	intronic	2.05E-1
rs657555	12837136	intronic	3.02E-9
rs908579	12881550	intronic	1.50E-1
rs11080606	12857969	intronic	4.33E-1

Search gene: TTN, CD27

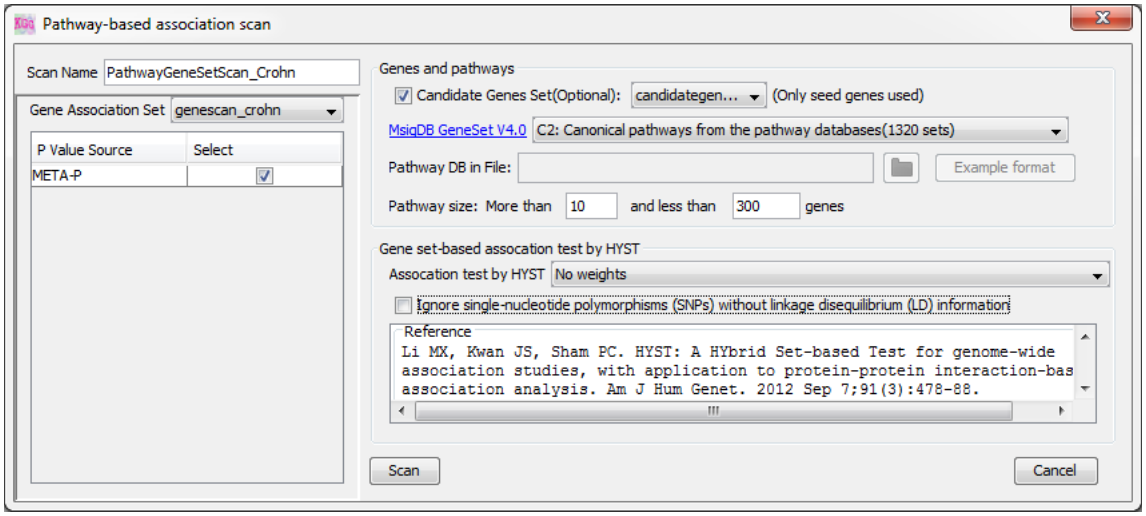
Multiple Test
Method: Benjamini & Hochberg (1995)
Error Rate: 0.05

Export Setting
Content: Variants inside genes
Format: Excel(.xlsx)
Gene p values <= 5E-2
Variants p values <= 5E-2

```
INFO 2014-11-28 11:30:24 - Variants-->Benjamini & Hochberg (1995) FDR-->The threshold is: 4.709124E-5
```

Figure 5.6 Function of displaying the gene-based association scan result

Step 7: perform pathway enrichment exploration both by gene p-values; settings as Figure 5.7.1 and the output as Figure 5.7.2.


Pathway-based association scan
Genes and pathways
Gene set-based assocation test by HYST
Reference
Scan
Cancel
Example format

Scan Name: PathwayGeneSetScan_Crohn

Gene Association Set: genescan_crohn

P Value Source	Select
META-P	☑

☑ Candidate Genes Set(Optional): candidategen... (Only seed genes used)

MsigDB GeneSet V4.0: C2: Canonical pathways from the pathway databases(1320 sets)

Pathway DB in File:

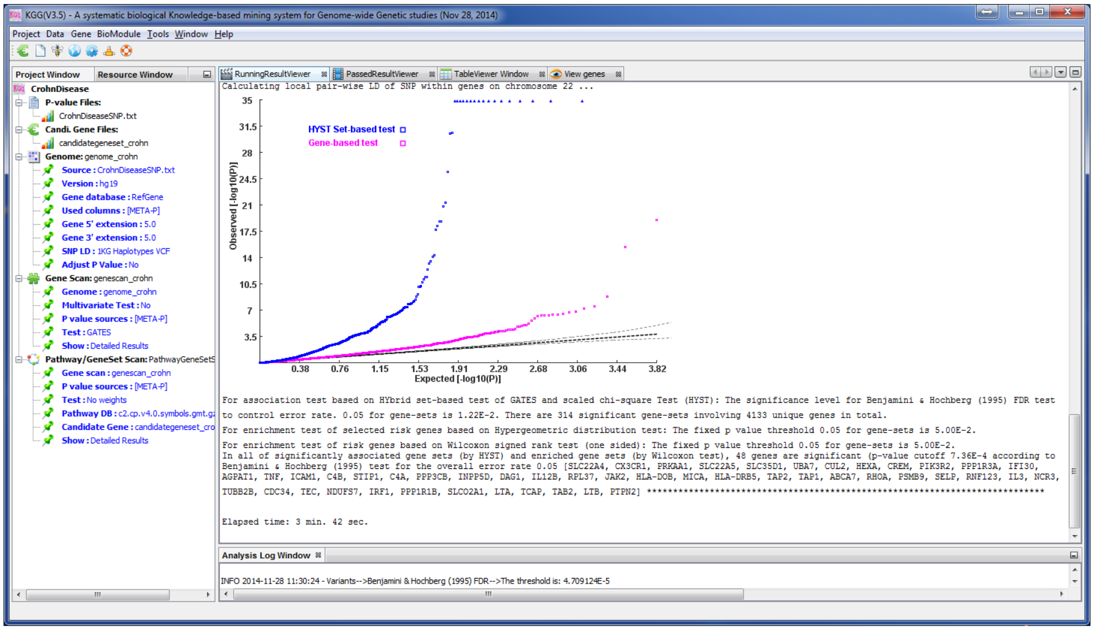
Pathway size: More than 10 and less than 300 genes

Assocation test by HYST: No weights

☐ Ignore single-nucleotide polymorphisms (SNPs) without linkage disequilibrium (LD) information

```
Li MX, Kwan JS, Sham PC. HYST: A HYbrid Set-based Test for genome-wide
association studies, with application to protein-protein interaction-bas
association analysis. Am J Hum Genet. 2012 Sep 7;91(3):478-88.
```

Figure 5.7.1 Pathway enrichment exploration by gene p-values



- CrohnDisease
  - P-value Files:
    - CrohnDiseaseSNP.txt
  - Candi. Gene Files:
    - candidategeneset_crohn
  - Genome: genome_crohn
    - Source : CrohnDiseaseSNP.txt
    - Version : hg19
    - Gene database : RefGene
    - Used columns : [META-P]
    - Gene 5' extension : 5.0
    - Gene 3' extension : 5.0
    - SNP LD : 1KG Haplotypes VCF
    - Adjust P Value : No
  - Gene Scan: genescan_crohn
    - Genome : genome_crohn
    - Multivariate Test : No
    - P value sources : [META-P]
    - Test : GATES
    - Show : Detailed Results
  - Pathway/GeneSet Scan: PathwayGeneSetS
    - Gene scan : genescan_crohn
    - P value sources : [META-P]
    - Test : No weights
    - Pathway DB : c2.cp.v4.0.symbols.gmt.gz
    - Candidate Gene : candidategeneset_cro
    - Show : Detailed Results

```
Calculating local pair-wise LD of SNP within genes on chromosome 22 ...
```

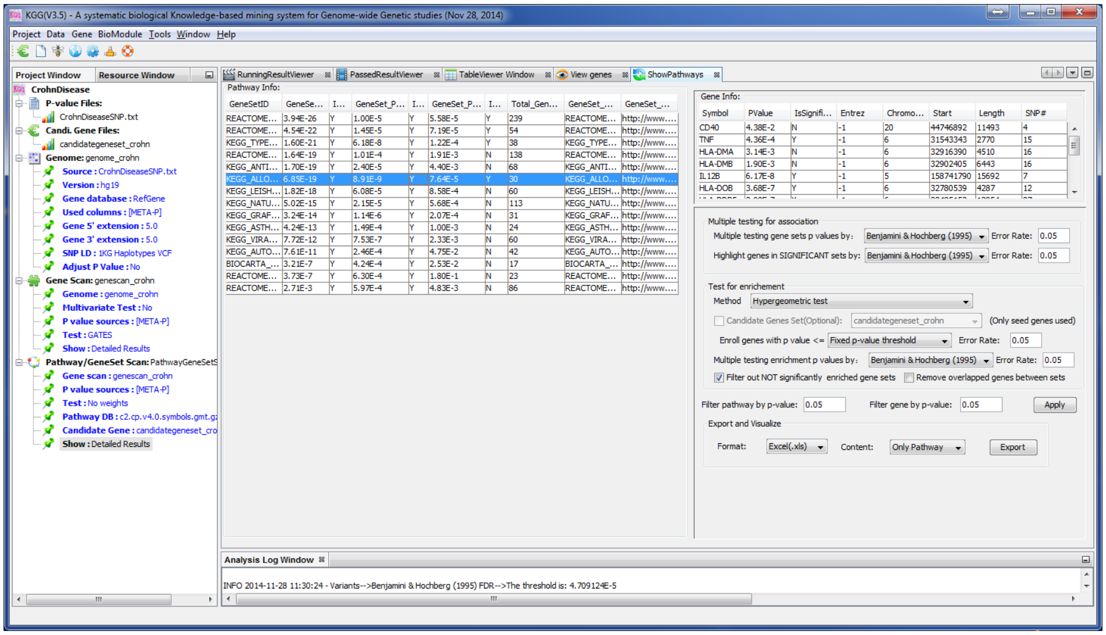
```
For association test based on HYbrid set-based test of GATES and scaled chi-square Test (HYST): The significance level for Benjamini & Hochberg (1995) FDR test
to control error rate. 0.05 for gene-sets is 1.22E-2. There are 314 significant gene-sets involving 4133 unique genes in total.
For enrichment test of selected risk genes based on Hypergeometric distribution test: The fixed p value threshold 0.05 for gene-sets is 5.00E-2.
For enrichment test of risk genes based on Wilcoxon signed rank test (one sided): The fixed p value threshold 0.05 for gene-sets is 5.00E-2.
In all of significantly associated gene sets (by HYST) and enriched gene sets (by Wilcoxon test), 48 genes are significant (p-value cutoff 7.36E-4 according to
Benjamini & Hochberg (1995) test for the overall error rate 0.05 [SLC22A4, CX3CR1, PRKAA1, SLC22A5, SLC35D1, UBA7, CUL2, HEXA, CREM, PIK3R2, PPP1R3A, IFI30,
AGPAT1, TNF, ICAM1, C4B, STIP1, C4A, PPP3CB, INPP5D, DAG1, IL12B, RPL37, JAK2, HLA-DOB, MICA, HLA-DRB5, TAP2, TAP1, ABCA7, RHOA, PSMB9, SELP, RNF123, IL3, NCR3,
TUBB2B, CDC34, TEC, NDUFS7, IRF1, PPP1R1B, SLCO2A1, LTA, TCAP, TAB2, LTB, PTPN2] ************************************************************************


Elapsed time: 3 min. 42 sec.
```

```
INFO 2014-11-28 11:30:24 - Variants-->Benjamini & Hochberg (1995) FDR-->The threshold is: 4.709124E-5
```

Figure 5.7.2 The display after pathway-based association scan

Step 8: for more detailed information of the result, you can click the node “Show Detailed Result” (Figure 5.8). You can also change the multiple test methods and export the results you want in this tab.


KGG(V3.5) - A systematic biological Knowledge-based mining system for Genome-wide Genetic studies (Nov 28, 2014)
Project Data Gene BioModule Tools Window Help
Project Window
Resource Window
RunningResultViewer
PassedResultViewer
TableViewer Window
View genes
ShowPathways
Apply
Export
Analysis Log Window

- CrohnDisease
  - P-value Files:
    - CrohnDiseaseSNP.txt
  - Candi. Gene Files:
    - candidategeneset_crohn
  - Genome: genome_crohn
    - Source : CrohnDiseaseSNP.txt
    - Version : hg19
    - Gene database : RefGene
    - Used columns : [META-P]
    - Gene 5' extension : 5.0
    - Gene 3' extension : 5.0
    - SNP LD : 1KG Haplotypes VCF
    - Adjust P Value : No
  - Gene Scan: genescan_crohn
    - Genome : genome_crohn
    - Multivariate Test : No
    - P value sources : [META-P]
    - Test : GATES
    - Show : Detailed Results
  - Pathway/GeneSet Scan: PathwayGeneSetS
    - Gene scan : genescan_crohn
    - P value sources : [META-P]
    - Test : No weights
    - Pathway DB : c2.cp.v4.0.symbols.gmt.gz
    - Candidate Gene : candidategeneset_cro
    - Show : Detailed Results

Pathway Info:

GeneSetID	GeneSe...	I...	GeneSet_P...	I...	GeneSet_P...	I...	Total_Gen...	GeneSet_...	GeneSet_...
REACTOME...	3.94E-26	Y	1.00E-5	Y	5.58E-5	Y	239	REACTOME...	http://www....
REACTOME...	4.54E-22	Y	1.45E-5	Y	7.19E-5	Y	54	REACTOME...	http://www....
KEGG_TYPE...	1.60E-21	Y	6.18E-8	Y	1.22E-4	Y	38	KEGG_TYPE...	http://www....
REACTOME...	1.64E-19	Y	1.01E-4	Y	1.91E-3	N	138	REACTOME...	http://www....
KEGG_ANTI...	1.70E-19	Y	2.40E-5	Y	4.40E-3	N	68	KEGG_ANTI...	http://www....
KEGG_ALLO...	6.85E-19	Y	8.91E-9	Y	7.64E-5	Y	30	KEGG_ALLO...	http://www....
KEGG_LEISH...	1.82E-18	Y	6.08E-5	Y	8.58E-4	N	60	KEGG_LEISH...	http://www....
KEGG_NATU...	5.02E-15	Y	2.15E-5	Y	5.68E-4	N	113	KEGG_NATU...	http://www....
KEGG_GRAF...	3.24E-14	Y	1.14E-6	Y	2.07E-4	N	31	KEGG_GRAF...	http://www....
KEGG_ASTH...	4.24E-13	Y	1.49E-4	Y	1.00E-3	N	24	KEGG_ASTH...	http://www....
KEGG_VIRA...	7.72E-12	Y	7.53E-7	Y	2.33E-3	N	60	KEGG_VIRA...	http://www....
KEGG_AUTO...	7.61E-11	Y	2.46E-4	Y	4.75E-2	N	42	KEGG_AUTO...	http://www....
BIOCARTA_...	3.21E-7	Y	4.24E-4	Y	2.53E-2	N	17	BIOCARTA_...	http://www....
REACTOME...	3.73E-7	Y	6.30E-4	Y	1.80E-1	N	23	REACTOME...	http://www....
REACTOME...	2.71E-3	Y	5.97E-4	Y	4.83E-3	N	86	REACTOME...	http://www....

Gene Info:

Symbol	PValue	IsSignifi...	Entrez	Chromo...	Start	Length	SNP#
CD40	4.38E-2	N	-1	20	44746892	11493	4
TNF	4.36E-4	Y	-1	6	31543343	2770	15
HLA-DMA	3.14E-3	N	-1	6	32916390	4510	16
HLA-DMB	1.90E-3	N	-1	6	32902405	6443	16
IL12B	6.17E-8	Y	-1	5	158741790	15692	7
HLA-DOB	3.68E-7	Y	-1	6	32780539	4287	12

Multiple testing for association

Multiple testing gene sets p values by: Benjamini & Hochberg (1995) Error Rate: 0.05

Highlight genes in SIGNIFICANT sets by: Benjamini & Hochberg (1995) Error Rate: 0.05

Test for enrichement

Method Hypergeometric test

☐ Candidate Genes Set(Optional): candidategeneset_crohn (Only seed genes used)

Enroll genes with p value <= Fixed p-value threshold Error Rate: 0.05

Multiple testing enrichment p values by: Benjamini & Hochberg (1995) Error Rate: 0.05

☑ Filter out NOT significantly enriched gene sets ☐ Remove overlapped genes between sets

Filter pathway by p-value: 0.05 Filter gene by p-value: 0.05

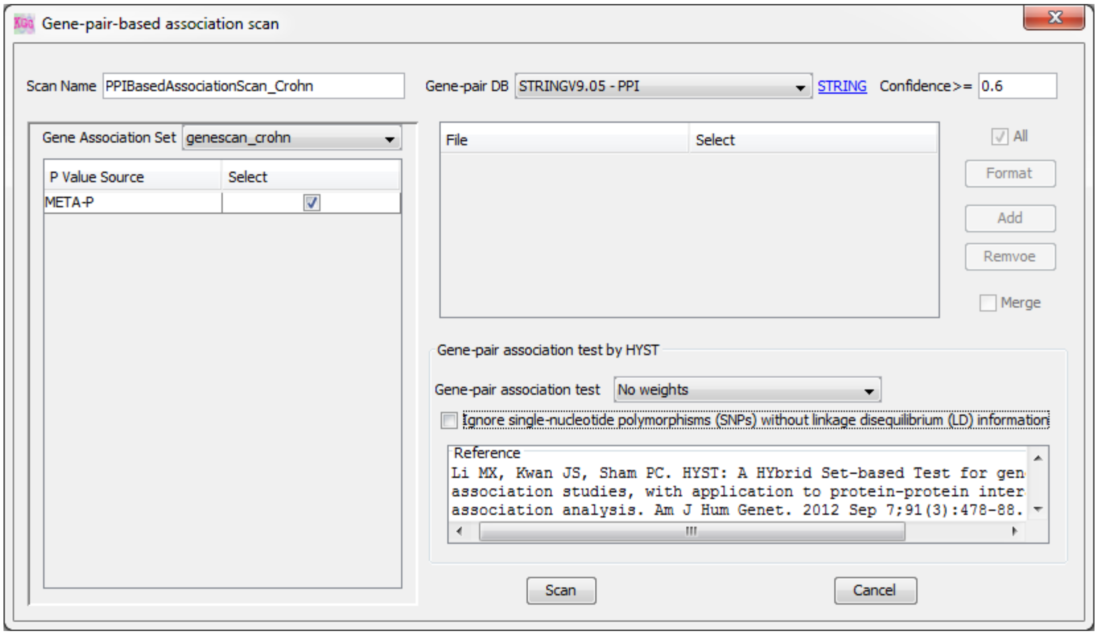
Export and Visualize

Format: Excel(.xls) Content: Only Pathway

```
INFO 2014-11-28 11:30:24 - Variants-->Benjamini & Hochberg (1995) FDR-->The threshold is: 4.709124E-5
```

Figure 5.8 Function of displaying the results of pathway-based analysis

Step9: search PPIs between significant genes. The significant genes can be picked up according to the gene p-values and SNP p-values; set as Figure 5.9.1; output as Figure 5.9.2



Gene-pair-based association scan

Scan Name: PPIBasedAssociationScan_Crohn

Gene-pair DB: STRINGV9.05 - PPI

Confidence>= 0.6

Gene Association Set: genescan_crohn

P Value Source	Select
META-P	☑

File	Select

Gene-pair association test by HYST

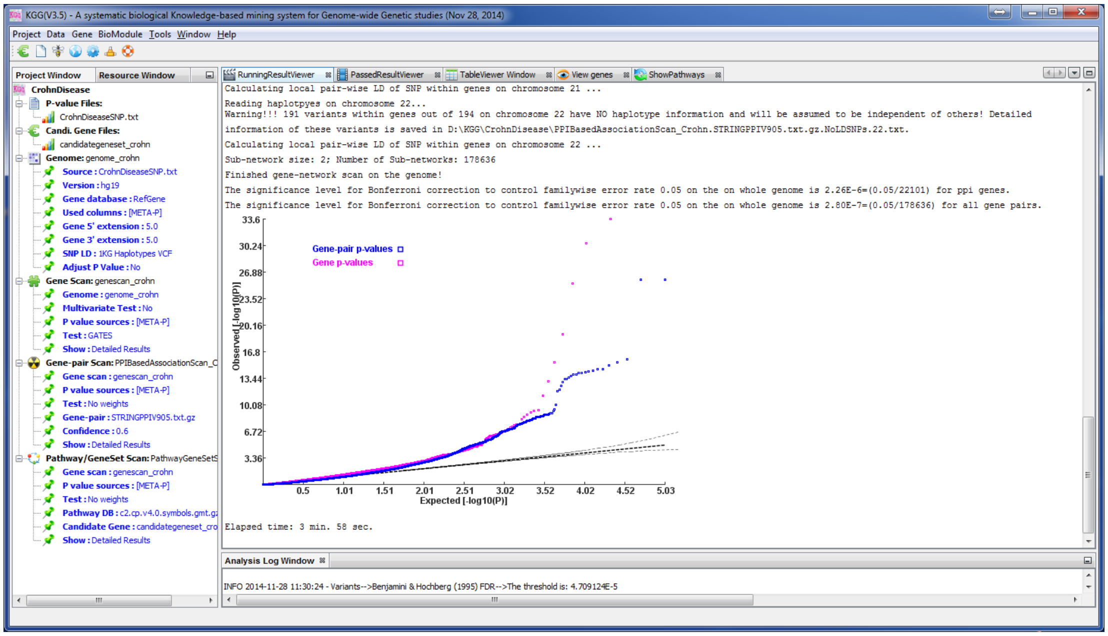
Gene-pair association test: No weights

□ Ignore single-nucleotide polymorphisms (SNPs) without linkage disequilibrium (LD) information

Reference

```
Li MX, Kwan JS, Sham PC. HYST: A HYbrid Set-based Test for gen
association studies, with application to protein-protein inter
association analysis. Am J Hum Genet. 2012 Sep 7;91(3):478-88.
```

Figure 5.9.1 PPI association scan by gene-based p-values



KGG(V3.5) - A systematic biological Knowledge-based mining system for Genome-wide Genetic studies (Nov 28, 2014)

- CrohnDisease
  - P-value Files:
    - CrohnDiseaseSNP.txt
  - Candi. Gene Files:
    - candidategeneset_crohn
  - Genome: genome_crohn
    - Source : CrohnDiseaseSNP.txt
    - Version : hg19
    - Gene database : RefGene
    - Used columns : [META-P]
    - Gene 5' extension : 5.0
    - Gene 3' extension : 5.0
    - SNP LD : 1KG Haplotypes VCF
    - Adjust P Value : No
  - Gene Scan: genescan_crohn
    - Genome : genome_crohn
    - Multivariate Test : No
    - P value sources : [META-P]
    - Test : GATES
    - Show : Detailed Results
  - Gene-pair Scan: PPIBasedAssociationScan_C
    - Gene scan : genescan_crohn
    - P value sources : [META-P]
    - Test : No weights
    - Gene-pair : STRINGPPIV905.txt.gz
    - Confidence : 0.6
    - Show : Detailed Results
  - Pathway/GeneSet Scan: PathwayGeneSetS
    - Gene scan : genescan_crohn
    - P value sources : [META-P]
    - Test : No weights
    - Pathway DB : c2.cp.v4.0.symbols.gmt.gz
    - Candidate Gene : candidategeneset_cro
    - Show : Detailed Results

```
Calculating local pair-wise LD of SNP within genes on chromosome 21 ...
Reading haplotpyes on chromosome 22...
Warning!!! 191 variants within genes out of 194 on chromosome 22 have NO haplotype information and will be assumed to be independent of others! Detailed
information of these variants is saved in D:\KGG\CrohnDisease\PPIBasedAssociationScan_Crohn.STRINGPPIV905.txt.gz.NoLDSNPs.22.txt.
Calculating local pair-wise LD of SNP within genes on chromosome 22 ...
Sub-network size: 2; Number of Sub-networks: 178636
Finished gene-network scan on the genome!
The significance level for Bonferroni correction to control familywise error rate 0.05 on the on whole genome is 2.26E-6=(0.05/22101) for ppi genes.
The significance level for Bonferroni correction to control familywise error rate 0.05 on the on whole genome is 2.80E-7=(0.05/178636) for all gene pairs.
```

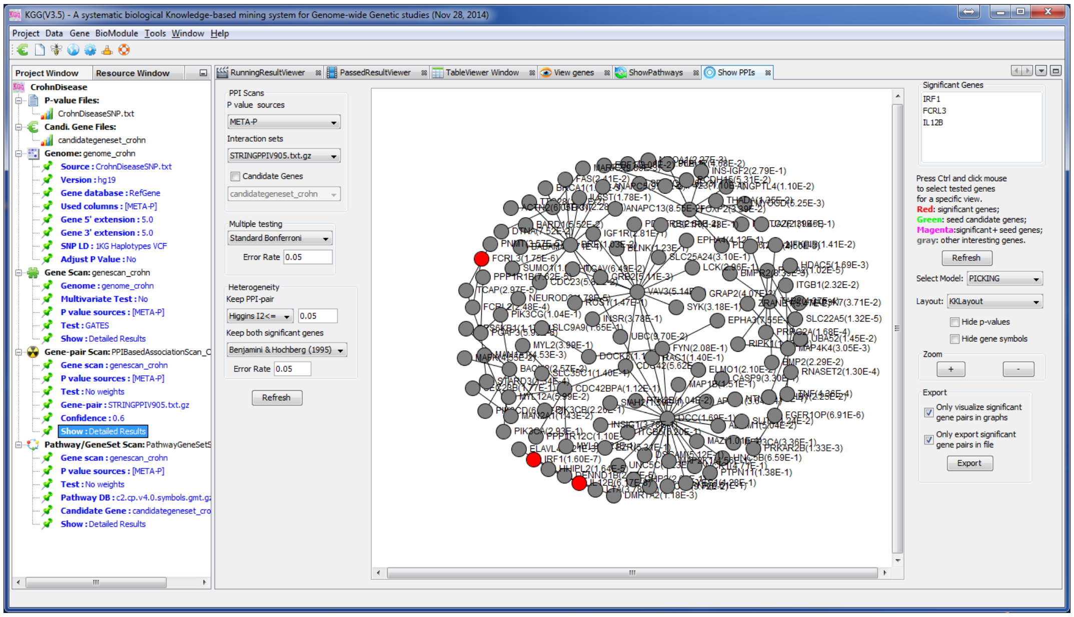
```
Elapsed time: 3 min. 58 sec.
```

Analysis Log Window

```
INFO 2014-11-28 11:30:24 - Variants-->Benjamini & Hochberg (1995) FDR-->The threshold is: 4.709124E-5
```

Figure 5.9.2 The display after running PPI-based association scan

Step 10: Click the node "Show: Detailed Results" and you will get the graph of PPI network. You can also export the results you want in this tab.


KGG(V3.5) - A systematic biological Knowledge-based mining system for Genome-wide Genetic studies (Nov 28, 2014)
Project Data Gene BioModule Tools Window Help
Project Window
Resource Window
RunningResultViewer
PassedResultViewer
TableViewer Window
View genes
ShowPathways
Show PPIs
Refresh
Zoom
+
-
Export

CrohnDisease
- P-value Files:
  - CrohnDiseaseSNP.txt
- Candi. Gene Files:
  - candidategeneset_crohn
- Genome: genome_crohn
  - Source : CrohnDiseaseSNP.txt
  - Version : hg19
  - Gene database : RefGene
  - Used columns : [META-P]
  - Gene 5' extension : 5.0
  - Gene 3' extension : 5.0
  - SNP LD : 1KG Haplotypes VCF
  - Adjust P Value : No
- Gene Scan: genescan_crohn
  - Genome : genome_crohn
  - Multivariate Test : No
  - P value sources : [META-P]
  - Test : GATES
  - Show : Detailed Results
- Gene-pair Scan: PPIBasedAssociationScan_C
  - Gene scan : genescan_crohn
  - P value sources : [META-P]
  - Test : No weights
  - Gene-pair : STRINGPPIV905.txt.gz
  - Confidence : 0.6
  - Show : Detailed Results
- Pathway/GeneSet Scan: PathwayGeneSetS
  - Gene scan : genescan_crohn
  - P value sources : [META-P]
  - Test : No weights
  - Pathway DB : c2.cp.v4.0.symbols.gmt.gz
  - Candidate Gene : candidategeneset_cro
  - Show : Detailed Results

PPI Scans
P value sources: META-P
Interaction sets: STRINGPPIV905.txt.gz
□ Candidate Genes: candidategeneset_crohn

Multiple testing: Standard Bonferroni
Error Rate: 0.05

Heterogeneity
Keep PPI-pair: Higgins I2<= 0.05
Keep both significant genes: Benjamini & Hochberg (1995)
Error Rate: 0.05

Significant Genes
IRF1
FCRL3
IL12B

Press Ctrl and click mouse to select tested genes for a specific view.
Red: significant genes;
Green: seed candidate genes;
Magenta:significant+ seed genes;
gray: other interesting genes.

Select Model: PICKING
Layout: KKLayout
□ Hide p-values
□ Hide gene symbols

Export
☑ Only visualize significant gene pairs in graphs
☑ Only export significant gene pairs in file

Figure 5.10 Function of displaying the results of PPI-based association scan

Step 12: View results of Crohn’s Disease

➢ By text file or Excel file

Open text or excel file for snp-based or gene-based analysis from local computer

➢ By Graphs

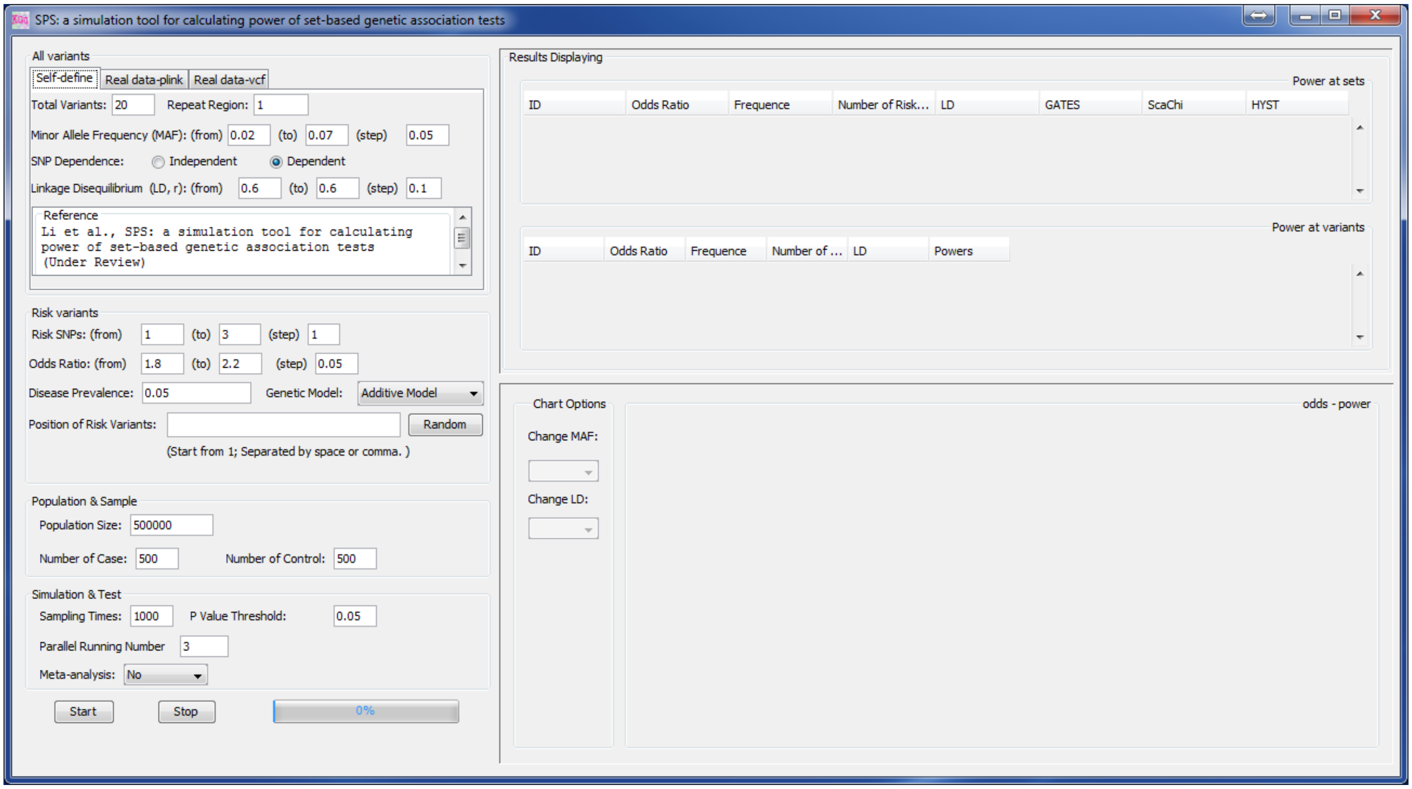
Check QQ plots and Manhattan plots saved in htmlLog folder

➢ By KGG Interface

Visualize pathway and PPI network output on KGG interface.

6. Power estimation of set-based tests by SPS

Step 1: Open the software and enter the main user interface on KGGV3.5 (Tools->Power Estimation). The interface is divided into two parts. The left one is used to set the basic parameters and the right one is to display the results.



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

All variants

Self-define | Real data-plink | Real data-vcf

Total Variants: 20 Repeat Region: 1

Minor Allele Frequency (MAF): (from) 0.02 (to) 0.07 (step) 0.05

SNP Dependence: ○ Independent ● Dependent

Linkage Disequilibrium (LD, r): (from) 0.6 (to) 0.6 (step) 0.1

Reference

Li et al., SPS: a simulation tool for calculating power of set-based genetic association tests (Under Review)

Risk variants

Risk SNPs: (from) 1 (to) 3 (step) 1

Odds Ratio: (from) 1.8 (to) 2.2 (step) 0.05

Disease Prevalence: 0.05 Genetic Model: Additive Model

Position of Risk Variants: Random

(Start from 1; Separated by space or comma.)

Population & Sample

Population Size: 500000

Number of Case: 500 Number of Control: 500

Simulation & Test

Sampling Times: 1000 P Value Threshold: 0.05

Parallel Running Number 3

Meta-analysis: No

Start Stop 0%

Results Displaying

Power at sets

ID	Odds Ratio	Frequence	Number of Risk...	LD	GATES	ScaChi	HYST

Power at variants

ID	Odds Ratio	Frequence	Number of ...	LD	Powers

Chart Options

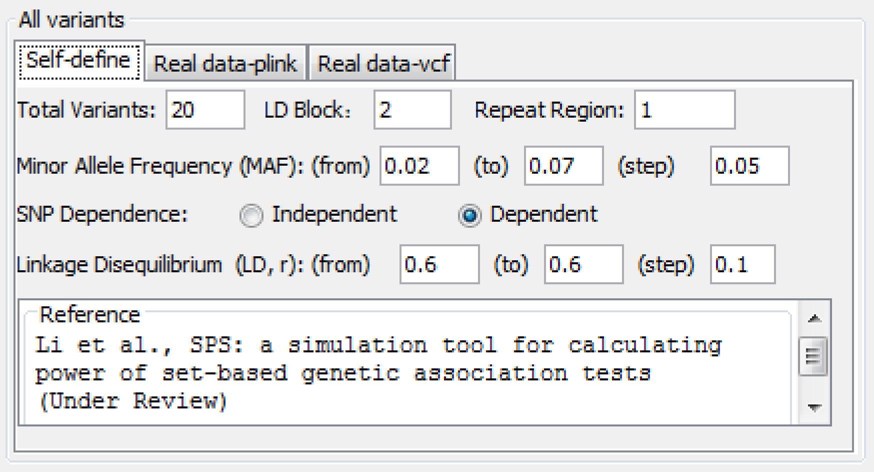
Change MAF:

Change LD:

odds - power

Figure 6.1 The Main interface of SPS.

Step 2: Set the parameters of all variants. The number of SNPs, the minor allele frequency (MAF) and LD information should be set here. When these SNP markers are divided into several LD blocks, the markers within the same LD block have the same LD as 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 replicate to make up of a larger marker set. Some of these parameters can vary within a certain region, such as MAF and LD, so that the users can investigate how the power will be affected by changing the critical parameters conveniently. Moreover, 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.



All variants

Self-define | Real data-plink | Real data-vcf

Total Variants: 20 LD Block: 2 Repeat Region: 1

Minor Allele Frequency (MAF): (from) 0.02 (to) 0.07 (step) 0.05

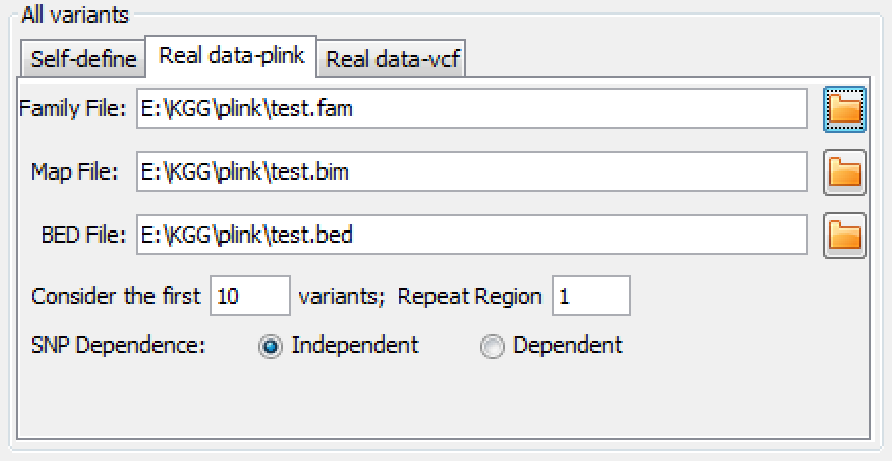
SNP Dependence: ○ Independent ● Dependent

Linkage Disequilibrium (LD, r): (from) 0.6 (to) 0.6 (step) 0.1

Reference

Li et al., SPS: a simulation tool for calculating power of set-based genetic association tests (Under Review)

Figure 6.2-1 Set parameters by users.



Family File: E:\KGG\plink\test.fam

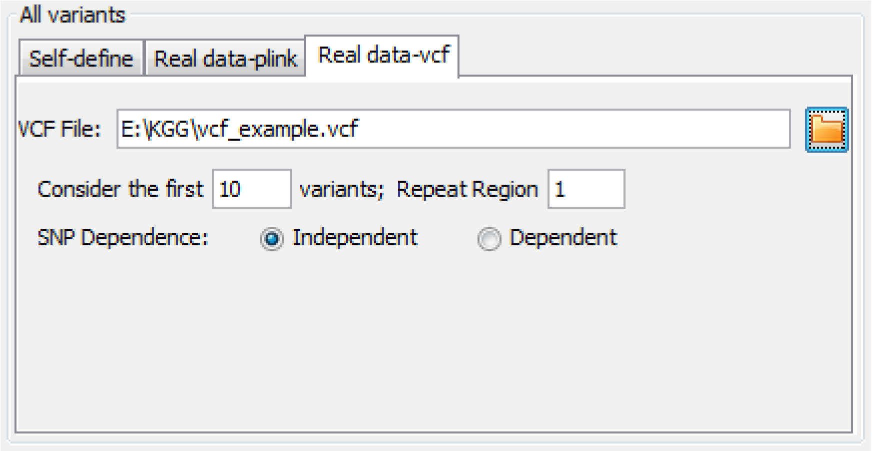
Map File: E:\KGG\plink\test.bim

BED File: E:\KGG\plink\test.bed

Consider the first 10 variants; Repeat Region 1

SNP Dependence: ● Independent ○ Dependent

Figure 6.2-2 Set parameters by plink file.



VCF File: E:\KGG\vcf_example.vcf

Consider the first 10 variants; Repeat Region 1

SNP Dependence: ● Independent ○ Dependent

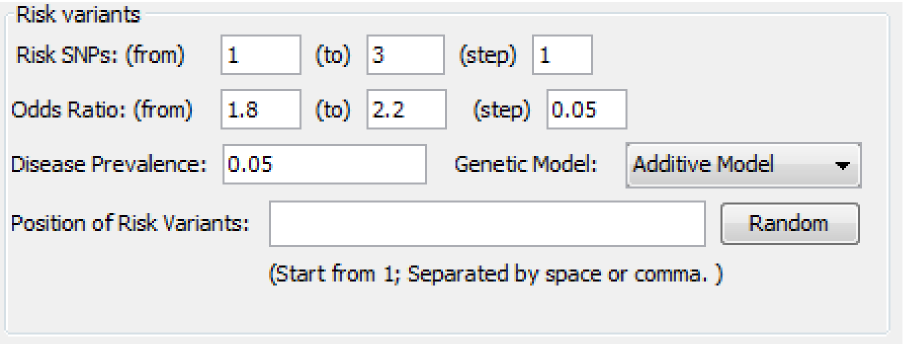
Figure 6.2-3 Set parameters by vcf file.

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.



Risk SNPs: (from) 1 (to) 3 (step) 1

Odds Ratio: (from) 1.8 (to) 2.2 (step) 0.05

Disease Prevalence: 0.05

Genetic Model: Additive Model

Position of Risk Variants:

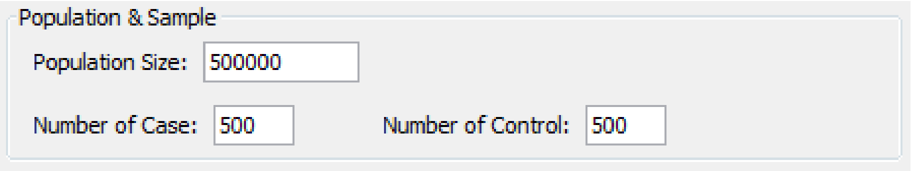
(Start from 1; Separated by space or comma.)

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.



Population Size: 500000

Number of Case: 500 Number of Control: 500

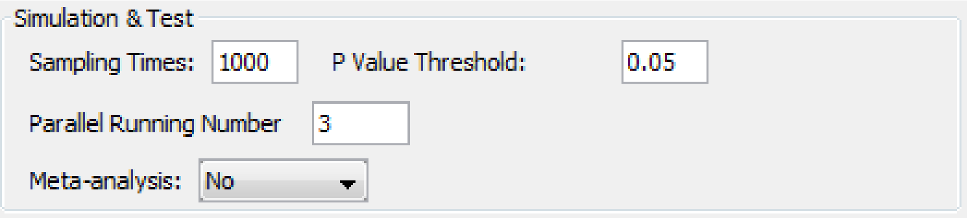
Figure 6.4. Set population and sample.

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 are subject to calculate the p value of the set-based test. The number of p values that pass the threshold will be counted to calculate the power. In order to speed up the simulation process, the user 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. When at variant level, the p values of variants in different studies will be combined using Fisher's Combination Test and these meta-p values in a set 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.



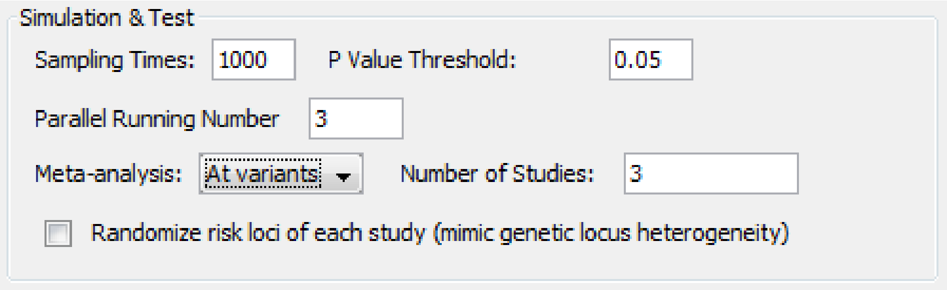
Sampling Times: 1000

P Value Threshold: 0.05

Parallel Running Number 3

Meta-analysis: No

Figure 6.5-1 Set simulation without meta-analysis.



Sampling Times: 1000

P Value Threshold: 0.05

Parallel Running Number 3

Meta-analysis: At variants

Number of Studies: 3

□ Randomize risk loci of each study (mimic genetic locus heterogeneity)

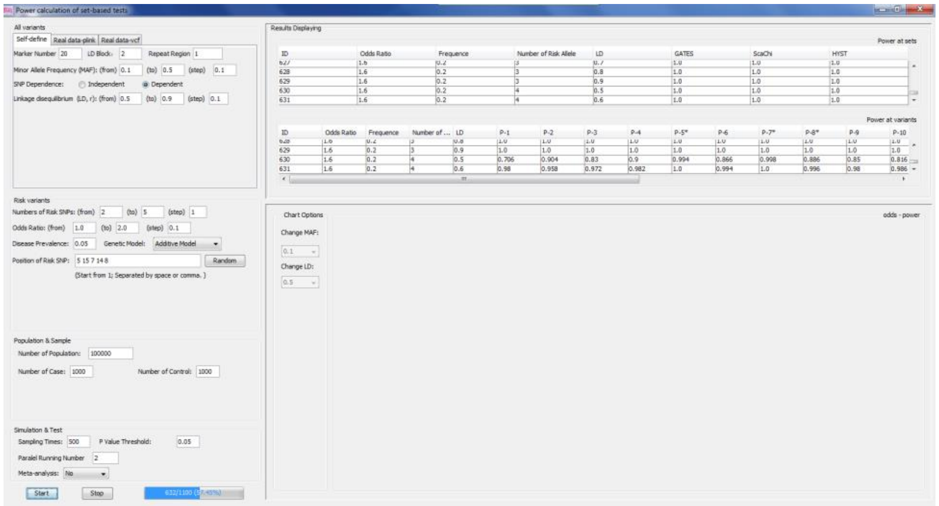
Figure 6.5-2 Set simulation with meta-analysis.

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 user can check the results from tables in the right part immediately. The progress bar can also provide the running information in a real time. If the user wants to stop the running program, just click the “stop” button.



All variants

Marker Number: 20 LD Block: 2 Repeat Region: 1

Minor Allele Frequency (MAF): (from) 0.1 (to) 0.5 (step) 0.1

SNP Dependence: ○ Independent ● Dependent

Linkage disequilibrium (LD, r): (from) 0.5 (to) 0.9 (step) 0.1

Risk variants

Numbers of Risk SNPs: (from) 2 (to) 5 (step) 1

Odds Ratio: (from) 1.0 (to) 2.0 (step) 0.1

Disease Prevalence: 0.05 Genetic Model: Additive Model

Position of Risk SNP: 5 15 7 14 8

(Start from 1; Separated by space or comma.)

Population & Sample

Number of Population: 100000

Number of Case: 1000 Number of Control: 1000

Simulation & Test

Sampling Times: 500 P Value Threshold: 0.05

Paralel Running Number 2

Meta-analysis: No

Results Displaying

Power at sets

ID	Odds Ratio	Frequence	Number of Risk Allele	LD	GATES	ScaChi	HYST
627	1.6	0.2	3	0.7	1.0	1.0	1.0
628	1.6	0.2	3	0.8	1.0	1.0	1.0
629	1.6	0.2	3	0.9	1.0	1.0	1.0
630	1.6	0.2	4	0.5	1.0	1.0	1.0
631	1.6	0.2	4	0.6	1.0	1.0	1.0

Power at variants

ID	Odds Ratio	Frequence	Number of ...	LD	P-1	P-2	P-3	P-4	P-5*	P-6	P-7*	P-8*	P-9	P-10
628	1.6	0.2	3	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
629	1.6	0.2	3	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
630	1.6	0.2	4	0.5	0.706	0.904	0.83	0.9	0.994	0.866	0.998	0.886	0.85	0.816
631	1.6	0.2	4	0.6	0.98	0.958	0.972	0.982	1.0	0.994	1.0	0.996	0.98	0.986

Chart Options

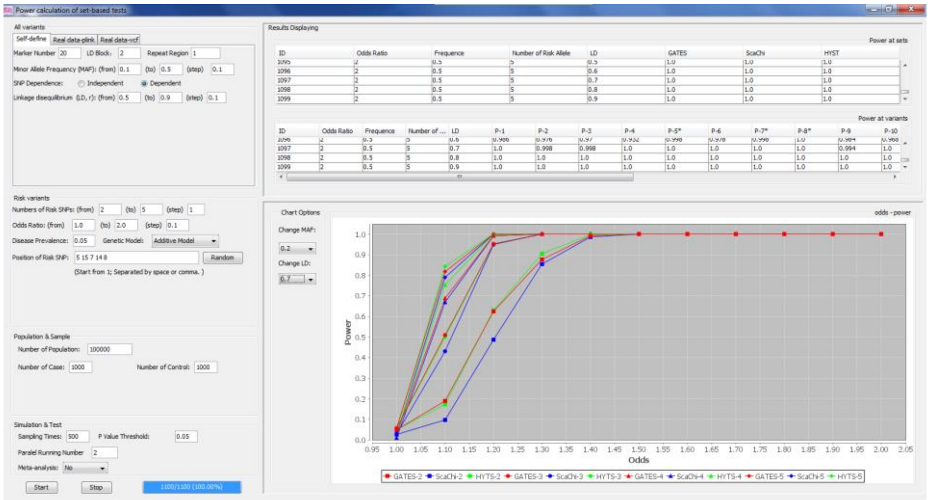
Change MAF: 0.1

Change LD: 0.5

odds - power

Figure 6.6 Run the program.

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



All variants

Marker Number: 20 LD Block: 2 Repeat Region: 1

Minor Allele Frequency (MAF): (from) 0.1 (to) 0.5 (step) 0.1

SNP Dependence: ○ Independent ● Dependent

Linkage disequilibrium (LD, r): (from) 0.5 (to) 0.9 (step) 0.1

Risk variants

Numbers of Risk SNPs: (from) 2 (to) 5 (step) 1

Odds Ratio: (from) 1.0 (to) 2.0 (step) 0.1

Disease Prevalence: 0.05 Genetic Model: Additive Model

Position of Risk SNP: 5 15 7 14 8

(Start from 1; Separated by space or comma.)

Population & Sample

Number of Population: 100000

Number of Case: 1000 Number of Control: 1000

Simulation & Test

Sampling Times: 500 P Value Threshold: 0.05

Paralel Running Number 2

Meta-analysis: No

Results Displaying

Power at sets

ID	Odds Ratio	Frequence	Number of Risk Allele	LD	GATES	ScaChi	HYST
1095	2	0.5	5	0.5	1.0	1.0	1.0
1096	2	0.5	5	0.6	1.0	1.0	1.0
1097	2	0.5	5	0.7	1.0	1.0	1.0
1098	2	0.5	5	0.8	1.0	1.0	1.0
1099	2	0.5	5	0.9	1.0	1.0	1.0

Power at variants

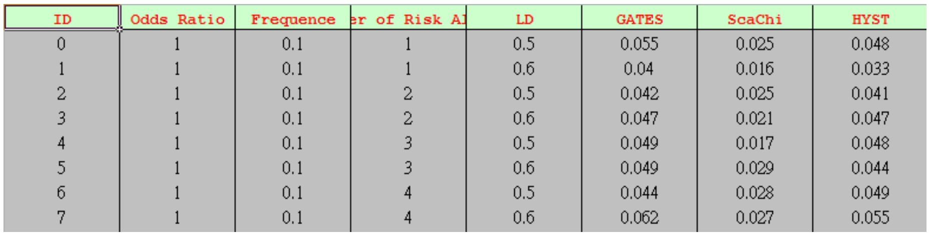
ID	Odds Ratio	Frequence	Number of ...	LD	P-1	P-2	P-3	P-4	P-5*	P-6	P-7*	P-8*	P-9	P-10
1096	2	0.5	5	0.6	0.986	0.976	0.97	0.932	0.998	0.978	0.998	1.0	0.984	0.968
1097	2	0.5	5	0.7	1.0	0.998	0.998	1.0	1.0	1.0	1.0	1.0	0.994	1.0
1098	2	0.5	5	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1099	2	0.5	5	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Chart Options

Change MAF: 0.2

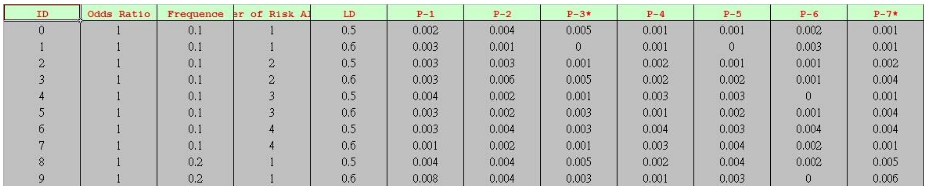
Change LD: 0.7

Figure 6.7-1. The output of SPS.



ID	Odds Ratio	Frequence	er of Risk Al	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 Al	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.

7. Update from KGG 3.0 to KGG 3.5

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

1) Multivariate gene-based association analysis;
2) Direct link to multiple bioinformatics annotation databases;
3) Simplified operation; better plotting function;
4) Integrate regulatory information to prioritize risk genes (under development).
5) SPS plug-in is included.