# IPGWAS: An Integrated Pipeline for Genome-Wide Association Studies

**User Manual**

**Yan-Hui Fan, You-Qiang Song**
**nolanfyh@gmail.com, songy@hkucc.hku.hk**

**Department of Biochemistry**
**Li Ka Shing Faculty of Medicine**
**The University of Hong Kong**
**21 Sassoon Road, Pokfulam, Hong Kong SAR, China**

Citation: Fan YH, Song YQ. IPGWAS: An integrated pipeline for rational quality control and association analysis of genome-wide genetic studies. Biochemical and Biophysical Research Communications, 2012, 422(3):363-368

- Jan 2014 -

# Content

# 1 Introduction

IPGWAS is an integrated pipeline for Genome-Wide Association Studies. It integrated quality control, combining different dataset, Manhattan and QQ plot, Cochran-Armitage trend test, and format converting for downstream analysis.

# 2 Installation

## 2.1 Windows OS

For Windows, download ActivePerl and install it use the default opinions, Perl will be installed under C:\Perl. The newest version of ActivePerl is available at http://www.activestate.com/activeperl/downloads. Open an MS-DOS window and type the following commands to install Perl/Tk and other Perl modules which were used in IPGWAS.

**ppm install Tk**
**cpan install Math::CDF**

## 2.2 Linux OS and MAC OS

For Linux and Mac OS, Perl is usually already installed. Type perl -v at the command line to find out which version. You can get the source of the latest stable realease of Perl from http://www.perl.org/get.html.

(1) Install Perl/Tk
For Linux OS Ubuntu 10.04, you may need to install X11 first.
**sudo apt-get install libx11-dev**
**sudo cpan**
**install Tk**

For Mac OS, you must have X11 installed. X11 is available at http://xquartz.macosforge.org/trac/wiki/X112.5.3 for both Leopard and SnowLeopard. Then install precompiled 1.2.27 Mac OS X binaries http://ethan.tira-thompson.org/Mac_OS_X_Ports.html

**sudo cpan**
**force install Tk**

(2) Install other Perl modules (Linux and Mac)
**sudo cpan**
**install Math::CDF**
**install Archive::Extract**

**install Archive::Tar**
**install LWP::Simple**

(3) Install GD::Graph::histogram module

For Linux OS Ubuntu 10.04,
**sudo apt-get install libgd2-xpm-dev**
**sudo cpan**
**install GD::Graph::histogram**
For Mac OS, it is little complicated to compile libgd and dependent libs before install GD::Graph::histogram module. First, download all source packages.

download libpng from http://www.libpng.org/pub/png/libpng.html
download zlib from http://www.zlib.net/
download libjpeg from http://www.ijg.org/
download freetype2 from http://sourceforge.net/projects/freetype/
download libgd from http://www.libgd.org/Downloads

Then, extract and Compile the dependent libraries. The compile step is in most cases:

**./configure**
**make**
**make install**

After finished all these steps, you can install GD::Graph::histogram by CPAN [1] now.

**sudo cpan**
**install GD::Graph**
**install GD::Graph::histogram**

**Note:**
The precompiled PLINK[2] (http://pngu.mgh.harvard.edu/~purcell/plink/) and Gnuplot[3] (http://www.gnuplot.info/) will be downloaded to the bin directory. If the precompiled PLINK and/or Gnuplot do not work on your OS, please download the source code and recompile it.

# 3 Run IPGWAS

For Windows OS, you can just double click **ipgwas_win.bat** to run IPGWAS or open an MS-DOS window and go to the bin directory and type **perl ipgwas_win.pl** to run IPGWAS.

For Linux OS, open a terminal and go to the ipgwas directory:

**chmod +x ipgwas_linux.sh**
**./ipgwas_linux.sh**

Or for MAC OS, open a terminal and go to the ipgwas directory:

**chmod +x ipgwas_mac.sh**
**./ipgwas_mac.sh**

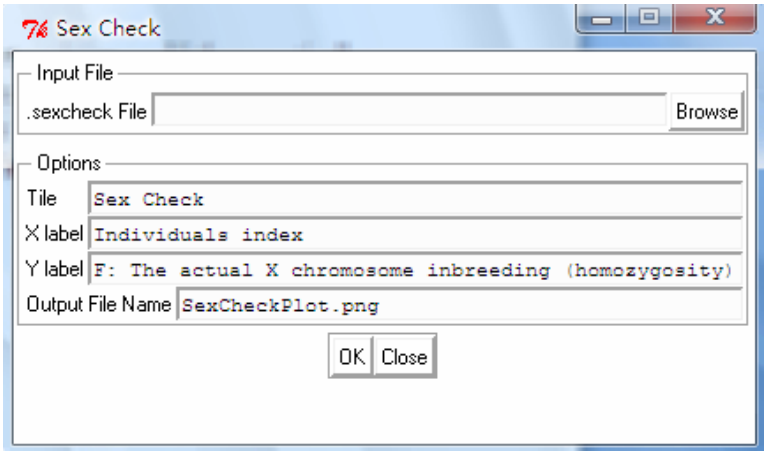
# 4 IGWAS Functions

## 4.1 Quality Control

This section introduces methods used to identify and remove individuals and SNPs.

### 4.1.1 Individual Quality Control

#### 4.1.1.1 Gender Check




PLINK use X chromosome data to determine sex. If the genetically gender is different from the phenotypically gender, the individual will be labeled as "PROBLEM". Gender mismatches reflect real rare medical conditions or labeling error. The gender mismatch individuals should be excluded.

Gender Check use the output of section 3.1.2.5 as input to plot the actual X chromosome inbreeding (homozygosity) estimate F, and write the FID and IID of "PROBLEM" individuals to "misMatchSexIndividuals.txt", then you can remove these individuals by using the function described in section 3.1.1.12. Figure 1 indicates 2 males were labeled as females.

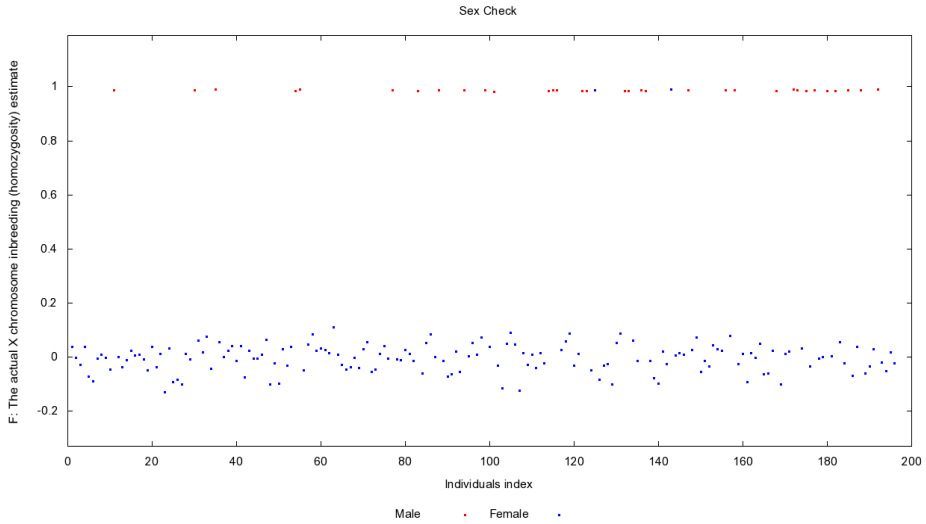
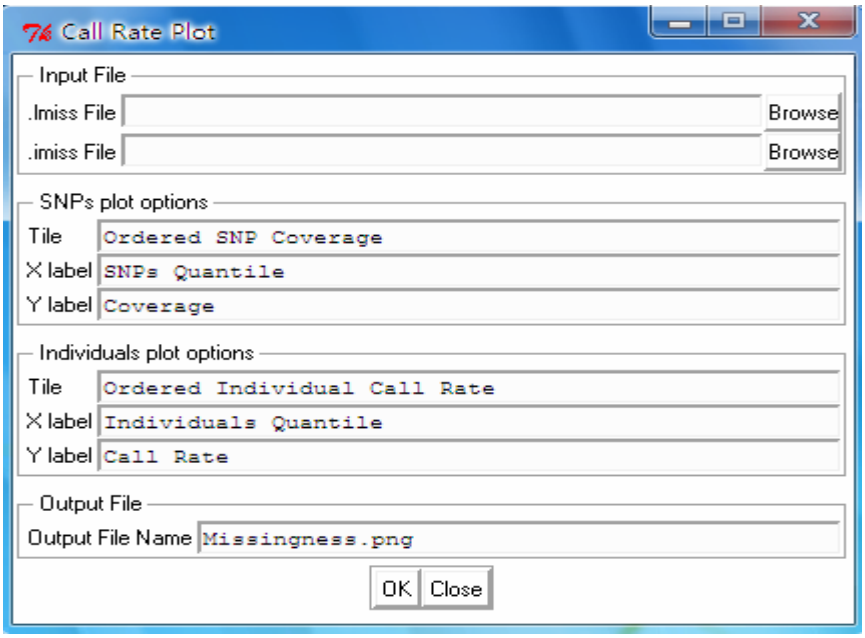



Figure 1 Example of gender check plot.

## 4.1.1.2 Missingness and/or Heterozygosity

### 4.1.1.2.1 Missingness Check




Missingness can cause both false positives and false negatives. This section plots the call rate/ coverage (1-missingness) against cumulative frequency of individuals and SNPs by using the output of section 3.1.2.1. The “elbow” in these figures denotes the QC threshold.

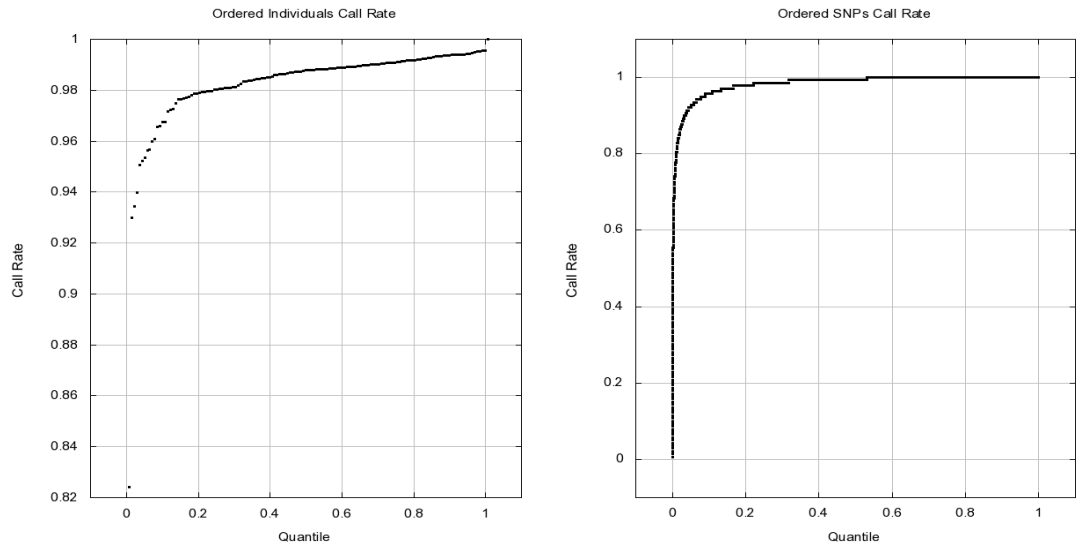
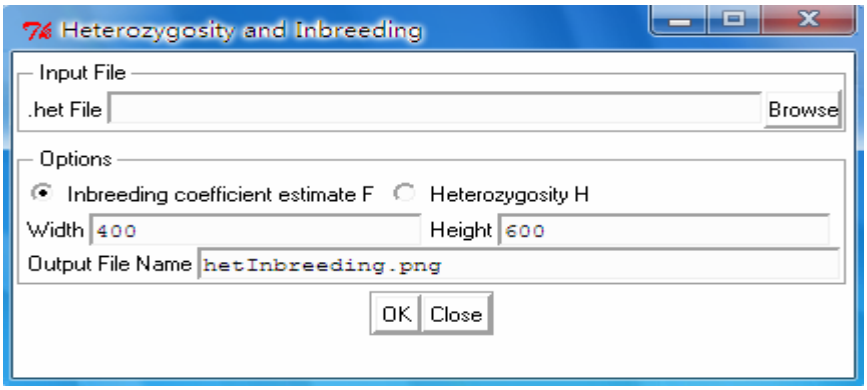
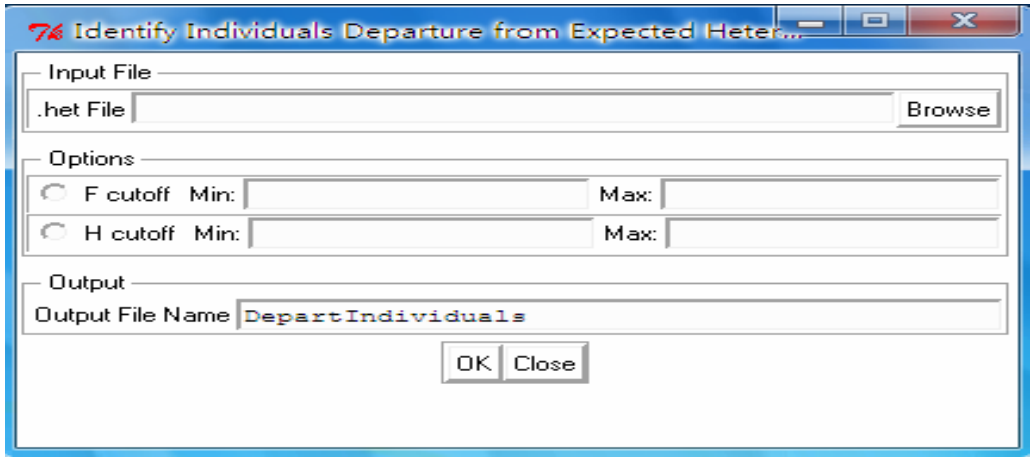



Figure 2 Example of call rate plot.

## 4.1.1.2.2 Heterozygosity and Inbreeding




This section plots the histograms of heterozygosity (H) and an inversely related value F (inbreeding coefficient estimate). Individuals with unusually high H and low F suggest possible sample contamination. Individuals with unusually low H and high F suggest a sample that does not belong in this population. Such individuals can be identified by using the cutoff of H or F.



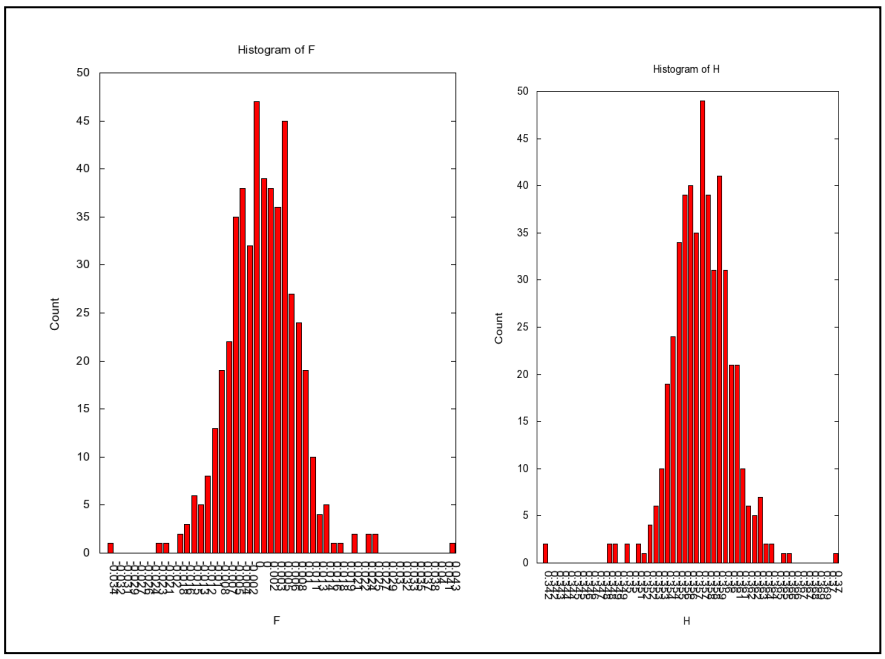
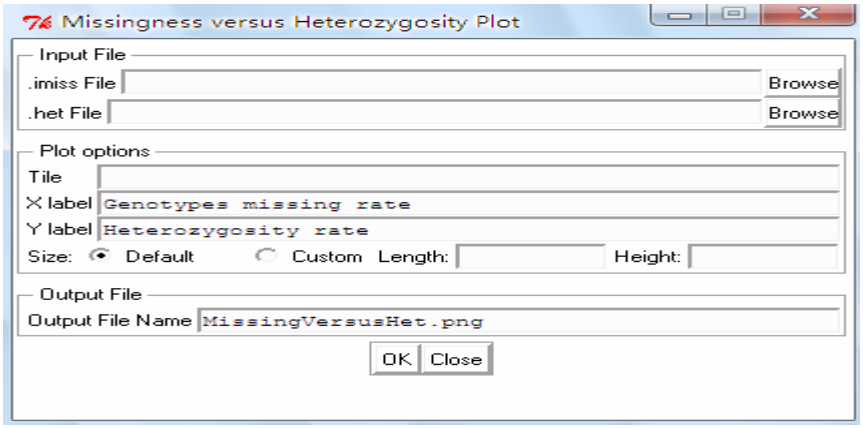




Figure 3 Example of histograms of F and H.

## 4.1.1.2.3 Missingness versus Heterozygosity




Create a graph in which the proportion of missing rate of each individual is plotted on

the x axis and the observed heterozygosity rate of each individual is plotted on y axis (Figure 4). Examine the plot to decide the reasonable missing rate and heterozygosity rate threshold to exclude individuals with high missing rate and extreme heterozygosity rate.

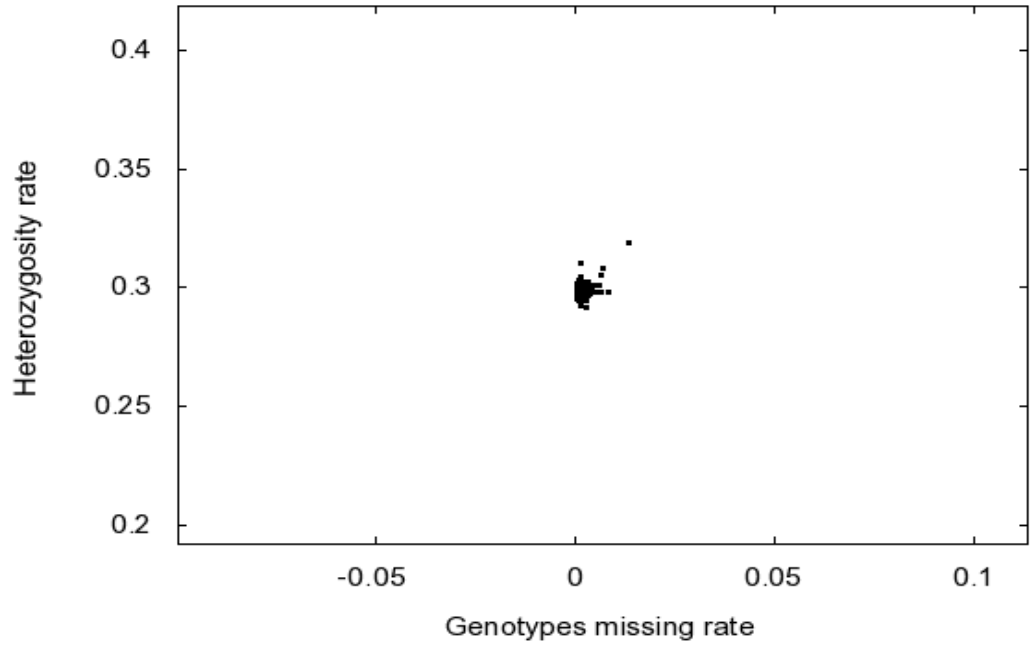
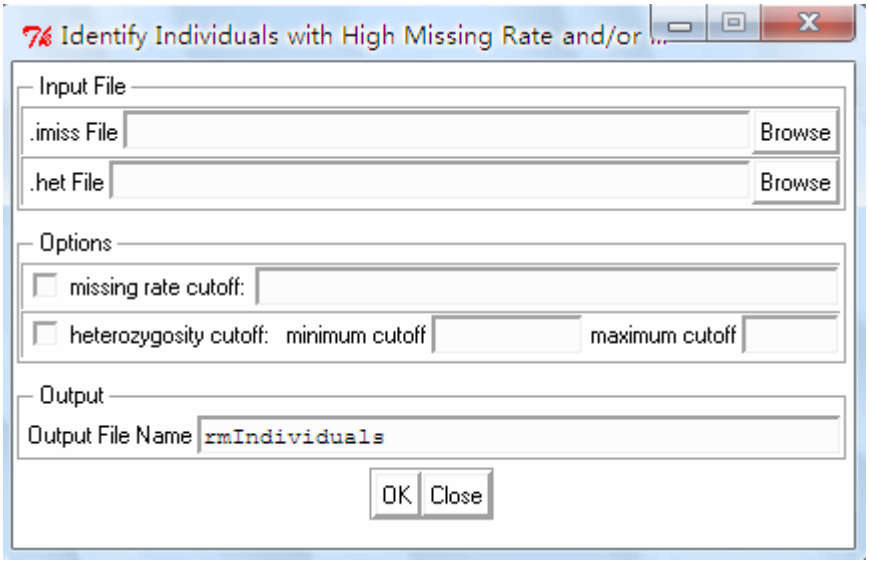
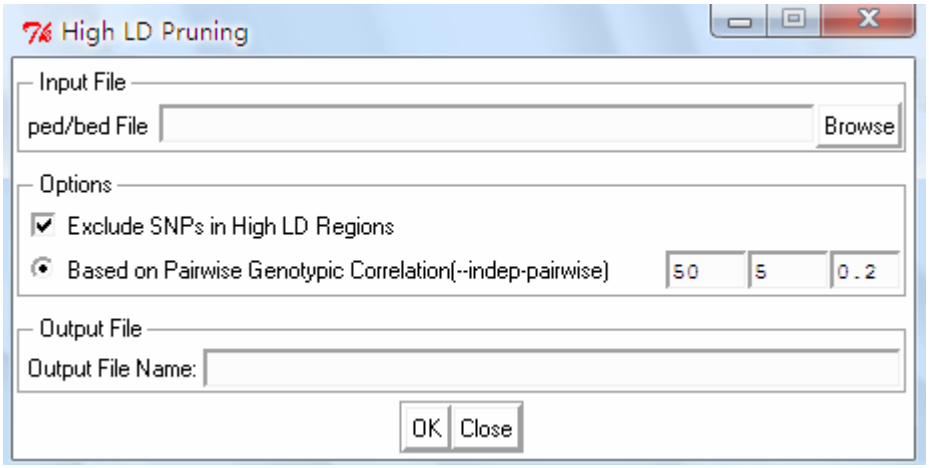



Figure 4 Example of genotypes missing rate versus heterozygosity rate plot



### 4.1.1.3 High LD Pruning



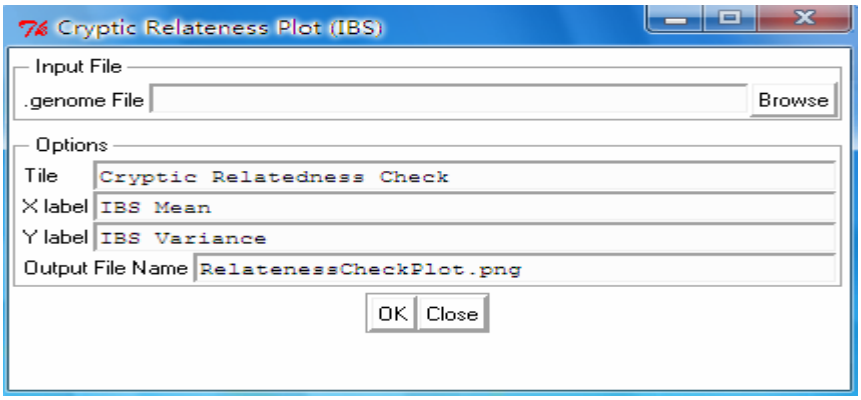

Since the procedures for identifying cryptic relatedness and population outliers work best under an assumption of no LD among SNPs, this section run "--indep-pairwise" option of PLINK [2] to thin the data and also exclude SNPs in high LD regions [4].

### 4.1.1.4 Cryptic relatedness Check

Cryptic relatedness between individuals induces a correlation structure that may introduce false positive and/or false negative results on downstream association analysis. This section provides two methods to identify the Cryptic related individuals.

**Note:** Since the cryptic relatedness procedures work best under an assumption of no LD among SNPs, it is useful to prepare an LD-pruned dataset (see 3.2.1.3 use the --indep-pairwise option with window size 1500, step size 150 and pairwise $r^2$ value 0.2) before run the --genome option.

#### 4.1.1.4.1 Identify Cryptic relatedness by mean-variance pair of IBS



To calculate the mean and variance of identical-by-state (IBS), when run PLINK with --genome option, the --genome-full option should also be selected. If the sample is homogeneous and independent, we expect each individual pair display similar pattern of allele sharing. In the mean-variance plot, the individual pairs will grouped together as a cluster, the individual pairs out of this cluster display different pattern of allele sharing [12]. Figure 5 shows several individual pairs have higher IBS means, which means that these individual pairs are more similar than other individual pairs.

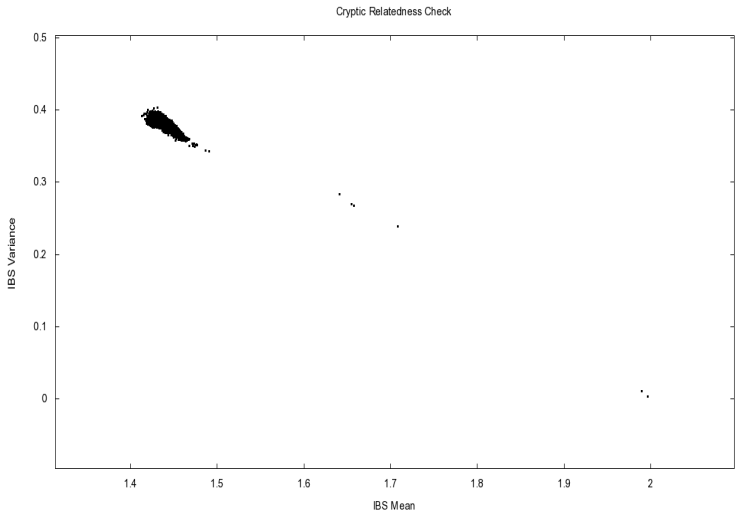
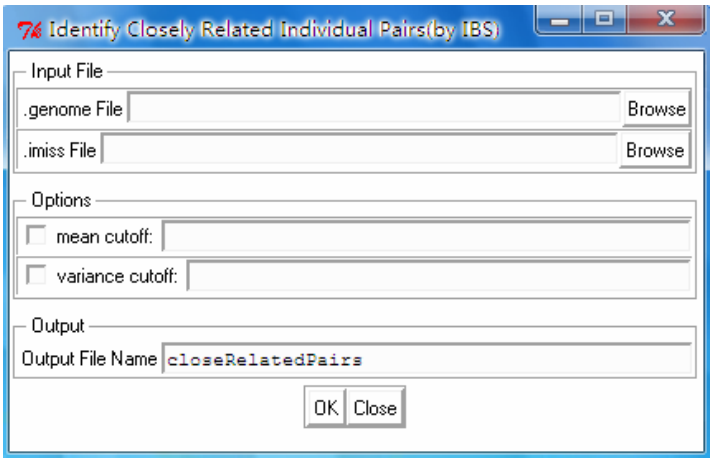



Figure 5 Example of mean-variance of IBS plot.

Based on the mean-variance of IBS plot, the cutoff for IBS mean and/ or variance can be used to identify which individual should be removed in each too similar individual pairs (remove individual with higher missing rate). The individual missing file *.imiss (See 3.1.2.1) should be provided to identify the individual with higher missing rate in the too similar individual pairs.



### 4.1.1.4.2 Identify Cryptic relatedness by IBD estimates (PI_HAT)

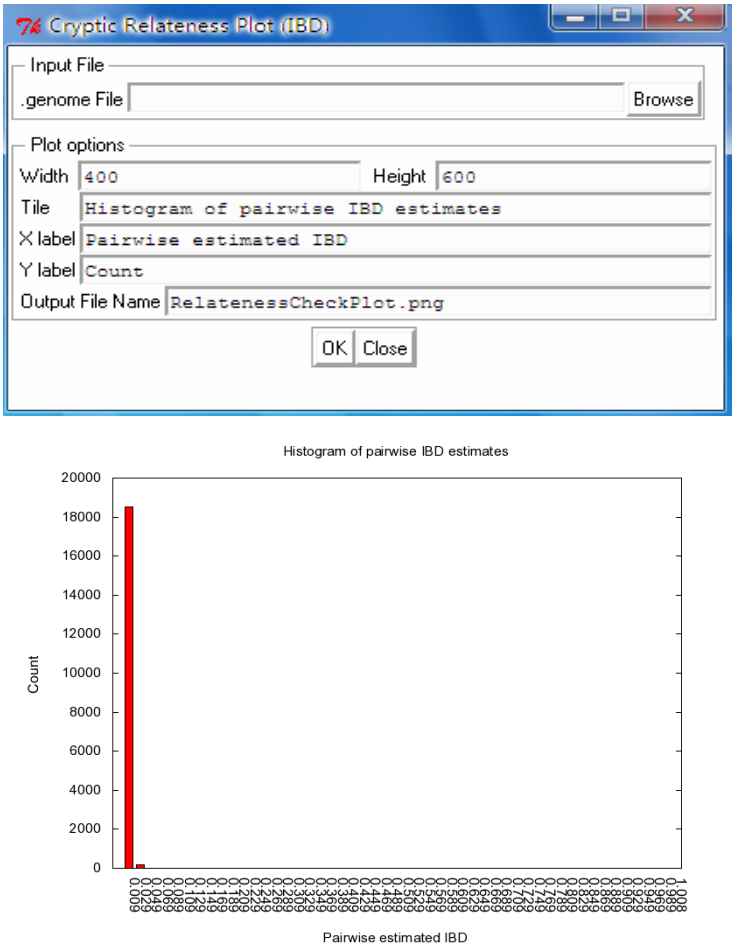
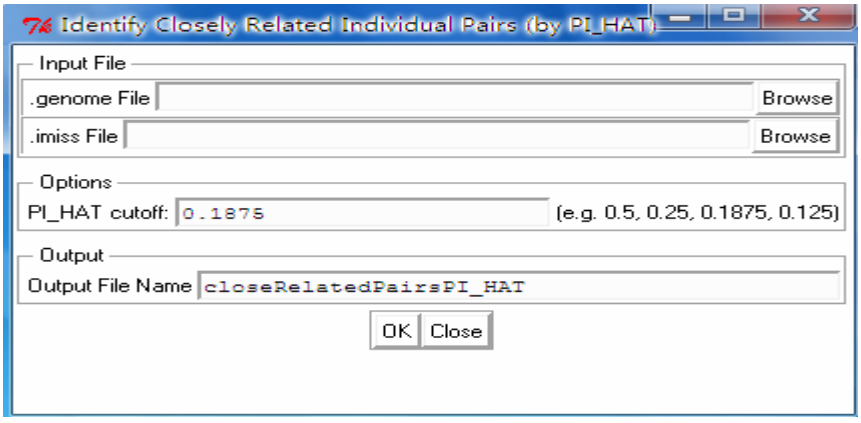



Figure 6 Example of histogram of pairwise IBD estimates.

If most individuals are unrelated but from a random homogenous population, PI_HAT close to 1 indicates a sample duplicate or MZ twins, PI_HAT close to 0.5 indicates 1st degree relatives, PI_HAT close to 0.25 indicates 2nd degree relatives, PI_HAT close to 0.125 indicates 3rd degree relatives. A good cutoff to use would be the half-way point between 2nd and 3rd degree relatives (PI_HAT>0.1875).



#### 4.1.1.5 Identification of Population Outliers

Genome-wide association studies assume samples are unrelated, so the individuals appear to have population ancestry from outside the population ancestry of most of the dataset should be removed. The most common method for indentifying population outliers is principal component analysis (PCA), for details please refer to EIGENSTRAT software [5].

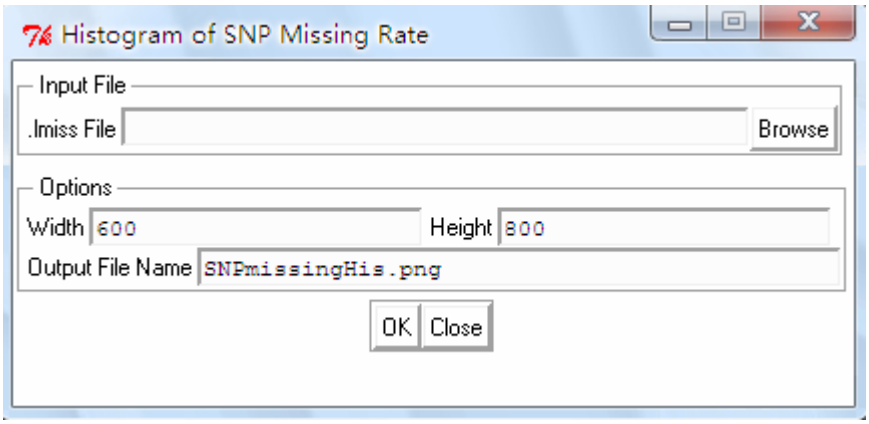
### 4.1.2 SNP Quality Control

#### 4.1.2.1 Missingness Genotype

##### 4.1.2.1.1 Missingness Check

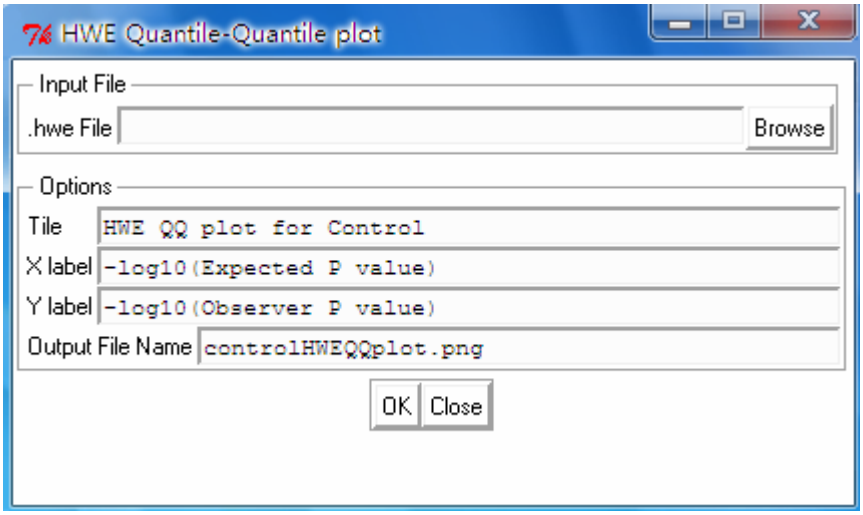
See 3.2.1.2.1 section.

##### 4.1.2.1.2 Histogram of SNPs Missing Genotype Rate




Plot a histogram of the SNP missing genotype rate by using the *.lmiss file, which was generated by "--missing" option of PLINK (see 3.1.2.1 section), to identify the threshold for extreme genotype missing rate.

### 4.1.2.2 Hardy-Weinberg Equilibrium




SNP departure from Hardy-Weinberg equilibrium (HWE) can indicates problems with genotype calling or a real strong signal of association. In controls, such SNPs usually were excluded from the downstream association analysis. While in cases, these SNPs were included and the HWE information was used to refer back for post-association QC.

This section generate a QQ plot of HWE p value (-log(p)) by using the output from section 3.1.3.4. Where the points lift above the 1:1 diagonal suggests a good cutoff. Figure 7 shows that –log(p)=2.3 (this corresponds to p value below 0.005) should be used as threshold to exclude SNPs that departure from HWE.

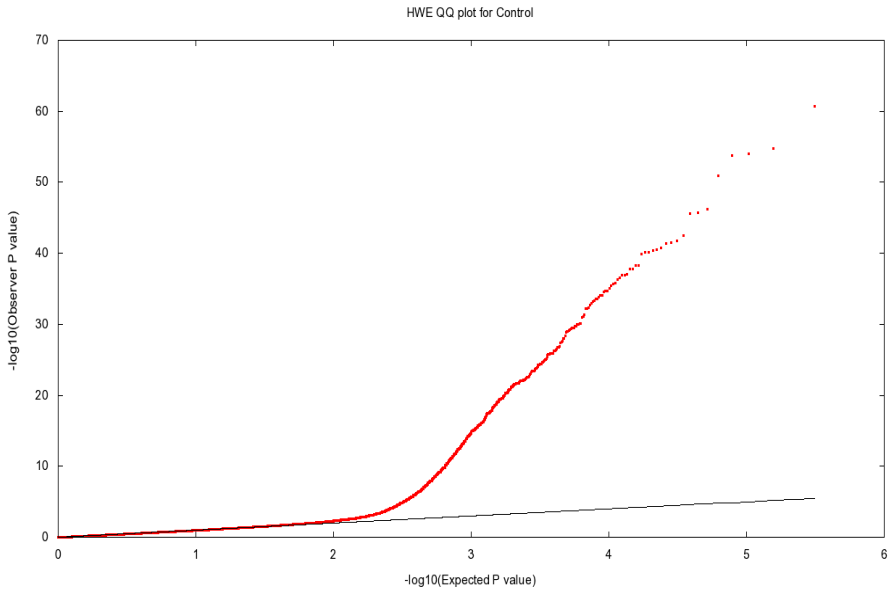
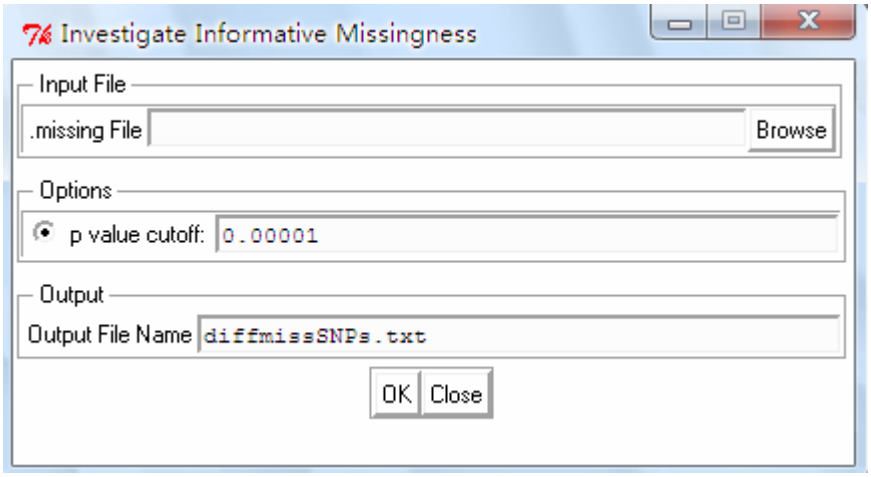



Figure 7 Example of HWE p values QQ plot.

### 4.1.2.3 Informative Missingness




The informative (non-random) missingness respect to phenotype (e.g. more missing data in cases than in controls, vice versa) can lead to false signal of association. Especially in studies which cases and/or controls have been drawn from different sources. The "--test-missing" option of PLINK (see 3.1.2.1 section) giving the p value for chi-square test of missingness difference between cases and controls. The default p value cutoff is $1x10^{-5}$.

### 4.1.2.4 Minor Allele Frequency (MAF)

Typically, a MAF threshold of 1%-5% is applied. However studies with small sample size may require a higher threshold. MAF > 10/n, where n is the number of samples, was a reasonable threshold.

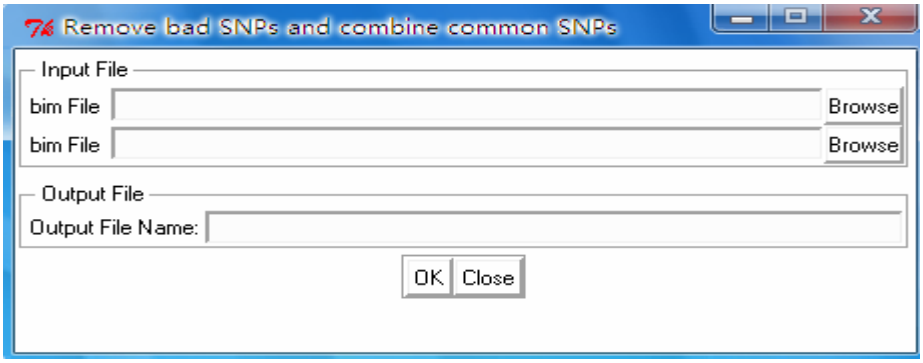
**QC note:**
Checking the genotype cluster plots manually is the best way to ensure the genotype calls are robust. It is essential to inspect the cluster plots for the association SNPs, especially, the singleton association SNPs, before the replication studies.

## 4.2 Combine datasets

It is necessary to combine GWAS datasets when cases and controls were genotyped on different genotyping platforms, when adjusting population stratification by using EIGENSTART with reference population and carrying out imputation of SNPs genotype with reference populations. There are two ways to combine datasets: (1) throw away all symmetric SNPs (A/T and C/G), and (2) align the SNPs in different datasets to a reference human genome, thence infer strand information.

**Note**: assuming all alleles of SNP were code as A/C/G/T format.

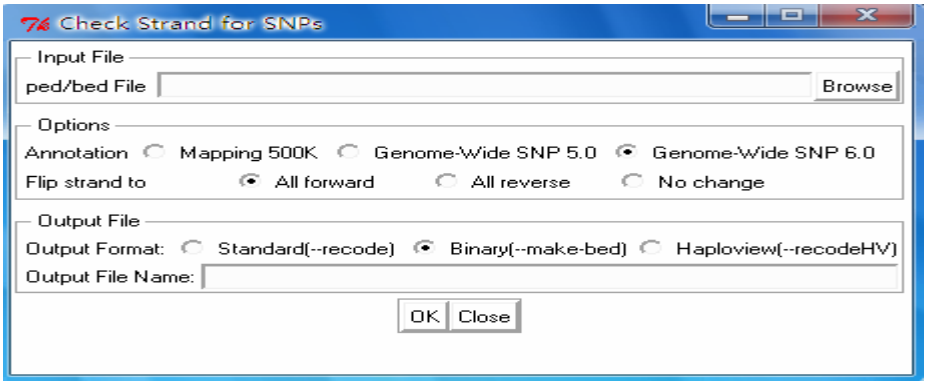
### 4.2.1 Remove bad SNPs



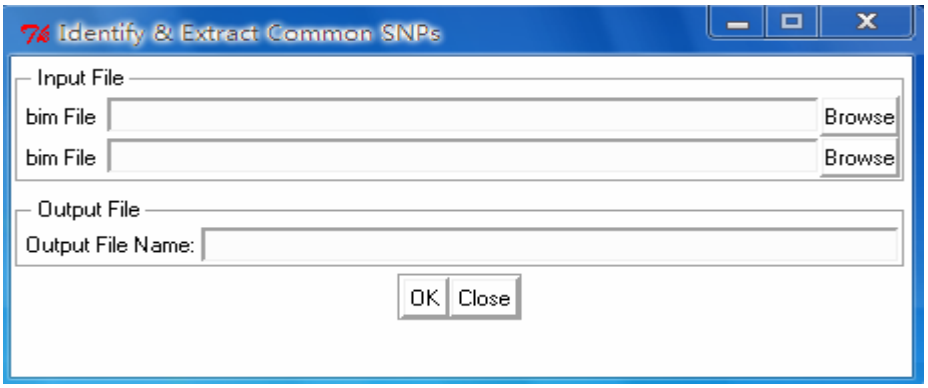

First, throw away all symmetric SNPs. Then identify the common SNPs in the two datasets, extract and merge the common SNPs. Symmetric SNPs list will be stored in "yourdataname_badSNPs.txt", and the common SNPs list will be stored in "yourdataname1. yourdataname2_commonSNPs.txt".

**Note:** make sure SNPs in two datasets have the same genomic physical locations.

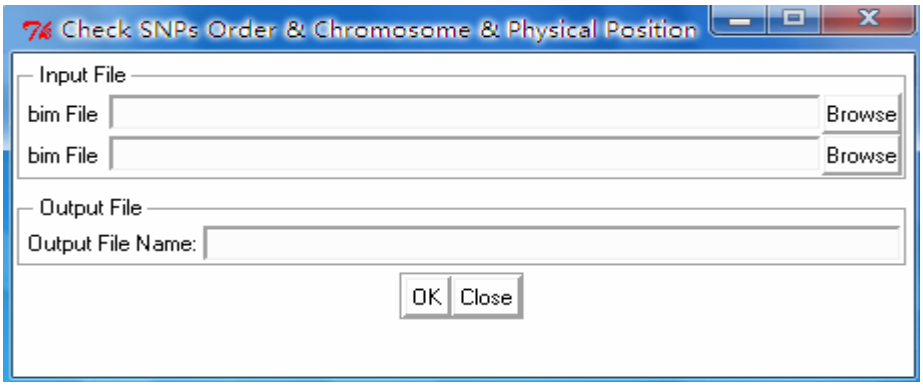
### 4.2.2 Check Strand




For Affymetrix datasets, the annotation files can be used to assign all SNPs to "forward" or "reverse" strand in the reference genome. Then, identify and extract the common SNPs in both datasets.

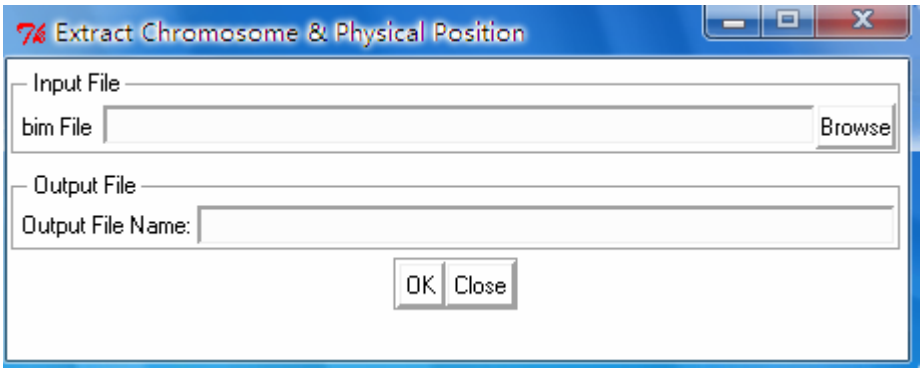


### 4.2.3 Check SNP information




Before combine the common SNPs datasets, check that whether SNPs in both datasets are in exactly the same order. If not, the SNPs list with different chromosome and physical position will be stored in “outname_chr.txt” and “outname_phy.txt”, respectively.

### 4.2.4 Extract chromosome and physical position

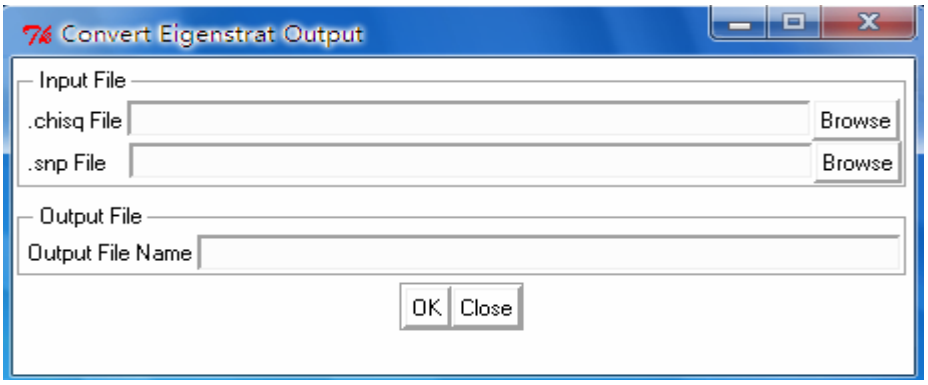



If SNPs in two datasets are not in the same order, you need to use the chromosome and physical position of one dataset as reference to update the chromosome and physical position of another dataset. The chromosome and physical position information will be stored in “outname.chr.txt” and “outname.phy.txt”, respectively. Update SNP information (see 3.1.1.6) in another dataset by using these files. Then merge the two datasets (see 3.1.1.3).

## 4.3 Convert

Convert file format for different software.

### 4.3.1 EIGENSTRAT (chi-square to p-value)




Convert the uncorrected Cochran-Armitage chi-square statistic and the corrected "EIGENSTRAT" chi-square statistic in the output file of EIGENSTRAT to p-values. The output file can be used for QQ plot (see 3.5.1) and Manhattan plot (see 3.5.2).

### 4.3.2 MACH

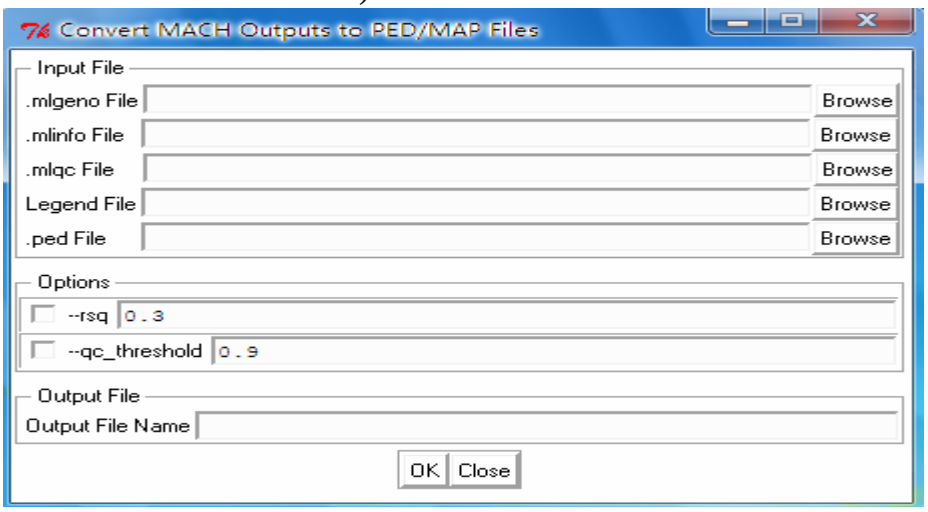
Convert the files generated by the MACH imputation program to PED/MAP format and to SNPTEST gen/sample format by using PERL script from GENGEN [6].

#### 4.3.2.1 MACH2PLINK

Convert MACH imputation results (mlgeno, mlinfo and mlqc file) to standard PED and MAP files for association analysis. HapMap legend file for the chromosome and the first 6 columns of ped information file or PLINK fam file will be used.

--rsq <float>: threshold to define high-qulaity SNP and write to extract file (default: 0.3).

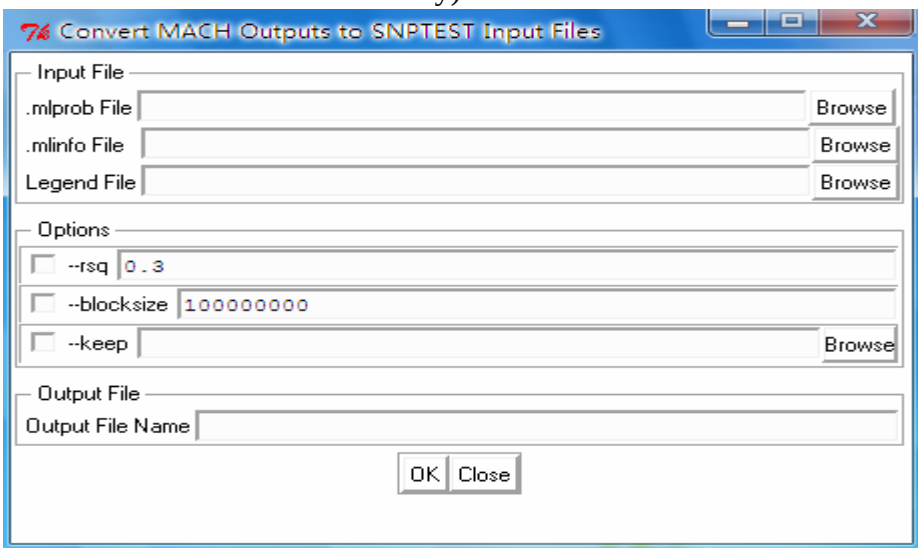
--qc_threshold <float>: genotype posterior probability threshold to output (default: 0.9).



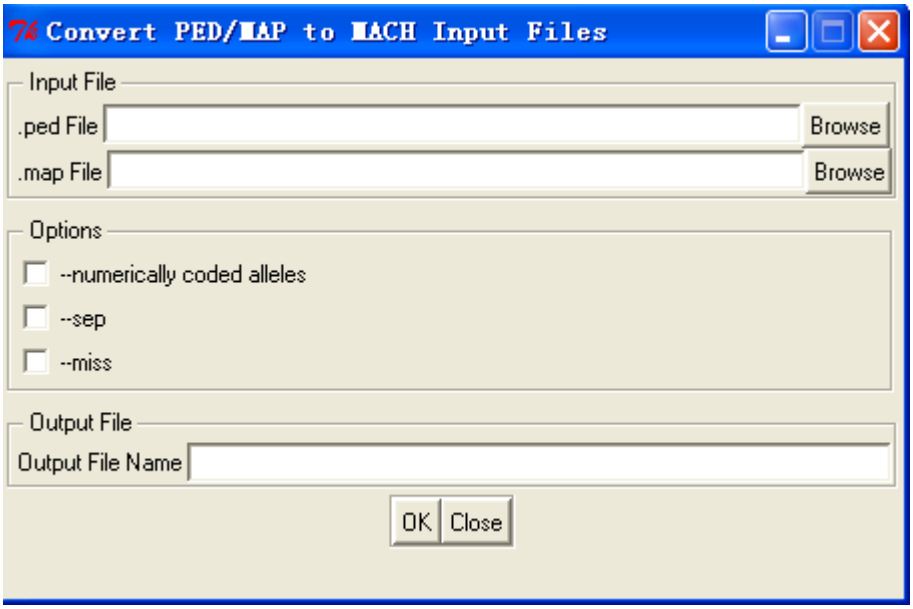
### 4.3.2.2 MACH2SNPTEST

Convert MACH imputation results (mlprob and mlinfo file) to SNPTEST input files (a gen file, a sample file, and an exclude file). This takes into account of genotype imputation uncertainty when performing association test.

--keep <file>: a two-column file specifying subjects to output.
--blocksize <int>: size of each reading block (default=100,000,000, reduce if out-of-memory).




### 4.3.2.3 PLINK2MACH




Convert standard PED/MAP files to PED/DAT files used by MACH (Merlin/QTDT format).

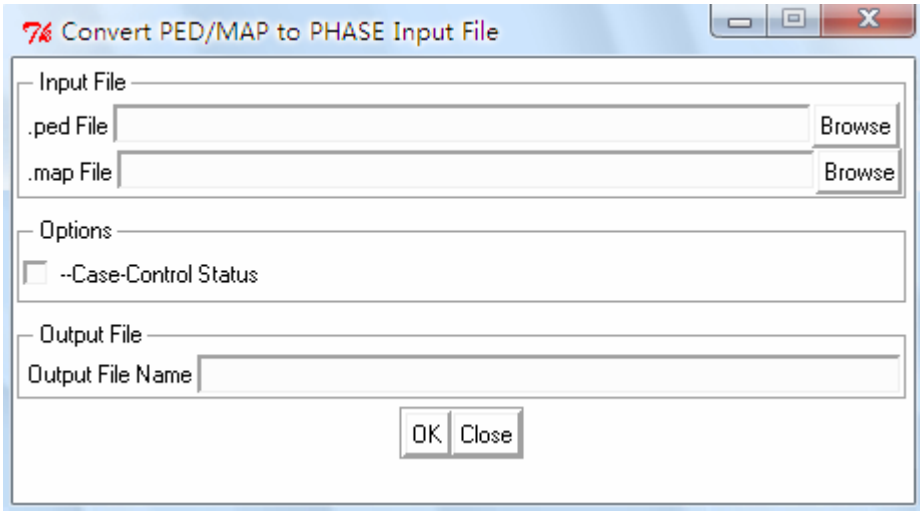
```
<Example of a simple data file>
 M marker1
 M marker2
 ...
<End of simple data file>

<Example of a pedigree file with base-pair coded alleles>
FAM1001    ID1234  0    0    M    A A    A C    C C
FAM1002    ID5678  0    0    F    A C    C C    G G
...
<End of pedigree file>
```

If "--numerically coded alleles" is selected, alleles will be coded as 1, 2, 3, 4 instead of the default A, C, G, T.

If "--sep" is selected , it will add a "/" to separate alleles. E.g. "A/A" (default, "A A").

If "--miss" is selected, missing geneotype will be coded as "." (default, "0").

### 4.3.3 PHASE




Convert PED/MAP files to the default input format of PHASE [7-9]. The default structure for the input file can be represented as follows:

**NumberOfIndividuals**
**NumberOfLoci**
**P Position(1) Position(2) Position(NumberOfLoci)**
**LocusType(1) LocusType(2) ... LocusType(NumberOfLoci)**
**ID(1)**
**Genotype(1)**
**ID(2)**

**Genotype(2)**
**…**
**ID(NumberOfIndividuals)]**
**Genotype(NumberOfIndividuals)**

where the quantities above are as follows:
1. NumberOfIndividuals: An integer specifying the number of individuals who have been genotyped.
2. NumberOfLoci: An integer specifying the number of loci or sites at which each individual has been typed.
3. P: The character 'P' (upper case, without quotation marks).
4. Position(i): A number indicating the position of locus i, relative to some arbitray reference point. The units of base pairs in MAP file were used for this, and the loci were in the same order as in MAP file (their physical increasing order along the chromosome).
5. LocusType(i): A letter indicating the type of locus i. The options are (a) S for a biallelic (SNP) locus, or biallelic site in sequence data. (b) M for microsatellite, or other multi-allelic locus (eg tri-allelic SNP, or HLA allele).
6. ID(i): A string, giving a label for individual i (both the FID and IID in the PED file were used for individual ID).
7. Genotype(i) The genotypes for the ith individual. This is given on two consecutive rows. At each locus, one allele is entered on the first row, and one on the second row. It does not matter which allele is entered on each row. Missing alleles at SNP loci were entered as "?".

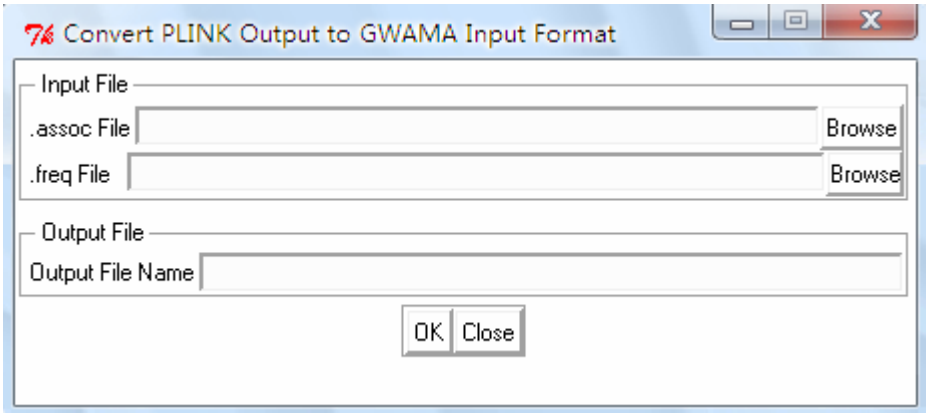
If you want to use PHASE for performing a permutation test for significant differences in haplotype frequencies in case and control groups, select the --Case-Control Status" option. This option specify the case-control status of each individual in the input file by putting a "0" or a "1", followed by a space, just before the individual's identifier "ID(i)".

The output file can be used by PHASE directly (e.g. **./PHASE test.inp test.out** or performing a case-control permutation test, **./PHASE -c test.inp test.out**).

### 4.3.4 GWAMA

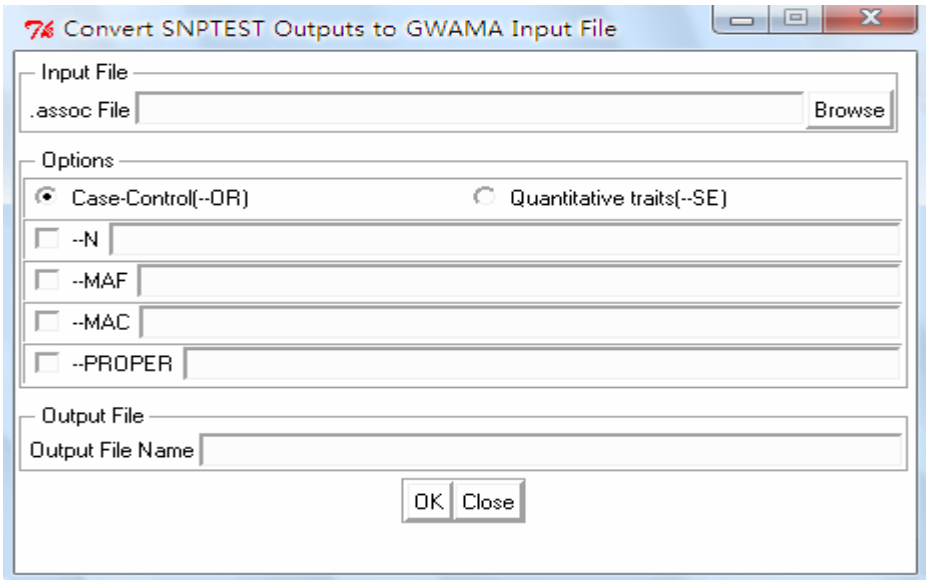
Reformat SNPTEST and PLINK output to GWAMA [18] input format. (For details, please refer to http://www.well.ox.ac.uk/gwama/index.shtml)

### 4.3.4.1 PLINK2GWAMA



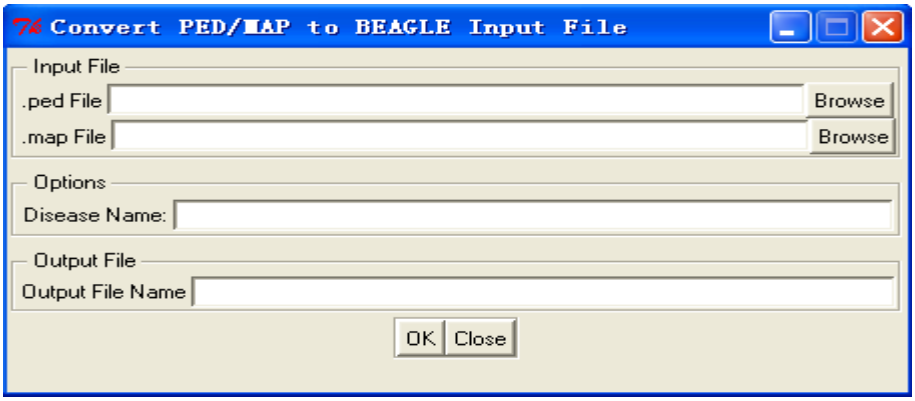

Script for creating GWAMA input file from PLINK association results file. The allele frequency file must also be used for generating GWAMA file. If PLINK association file contains data of covariate effects or multiple models then please remove unnecessary rows prior using this script.

### 4.3.4.2 SNPTEST2GWAMA




Script for creating GWAMA input file from SNPTEST association results file. The BETA and SE values should be in last columns of file. If multiple analyses models were used then please edit the output prior using this script. Script expects that all markers are from positive strand. If not, Strand column must be modified with correct strand information. Data can be filtered according to minimum number of samples (--N), minor allele frequency (--MAF), and minimum number of allele count (--MAC = MAF*N). For example: N=100 MAF=0.01 MAC=10 PROPER=0.4, will remove markers with less than 100 individuals, MAF<1% and MAC<10 and properinfo<0.4.

### 4.3.5 BEAGLE




Convert PED/MAP files of PLINK to unphased genotype file of BEAGLE [20], which is a software program for imputing genotypes, inferring haplotype phase, and performing genetic association analysis. A Beagle genotypes file has a simple format: rows are variables and columns are individuals. Here is an example of a Beagle genotypes file with three individuals and three genotyped markers:

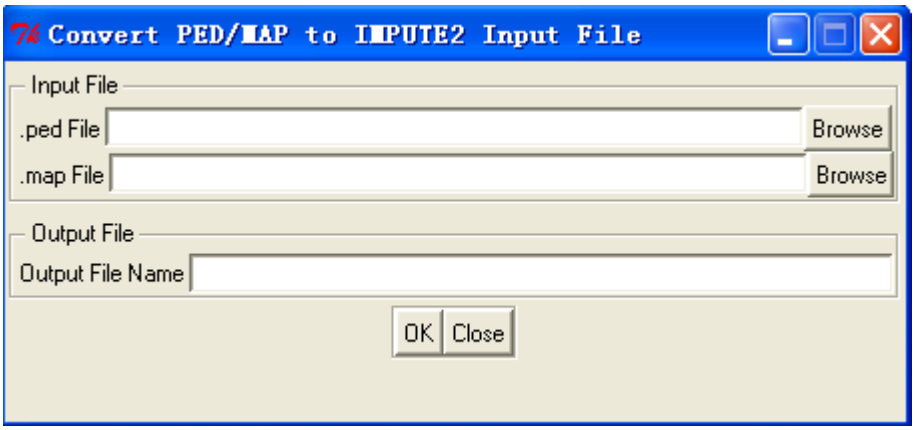
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I | id | 1001 | 1001 | 1002 | 1002 | 1003 | 1003 |
| A | diabetes | 1 | 1 | 2 | 2 | 2 | 2 |
| M | rs228911 | A | G | G | G | A | G |
| M | rs1248628 | T | T | T | C | T | T |
| M | rs10762764 | G | T | T | T | G | T |

In a Beagle genotypes file, the first column describes the data on each line. In the example above, an “I” denotes sample identifier data, an “A” denotes affection status data, and an “M” denotes marker data. The second column contains the name of the variable whose data is given on each line. Variable names should be unique. In Example 1 there are two columns for each individual: columns 3-4 give data for the first individual, columns 5-6 give data for the second individual, and so on [21].

### 4.3.6 IMPUTE2

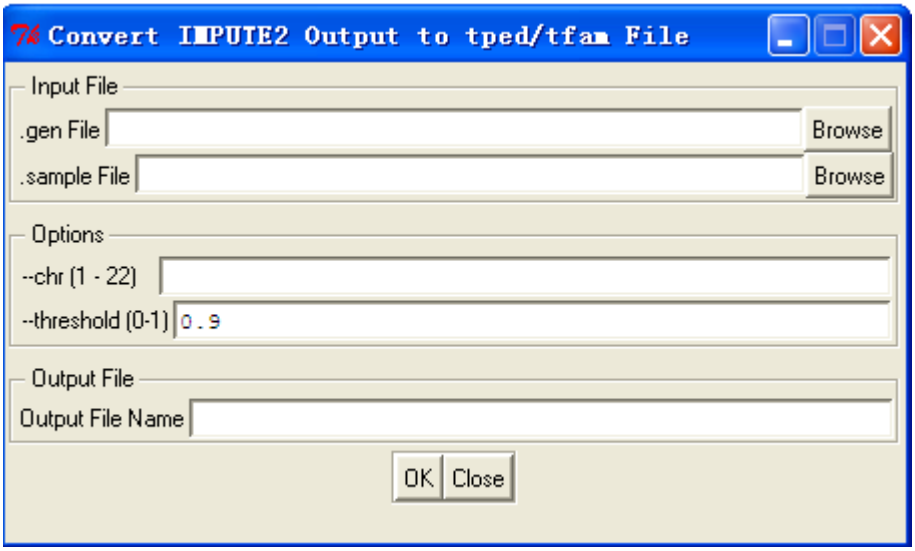
Convert PED/MAP files of PLINK to .geno/.sample as input of IMPUTE2 [24] and convert the output of IMPUTE2 to tped/tfam files of PLINK.

### 4.3.6.1 PLINK2IMPUTE




Convert PED/MAP files of PLINK to .geno/.sample/.strand as input of IMPUTE2. Make sure all SNPs are on the "+" strand before using this function. You can use the "--flip" option to flip the strand.

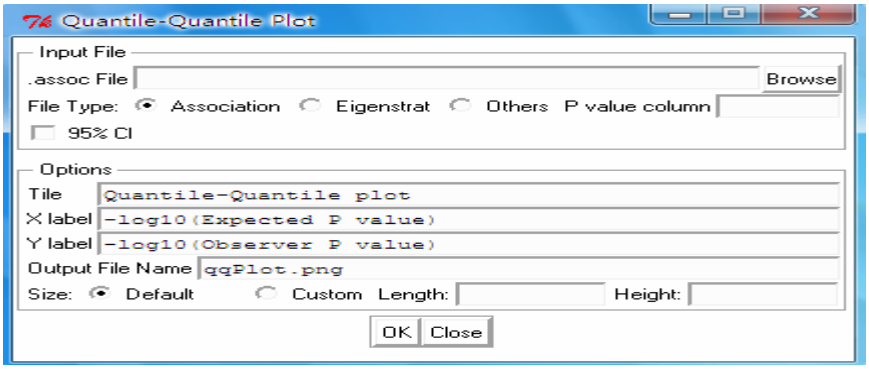
### 4.3.6.2 IMPUT2PLINK




Convert the output of IMPUTE2 to tped/tfam files of PLINK.
.gen file: the output genotype file of IMPUTE2
.sample file: the pair file of the gen file (one of the output file of 4.3.6.1)
--chr: chromosome information of the gen file (usually: 1-22)
--threshold: threshold to call genotype (default: 0.9)

# 4.4 Plot

## 4.4.1 Quantile-Quantile (QQ) Plot




P-values generated under the Null hypothesis should be drawn from a Uniform distribution between zero and one. It follows that if a set of m p-values are ordered from lowest to highest, then the observed quantile of the jth ordered item should on average be equal to the corresponding expected quantile for m items drawn from a Uniform(0,1) distribution (and this expected value can be shown to be j/(m + 1)). Thus if the observed quantiles are plotted against the expected ones, one would expect to see a roughly straight line through the origin with a unit slope, albeit with some random variation. A log scaling will emphasise these low p-values more, and so it is a common practice to plot p-value Q-Q plots on a negative logarithmic scale. SNPs departing from the Null will now appear as points rose above the unit line towards the top right of the plot [10].

The input file can be the association results of PLINK [2], the converted results of EIGENSTRAT (see 3.4.1) and any file contains a column of p-values (a head line is needed to specify the p-values column). The default size of the plot is 1260 × 689 (Figure 8).

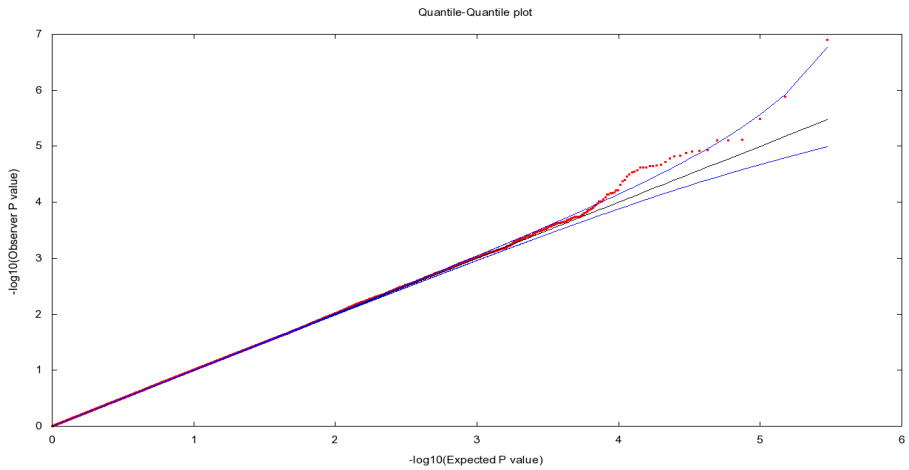



Figure 8 Example of QQ plot of the association result of PLINK.

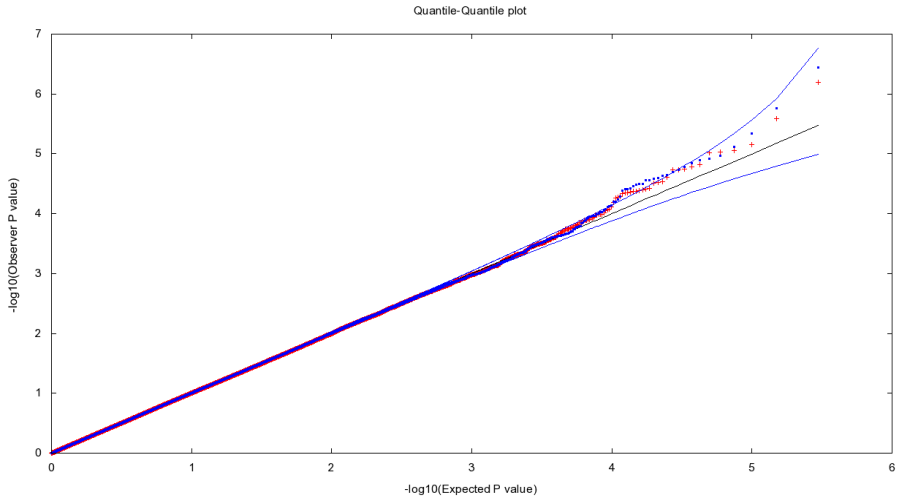
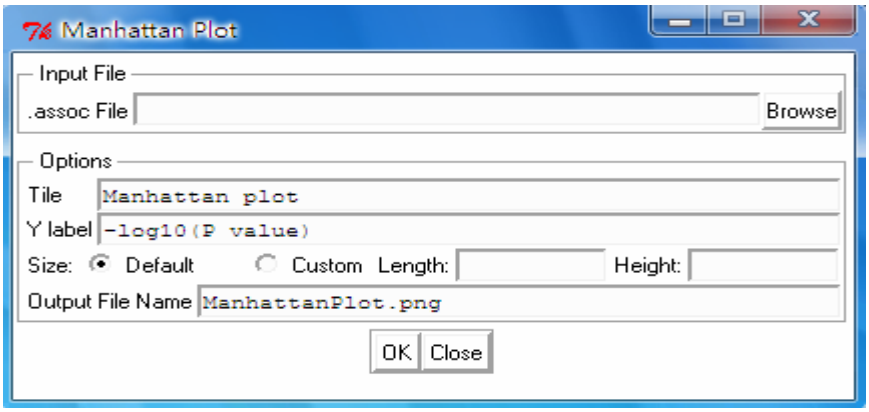



Figure 9 Example of QQ plot of the converted p-value of EIGENSTRAT.
Blue dots denote uncorrected Cochran-Armitage p-values, red dots denote EIGENSTRAT corrected p-values, and black line denotes the NULL line.

## 4.4.2 Manhattan Plot




A Manhattan plot is a type of scatter plot, usually used to display data with a large number of data-points. In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with the negative logarithm of the association P-value for each single nucleotide polymorphism displayed on the Y-axis [11].

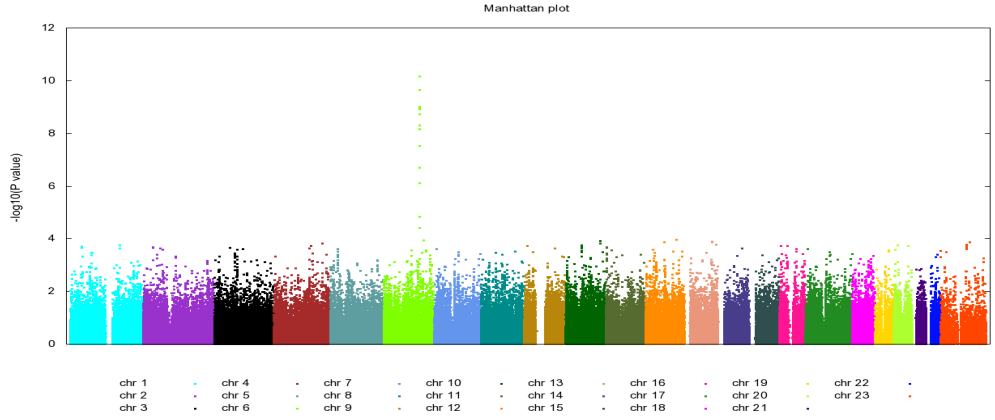
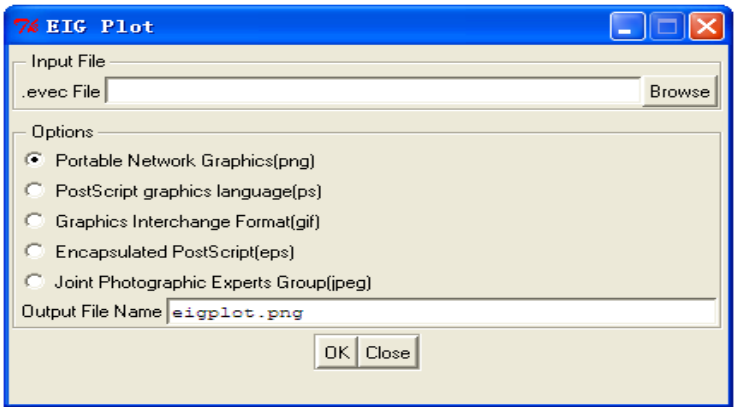


Figure 10 Example of Manhattan plot.

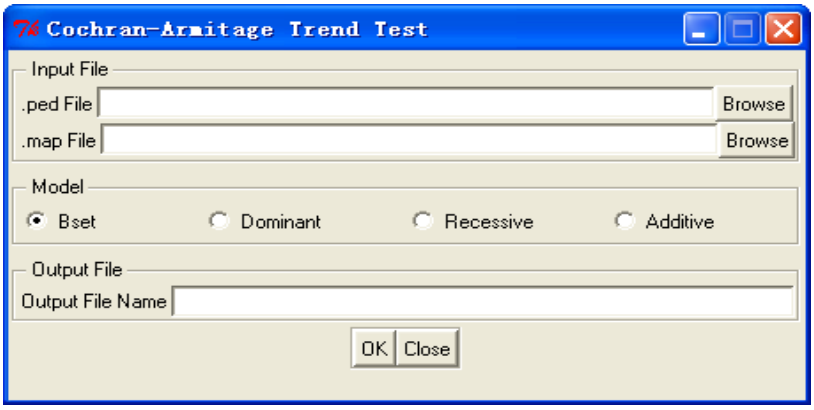
### 4.4.3 ploteig




ploteig constructs a plot of the top two principal components by using uses the gnuplot utility.

## 4.5 Statistics

### 4.5.1 Cochran-Armitage Trend Test




Cochran-Armitage trend test [13-14] modifies the chi-square test to incorporate a suspected ordering in the effects of the k categories of the second variable. In genetics applications, the weights are selected according to the suspected mode of inheritance. For example, the genotype from a case-control study are represented in Table 1, B is risk allele.

Table 1. Genotype distribution

| | **AA** | **AB** | **BB** | **Total** |
|---|---|---|---|---|
| **Cases** | $r_0$ | $r_1$ | $r_2$ | R |
| **Controls** | $s_0$ | $s_1$ | $s_2$ | S |
| **Total** | $n_0$ | $n_1$ | $n_2$ | N |

Then the Cochran-Armitage trend test can be written as

$$Z = \frac{U}{\sqrt{\mathrm{var}_{H_0}(U)}}$$

Where

$$U = \frac{1}{N}\sum_{i=0}^{2} x_i (Sr_i - Rs_i)$$

and

$$\mathrm{var}_{H_0}(U) = \frac{RS}{N^3}[N\sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2]$$

In order to test whether the risk allele is dominant or recessive over the other allele, or the risk allele and the other allele are codominant, the scores choice t = (0, 1, 1), or t = (0, 0, 1), or t = (0, 1, 2) is optimal, respectively. Then

$$U_{dom} = \frac{[N(r_1 + r_2) - R(n_1 + n_2)]}{N}$$

$$U_{rec} = \frac{Nr_2 - Rn_2}{N}$$

$$U_{add} = \frac{[N(r_1 + 2r_2) - R(n_1 + 2n_2)]}{N}$$

and

$$\mathrm{var}_{H_0}(U_{dom}) = \frac{R(N-R)n_0(n_1 + n_2)}{N^3}$$

$$\mathrm{var}_{H_0}(U_{rec}) = \frac{R(N-R)n_2(n_0 + n_1)}{N^3}$$

$$\mathrm{var}_{H_0}(U_{add}) = \frac{R(N-R)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}{N^3}$$

Then

$$\chi^2_{dom} = \frac{N[N(r_1 + r_2) - R(n_1 + n_2)]^2}{R(N-R)n_0(n_1 + n_2)} \sim \chi^2_1$$

$$\chi^2_{rec} = \frac{N(Nr_2 - Rn_2)^2}{R(N-R)n_2(n_0 + n_1)} \sim \chi^2_1$$

$$\chi^2_{add} = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{R(N-R)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]} \sim \chi^2_1$$

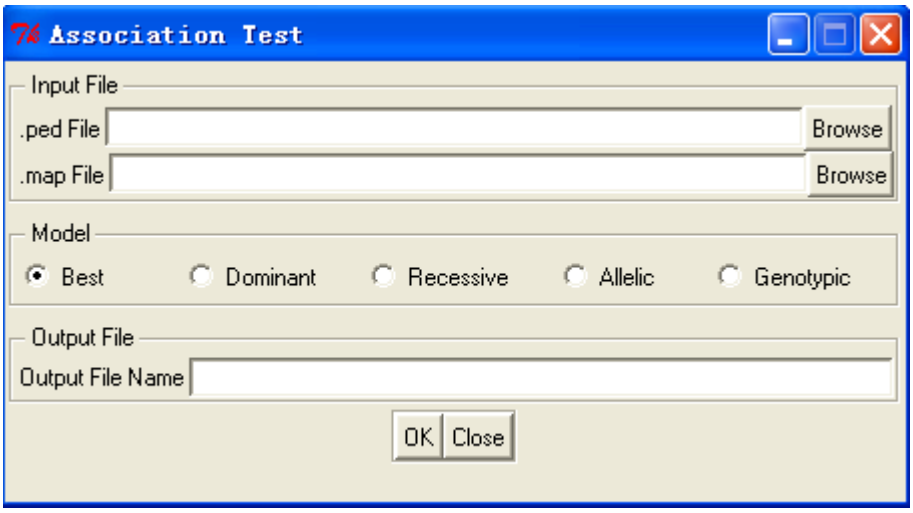
The output will contain fourteen columns:

| | |
|---|---|
| CHR | Chromosome number |
| SNP | SNP identifier |
| BP | Physical position (base-pair) |
| A1 | Risk allele |

| | |
|---|---|
| A2 | Other allele |
| AFF | Genotypes in cases (homozygous risk allele/heterozygous/ homozygous other allele) |
| UNAFF | Genotypes in controls (homozygous risk allele/heterozygous/ homozygous other allele) |
| OR | Estimated allelic odds ratio |
| SE | Standard error of odds ratio |
| L95 | Lower bound of 95% confidence interval for odds ratio |
| U95 | Upper bound of 95% confidence interval for odds ratio |
| Model | Model of test (Dominant, Recessive, and Additive) |
| N | Number of individuals with non-missing genotypes |
| F_A | Frequency of minor allele in cases |
| F_U | Frequency of minor allele in controls |
| Freq | Frequency of minor allele in all data set |
| Chisq | Chi-square statistic |
| P | Asymptotic p-value |

If the model “best” was selected, the p-value for each model will be calculated, only the smallest p-value and the corresponding model will be list out in the results.

## 4.5.2 Association Test




Similar with “--assoc” and “--model” option of PLINK but provides more detail information for genotypic, allelic, dominant and recessive model. If the model “best” was selected, the p-value for each model will be calculated, only the smallest p-value and the corresponding model will be list out in the results.

For example, the genotype from a case-control study are represented in Table 2 and the alleles distribution are represented in Table 3. A is minor allele.

Table 2. Genotype distribution

| | **AA** | **AB** | **BB** | **Total** |
|---|---|---|---|---|
| **Cases** | $r_0$ | $r_1$ | $r_2$ | R |
| **Controls** | $s_0$ | $s_1$ | $s_2$ | S |
| **Total** | $n_0$ | $n_1$ | $n_2$ | N |

Table 3. Allele distribution

| | **Class A** | **Class B** | **Total** |
|---|---|---|---|
| **Cases** | a | b | a+b |
| **Controls** | c | d | c+d |
| **Total** | a+c | b+d | a+b+c+d |

Then $\chi^2$ is calculated from the following formula:

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

For allelic model:

$a=2r_0+r_1$
$b=2r_2+r_1$
$c=2s_0+s_1$
$d=2s_2+s_1$

For dominant model:

$a=r_0+r_1$
$b=r_2$
$c=s_0+s_1$
$d=s_2$

For recessive model:

$a=r_0$
$b=r_2+r_1$
$c=s_0$
$d=s_2+s_1$

While the genotypic test provides a general test of association in the 2-by-3 table of disease-by-genotype, the $\chi^2$ is calculated from the following formula:

$$\chi^2 = \sum_{i=0}^{2} \frac{(r_i - E(r_i))^2}{E(r_i)} + \sum_{i=0}^{2} \frac{(s_i - E(s_i))^2}{E(s_i)}$$

where

$$E(r_i) = \frac{Rn_i}{N}$$

and

$$E(s_i) = \frac{Sn_i}{N}$$

The output can be used as input for IPGWAS for QQ plot and Manhattan plot, and can

also be used by PLINK [2] and Metal [16] for meta-analysis. The output will contain 18 columns:

| | |
|---|---|
| CHR | Chromosome number |
| SNP | SNP identifier |
| BP | Physical position (base-pair) |
| A1 | Minor allele |
| A2 | Major allele |
| AFF | Genotypes in cases (homozygous risk allele/heterozygous/ homozygous other allele) |
| UNAFF | Genotypes in controls (homozygous risk allele/heterozygous/ homozygous other allele) |
| Model | Model of test (Dominant, Recessive, and Additive) |
| N | Number of individuals with non-missing genotypes |
| F_A | Frequency of minor allele in cases |
| F_U | Frequency of minor allele in controls |
| Freq | Frequency of minor allele in cases and controls |
| Chisq | Chi-square statistic |
| P | Asymptotic p-value |
| OR | Estimated odds ratio |
| SE | Standard error of odds ratio |
| L95 | Lower bound of 95% confidence interval for odds ratio |
| U95 | Upper bound of 95% confidence interval for odds ratio |

For allelic, dominant and recessive modle,

$$OR = \frac{ad}{bc}$$

$$SE(\log OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

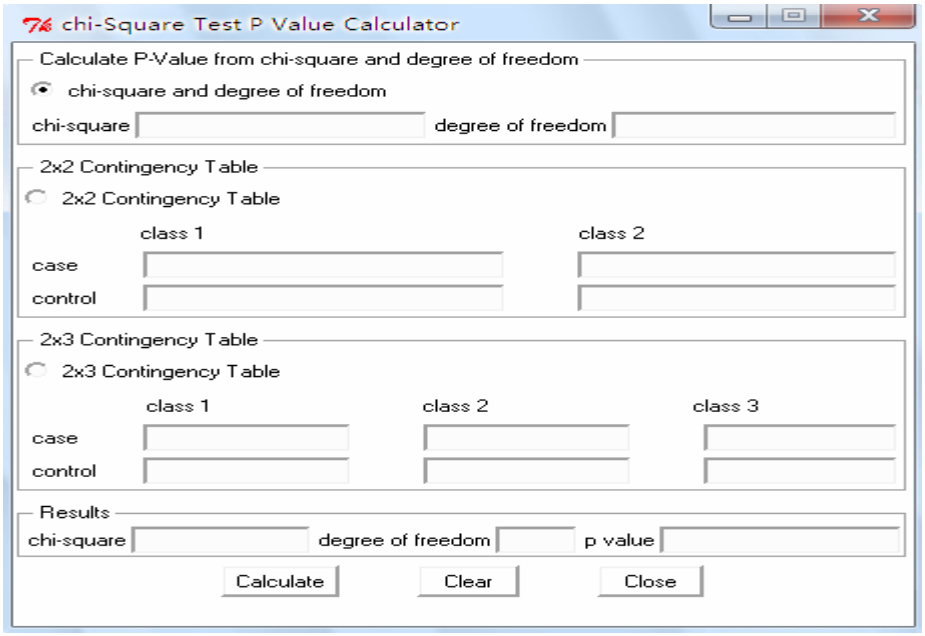
$$U95 = \exp(\log OR + 1.96 SE(\log OR))$$

$$L95 = \exp(\log OR - 1.96 SE(\log OR))$$

For genotypic model, the allelic OR and CI will be used.

## 4.5.3 P-Value Calculator

### 4.5.3.1 Chi-square test



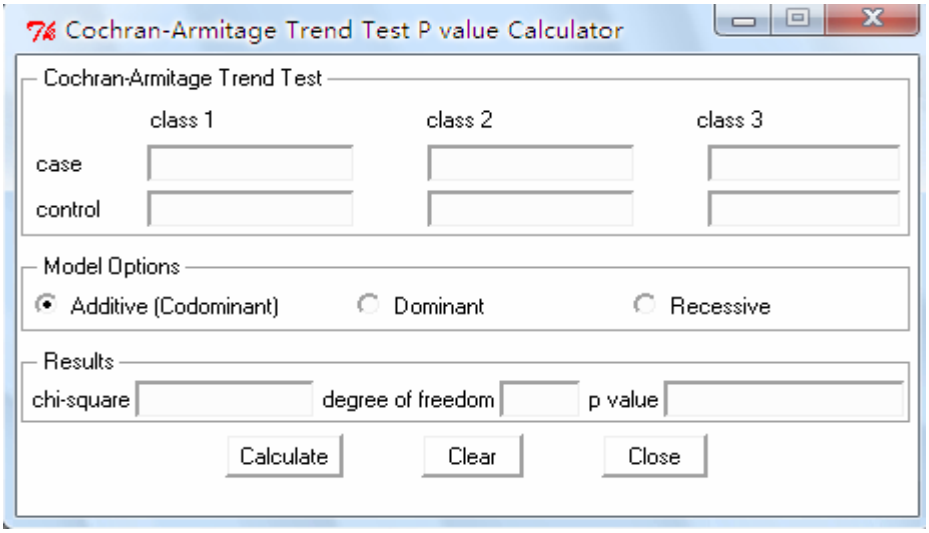

The chi-square statistic is calculated by the following formula [15]:

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where

$\mathrm{X}^2$ = Pearson's cumulative test statistic,
$O_i$ = an observed frequency;
$E_i$ = an expected (theoretical) frequency, asserted by the null hypothesis;
$n$ = the number of cells in the table.

### 4.5.3.2 Cochran-Armitage trend test



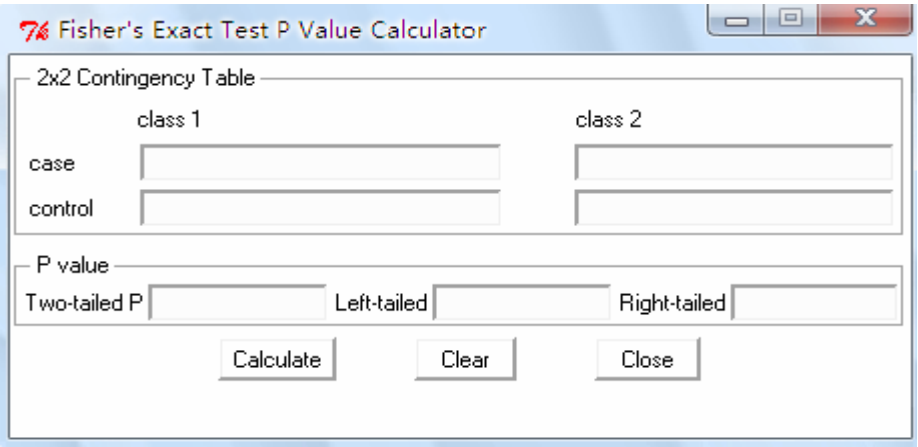

Suppose we have R cases and S controls, and the genotypes distribution as the following table [14].

| | Number of disease alleles | | | |
|---|---|---|---|---|
| | 0 | 1 | 2 | Totals |
| Cases | $r_0$ | $r_1$ | $r_2$ | R |
| Controls | $s_0$ | $s_1$ | $s_2$ | S |
| Totals | $n_0$ | $n_1$ | $n_2$ | N |

If the disease allele and the other allele are codominant, the score t = (0, 1, 2) will be used to weight which genotype, then the statistic $X^2_G$ asymptotically approaches a $\chi^2$ distribution.

$$\chi^2_G = \frac{N\left[N(r_1+2r_2)-R(n_1+2n_2)\right]^2}{(N-R)R\left[N(n_1+4n_2)-(n_1+2n_2)^2\right]} \sim \chi^2_1.$$

#### 4.5.3.3 Fisher's exact test



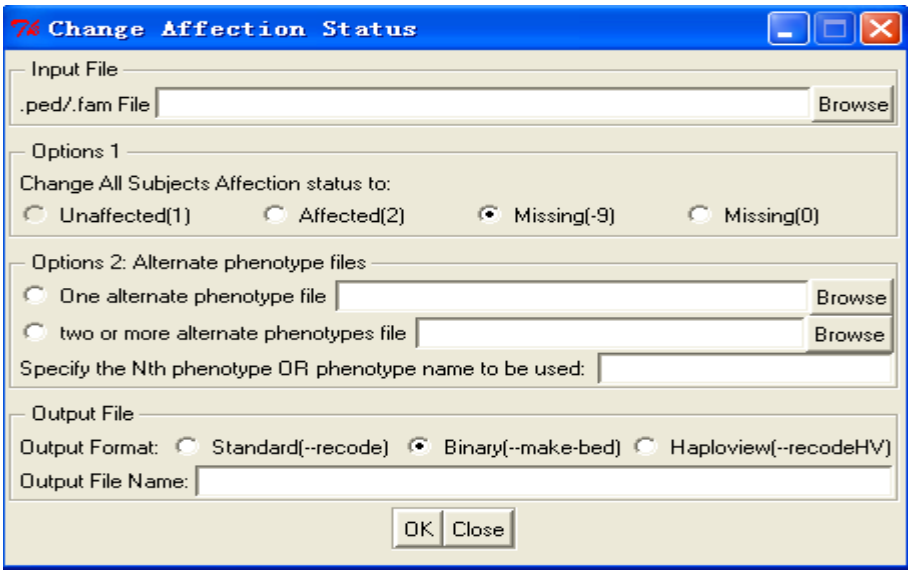

The PERL module Text-NSP [18] was used to calculate the one-tailed and two-tailed P value for Fisher's exact test [17].

## 4.6 Manipulation

Data manipulation functions.

### 4.6.1 Change affection status

Affection status, by default, should be coded as: -9 = missing, 0 = missing, 1 = unaffected, 2 = affected. The affection status of one individual could be changed in different analysis. In the imputation, the affection status of references should be code as 0.




To specify an alternate phenotype for analysis, use the --pheno option:

plink --file mydata --pheno pheno.txt

where pheno.txt is a file that contains 3 columns (one row per individual):

Family ID
Individual ID
Phenotype

If the phenotype file contains more than one phenotype, then use the --mpheno N option to specify the Nth phenotype is the one to be used:

plink --file mydata --pheno pheno2.txt --mpheno 4

where pheno2.txt contains 5 different phenotypes (i.e. 7 columns in total), this command will use the 4th for analysis (phenotype D):

Family ID
Individual ID
Phenotype A
Phenotype B
Phenotype C
Phenotype D
Phenotype E

Alternatively, your alternate phenotype file can have a header row, in which case you can use variable names to specify which phenotype to use. If you have a header row, the first two variables must be labelled FID and IID. All subsequent variable names cannot have any whitespace in them. For example,

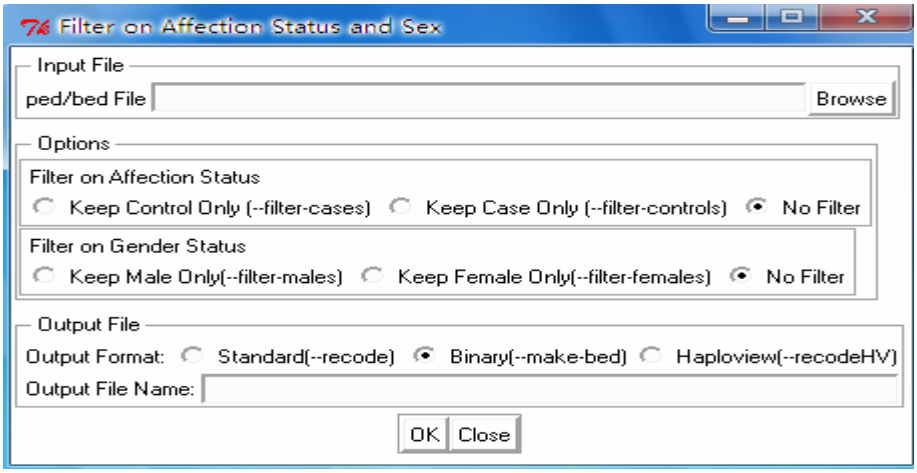
| FID | IID | qt1 | bmi | site |
|---|---|---|---|---|
| F1 | 1110 | 2.3 | 22.22 | 2 |
| F2 | 2202 | 34.12 | 18.23 | 1 |
| ... | | | | |

then

plink --file mydata --pheno pheno2.txt --pheno-name bmi
(http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#pheno)

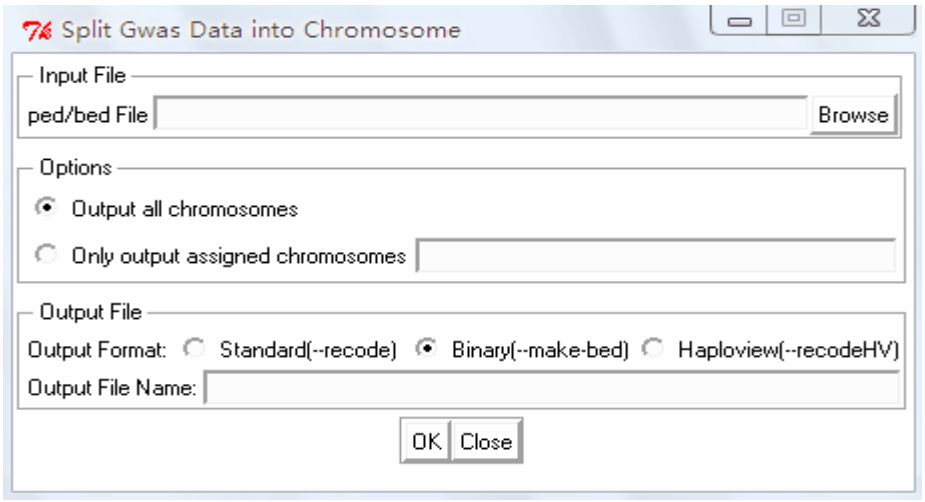
IPGWAS can detect whether the alternate phenotype file has a header row or not, automatically.

## 4.6.2 Subjects filter



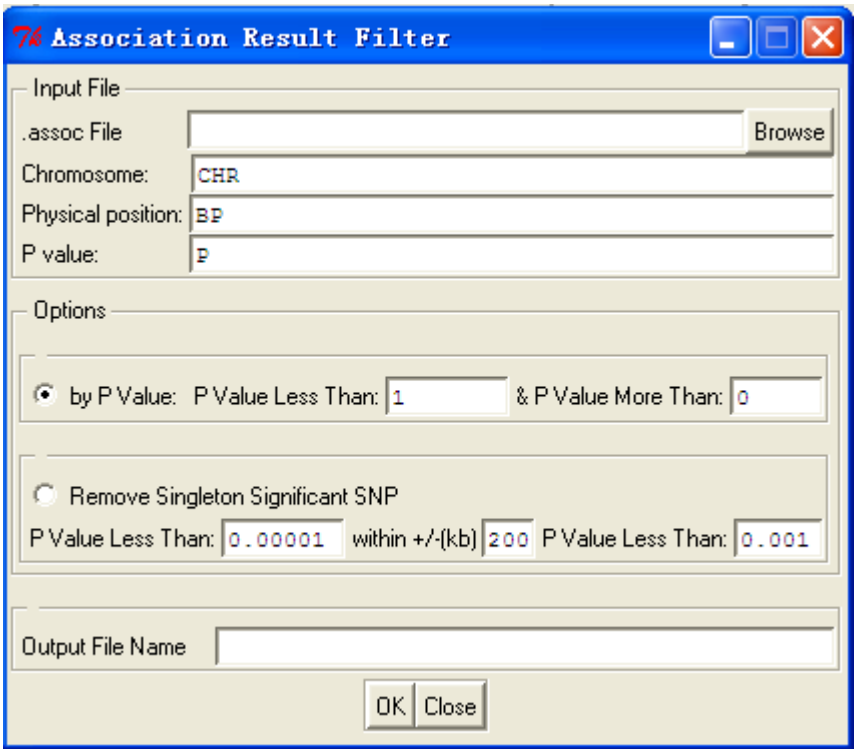

Keep only case or control, and/ or male or female of your dataset.

## 4.6.3 Split Gwas Data by Chromosome




This function can split up a GWAS dataset (PED/BED) by chromosome. The "Output all chromosomes" will split up the GWAS dataset and output the 22 autosomes, each chromosome in a separated PED/BED file. You can also assign which chromosome(s) to be output by using "Only output assigned chromosomes", if two or more chromosomes were assigned, the number should be separated by comma (e.g. "1,5,22" will output the chromosome 1, 5 and 22). This function is useful for preparing data for imputation.

### 4.6.7 Association result filter




The first function only output SNPs with p value between the two p value threshold. The second function remove singleton significant SNPs, for example, if the p value of one SNP is less than 0.00001 and there is no SNP with p value less than 0.001 within ±200kb, this SNP will be removed.

## 4.7 GUI for Plink

This section provides GUI for the basic usage of Plink.

### 4.7.1 Data Management

Refer to: http://pngu.mgh.harvard.edu/~purcell/plink/dataman.shtml

#### 4.7.1.1 Recode

Plink commands involved: **--recode, --make-bed, --recodeHV**.

#### 4.7.1.2 Flip Strand

Plink command involved: **--flip**.

#### 4.7.1.3 Merge two filesets

Plink commands involved: **--merge, --bmerge**.

#### 4.7.1.4 Merge multiple filesets

Plink command involved: **--merge-list**.

#### 4.7.1.5 Write SNP list files

Plink command involved: **--write-snplist**.

#### 4.7.1.6 Update SNP information

Plink commands involved: **--update-map, --update-chr, --update-name**

#### 4.7.1.7 Update allele information

Plink command involved: **--update-alleles.**

#### 4.7.1.8 Update individual information

Plink commands involved: **--update-ids, --update-sex, --update-parents.**

#### 4.7.1.9 Extract a subset of SNPs

Plink commands involved: **--extract; --chr, --from-kb and --to-kb; --from and --to; --snp and --window.**

#### 4.7.1.10 Remove a subset of SNPs

Plink command involved: **--exclude.**

#### 4.7.1.11 Extract a subset of individuals

Plink command involved: **--keep.**

#### 4.7.1.12 Remove a subset of individuals

Plink command involved: **--remove.**

### 4.7.2 Summary Statistics

Refer to: http://pngu.mgh.harvard.edu/~purcell/plink/summary.shtml

#### 4.7.2.1 Missingness

Plink commands involved: **--missing, --test-missing, --test-mishap.**

#### 4.7.2.2 Hardy-Weinberg Equilibrium

Plink command involved: **--hardy.**

#### 4.7.2.3 Allele frequency

Plink command involved: **--freq.**

#### 4.7.2.4 Mendel errors

Plink command involved: **--mendel.**

#### 4.7.2.5 Sex check

Plink command involved: **--check-sex.**

#### 4.7.2.6 Sex impute

Plink command involved: **--impute-sex.**

#### 4.7.2.7 Linkage disequilibrium based SNP pruning

Plink commands involved: **--indep, --indep-pairwise.**

### 4.7.3 Filters

Refer to: http://pngu.mgh.harvard.edu/~purcell/plink/thresh.shtml

#### 4.7.3.1 Missingness per individual

Plink command involved: **--mind.**

#### 4.7.3.2 Allele frequency

Plink command involved: **--maf.**

#### 4.7.3.3 Missingness per marker

Plink command involved: **--geno.**

#### 4.7.3.4 Hardy-Winberg equilibrum

Plink commands involved: **--hwe, --hwe-all, --nonfounders.**

#### 4.7.3.5 Mendel error rates

Plink command involved: **--me.**

### 4.7.4 IBS/IBD Estimation

Refer to: http://pngu.mgh.harvard.edu/~purcell/plink/ibdibs.shtml

#### 4.7.4.1 Pairwise IBD estimation

Plink commands involved: **--genome, --genome-full, --min, --max.**

#### 4.7.4.2 Inbreeding codfficients

Plink command involved: **--het.**

#### 4.7.4.3 Runs of homozygosity

Plink commands involved: **--homozyg, --homozyg-snp, --homozyg-kb, homozyg-group, --homozyg-match, --homozyg-verbose.**

### 4.7.5 Association Analysis

Refer to: http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml

#### 4.7.5.1 Basic case/control association test

Plink command involved: **--assoc.**

#### 4.7.5.2 Full model association tests

Plink commands involved: **--model, --cell, --mperm, --perm, --model-gen, --model-trend, --model-dom, --model-rec, --fisher.**

#### 4.7.5.3 Linear and logistic models

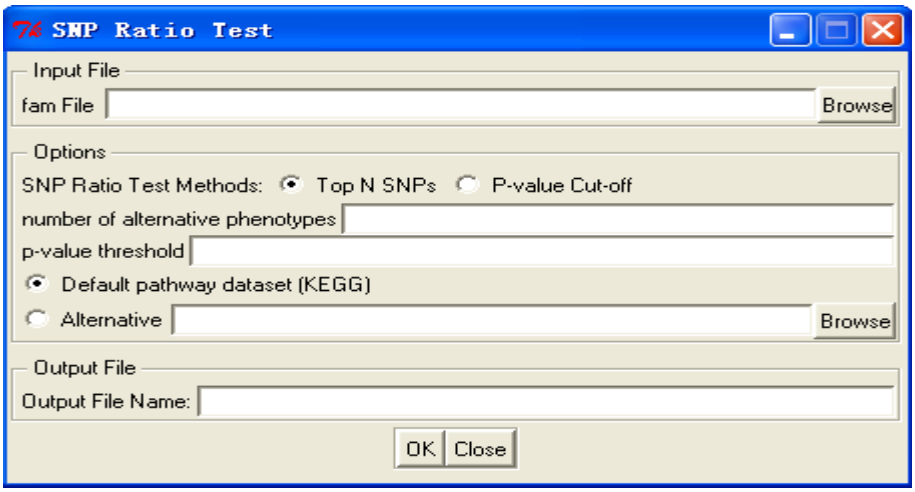
Plink commands involved: **--linear, --logistic.**

#### 4.7.5.4 Covariates and interactions

Plink commands involved: **--covar, --covar-name, --covar-number, --condition, --condition-list, --sex, interaction.**

# 4.8 Pathway based analysis

## 4.8.1 SNP Ratio Test (SRT)




The SNP ratio test (SRT) compares the proportion of significant to all SNPs within genes that are part of a pathway and computes an empirical P-value based on comparisons to ratios in datasets where the assignment of case/control status has been randomized [22].

The final output file contains (a) pathway ID, (b) the number of times a simulated (alt. pheno) ratio was greater than or equal to the original GWAS ratio, (c) the total number of simulations, (d) the empirical p-value. Refer to https://sourceforge.net/projects/snpratiotest/ for details.

# 4.9 Downloads

Download resources which will be used by IPGWAS during the analysis.

# 4.10 Help

## 4.10.1 Version

Release version information.

## 4.10.2 Manual

Refer to this document.

### 4.10.3 Home Page

http://ipgwas.sourceforge.net/,
http://sourceforge.net/projects/ipgwas/

## 4.11 Citation

Yan-Hui Fan, You-Qiang Song. (2012) IPGWAS: An integrated pipeline for rational quality control and association analysis of genome-wide genetic studies. Biochem. Biophys. Res. Commun. 422(3):363-368 http://dx.doi.org/10.1016/j.bbrc.2012.04.117

# 5 References

[1] CPAN, Comprehensive Perl Archive Network. http://www.cpan.org/ (Accessed on 26 Feb 2011)
[2] PLINK http://pngu.mgh.harvard.edu/~purcell/plink/ (Accessed on 26 Feb 2011)
[3] Gnuplot http://www.gnuplot.info/ (Accessed on 26 Feb 2011)
[4] Price AL, et al. (2008) Long-Range LD Can Confound Genome Scans in Admixed Populations. Am J Hum Genet 83:132-135.
[5] Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet 38:904-909
[6] GenGen: Genetic Genomics Analysis of Complex Data. http://www.openbioinformatics.org/gengen/ (Accessed on 26 Feb 2011)
[7] Stephens M, Donnelly P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162-1169.
[8] Stephens M, Smith N, Donnelly P. (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Geneti 68: 978-989.
[9] Stephens M, Scheet P. (2005) Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. Am J Hum Genet 76:449-462.
[10] Weale ME (2010). Quality control for genome-wide association studies. Methods Mol Biol 628:341-372
[11] Wikipedia. Manhattan plot. http://en.wikipedia.org/wiki/Manhattan_plot (Accessed on 26 Feb 2011)
[12] Abecasis GR, et al. (2001) GRR: graphical representation of relationship errors. Bioinformatics. 17:742-743
[13] Wikipedia. Cochran-Armitage test for trend. http://en.wikipedia.org/wiki/Cochran-Armitage_test_for_trend (Accessed on 26 Feb 2011)
[14] statgen.org (2007). A derivation for Armitage's trend test for the 2 x 3 genotype table. http://www.statgen.org/main/images/www_statgen_org/downloads/Dana/armitage.pdf

(Accessed on 26 Feb 2011)
[15] Wikipedia. Pearson's chi-square test http://en.wikipedia.org/wiki/Pearson%27s_chi-square_test (Accessed on 7 March 2011)
[16] Willer CJ, Li Y, Abecasis GR. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190–2191.
[17] Wikipedia. Fisher's exact test. http://en.wikipedia.org/wiki/Fisher's_exact_test (Accessed on 7 March 2011)
[18] Ted Pedersen. Perl module that provides methods to compute the Fishers exact tests. http://search.cpan.org/dist/Text-NSP/ (Accessed on 7 March 2011)
[19] Magi, R. & Morris, A. (2010) GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics. 11:288
[20] B L Browning and S R Browning (2011) High-resolution detection of identity by descent in unrelated individuals. The American Journal of Human Genetics 88:173-182.
[21] The documentation for BEAGLE (Accessed on 14 April 2011) http://faculty.washington.edu/browning/beagle/beagle_3.3_12Feb11.pdf
[22] O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, et al. (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. Bioinformatics 25: 2762-2763.
[23] http://www.sph.umich.edu/csg/abecasis/MaCH/
[24] http://mathgen.stats.ox.ac.uk/impute/impute_v2.html