*Version 1.1 (Jan, 2009)*

# User guide of software KBAT

**Hsin-Chou Yang, Hsin-Yi Hsieh and Cathy S. J. Fann**

## Table of Contents:

## 1. KBAT LICENSE

 If publications are based on the results from the use of **KBAT**, please cite the following reference: Hsin-Chou Yang, Hsin-Yi Hsieh & Cathy SJ Fann (2008) KBAT: Kernel-based association test. *Genetics* **179**, 1057-1068.

# 2. INTRODUCTION

**KBAT** (Kernel-Based Association Test) is a convenient analysis tool for disease gene association mapping. Several powerful association tests in **KBAT** are developed based on the integrated concept of moving average, p-value weighting, p-value truncation and p-value combination. The methods provide systematic genome-wide and candidate-region searches for disease susceptibility genes. Numerical results and graphic results provide insight into the disease-marker association in study regions. The detailed formulae of the test statistics are provided in Appendix.

## 3. SOFTWARE DOWNLOAD AND INSTALLATION

**KBAT** was written in language R and R-GUI, which are publicly available software. Execution of **KBAT** requires installation of four programs. They are (1) Program **KBAT**, (2) program R, (3) program Genetics, and (4) program ActiveTcl.

1. Download program **KBAT**:

   Software **KBAT** is available at the **KBAT** website at http://www.stat.sinica.edu.tw/hsinchou/genetics/association/KBAT.htm. The zipped file "KBAT.zip" can be downloaded and then unzipped to obtain a directory "KBAT" containing the programs of **KBAT** and several illustrated data examples. Then, the directory "KBAT" can be saved as a working directory, such as "C:\KBAT".

2. Download program R:

   Users should download language R from the website of "The R Project for Statistical Computing" at http://www.r-project.org/. Users click "CRAN" (Comprehensive R Archive Network) in the left of the page and then select a suitable mirror site to download software R. Select a platform (Linux, MacOS X, Windows (95 and later)) for R execution in your end. Click the hyperlink "base", select "R-2.6.0-win32.exe" and then execute the file "C:\Program Files\R\R-2.6.0" to install program R. After finishing the installation of program R, doubly click the icon "R-2.6.0" to initialize program R, a window "RGui" with a sub-window "R Console" jumps up await for the subsequent analysis action. Users are suggested to update packages in R. They can select "Packages" in the tool bar, click "Update packages" and then select a suitable mirror site to update packages. A window "CRAN mirror" jumps up and the icon "OK" is clicked to update packages.

3. Download program "Genetics":

   Initialize program R by clicking the icon "R-2.6.0". Then check "Packages" in the tool bar and then select "Install package(s)". A window "Packages" in R jumps up and then the program "Genetics" is selected to include.

4. Download program "ActiveTcl":

   Program ActiveTcl can be downloaded from the website of ActiveState at http://www.activestate.com/store/productdetail.aspx?prdGuid=f0cd6399-fefb-466e-ba17-220dcd6f4078. Click "Download" to enter a registration site. After finishing the registration, users will be brought to a new page for software download. Users can select the suitable system for their platform, e.g., Windows, and then download and execute the executable file "ActiveTcl8.4.16.0.282109-win32-ix86-threaded.exe" to install Program ActiveTcl.

# 4. KBAT INITIALIZATION

Once the software mentioned in the previous section is installed, **KBAT** can be initialized by the following procedures. Here, we suppose that programs of **KBAT** are saved in the destination directory "C:\KBAT".

1. Initialize software R by doubly clicking the icon "R-2.6.0".
2. Key in "KBAT<-c("C:\\KBAT\\Real\\KBAT_R.r")" in the command line in the window "R Console" and press the Enter key.
3. Type "source(KBAT)" in the command line to initialize **KBAT**. The **KBAT** interface jumps up and waits for the data entry after pressing the Enter key.

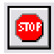
## 5. DESCRIPTION OF WORKING DIRECTORIES

The main directory “KBAT” contains two sub-directories “Real” and “Sim”. The directory “Real” is designed for real data analysis and directory “Sim” is designed for simulation data analysis (The subdirectory “Sim” is still empty and the functions for simulation data analysis is under development). The directory “Real” consists of four directories and a program file. The program file “KBAT_R.r” is the main program of **KBAT**. The four directories are “Example”, “Program”, “Input” and “Output”. The directory “Example” consists of four subdirectories for four real examples respectively. The directory “Program” consists of several programs of **KBAT**. The directory “Input” is the defaulted data input directory in **KBAT** (as shown in the **KBAT** interface in **Figure 1**). The working directory can be changed by keying the target directory name in the **KBAT** interface. The directory “Output” is the defaulted result output directory in **KBAT**. Results will be automatically saved in this directory. However, users can also change the output directory by keying the target directory name in the **KBAT** interface.

# 6. KABT INTERFACE AND FUNCTIONS

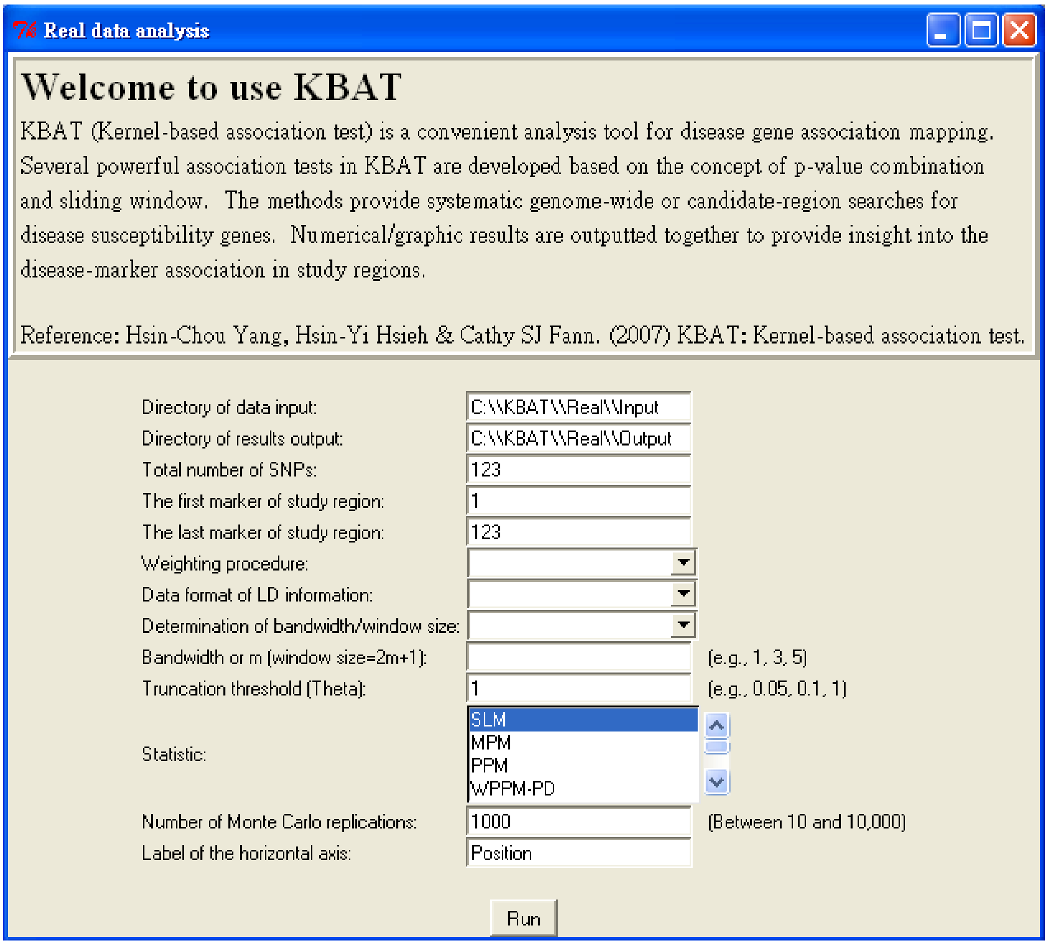
**KBAT** has a user friendly interface developed by R-GUI (See **Figure 1**). The interface contains a preface for a short introduction of **KBAT**. Thirteen item questions are designed for providing required/optional information for **KBAT** data analysis.

1. Directory of data input: Users should provide the working directory where their data are saved.
2. Directory of results output: Users should provide the working directory where their output should be saved. Note that the output directory must exist before executing **KBAT**.
3. Total number of SNPs: Users should provide the total number of SNPs in analysis.
4. The first marker of study region: Users should provide the first marker in the study.
5. The last marker of study region: Users should provide the last marker in the study.
6. Weighting procedure: Users should determine which type of weighting procedure will be used, including “Distance”, “LD”, and “LD and/or distance”.
7. Data format of LD information: Users should determine which type of data format of LD information will be provided, including “Not available”, “LD measure”, and “Genotype data”.
8. Determination of bandwidth/window size: Users should determine either bandwidth or window size to be used for window construction.
9. Bandwidth or m (window size = 2m+1): Bandwidth or window size should be inputted. Input of multiple bandwidths or window sizes in an analysis is admittable.
10. Truncation threshold (Theta): Truncation threshold should be provided. A threshold of 1 signifies that no truncation is applied. Input of multiple truncation thresholds in an analysis is admittable.
11. Statistic: Users should choose statistics which will be calculated in the analysis. Selection of more than one statistic in a run is allowed.
12. Number of Monte Carlo replications: The number of Monte Carlo replications should be provided. The number of Monte Carlo replications should be between 10 and 10,000.
13. Label of the horizontal axis: Users can provide a label of the horizontal axis.

Once all item questions are answered, icon “RUN” can be clicked to submit computational job. Then, **KBAT** will check the inputted data information and data files. If the inputted information is invalid or the data files are ill-format, **KBAT** shows warning message or error message, which provides users to make corrections. If the inputted data pass the examination, **KBAT** starts to perform analysis and a message “Please wait a while, KBAT is running…” will be shown in the command line. A prompt sign will appear immediately but the computation is proceeding. Please wait until a new window with the message “Computation of KBAT is finished.” jumps up to acknowledge users the completion of KBAT computation. Note that users can interrupt the execution of KBAT anytime by clicking the button  in the tool bar of RGUI window.

Once the execution of KBAT is finished, the numerical results and graphic outputs will be automatically saved in the output directory that users provide. The numerical results will be saved as a filename “output.txt” and the file will be automatically replaced if the next analysis is performed and the same output directory is set. The graphic results will be saved with a filename with respective to the inputted conditions (bandwidth/m and threshold). We suggest that users should remove figure files from a previous analysis before a new analysis in case of the confusion of multiple figure files from old and new analyses.

**Figure 1.** Interface of software **KBAT**



# 7. DATA INPUT FORMAT

Several data files should be provided.

1. The p-value data file with a filename "pv.txt": This file contains only one column. P-values from the single locus association tests are arranged in the order of marker position.
2. The map data file with a filename "map.txt": This file contains only one column. Physical or genetic positions of markers are recorded and arranged in the only column in this file. Note that the order of marker position in this file must match in the order of p-value in the file "pv.txt".
3. The LD data file with a filename "ld.txt": This file contains three columns, which provide intermarker LD coefficients of any two markers. The first two columns are the labels of two markers. The third column records the pair-wise LD coefficient. This file is optional and only should be supplied while users would like to calculated LD-based weights (the other type of data format for LD calculation is shown below).
4. The genotype data file with a filename "geno.txt": This file contains n rows and 2p columns. The n rows stand for n study individuals. The 2p columns are used to denote p SNP markers, where two columns are used to present a pair of alleles of a SNP marker. **KBAT** only analyzes SNP markers which are diallelic. Therefore, each column contains at most two numerical values. Missing data can be handled by inputting "NA". This file is optional and only should be provided while users would like to calculated LD-based weights.

# 8. EXAMPLES

In this section, we illustrated the execution of **KBAT** by using two examples.

**Example 1: Psoriasis data analysis**

This example examined the disease association between psoriasis and 123 SNP markers. Single locus association tests were performed by using TDT-AE. In total, 123 SNP markers on 17q25 (Helms et al., 2003) were analyzed. The p-values and physical positions for the 123 SNPs were saved in directory "C:\KBAT\Real\Example\Psoriasis". Here, we reanalyze this data with **KBAT**.

We copy files "pv.txt" and "map.txt" to the working directory "C:\KBAT\Real\Input". In the analysis, statistics SLM, PPM, WPPM-PD and KBAT-PD were calculated under the two window sizes of 3 (m=1) and 7 (m=3). Truncation was not considered in the analysis. Because LD information was not available in this example, only distance-based weight was used. The operating procedures are listed below and also shown in **Figure 2**:

(1) Directory of data input: "C:\\KBAT\\Real\\Input" was keyed in.

(2) Directory of results output: "C:\\KBAT\\Real\\Output" was keyed in.

(3) Total number of SNPs: "123" was inputted.

(4) The first marker of study region: "1" was inputted.

(5) The last marker of study region: "123" was inputted.

(6) Weighting procedure: "Distance" was selected.

(7) Data format of LD information: "Not available" was selected.

(8) Determination of bandwidth/window size: "Window" was selected.

(9) Bandwidth or m (window size = 2m+1): "1,3" was inputted.

(10) Truncation threshold (Theta): "1" was inputted.

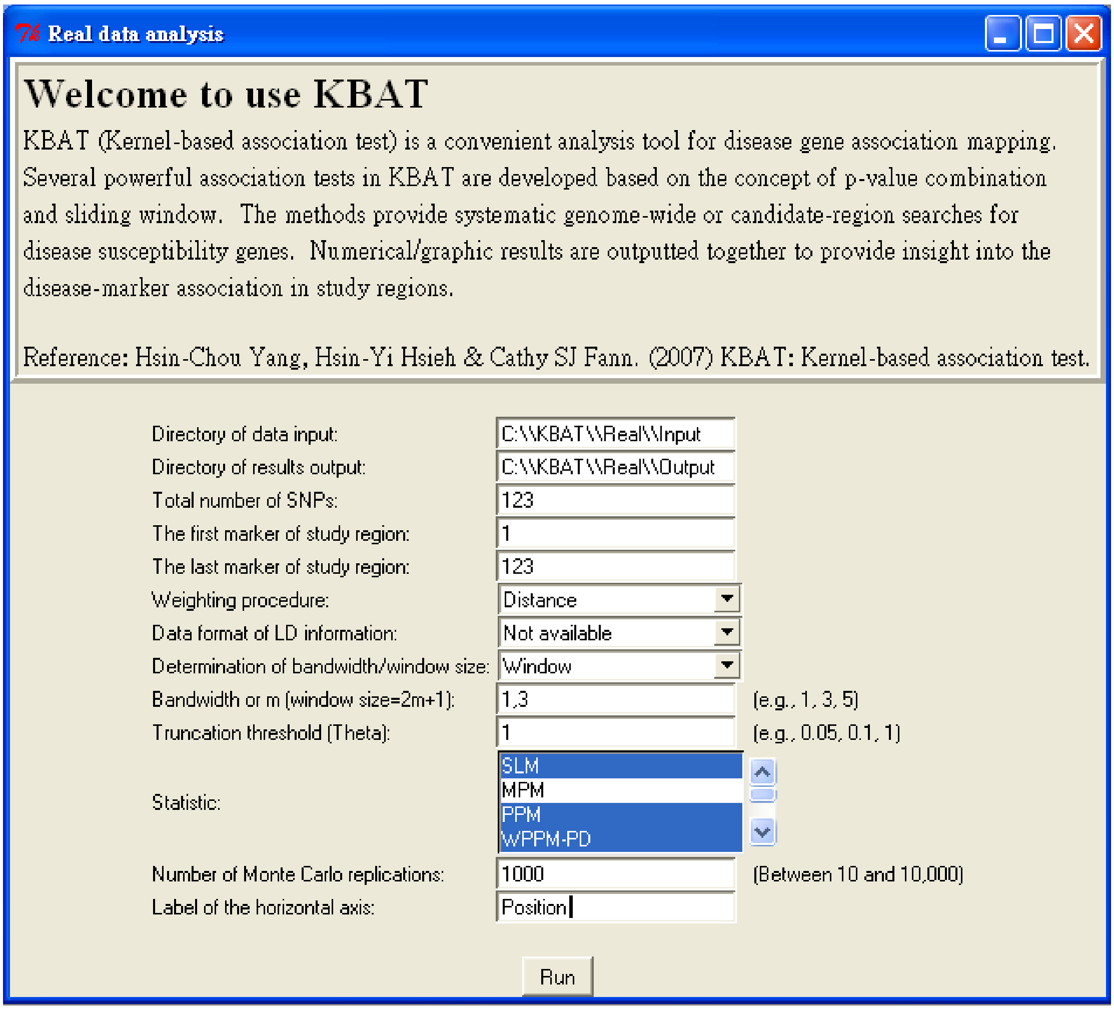
(11) Statistic: "SLM", "PPM", "WPPM-PD" and "KBAT-PD" were selected.

(12) Number of Monte Carlo replications: "1000" was inputted.

(13) Label of the horizontal axis: "Position" was keyed in.

(14) The icon "RUN" was pressed to execute **KBAT**.

**Figure 2.** Interface for the example of psoriasis data analysis




In this example, computation takes ~ 2 minutes with a PC having a CPU of Intel P4 3GHZ and 2GB RAM. Once the execution is finished, **KBAT** saves the numerical output in a file “Output.txt” (See **Table 1**) and the figure files (See **Figure 3**) in “C:\KBAT\Real\Output”. In **Table 1**, empirical p-values of 123 SNP markers based on the four test statistics are shown by m=1 and m=3 in order. Table title is shown first and followed by the names of variables. The first column is the index of SNP marker. The second to the fifth columns are empirical p-values of the four test statistics, SLM, PPM, WPPM-PD and KBAT-PD.

In **Figure 3**, empirical p-values are drawn. The vertical axis is the empirical p-values in a scale of $-\log_{10}$ and the horizontal axis is physical position. The titles of the subfigures demonstrate which test statistic and window size were used.

**Table 1.** Numerical output in the psoriasis data analysis

Table: P-values of all statistics for each marker while the m is 1 and truncation threshold is 1

| Marker | SLM | PPM | WPPM-PD | KBAT-PD |
|---|---|---|---|---|
| 1 | 1.000000e+00 | 8.030000e-01 | 1.000000e+00 | 8.370000e-01 |
| 2 | 5.153000e-01 | 8.230000e-01 | 5.400000e-01 | 8.130000e-01 |
| 3 | 6.685000e-01 | 6.940000e-01 | 6.440000e-01 | 7.150000e-01 |
| 4 | 6.510000e-01 | 4.970000e-01 | 6.470000e-01 | 4.810000e-01 |
| 5 | 1.809000e-01 | 1.640000e-01 | 2.050000e-01 | 1.470000e-01 |
| 6 | 5.780000e-02 | 2.060000e-01 | 6.500000e-02 | 1.330000e-01 |
| 7 | 1.000000e+00 | 3.740000e-01 | 1.000000e+00 | 6.190000e-01 |
| 8 | 7.457000e-01 | 6.890000e-01 | 7.640000e-01 | 6.520000e-01 |
| 9 | 2.511000e-01 | 4.520000e-01 | 2.430000e-01 | 4.440000e-01 |
| 10 | 3.488000e-01 | 4.740000e-01 | 3.400000e-01 | 3.970000e-01 |
| 11 | 7.760000e-01 | 4.210000e-01 | 7.540000e-01 | 4.690000e-01 |
| 12 | 2.110000e-01 | 2.560000e-01 | 2.240000e-01 | 3.640000e-01 |
| 13 | 9.740000e-02 | 1.980000e-01 | 1.050000e-01 | 2.010000e-01 |
| 14 | 4.564000e-01 | 2.990000e-01 | 4.260000e-01 | 2.560000e-01 |
| 15 | 5.561000e-01 | 7.600000e-01 | 5.480000e-01 | 7.900000e-01 |
| 16 | 1.000000e+00 | 8.450000e-01 | 1.000000e+00 | 8.600000e-01 |
| 17 | 6.723000e-01 | 6.650000e-01 | 6.700000e-01 | 7.250000e-01 |
| 18 | 2.607000e-01 | 4.350000e-01 | 2.570000e-01 | 3.860000e-01 |
| 19 | 3.438000e-01 | 4.880000e-01 | 3.310000e-01 | 4.940000e-01 |
| 20 | 9.557000e-01 | 7.970000e-01 | 9.450000e-01 | 8.520000e-01 |
| : | : | : | : | : |
| : | : | : | : | : |
| 116 | 6.339000e-01 | 7.270000e-01 | 6.420000e-01 | 7.840000e-01 |
| 117 | 9.656000e-01 | 7.800000e-01 | 9.700000e-01 | 8.570000e-01 |
| 118 | 4.732000e-01 | 8.800000e-01 | 4.610000e-01 | 7.580000e-01 |
| 119 | 1.000000e+00 | 7.520000e-01 | 1.000000e+00 | 8.150000e-01 |
| 120 | 5.403000e-01 | 7.700000e-01 | 5.350000e-01 | 6.800000e-01 |
| 121 | 5.220000e-01 | 5.930000e-01 | 5.270000e-01 | 5.920000e-01 |
| 122 | 4.597000e-01 | 6.870000e-01 | 5.000000e-01 | 5.860000e-01 |
| 123 | 7.069000e-01 | 6.180000e-01 | 6.700000e-01 | 6.790000e-01 |

Table: P-values of all statistics for each marker while the m is 3 and truncation threshold is 1

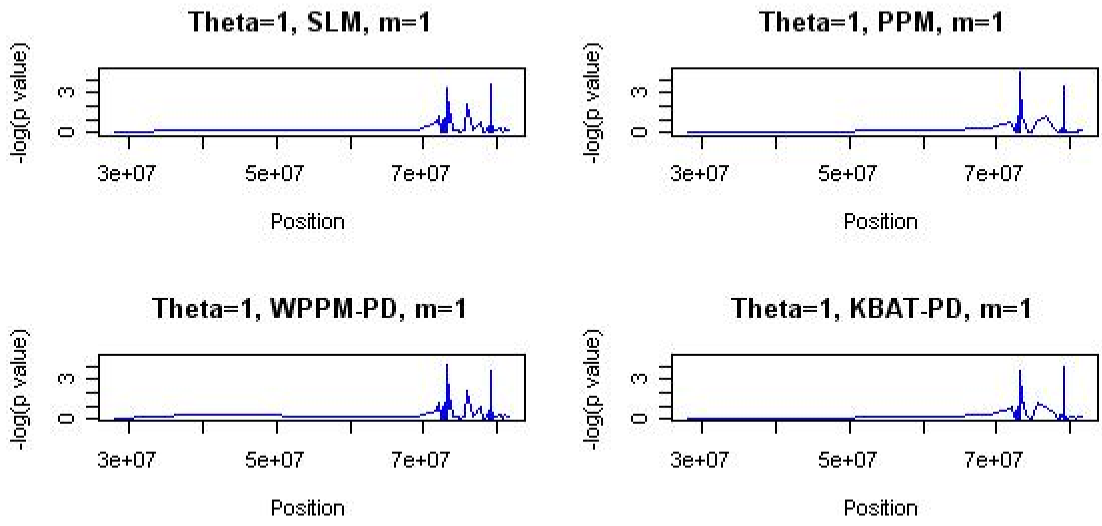
| Marker | SLM | PPM | WPPM-PD | KBAT-PD |
|---|---|---|---|---|
| 1 | 1.000000e+00 | 8.370000e-01 | 1.000000e+00 | 8.450000e-01 |
| 2 | 5.153000e-01 | 6.930000e-01 | 5.280000e-01 | 8.420000e-01 |
| 3 | 6.685000e-01 | 4.160000e-01 | 6.600000e-01 | 3.000000e-01 |
| 4 | 6.510000e-01 | 4.790000e-01 | 6.490000e-01 | 3.640000e-01 |
| 5 | 1.809000e-01 | 4.790000e-01 | 1.820000e-01 | 4.520000e-01 |
| 6 | 5.780000e-02 | 4.320000e-01 | 7.300000e-02 | 4.230000e-01 |
| 7 | 1.000000e+00 | 3.410000e-01 | 1.000000e+00 | 3.240000e-01 |
| 8 | 7.457000e-01 | 3.510000e-01 | 7.710000e-01 | 3.760000e-01 |
| 9 | 2.511000e-01 | 4.110000e-01 | 2.600000e-01 | 5.510000e-01 |
| 10 | 3.488000e-01 | 4.390000e-01 | 3.300000e-01 | 4.190000e-01 |
| 11 | 7.760000e-01 | 3.380000e-01 | 7.460000e-01 | 3.550000e-01 |
| 12 | 2.110000e-01 | 3.350000e-01 | 2.120000e-01 | 3.090000e-01 |
| 13 | 9.740000e-02 | 4.460000e-01 | 1.050000e-01 | 3.240000e-01 |
| 14 | 4.564000e-01 | 5.390000e-01 | 4.560000e-01 | 4.400000e-01 |
| 15 | 5.561000e-01 | 4.040000e-01 | 5.560000e-01 | 5.060000e-01 |
| 16 | 1.000000e+00 | 4.880000e-01 | 1.000000e+00 | 5.700000e-01 |
| 17 | 6.723000e-01 | 7.060000e-01 | 6.940000e-01 | 7.100000e-01 |
| 18 | 2.607000e-01 | 8.060000e-01 | 2.590000e-01 | 7.790000e-01 |
| 19 | 3.438000e-01 | 8.430000e-01 | 3.600000e-01 | 7.320000e-01 |
| 20 | 9.557000e-01 | 7.780000e-01 | 9.560000e-01 | 7.610000e-01 |
| : | : | : | : | : |
| : | : | : | : | : |
| 116 | 6.339000e-01 | 8.930000e-01 | 6.230000e-01 | 8.760000e-01 |

| | | | | |
|---|---|---|---|---|
| 117 | 9.656000e-01 | 8.630000e-01 | 9.670000e-01 | 8.520000e-01 |
| 118 | 4.732000e-01 | 7.950000e-01 | 4.670000e-01 | 7.800000e-01 |
| 119 | 1.000000e+00 | 8.200000e-01 | 1.000000e+00 | 7.750000e-01 |
| 120 | 5.403000e-01 | 8.430000e-01 | 5.610000e-01 | 7.760000e-01 |
| 121 | 5.220000e-01 | 7.600000e-01 | 5.190000e-01 | 7.520000e-01 |
| 122 | 4.597000e-01 | 7.950000e-01 | 4.470000e-01 | 7.060000e-01 |
| 123 | 7.069000e-01 | 6.670000e-01 | 7.020000e-01 | 6.900000e-01 |

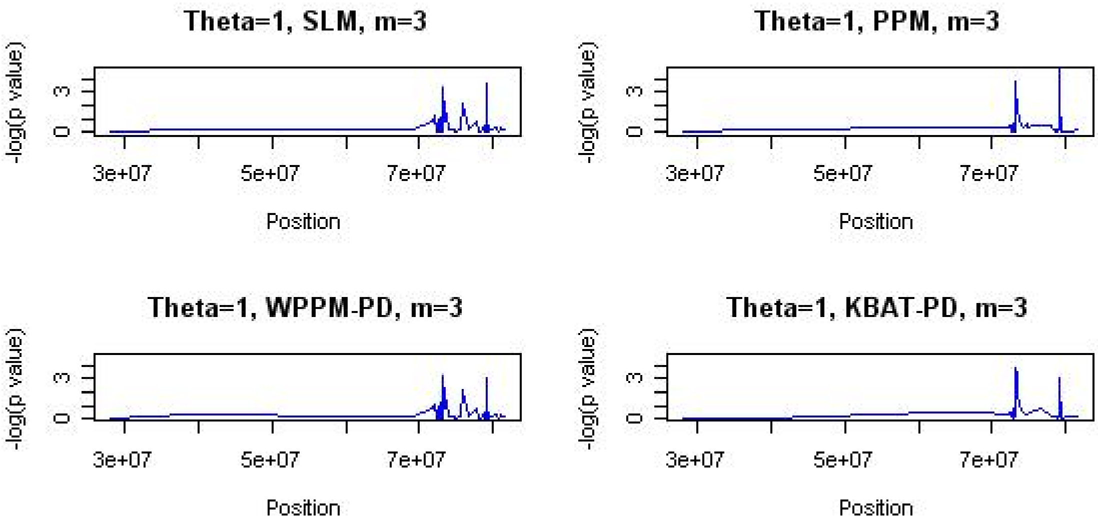
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

**Figure 3.** Graphic output in the psoriasis data analysis. Results will be saved as PDF files by bandwidths or window sizes. (A) Results for m=1. (B) Results for m=3.

(A)




(B)




**Example 2: Simulation data analysis**

In order to illustrate different types of data formats that KBAT can handle, we generated a data set which contained genotyped data of 31 SNP markers for 500 cases and 500 controls. Genotype data of 31 SNP markers were generated based on a disease model with penetrance PV=(0.1, 0.3, 0.4) for genotype (*dd*, *dD* and *DD*), where *D* was the disease allele. The intermarker recombination was set following a flat U recombination function, which can refer to (Yang et al., 2006). The true disease locus was arranged closed to the $16^{th}$ SNP. All files of genotype data, map data, LD data and p-value data were provided in the directory "C:\KBAT\Real\Example\Sim". Genotype data of the 31 SNP markers can refer to file "geno.txt". Map data of the 31 SNPs can refer to file "map.txt". Intermarker LD data of the 31 SNPs can refer to file "ld.txt". And, p-value data of single locus association tests can refer to file "pv.txt". Here, we reanalyze this data with **KBAT**.

We copied all of the files to the working directory "C:\KBAT\Real\Input". In the analysis, statistics WPPM-PD, WPPM-PDLD, KBAT-PD and KBAT-PDLD were calculated under the two window sizes of 3 (m=1) and 11 (m=5). Truncation was not considered in the analysis. Calculation of the four statistics requires both LD and position information. Therefore, in addition to files "pv.txt" and "map.txt", LD information should be provided. Users can do that by providing the LD file "ld.txt", and then KBAT can directly use the information to calculate p-value weights. In this case, "LD measure" in the item "Data format of LD information" should be selected. Or, users can do that by providing the genotype file "geno.txt", and then KBAT can help calculate LD. In this case, "Genotype data" in the item "Data format of LD information" should be selected. We illustrate the former situation in the following operating procedure (See **Figure 4**).

(1) Directory of data input: "C:\\KBAT\\Real\\Input" was keyed in.
(2) Directory of results output: "C:\\KBAT\\Real\\Output" was keyed in.
(3) Total number of SNPs: "31" was inputted.
(4) The first marker of study region: "1" was inputted.
(5) The last marker of study region: "31" was inputted.
(6) Weighting procedure: "LD and/or distance" was selected.
(7) Data format of LD information: "LD measure" was selected.
(8) Determination of bandwidth/window size: "Window" was selected.
(9) Bandwidth or m (window size = 2m+1): "1,5" was inputted.
(10) Truncation threshold (Theta): "1" was inputted.

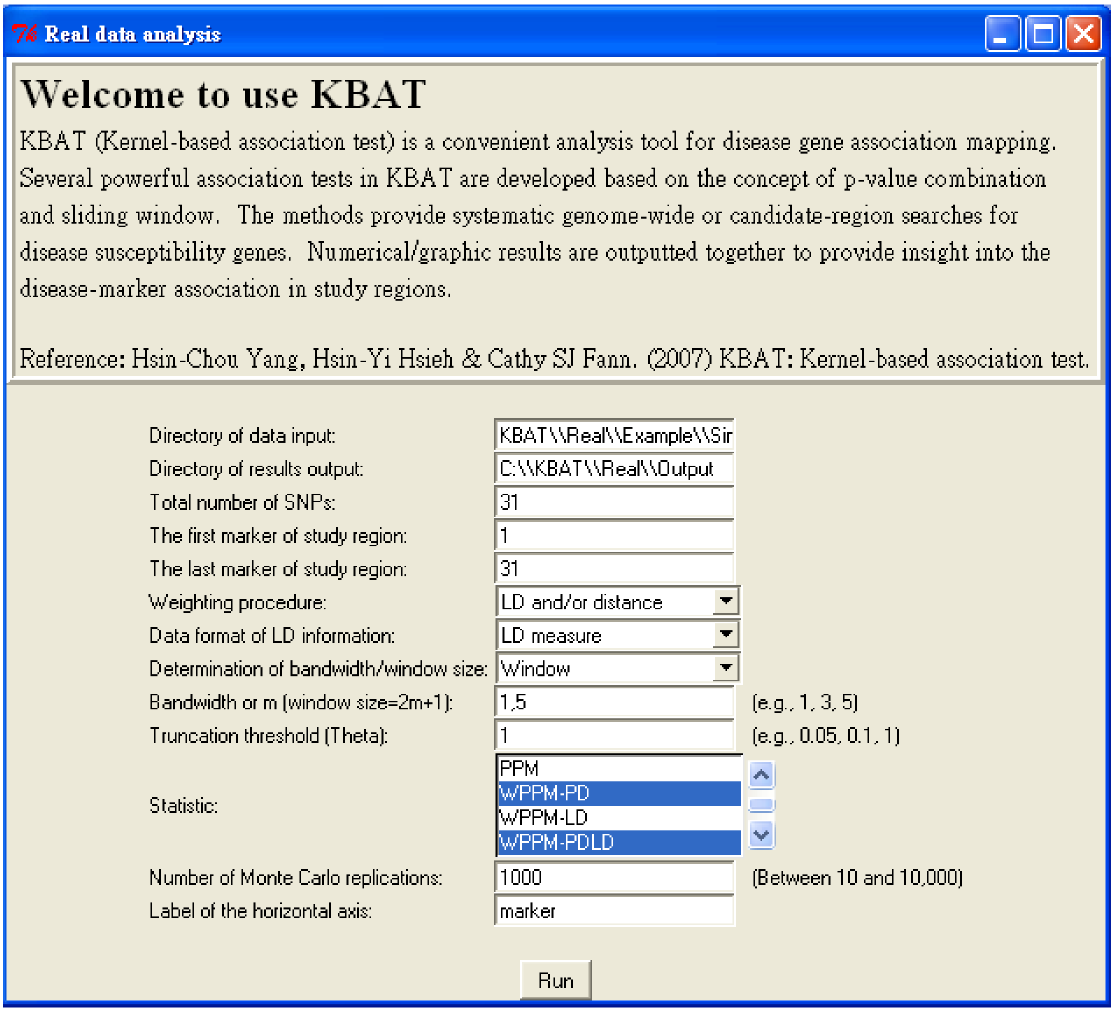
(11) Statistic: “WPPM-PD”, “WPPM-PDLD”, “KBAT-PD” and “KBAT-PDLD” were selected.

(12) Number of Monte Carlo replications: “1000” was inputted.

(13) Label of the horizontal axis: “Marker” was keyed in.

(14) The icon “RUN” was pressed to execute **KBAT**.

**Figure 4.** Interface for the example of simulation data analysis




In this example, computation will take ~ 2 minutes with a PC having a CPU of Intel P4 3GHZ and 2GB RAM. Once the execution is finished, **KBAT** saves the numerical output “Output.txt” (See **Table 2**) and the figure file (See **Figure 5**) in “C:\KBAT\Real\Output”. In **Table 2**, empirical p-values of 31 SNP markers based on the four test statistics are shown by different window sizes of m=1 and m=5 in order. Table title is shown first and followed by the names of variables. The first column is the index of SNP marker. The second to the fifth columns are empirical p-values of the four test statistics, WPPM-PD, WPPM-PDLD, KBAT-PD and KBAT-PDLD.

In **Figure 5**, empirical p-values are drawn. The vertical axis is the empirical p-values and the horizontal axis is physical position. The titles of the subfigures demonstrate which test statistic and window size were used.

**Table 2.** Numerical output in the simulation data analysis

Table: P-values of all statistics for each marker while the m is 1 and truncation threshold is 1

| Marker | WPPM-PD | WPPM-PDLD | KBAT-PD | KBAT-PDLD |
|---|---|---|---|---|
| 1 | 7.240000e-01 | 7.210000e-01 | 6.560000e-01 | 6.520000e-01 |
| 2 | 4.250000e-01 | 7.270000e-01 | 4.250000e-01 | 7.270000e-01 |
| 3 | 4.870000e-01 | 2.360000e-01 | 3.140000e-01 | 1.690000e-01 |
| 4 | 5.360000e-01 | 6.210000e-01 | 5.940000e-01 | 6.630000e-01 |
| 5 | 9.650000e-01 | 9.650000e-01 | 9.790000e-01 | 9.790000e-01 |
| 6 | 9.990000e-01 | 9.870000e-01 | 9.950000e-01 | 9.760000e-01 |
| 7 | 9.430000e-01 | 9.940000e-01 | 9.580000e-01 | 9.950000e-01 |
| 8 | 5.840000e-01 | 6.030000e-01 | 6.080000e-01 | 5.880000e-01 |
| 9 | 1.470000e-01 | 1.480000e-01 | 1.450000e-01 | 1.460000e-01 |
| 10 | 1.570000e-01 | 1.050000e-01 | 1.220000e-01 | 1.020000e-01 |
| 11 | 2.640000e-01 | 5.120000e-01 | 3.210000e-01 | 5.130000e-01 |
| 12 | 6.350000e-01 | 5.850000e-01 | 5.850000e-01 | 5.350000e-01 |
| 13 | 7.810000e-01 | 7.790000e-01 | 7.480000e-01 | 7.440000e-01 |
| 14 | 8.450000e-01 | 8.470000e-01 | 8.450000e-01 | 8.470000e-01 |
| 15 | 1.900000e-02 | 7.380000e-01 | 1.900000e-02 | 7.380000e-01 |
| 16 | 1.000000e-03 | 2.381095e-04 | 1.000000e-03 | 2.348265e-04 |
| 17 | 1.945080e-04 | 4.316122e-04 | 3.251567e-04 | 7.780396e-05 |
| 18 | 4.540000e-01 | 5.470000e-01 | 4.540000e-01 | 5.470000e-01 |
| 19 | 2.170000e-01 | 1.250000e-01 | 2.390000e-01 | 1.760000e-01 |
| 20 | 1.900000e-01 | 1.900000e-01 | 1.360000e-01 | 1.350000e-01 |
| 21 | 3.200000e-02 | 3.200000e-02 | 5.900000e-02 | 6.000000e-02 |
| 22 | 1.070000e-01 | 1.060000e-01 | 3.900000e-02 | 3.800000e-02 |
| 23 | 6.200000e-02 | 5.800000e-02 | 1.350000e-01 | 1.410000e-01 |
| 24 | 1.470000e-01 | 1.310000e-01 | 1.420000e-01 | 1.340000e-01 |
| 25 | 1.670000e-01 | 1.640000e-01 | 1.110000e-01 | 1.090000e-01 |
| 26 | 2.590000e-01 | 2.630000e-01 | 3.840000e-01 | 3.900000e-01 |
| 27 | 9.100000e-01 | 8.990000e-01 | 8.130000e-01 | 7.910000e-01 |
| 28 | 8.290000e-01 | 7.780000e-01 | 8.830000e-01 | 8.450000e-01 |
| 29 | 9.080000e-01 | 7.660000e-01 | 8.880000e-01 | 7.270000e-01 |
| 30 | 7.450000e-01 | 7.040000e-01 | 7.450000e-01 | 7.040000e-01 |

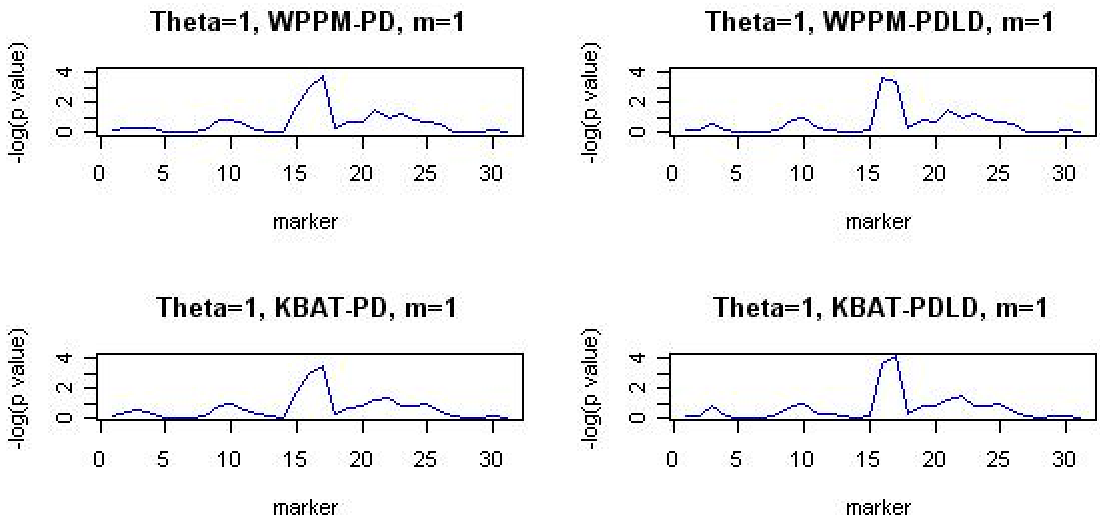
| | | | | |
|---|---|---|---|---|
| 31 | 8.650000e-01 | 8.650000e-01 | 8.650000e-01 | 8.650000e-01 |

Table: P-values of all statistics for each marker while the m is 5 and truncation threshold is 1

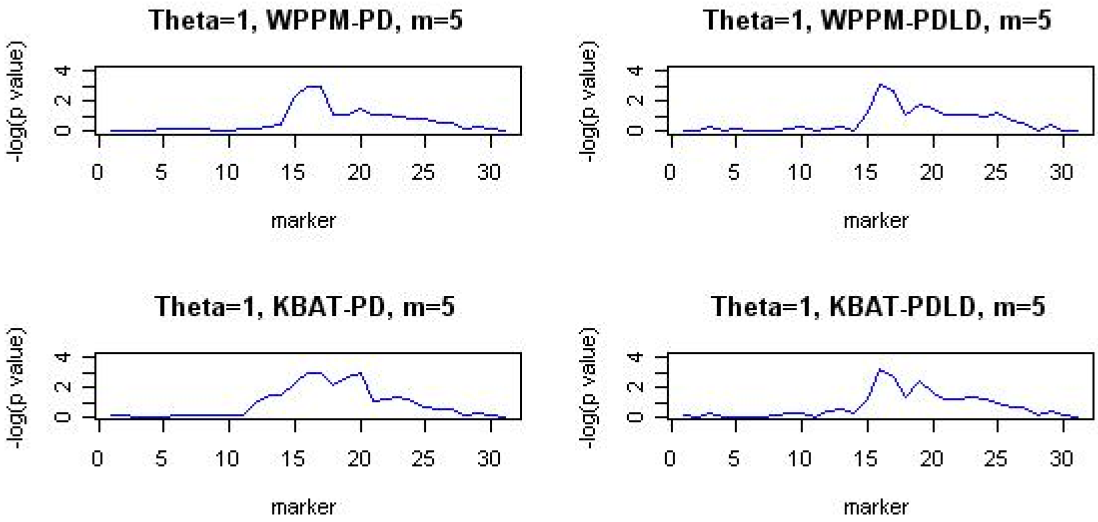
| Marker | WPPM-PD | WPPM-PDLD | KBAT-PD | KBAT-PDLD |
|---|---|---|---|---|
| 1 | 0.8220000000 | 0.8300000000 | 0.6680000000 | 0.6840000000 |
| 2 | 0.9140000000 | 0.9880000000 | 0.7640000000 | 0.9580000000 |
| 3 | 0.9310000000 | 0.5260000000 | 0.8650000000 | 0.4970000000 |
| 4 | 0.8190000000 | 0.9360000000 | 0.8950000000 | 0.9460000000 |
| 5 | 0.6400000000 | 0.6800000000 | 0.8580000000 | 0.8810000000 |
| 6 | 0.6860000000 | 0.9570000000 | 0.6860000000 | 0.9570000000 |
| 7 | 0.6950000000 | 0.9850000000 | 0.7220000000 | 0.9860000000 |
| 8 | 0.6310000000 | 0.7600000000 | 0.6670000000 | 0.7000000000 |
| 9 | 0.8510000000 | 0.7360000000 | 0.6260000000 | 0.4930000000 |
| 10 | 0.8030000000 | 0.5110000000 | 0.8030000000 | 0.5110000000 |
| 11 | 0.7000000000 | 0.8450000000 | 0.7000000000 | 0.8450000000 |
| 12 | 0.6140000000 | 0.7340000000 | 0.1270000000 | 0.4180000000 |
| 13 | 0.4530000000 | 0.4530000000 | 0.0380000000 | 0.2580000000 |
| 14 | 0.4050000000 | 0.8320000000 | 0.0320000000 | 0.4920000000 |
| 15 | 0.0050000000 | 0.0610000000 | 0.0050000000 | 0.0610000000 |
| 16 | 0.0010000000 | 0.0008103583 | 0.0010000000 | 0.0005896656 |
| 17 | 0.0010000000 | 0.0020000000 | 0.0010000000 | 0.0020000000 |
| 18 | 0.0830000000 | 0.0810000000 | 0.0070000000 | 0.0440000000 |
| 19 | 0.0690000000 | 0.0190000000 | 0.0020000000 | 0.0040000000 |
| 20 | 0.0350000000 | 0.0320000000 | 0.0010000000 | 0.0250000000 |
| 21 | 0.0720000000 | 0.0690000000 | 0.0720000000 | 0.0690000000 |
| 22 | 0.0690000000 | 0.0680000000 | 0.0690000000 | 0.0680000000 |
| 23 | 0.1050000000 | 0.0820000000 | 0.0410000000 | 0.0410000000 |
| 24 | 0.1370000000 | 0.1040000000 | 0.0790000000 | 0.0640000000 |
| 25 | 0.1590000000 | 0.0640000000 | 0.1780000000 | 0.0990000000 |
| 26 | 0.2500000000 | 0.2100000000 | 0.2500000000 | 0.2100000000 |
| 27 | 0.2410000000 | 0.2330000000 | 0.2410000000 | 0.2330000000 |
| 28 | 0.6340000000 | 0.8010000000 | 0.6340000000 | 0.8010000000 |
| 29 | 0.5080000000 | 0.3450000000 | 0.5080000000 | 0.3450000000 |
| 30 | 0.6450000000 | 0.7590000000 | 0.6450000000 | 0.7590000000 |
| 31 | 0.8770000000 | 0.8770000000 | 0.8770000000 | 0.8770000000 |

**Figure 5.** Graphic output in the simulation data analysis. The results will be saved as PDF files by the setting of bandwidth or window size. (A) Result for m=1; (B) Results for m=5.

(A)




(B)



# 9. KBAT VERSION UPGRADE

Versions:

KBAT Version 1.0: Oct. 2007

KBAT Version 1.1: Dec. 2008

**What are the new features in KBAT Version 1.1 compared to Version 1.0?**

(1) KBAT calculates combination statistic(s) of p-values from single-locus association tests. However, if a p-value of single-locus association test is very significant statistically, a combination of p-values may not gain statistical power. At the situation, KBAT will show a single-locus p-value instead of calculating an empirical p-value of p-value combination. The threshold is $10^{-5}$ in Version 1.0 and changed to a value of Bonferroni's level, i.e., $\alpha/M$, in Version 1.1, where $\alpha$ is test size and $M$ is the total number of single-locus association tests.

(2) KBAT calculates empirical p-values by using a Monte Carlo procedure. Monte Carlo may not observe any least probable outcomes, implying that the empirical p-value is $< 1/n$, where $n$ is the number of Monte Carlo replications. How to assign a value for the empirical p-value? In Version 1.0, empirical p-values are estimated by drawing a real number from a uniform distribution with a lower bound 0 and an upper bound $1/n$. In Version 1.1, the procedure is modified as follows: (a) if $n \geq$ the ceiling of $M/\alpha$, then the empirical p-value is assigned a value of Bonferroni's level, i.e., $\alpha/M$; (b) if $n <$ the ceiling of $M/\alpha$, then KBAT automatically supplements the number of Monte Carlo replications to the ceiling of $M/\alpha$. If still no least probable outcomes are observed, the empirical p-value is assigned a value of Bonferroni's level.

(3) How to assign a value to a window containing only a single p-value after considering the p-value truncation? In Version 1.0, an empirical p-value from a non-truncated statistic will be assigned to this window. In Version 1.1, a p-value of an anchor marker will be assigned to this window.

## 10. REFERENCE

1. Helms C, Cao L, Krueger JG, Wijsman EM, Chamian F, Gordon D, Heffernan M, Daw JAW, Robarge J, Ott J, Kwok PY, Menter A, Bowcock AM. 2003. A putative RUNX1 binding site variant between *SLC9A3R1* and *NAT9* is associated with susceptibility to psoriasis. *Nature Genetics* **35**: 349-256.
2. Yang HC, Lin CY, Fann CSJ. 2006. A sliding-window weighted linkage disequilibrium test. *Genetic Epidemiology* **30**: 531-545.

# 11. APPENDIX – TEST STATISTICS

- **Single locus method (SLM):**

$$Q_{i,m} = p_i, \forall i = 1, \cdots, N.$$

- **Minimum p-value method (MPM):**

$$Q_{i,m} = \min_{j \in \Im(i,m)} \{p_j\}, \forall i = 1, \cdots, N.$$

- **Product p-value method (PPM):**

$$Q_{i,m} = \sum_{j \in \Im(i,m)} \ln(p_j) I[p_j < \tau], \ \forall i = 1, \cdots, N.$$

- **Distance-weight product p-value method (WPPM-PD):**

$$Q_{i,m} = \sum_{j \in \Im(i,m)} w_j(i,m) \ln(p_j) I[p_j < \tau], \ \forall i = 1, \cdots, N, \quad w_j(i,m) = h_{i,j}^*, \text{ where}$$

$$h_{i,j}^* = h_{i,j} \Big/ \sum_{k \in \Im(i,m)} h_{i,k} \quad \text{and} \quad h_{i,j} = 1/(1 + d_{i,j}).$$

- **LD-weight product p-value method (WPPM-LD):**

$$Q_{i,m} = \sum_{j \in \Im(i,m)} w_j(i,m) \ln(p_j) I[p_j < \tau], \ \forall i = 1, \cdots, N, \quad w_j(i,m) = \hat{\rho}_{i,j}^*, \text{ where}$$

$$\hat{\rho}_{i,j}^* = \hat{\rho}_{i,j} \Big/ \sum_{k \in \Im(i,m)} \hat{\rho}_{i,k} \quad \text{and} \quad \hat{\rho}_{i,j} = \left| \hat{\lambda}_{11}(i,j) \hat{\lambda}_{22}(i,j) - \hat{\lambda}_{12}(i,j) \hat{\lambda}_{21}(i,j) \right| \Big/ \left| \hat{\lambda}_{1+}(i,j) \hat{\lambda}_{+2}(i,j) \right|.$$

- **Distance-LD-weight product p-value method (WPPM-PDLD):**

$$Q_{i,m} = \sum_{j \in \Im(i,m)} w_j(i,m) \ln(p_j) I[p_j < \tau], \ \forall i = 1, \cdots, N, \quad w_j(i,m) = \frac{h_{i,j}^* \times \hat{\rho}_{i,j}^*}{\sum_{k \in \Im(i,m)} h_{i,k}^* \times \hat{\rho}_{i,k}^*},$$

$$\text{where} \quad h_{i,j}^* = h_{i,j} \Big/ \sum_{k \in \Im(i,m)} h_{i,k} \quad \text{and} \quad \hat{\rho}_{i,j}^* = \hat{\rho}_{i,j} \Big/ \sum_{k \in \Im(i,m)} \hat{\rho}_{i,k}.$$

- **Kernel-based association test – physical distance (KBAT-PD):**

$$G_{t,h} = \sum_{i \in W(t,h)} \left( a_i \times \ln(p_i) \right), \text{ where } \quad a_i = \frac{\left( \mathrm{K}((t - t_i)/h) \right)}{\sum_{j \in W(t,h)} \left( \mathrm{K}((t - t_j)/h) \right)}.$$

- **Kernel-based association test – distance-LD (KBAT-PDLD):**

$$G_{t,h} = \sum_{i \in W(t,h)} \left( a_{i,j} \times \ln(p_i) \right), \quad \text{where} \quad a_{i,j} = \frac{\hat{h}_i^*(t) \times \hat{\rho}_{i,j}^*}{\sum_{j \in W(t,h)} \hat{h}_i^*(t) \times \hat{\rho}_{i,j}^*},$$

$$\hat{h}_i^*(t) = \frac{\left( \mathrm{K}((t_j - t_i)/h) \right)}{\sum_{k \in W(t,h)} \left( \mathrm{K}((t_k - t_i)/h) \right)} \quad \text{and} \quad \hat{\rho}_{i,j}^* = \hat{\rho}_{i,j} \Big/ \sum_{k \in \Im(i,m)} \hat{\rho}_{i,k}.$$