Version 1.0 (May, 2016)

OPATs User Guide

Chia-Wei Chen and Hsin-Chou Yang[†]

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

[†] hsinchou@stat.sinica.edu.tw

Table of Contents:

1. OPATs LICENSE

All copyright are reserved by authors of **OPATs**. **OPATs** are released under GPL_v2 license. We welcome any noncommercial use of **OPATs** for your own research. Commercial use of **OPATs** should be directed to hsinchou@stat.sinica.edu.tw. For free software **OPATs**, we assume no warranty and no responsibility for the results of analyses. If publications are based on the results from the use of **OPATs**, please cite the following reference:

Chia-Wei Chen and Hsin-Chou Yang (2017). OPATs: Omnibus P-Value Association Tests. *Briefings in Bioinformatics*. Under revision.

2. INTRODUCTION

OPATs (**O**mnibus ***P***-value **A**ssociation **T**ests), written in R and R graphical user interface (GUI), is an integrated analysis tool for identifying disease-associated genomic segments, functional genes, and biological pathways in genome-wide association studies. In addition to analysis modules for data quality control and single-locus association tests, **OPATs** provides three types of set-based association test: window-based association test (WBAT), gene-based association test (GBAT), and biopathway-based association test (BBAT). The significance of a set-based association test is evaluated by using resampling procedures.

3. SOFTWARE DOWNLOAD AND INSTALLATION

Execution of **OPATs** requires the installation of **OPATs** program and R program. In addition, several R packages should be installed. For data management, packages gdata, gtools, and R.utils should be installed. For calculation acceleration and memory allocation, packages matrixStats, inline, Rcpp[1, 2], RcppArmadillo [3], and bigmemory [4] should be installed. For interface display, packages tcltk and tcltk2 should be installed. In addition, three external tools should be installed. For execution of the inline package, Windows users should install Rtools (https://cran.r-project.org/bin/windows/Rtools/). For interactive visualization, the Tcl tool ActiveTcl8.5 (http://www.activestate.com/activetcl) should be installed. For phenotype-based permutation, PLINK should be saved in the same directory where **OPATs** is installed.

Procedures for downloading and installing the two programs are described as follows:

1. **OPATs**:

   OPATs program is available at the **OPATs** website at http://www.stat.sinica.edu.tw/hsinchou/genetics/association/OPATs.htm. The zipped file "OPATs (Version 1.0).rar" can be downloaded and then unzipped to obtain a directory "OPATs" containing the program codes of **OPATs** and two illustrated examples.

2. R:

   Users can download R program "R-3.3.3-win.exe" from the **OPATs** website. Or users can download 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 R. Select a platform (Linux, MacOS X, or Windows) for R execution in your end. Click the hyperlink "base" and select "Download R 3.3.2 for Windows". Then execute the file to install R to "C:\Program Files\R\R-3.3.2". After finishing the installation of R, doubly click the icon "R i386 3.3.2" or "R x64 3.3.2" to initialize R in a 32-bit or 64-bit system, respectively. 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. Note that the analyses provided by **OPATs** require some additional R packages: gdata, gtools, R.utils, matrixStats, inline, Rcpp, RcppArmadillo, bigmemory, tcltk, and tcltk2. These packages will be automatically downloaded when **OPATs** is initializing. **Note that users are suggested to use program**

R-3.3.0 or later for execution of OPATs. OPATs does not work with older R versions (e.g., R-3.0.X).

3. Rtools:

Windows users can download Rtools from the website of "Building R for Windows" at https://cran.r-project.org/bin/windows/Rtools/. According to the R version installed, users should choose the corresponding Rtools version to download (e.g., Rtools34.exe is for R-3.3.0 and later). Then execute the file to install Rtools to "C:\Rtools".

4. ActiveTcl8.5:

Program ActiveTcl Community Edition can be downloaded from the website of ActiveState. For 32-bit Windows system users, please download ActiveTcl from the following hyperlink http://www.activestate.com/activetcl/downloads/thank-you?dl=http://downloads.activestate.com/ActiveTcl/releases/8.5.18.0/ActiveTcl8.5.18.0.298892-win32-ix86-threaded.exe. For 64-bit Windows system users, please download ActiveTcl from the following hyperlink http://www.activestate.com/activetcl/downloads/thank-you?dl=http://downloads.activestate.com/ActiveTcl/releases/8.5.18.0/ActiveTcl8.5.18.0.298892-win32-x86_64-threaded.exe.

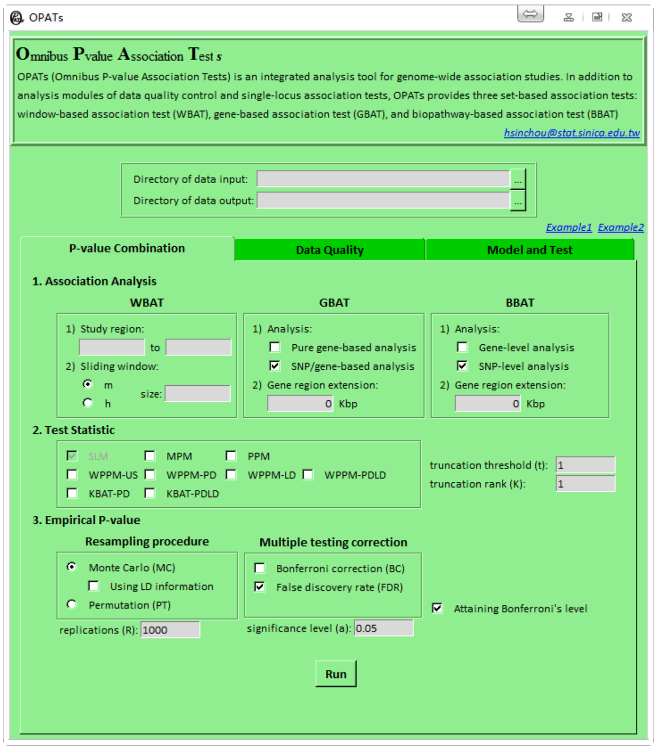
5. PLINK:

Users can download PLINK binary file from the website of "PLINK 1.90 beta" at https://www.cog-genomics.org/plink2/. Users click the "download" in the stable build according to the operating system. Next, users unzip the downloaded file, e.g., "plink_win64.zip", and then copy the "plink.exe" to the **OPATs** folder.

4. OPATs INITIALIZATION AND IMPLEMENTATION

Once **OPATs** is downloaded and unzipped, all files must be saved in the same destination directory, such as "D:/OPATs." Unzip OPATs.rar to obtain a folder named **OPATs** that contains programs and examples. Note that the folder can be saved anywhere. However, Windows users are recommended not to save the folder in the system disk in case of errors while saving output results. **OPATs** can be initialized in two ways: double click the executable file **OPATs.bat** or run R and execute the following two command lines in the R console; then, the **OPATs** interface is activated, as shown in Figure 1.

```
R> OPATs <- "D:/OPATs/OPATs.R"
R> source(OPATs)
```



OPATs

Omnibus Pvalue Association Test s

OPATs (Omnibus P-value Association Tests) is an integrated analysis tool for genome-wide association studies. In addition to analysis modules of data quality control and single-locus association tests, OPATs provides three set-based association tests: window-based association test (WBAT), gene-based association test (GBAT), and biopathway-based association test (BBAT)

hsinchou@stat.sinica.edu.tw

Directory of data input: ...
Directory of data output: ...

Example1 Example2

P-value Combination | Data Quality | Model and Test

1. Association Analysis

WBAT
1) Study region: to
2) Sliding window:
● m
○ h
size:

GBAT
1) Analysis:
□ Pure gene-based analysis
☑ SNP/gene-based analysis
2) Gene region extension: 0 Kbp

BBAT
1) Analysis:
□ Gene-level analysis
☑ SNP-level analysis
2) Gene region extension: 0 Kbp

2. Test Statistic

☑ SLM □ MPM □ PPM
□ WPPM-US □ WPPM-PD □ WPPM-LD □ WPPM-PDLD
□ KBAT-PD □ KBAT-PDLD

truncation threshold (t): 1
truncation rank (K): 1

3. Empirical P-value

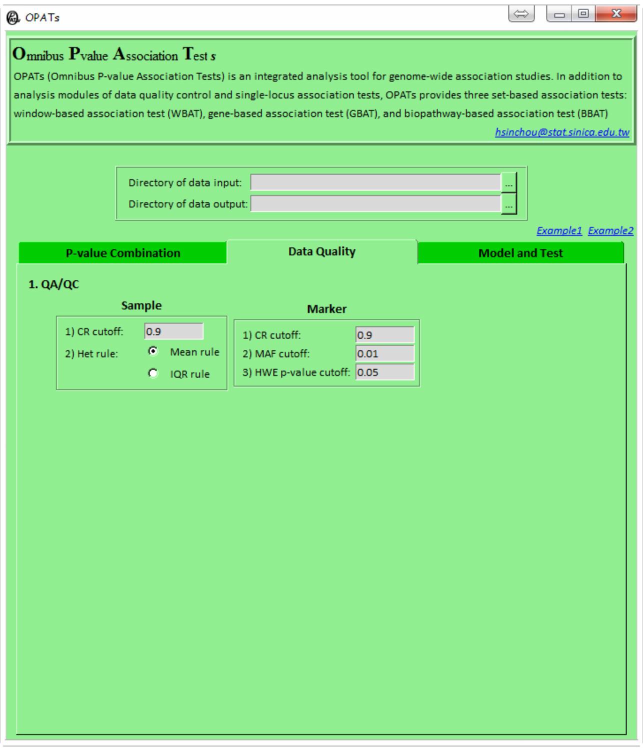
Resampling procedure
● Monte Carlo (MC)
□ Using LD information
○ Permutation (PT)
replications (R): 1000

Multiple testing correction
□ Bonferroni correction (BC)
☑ False discovery rate (FDR)
significance level (a): 0.05

☑ Attaining Bonferroni's level

Figure 1: Initial **OPATs** interface (the *P*-value Combination tab)

As shown in Figure 1, the **OPATs** interface comprises three parts. The first part is a preface to introduce **OPATs**. The second part contains the directories of data input and output. Users can either directly type the paths of directories into the edit boxes or press the Browse button to select directories. The input directory must be specified. An output directory named Output under the input directory is automatically generated if not specified. The final part contains the three function tabs: (1) *P*-value Combination (Figure 1); (2) Data Quality (Figure 2); (3) Model and Test (Figure 3). The details can refer to our publication paper.



OPATs (Omnibus P-value Association Tests) is an integrated analysis tool for genome-wide association studies. In addition to analysis modules of data quality control and single-locus association tests, OPATs provides three set-based association tests: window-based association test (WBAT), gene-based association test (GBAT), and biopathway-based association test (BBAT)

hsinchou@stat.sinica.edu.tw

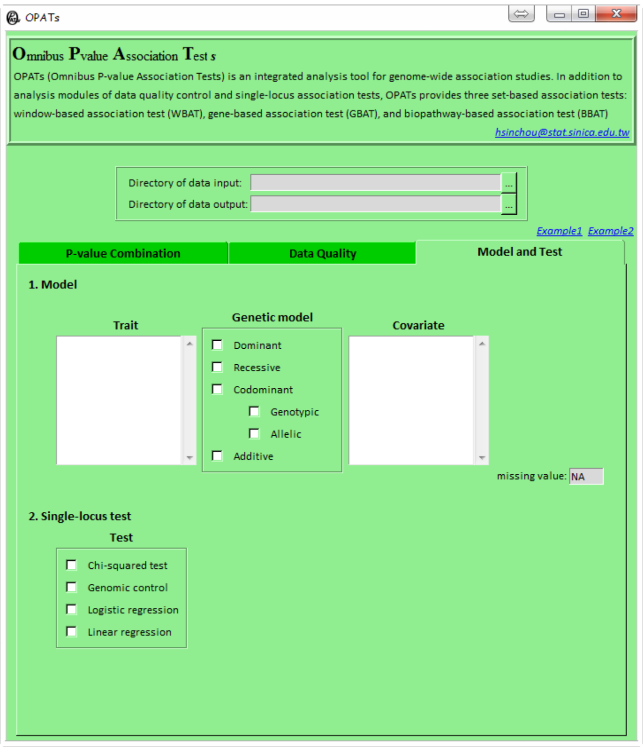
Directory of data input:

Directory of data output:

1) CR cutoff: 0.9
2) Het rule: ● Mean rule ○ IQR rule

1) CR cutoff: 0.9
2) MAF cutoff: 0.01
3) HWE p-value cutoff: 0.05

Figure 2: Second **OPATs** interface (the Data Quality tab)



OPATs (Omnibus P-value Association Tests) is an integrated analysis tool for genome-wide association studies. In addition to analysis modules of data quality control and single-locus association tests, OPATs provides three set-based association tests: window-based association test (WBAT), gene-based association test (GBAT), and biopathway-based association test (BBAT)

hsinchou@stat.sinica.edu.tw

Directory of data input:

Directory of data output:

□ Dominant
□ Recessive
□ Codominant
□ Genotypic
□ Allelic
□ Additive

missing value: NA

□ Chi-squared test
□ Genomic control
□ Logistic regression
□ Linear regression

Figure 3: Third **OPATs** interface (the Model and Test tab)

In addition to the Windows graphical user interface (GUI) environment, **OPATs** can be executed using command lines under the Windows and Linux environments. The commands and options are written as follows.

```
>R CMD BATCH --no-restore --arg [options] D:/OPATs/OPATs.R [Rout file]
```

Table 1 provides some options and their descriptions. By using this batch mode, users can analyze multiple data sets simultaneously.

Table 1: **OPATs** command line options

Option	Description	Value
--indir	Specify input directory of data	String
--outdir	Specify output directory of results	String
--assoc	Specify the type of association study	1 (WBAT), 2 (GBAT), or 3 (BBAT)
--seq	Specify which p-value sequence is analyzed	Integer
--st	For WBAT, specify the start of study region	Integer
--ed	For WBAT, specify the end of study region	Integer
--wintype	For WBAT, specify how to construct a window	1 (m) or 2 (h)
--winsize	For WBAT, specify a half size of a window	Integer
--anal	For GBAT or BBAT, specify the type of analysis	1 (pure gene GBAT or gene-level BBAT) or 2 (SNP/gene GBAT or SNP-level BBAT)
--ext	For GBAT or BBAT, specify how large gene regions to be extended	Numeric
--tau	Specify the truncation threshold	(0, 1]
--stat	Specify the test statistic	SLM, MPM, PPM, WPPM-US, WPPM-PD, WPPM-LD, WPPM-PDLD, KBAT-PD, KBAT-PDLD
--mtc	Specify the method of multiple testing correction	1 (Bonferroni) or 2 (FDR)
--alpha	Specify the significance level	(0, 1]
--tobl	Run more replications up to Bonferroni's level when there are no extreme results in specified replication times	0 (No) or 1 (Yes)
--proc	Specify the procedure for calculation of empirical *p*-values	1 (Monte Carlo), 2 (Monte Carlo with LD information), or 3 (permutation)

--b	Specify the number of replications/permutations	Integer
--ytype	Specify the phenotype variable for single-locus association analysis	String
--xtype	Specify the disease model for single-locus association analysis	1 (dominant), 2 (recessive), 3 (codominant-genotypic), 4 (codominant-allelic), and 5 (additive)
--test	Specify the test for single-locus association analysis	1 (chi-squared test), 2 (genomic control), 3 (logistic regression), or 4 (linear regression)
--icr	Specify the CR cutoff for individuals	(0, 1]
--het	Specify the Het. rule for individuals	1 (mean rule) or 2 (IQR rule)
--mcr	Specify the CR cutoff for markers	(0, 1]
--maf	Specify the MAF cutoff for markers	(0, 1]
--hwe	Specify the *p*-value cutoff of HWE test for markers	(0, 1]

5. DATA INPUT

The WBAT and GBAT analyze *p*-value data (or genotypic data) and annotation data. The BBAT analyzes *p*-value data (or genotypic data), annotation data, and gene set data. The input data formats are introduced below.

P-value data (.pv)

This is a white-space or comma-delimited file. Each row represents a marker and each column represents the *p*-value of a marker from different kinds of single-locus association tests. A header row is optional and used to indicate the genetic model that was used to obtain the *p*-value sequence data. A missing value is expressed as “NA.” An example of three markers, whose *p*-values were obtained on the basis of additive and genotypic models, is as follows.

Additive	Genotypic
0. 4306	0.3302
0. 2628	0.0928
NA	NA

Annotation data (.anno)

This is a white-space-delimited file. Each row describes the annotation information of a marker. The markers should be arranged in the same marker order as that in the *p*-value file. A header row is required and column names are fixed for the following information. (1) ID: marker id, (2) CHR: the chromosome in which a marker is located, (3) GID: gene id, (4) GSYM: gene symbol, (5) BP: physical position (unit: bp) of a marker, (6) CR: call rate, (7) MAF: minor allele frequency, (8) HWE: *p*-value of HWE test, (9) WS: the weight of a SNP marker in the GBAT and SNP-level BBAT, and (10) WG: the weight of a gene in the gene-level BBAT. For the GBAT and BBAT, the GSYM or GID is required. If a marker is located in more than one gene, multiple gene ids or symbols can be provided and separated by commas. If users provide the CR, MAF, and HWE in this file (optionally), **OPATs** can use the information to eliminate poor-quality markers according to the specified cutoffs in the second function tab (Data Quality). In addition, users have the option to provide weights (WS and/or WG) in this file. The weights are used to calculate WPPM-US in the Test Statistic option in the first function tab (*P*-value Combination). A missing value is expressed as “NA.” An example of three SNPs is given as follows.

ID	CHR	GID	GSYM	BP	CR	MAF	HWE	WS	WG
SNP_A-1938722	1	55092	TMEM51	15512757	1	0.457	0.6073	0.5	0.5
SNP_A-4217222	1	55083	KIF26B	245780760	1	0.3521	0.4447	0.5	0.2
SNP_A-2023862	1			194580718	1	0.001299	1	NA	NA

Gene set data (.gmt)

This is a tab-delimited file. Gmt is a popular format for a gene set biopathway analysis (e.g., Gene Set Enrichment Analysis; http://www.broadinstitute.org/gsea/index.jsp). No header is required, and each row represents a biopathway. The first two columns are the name and description of a biopathway and are followed by the gene symbols or ids in the biopathway. An example of two biopathways is given as follows.

hsa00010	Glycolysis / Gluconeogenesis	226	229	230	217
Citrate_cycle	NA	1738	4967	55753	1743

⋮

Genotypic data

OPATs adopts PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/), pedigree (*.ped and *.map), transposed pedigree (*.tped and *.tfam), long (*.lgen, *.map, and *.fam), and binary (*.bed, *.bim, and *.fam) file formats for genotypic data input. The genotype files are large for a large-scale genome-wide association studies. In such cases, partitioning of the file according to markers is recommended.

Phenotypic data (.pheno)

OPATs adopts the phenotype file format of PLINK. This is a white-space-delimited file that contains at least three columns. The first two columns are family and individual ids, followed by at least one phenotype. A header row is required. The first two column names are fixed as FID (family id) and IID (individual id), and the remaining headers are for phenotype data. **OPATs** identifies types of phenotype variables according to case-insensitive prefixes of header names; C and D indicate quantitative and dichotomous traits, respectively. **OPATs** uses the phenotypes listed in this file rather than the variable in the sixth column of the PLINK pedigree file. A missing value can be expressed as “-9.” An example including three phenotypes (status, trait1, and trait2) is given as follows.

FID	IID	dStatus	cTrait1	cTrait2
IND1	IND1	2	96.384	28
IND2	IND2	1	-9	17
IND3	IND3	1	89.2525	-9
⋮	⋮	⋮		⋮

Covariate data (.cov)

OPATs adopts the covariate file format of PLINK. This is a white-space-delimited file that contains at least three columns. The first two columns are family and individual ids, followed by at least one covariate. A header row is required. The first two column names are fixed as FID and IID, and the remaining headers are for covariate data. **OPATs** identifies types of covariates according to the case-insensitive prefixes of header names; C and D indicate quantitative and dichotomous traits, respectively. A missing value can be expressed as “-9.” An example of three covariates (age, gender, and BMI) is given as follows:

FID	IID	dAge	dGender	cBMI
IND1	IND1	2	2	34.0840
IND2	IND2	1	1	24.4649
IND3	IND3	1	2	19.3470
⋮	⋮	⋮	⋮	⋮

LD data (.ld)

There are two formats. One inherits the format provided in the KBAT (http://www.stat.sinica.edu.tw/hsinchou/genetics/association/KBAT.htm). This is a comma-delimited file which contains three columns. The first two columns are the indices of two markers according to the order in the annotation file, and the third column records their pairwise LD coefficients. The other one adopts the output file from PLINK with --r2 flag.

6. RESULT OUTPUT

When an analysis is complete, a message box reading “The analysis of **OPATs** is done.” appears to acknowledge the analysis completion. Numerical (.txt) and graphical (.pdf) results will be automatically generated and saved in the specified output directory. Numerical results contain three files: Description file (_NOTE.txt), Annotation file (_ANNO.txt), and Result files (_RESULT.txt). Table 2 summarizes the contents of each file depending on the selected set-based analysis. The details are illustrated in two examples in the Examples section. The graphical results are Manhattan plots of empirical *p*-values from different types of set-based association analyses, *p*-value sequences, truncation thresholds, test statistics, and resampling procedures. In a Manhattan plot, the vertical axis represents the empirical *p*-value in a $-\log_{10}$ scale. The horizontal axis represents the physical position of the genetic markers for the WBAT, a list of ordered genes for the GBAT, and a list of biopathways for the BBAT. If different multiple-testing corrections are performed, additional plots for the adjusted empirical *p*-values in a $-\log_{10}$ scale are provided. Meanwhile, graphical results are visually represented in the Output Viewer of **OPATs** (Figure 4).

Table 2. Contents of numerical results in text formats

Output	Information	Analysis		
		WBAT	GBAT	BBAT
Description file	**analysis_NOTE.txt**			
	parameter settings	O	O	O
	# of individuals/markers in the raw data	O	O	O
	# of individuals/markers excluded by QC	O$^+$	O$^+$	O$^+$
	# of individuals/markers in the validated data	O$^+$	O$^+$	O$^+$
	QC table	O$^+$	O$^+$	O$^+$
	# of markers after excluding missing *p*-values	O	O	O
	# of markers in the analysis	O	O	O
	summary table	O	O	O
Annotation file	**anlaysis_pvseq_ANNO.txt**			
	Pathway			O
	Gene		O	O
	SNP		O	O
	Chromosome		O	O
	Physical position		O	O
	single-locus *p*-value		O	O
Result file	**anlaysis_pvseq_(bandwidth)_PROC_RESULT.txt**			
	Pathway			O
	Gene		O	
	SNP	O		
	Chromosome	O	O	
	Physical position	O	O	
	# of markers in the region	O	O	O
	# of significant markers in the region	O	O	O
	min(single-locus *p*-value)	O	O	O
	marker attaining min(single-locus *p*-values)	O	O	O
	unadjusted/adjusted empirical *p*-value	O	O	O

$^+$ **:** Genotypic data or QC information (CR, MAF, or HWE) in the annotation data file (see the Data Input section).

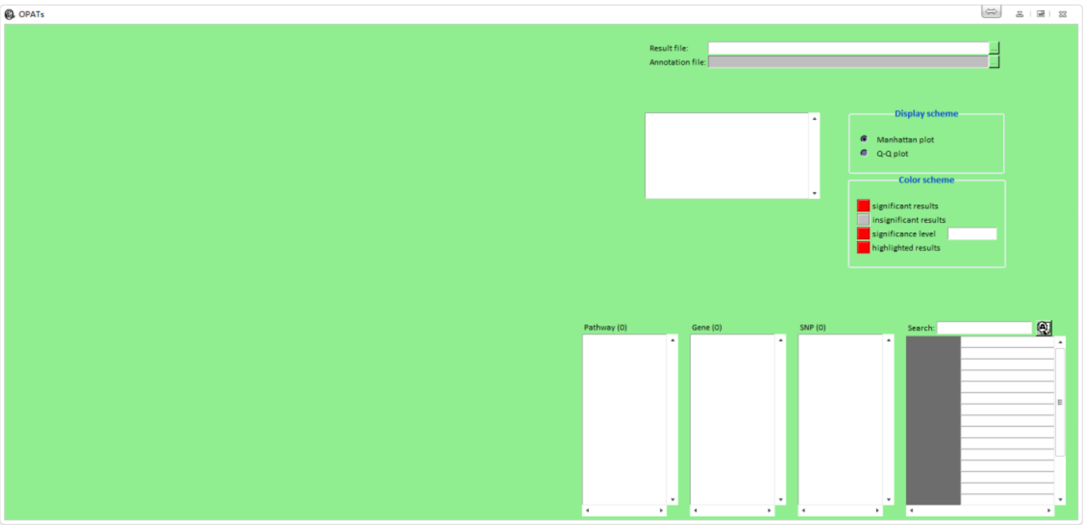


Figure 4. Output Viewer of **OPATs**

The Output Viewer can also be used to display graphical outputs of previous analyses from **OPATs**. Users can either directly type the paths of the result and annotation files into the edit boxes, or press the Browse button to select files. A Manhattan plot is displayed in the left-hand side panel. Users can select different color schemes to indicate significant and insignificant markers and the markers to be highlighted. Users can input a significance level to draw a horizontal reference line for the *p*-value. The biopathways, genes, and SNPs analyzed in the data set are listed. If users click a specific gene in the "Gene" panel, all SNPs located on this gene and their corresponding annotations will be shown in the "SNP" and final panels, respectively. Users can also use the "Search" function to search for biopathways, genes, and SNPs of interest in the data set. Finally, if users are interested in any point in the Manhattan plot, they can move the mouse cursor to the point to show the annotation information of the point.

7. EXAMPLES

OPATs provides two real examples. The first example analyzes *p*-value sequence data from a case-control study. The second example analyzes genotypic data from a population genetics study. These examples can be easily executed through the *Example1* and *Example2* hyperlinks on the first function tab (*P*-value Combination).

***P*-value data from a Wellcome Trust Case Control Consortium study**

The Wellcome Trust Case Control Consortium (WTCCC) recruited 1,999 rheumatoid arthritis (RA) cases and 3,002 normal controls in the British population [5]. All samples were genotyped using Affymetrix Human Mapping 500K Array Set. The sample contained 490,032 autosomal SNPs. The asymptotic *p*-values of Armitage trend tests with genomic control for 31,439 SNPs on chromosome 6 were calculated (*p*-value file: WTCCC.pv). SNP annotations comprising chromosome, physical position, and gene information were prepared according to the National Center for Biotechnology Information (NCBI) 37.3 (annotation file: WTCCC.anno). Quality control information comprising the CR, MAF, and HWE were provided in the annotation file. In this analysis, the cutoffs for excluding poor-quality SNPs were assigned as follows: a genotype CR of <0.9, an MAF of <0.01, and a *p*-value for the HWE test of <0.05. WTCCC.pv and WTCCC.anno are provided in "D:/OPATs/Examples/RA_WTCCC." Users can click *Example1* on the *P*-value Combination tab and press the Run button to run the GBAT, or click the BBAT frame and press the Run button to run the BBAT. The default setting did not specify an output directory. Therefore, results in this analysis were saved in "D:/OPATs/Examples/RA_WTCCC/Output." The GBAT and BBAT analyses produced three text files with numerical results (_NOTE.txt, _ANNO.txt, and _RESULT.txt) and one pdf file with graphical results.

For the GBAT, the note file (GBAT_NOTE.txt) initially recorded the parameter setting and data summary in this analysis (Table 3). The parameter settings revealed that a pure gene-based GBAT analysis was performed; a truncated PPM with a truncation threshold of 0.05 was considered; the number of Monte Carlo simulations was 10,000; and FDRs were performed for a multiple-testing correction with a significance level of 0.05. The data summary displayed the following information: “[1] Raw data information” shows that the raw data contained 31,439 SNPs; “[2] Validated data information” shows that 25,244 SNPs remained after eliminating 3,567 SNPs with an MAF of <0.01 and 3,550 SNPs with a *p*-value for HWE of <0.05; and “[3] Data summary for *p*-value sequence ‘Additive’” shows that no further SNPs were removed because of missing *p*-values. Finally, 25,244 SNPs were included in the subsequent analysis.

Table 3. GBAT_NOTE.txt

```
Start time: 2017-04-27 14:08:23

● PARAMETER SETTING
------------------------------------------------------------------------------------------
[1] P-value Combination
   1. Association Analysis component
        1) Procedure                                          :: Pure gene-based GBAT
   2. Test Statistic component
        1) Test statistic                                     :: PPM
        2) Truncation threshold                               :: 0.05
   3. Empirical P-value component
        1) Resampling procedure (# of replications)           :: Monte Carlo (10000)
        2) Multiple testing correction (significance level)   :: FDR (0.05)
        3) Attaining Bonferroni's level                       :: Yes
------------------------------------------------------------------------------------------


● DATA SUMMARY

[1] Raw data information
        • 31439 markers in the raw data

[2] Validated data
        • 25244 marker(s) in the validated data
                0 marker(s) are excluded due to duplicated name(s)
                0 marker(s) are excluded due to missing/incorrect information about chromosome(s)
              172 marker(s) are excluded due to missing/incorrect information about physical position(s)
                0 marker(s) are excluded due to CR < 0.9
             3567 marker(s) are excluded due to MAF < 0.01
             3550 marker(s) are excluded due to pHWE < 0.05
       •Note: The above quality control procedures exclude marker(s) independently (not hierarchically)
```

```
         ----------------------------------------------------------------
         Excluded Marker      HOMO        MAF         pHWE
         ----------------------------------------------------------------
         SNP_A-2225079                    9e-04
         SNP_A-2307479                    0.001       0
         SNP_A-2064685                    6e-04
                 ⋮
         ----------------------------------------------------------------

[3] Data summary for p-value sequence "Additive":
    •  0 marker(s) are excluded due to missing p-value(s)::
    •  0 marker(s) have zero p-value(s)::
    •  1009 gene(s), 25244 marker(s) are remained in the analysis
         -----------------------------------------------------------------------------------
                                    Chromosome-wise frequency of markers
          Chromosome   ---------------------------------------------------------------------
                         Gene      Intragenic SNPs    Intergenic SNPs    Total SNPs
         -----------------------------------------------------------------------------------
         6              1009       10435              14809              25244
         -----------------------------------------------------------------------------------
         Total          1009       10435              14809              25244


End time: 2017-04-27 14:10:09
```

Second, the annotation file (GBAT_Additive_ANNO.txt) recorded the annotation information of intragenic SNPs and the corresponding *p*-values in the final column titled “SLM” (Table 4.). Among 25,244 SNPs, there were 10,435 intragenic SNPs. The SNPs were arranged according to the chromosome order and physical position.

Table 4. GBAT_Additive_ANNO.txt

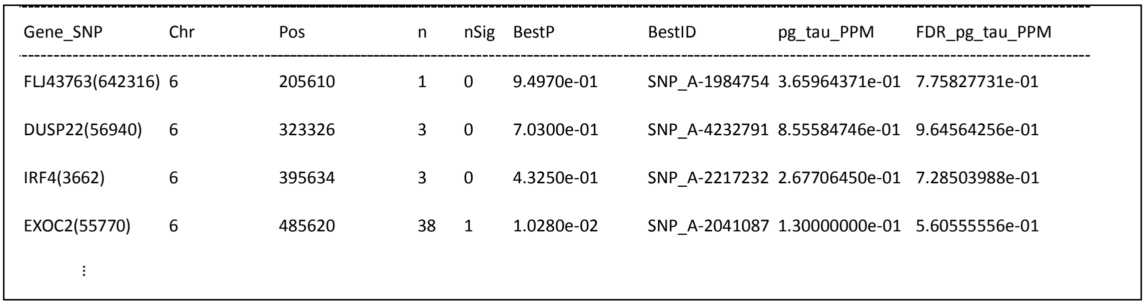
Table: single-locus *p*-values of intragenic SNPs

Gene_ID	Gene_Symbol	SNP	Chr	Pos	SLM
642316	FLJ43763	SNP_A-1984754	6	205610	9.497e-01
56940	DUSP22	SNP_A-4232791	6	323326	7.03e-01
56940	DUSP22	SNP_A-1937356	6	334825	9.881e-01
56940	DUSP22	SNP_A-2175990	6	349386	9.336e-01
⋮					

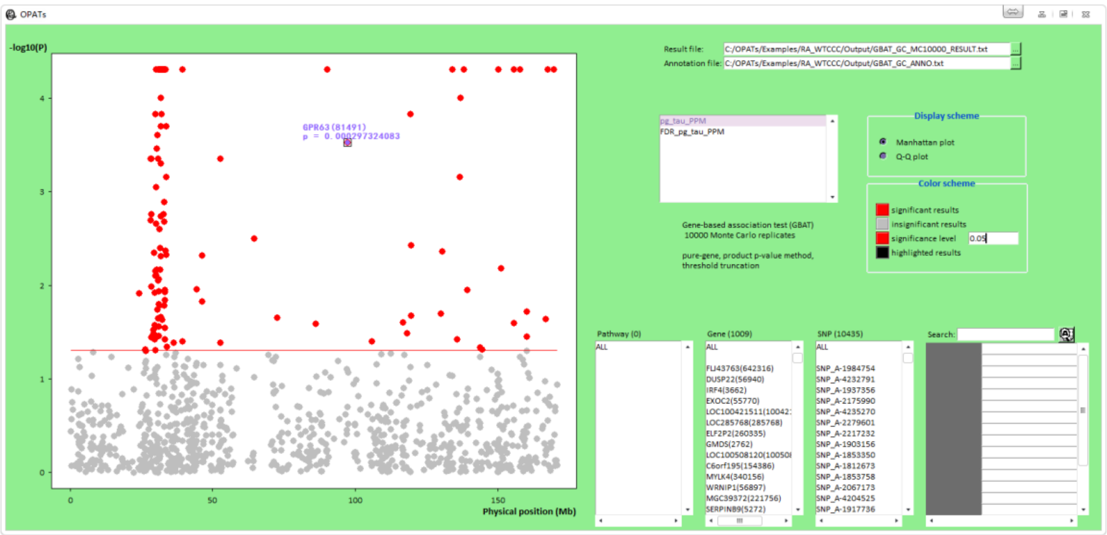
Finally, the results file (WTCCC_GBAT_Additive_MC10000_RESULT.txt) recorded the results of the GBAT combined with the basic information regarding the SNPs and genes (Table 5). For example, the fourth gene (gene symbol, *EXOC2* and gene id, 55770) was located at 485,620 bp (physical position) on chromosome 6 according to the first SNP on this gene in the annotation file. This gene contained 38 SNPs (n = 38) in these data. Among the 38 SNPs, SNP_A-2041087 (BestID) had the smallest *p*-value of 1.0280e-02 (BestP), and it was the only SNP with a *p*-value of <0.05 (nSig = 1) in this gene. The remaining columns recorded empirical *p*-values. The nomenclature of the header variables follows the Strategy_Truncation_Statistic rule. The first field "Strategy" illustrates the strategy ("pg" = "pure gene-based analysis" and "sg" = "composite of SNP- and gene-based analyses"). The second field "Truncation" illustrates the types of *p*-value truncation ("tau" = "threshold truncation" and "rank" = "rank truncation"). The third field "Statistic" presents the statistic, such as the MPM and PPM. If Bonferroni or FDR correction was performed for a multiple-testing correction, "Bonf" or "FDR" were added as a prefix, respectively.

Figure 5 shows the Manhattan plot for the first analysis (pg_tau_PPM). Red and gray points indicate significant and insignificant SNPs, respectively. A horizontal reference line indicates a *p*-value of 0.05. All the 1,009 genes, which contained 10,435 intragenic SNPs, are displayed. Moreover, users can display the Manhattan plot in the second analysis by clicking "FDR_pg_tau_PPM" in the upper-right window. This analysis identified some previously reported RA-associated genes, such as *BTNL2*, *TNFAIP3*, and a number of genes in the major histocompatibility complex region.

Table 5. GBAT_Additive_MC10000_RESULT.txt



```
Gene_SNP          Chr     Pos        n    nSig  BestP        BestID          pg_tau_PPM        FDR_pg_tau_PPM
------------------------------------------------------------------------------------------------------------
FLJ43763(642316)  6       205610     1    0     9.4970e-01   SNP_A-1984754   3.65964371e-01    7.75827731e-01

DUSP22(56940)     6       323326     3    0     7.0300e-01   SNP_A-4232791   8.55584746e-01    9.64564256e-01

IRF4(3662)        6       395634     3    0     4.3250e-01   SNP_A-2217232   2.67706450e-01    7.28503988e-01

EXOC2(55770)      6       485620     38   1     1.0280e-02   SNP_A-2041087   1.30000000e-01    5.60555556e-01
        ⋮
```



Result file: C:/OPATs/Examples/RA_WTCCC/Output/GBAT_GC_MC10000_RESULT.txt

Annotation file: C:/OPATs/Examples/RA_WTCCC/Output/GBAT_GC_ANNO.txt

pg_tau_PPM
FDR_pg_tau_PPM

Gene-based association test (GBAT)
10000 Monte Carlo replicates

pure-gene, product p-value method,
threshold truncation

Display scheme
● Manhattan plot
○ Q-Q plot

Color scheme
🟥 significant results
⬜ insignificant results
🟥 significance level 0.05
⬛ highlighted results

Pathway (0)
ALL

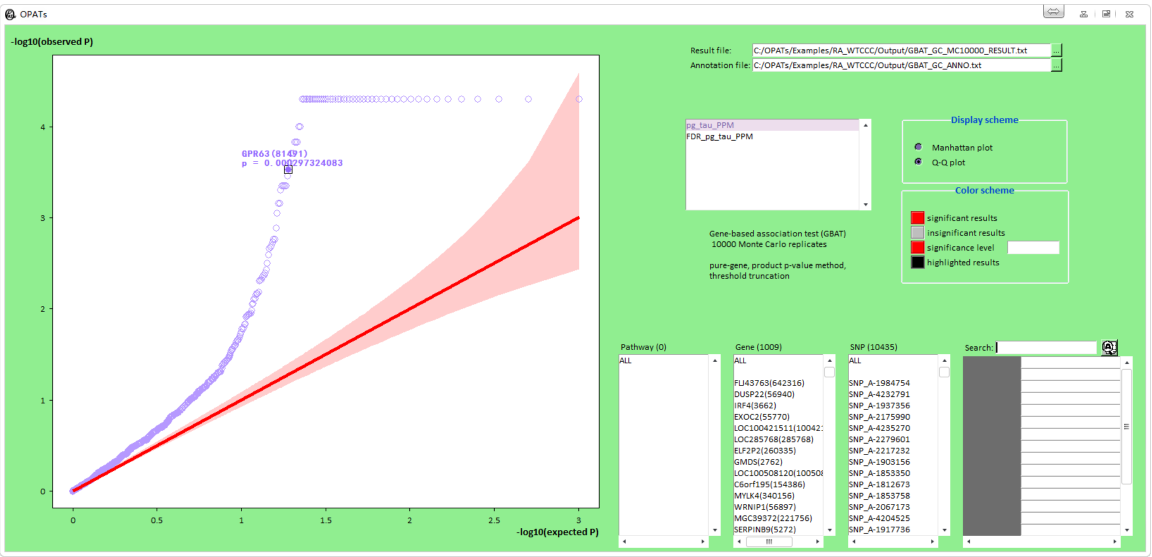
Gene (1009)
ALL
FLJ43763(642316)
DUSP22(56940)
IRF4(3662)
EXOC2(55770)
LOC100421511(10042
LOC285768(285768)
ELF2P2(260335)
GMDS(2762)
LOC100508120(10050
C6orf195(154386)
MYLK4(340156)
WRNIP1(56897)
MGC39372(221756)
SERPINB9(5272)

SNP (10435)
ALL
SNP_A-1984754
SNP_A-4232791
SNP_A-1937356
SNP_A-2175990
SNP_A-4235270
SNP_A-2279601
SNP_A-2217232
SNP_A-1903156
SNP_A-1853350
SNP_A-1812673
SNP_A-1853758
SNP_A-2067173
SNP_A-4204525
SNP_A-1917736

Search:

Figure 5. Manhattan plot and quantile–quantile (Q–Q) plot in the GBAT analysis of Example 1

Figure 6 shows the Q-Q plot. The vertical axis (Y axis) indicates the observed *p*-values (in a –$\log_{10}$ scale) and horizontal axis (X axis) indicates the expected *p*-values (in a –$\log_{10}$ scale). A red line of X = Y and the corresponding 95% confidence bands are provided.



Result file: C:/OPATs/Examples/RA_WTCCC/Output/GBAT_GC_MC10000_RESULT.txt

Annotation file: C:/OPATs/Examples/RA_WTCCC/Output/GBAT_GC_ANNO.txt

pg_tau_PPM
FDR_pg_tau_PPM

Gene-based association test (GBAT)
10000 Monte Carlo replicates

pure-gene, product p-value method,
threshold truncation

Display scheme
○ Manhattan plot
● Q-Q plot

Color scheme
🟥 significant results
⬜ insignificant results
🟥 significance level
⬛ highlighted results

Pathway (0)
ALL

Gene (1009)
ALL
FLJ43763(642316)
DUSP22(56940)
IRF4(3662)
EXOC2(55770)
LOC100421511(10042
LOC285768(285768)
ELF2P2(260335)
GMDS(2762)
LOC100508120(10050
C6orf195(154386)
MYLK4(340156)
WRNIP1(56897)
MGC39372(221756)
SERPINB9(5272)

SNP (10435)
ALL
SNP_A-1984754
SNP_A-4232791
SNP_A-1937356
SNP_A-2175990
SNP_A-4235270
SNP_A-2279601
SNP_A-2217232
SNP_A-1903156
SNP_A-1853350
SNP_A-1812673
SNP_A-1853758
SNP_A-2067173
SNP_A-4204525
SNP_A-1917736

Search:

Figure 6. Quantile-quantile (Q-Q) plot in the GBAT analysis of Example 1

A biopathway-based association test was conducted by using the BBAT. We used a gmt file (WTCCC.gmt) from Kyoto Encyclopedia of Genes and Genomes [6], which contains 303 biopathways. The note file results (BBAT_NOTE.txt) were similar to that of the GBAT (Table 6). The main difference was that biopathway information was added in "[3] Data summary for *p*-value sequence 'Additive'."

Table 6. BBAT_NOTE.txt

```
Start time: 2017-04-27 15:21:19

●  PARAMETER SETTING
------------------------------------------------------------------------------------------------------------
[1] P-value Combination
   1. Association Analysis component
        1) Procedure                                          :: Gene-level BBAT
   2. Test Statistic component
        1) Test statistic                                     :: PPM
        2) Truncation threshold                               :: 0.05
   3. Empirical P-value component
        1) Resampling procedure (# of replications)           :: Monte Carlo (10000)
        2) Multiple-test correction (significance level)      :: FDR (0.05)
------------------------------------------------------------------------------------------------------------

●  DATA SUMMARY

[1] Raw data
        ・31439 markers in the raw data

[2] Validated data information
        ・25244 marker(s) in the validated data
                 0 marker(s) are excluded due to duplicated name(s)
                 0 marker(s) are excluded due to missing/incorrect information about chromosome(s)
               172 marker(s) are excluded due to missing/incorrect information about physical position(s)
                 0 marker(s) are excluded due to CR < 0.9
              3567 marker(s) are excluded due to MAF < 0.01
              3550 marker(s) are excluded due to pHWE < 0.05
        ・Note: The above quality control procedures exclude marker(s) independently (not hierarchically)
             ------------------------------------------------------------------
              Excluded Marker      HOMO          MAF           pHWE
             ------------------------------------------------------------------
              SNP_A-2225079                      9e-04

              SNP_A-2307479                      0.001         0

              SNP_A-2064685                      6e-04

                        ⋮
             ------------------------------------------------------------------

[3] Data summary for p-value sequence "Additive":
        •  0 marker(s) are excluded due to missing p-value(s)::
        •  0 marker(s) have zero p-value(s)::
```

```
• 303 pathway(s), 1009 gene(s), 25244 marker(s) are remained in the analysis
--------------------------------------------------------------------------------------------------
                                                             Frequency
           Pathway                   -------------------------------------------------------------
                                     Genes in pathway   Genes in the data   SNPs in the data
--------------------------------------------------------------------------------------------------
2-Oxocarboxylic acid metabolism      17                 0                   0
ABC transporters                     44                 3                   18
Acute myeloid leukemia               57                 1                   12
               ⋮
--------------------------------------------------------------------------------------------------
                                     24036              1049                12423


End time: 2017-04-27 15:22:44
```

The annotation file (BBAT_Additive_ANNO.txt) listed biopathways that contained at least one SNP in the data set, followed by the gene and intragenic SNP ids, chromosome, physical position, and corresponding SLM *p*-values of the intragenic SNPs in the biopathways (Table 7).

Table 7. GBAT_Additive_ANNO.txt

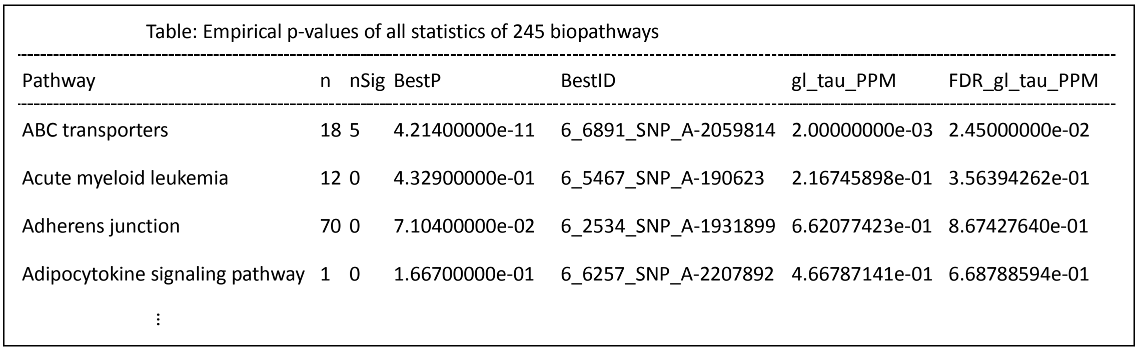
Table: single-locus *p*-values of intragenic SNPs

Pathway	Gene_ID	SNP	Chr	Pos	SLM
ABC_transporters	646104	SNP_A-1814092	6	151417571	2.053e-01
ABC_transporters	100507662	SNP_A-4278462	6	151424498	5.444e-01
ABC_transporters	100507662	SNP_A-1897473	6	151439698	7.349e-01
⋮					

The results file (BBAT_Additive_MC10000_RESULT.txt) recorded the results of the BBAT in addition to the basic information regarding SNPs, genes, and biopathways (Table 8). For example, the ABC transporter was the first biopathway and contained 18 SNPs (n = 18), among which seven (nSig = 5) were marginally significant. Of these seven SNPs, the most significant was SNP_A-2059814, which was located on *TAP2* (6891) on chromosome 6 (BestID = 6_6891_SNP_A-2059814).

The remaining columns recorded empirical *p*-values. The variable nomenclature followed the Strategy_Truncation_Statistic rule, in which "Strategy" was determined as "gl" = "gene-level biopathway analysis" or "sl" = "SNP-level biopathway analysis". In this example, 245 of the 303 biopathways provided data for analysis, and 91 biopathways were significant according to the FDR-adjusted *p*-values of <0.05. Some crucial biopathways associated with RA, such as gonadotropin-releasing hormone signaling and mitogen-activated protein kinase signaling pathways, were identified in this example.

Table 8. BBAT_Additive_MC10000.txt



```
                  Table: Empirical p-values of all statistics of 245 biopathways
-----------------------------------------------------------------------------------------------------------------------
Pathway                            n   nSig  BestP             BestID                  gl_tau_PPM        FDR_gl_tau_PPM
-----------------------------------------------------------------------------------------------------------------------
ABC transporters                   18  5     4.21400000e-11    6_6891_SNP_A-2059814    2.00000000e-03    2.45000000e-02
Acute myeloid leukemia             12  0     4.32900000e-01    6_5467_SNP_A-190623     2.16745898e-01    3.56394262e-01
Adherens junction                  70  0     7.10400000e-02    6_2534_SNP_A-1931899    6.62077423e-01    8.67427640e-01
Adipocytokine signaling pathway    1   0     1.66700000e-01    6_6257_SNP_A-2207892    4.66787141e-01    6.68788594e-01
                  ⋮
```

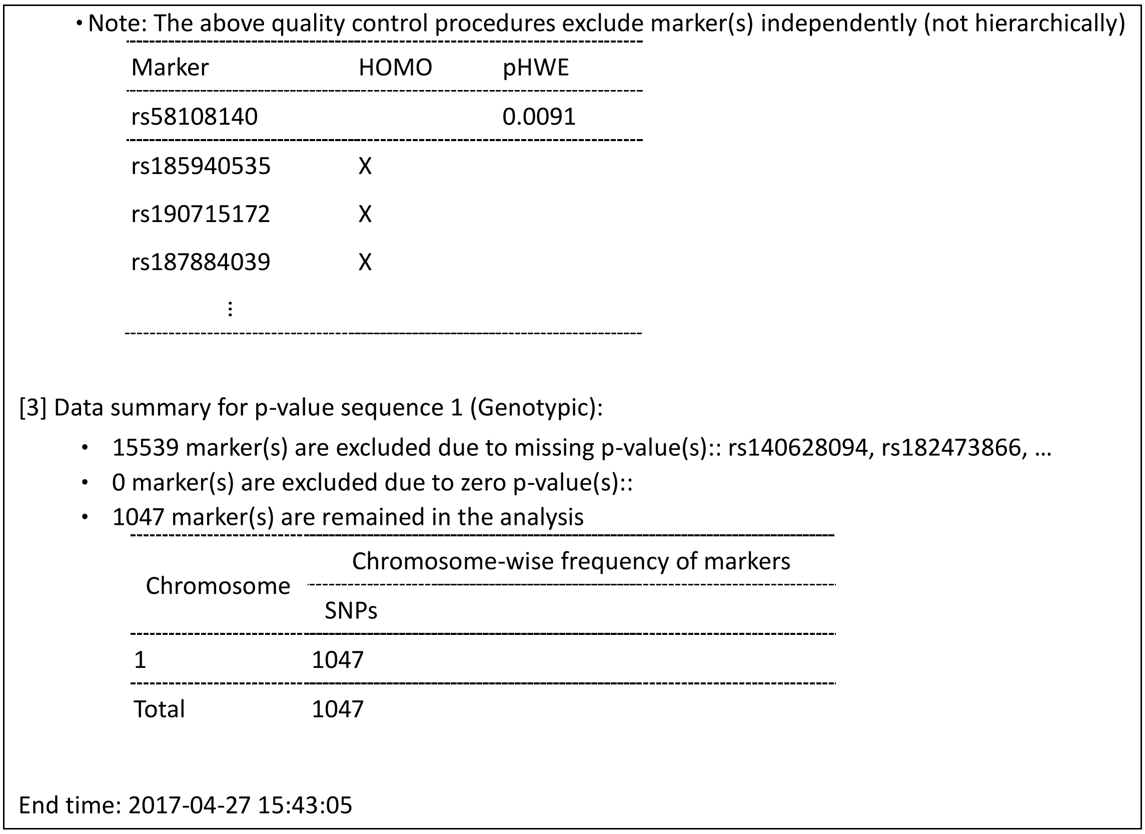
Sequencing data from the 1000 Genomes Project

The 1000 Genomes (1KG) Project [7] provided a comprehensive catalog of different human genetic variations by performing next-generation sequencing experiments. In this example, we investigated ancestry informative markers for European and African ancestry populations according to 85 CEU (CEPH in Utah with European ancestry) and 88 YRI (Yoruba from Ibadan, Nigeria with African ancestry). Variant call format (VCF) files from the website of the 1KG Project (http://www.1000genomes.org/) were converted to transposed pedigree format files (genotype files: chr01_CEU.tped, chr01_CEU.tfam, chr01_YRI.tped, and chr01_YRI.tfam) by using VCFtools. To help users download data, we reduced the marker data through variant thinning that drew only one in every 100 variants on chromosome 1; this procedure retained 28,970 variants for the study. The annotation file (chr01.anno) was prepared on the basis of NCBI 37.3. CEU and YRI were considered the case and control groups, respectively (phenotype file: chr01.pheno). All data files were saved in **OPATs** in "D:/OPATs/Examples/1KGP." Users can click *Example2* on the *P*-value Combination tab and press the Run button to run the WBAT.

The analysis results were saved in "D:/OPATs/Examples/1KGP/Output." The output folder included four text files of numerical results, one pdf file of graphical results, and two additional files (chr01.pv and chr01.anno) in the **OPATs** format. First, the note file (WBAT_NOTE.txt) recorded the parameter settings and data summary of the analysis (Table 9). The following parameter settings were recorded: [1] "P-value Combination" indicated the WBAT, a window size of 5 [i.e., an anchor marker in the middle and two additional variants on each side (m = 2)], a truncated PPM with a truncation threshold of 0.05, permutation with 10,000 replications, and an FDR with a significance level of 0.05; [2] "Data Quality" displayed an IQR rule for excluding outliers of heterozygosity and indicated that the cutoffs for the genotype CR, MAF, and p-value for the HWE test were <0.9, <0, and <0.05, respectively; and [3] "Model and Test" showed that the chi-squared test for genotype-based analysis was used. The following items were recorded in the data summary: [1] "Raw data information" showed that the raw data included 173 individuals and 28,970 variants; [2] "Validated data information" indicated that all individuals and 16,586 variants were retained after eliminating 12,157 nonpolymorphic variants and 227 variants with a p-value for HWE of <0.05; and [3] "Data summary for p-value sequence 1" showed that 15,539 of the 16,586 variants were eliminated because of missing p-values. Finally, 1,047 variants were analyzed by in WBAT.

Table 9. WBAT_NOTE.txt

```
Start time: 2017-04-27 15:40:15

●  PARAMETER SETTING
--------------------------------------------------------------------------------------------------------
[1] P-value Combination
   1. Association Analysis component
        1) Procedure                                        :: WBAT
        2) Study region                                     :: 1 to 28970
        3) Sliding window construction (bandwidth)          :: Marker(m) (2)
   2. Test Statistic component
        1) Test statistic                                   :: PPM
        2) Truncation threshold                             :: 0.05
   3. Empirical P-value component
        1) Resampling procedure (# of replications)         :: Permutation (10000)
        2) Multiple testing correction (significance level) :: FDR (0.05)
        3) Attaining Bonferroni's level                     :: Yes
[2] Data Quality
   1. Criteria of samples
        1) CR cutoff                                        :: 0
        2) Het rule                                         :: IQR rule: [Q1 - 1.5×IQR , Q3 + 1.5×IQR]
   2. Criteria of markers
        1) CR cutoff                                        :: 0.9
        2) MAF cutoff                                       :: 0
        3) HWE p-value cutoff                               :: 0.05
[3] Model and Test
   1. Model component
        1) Trait                                            :: dPOPU
        2) Genetic model                                    :: Codominant(Genotypic)
        3) Covariate                                        :: None
   2. Single-locus test component
        1) Test                                             :: Chi-squared test
--------------------------------------------------------------------------------------------------------



●  DATA SUMMARY

[1] Raw data
        •    173 individuals in the raw data
        • 28970 markers in the raw data

[2] Validated data
        •    173 individual(s) in the validated data
                 0 individual(s) are excluded due to unknown phenotype(s)
                 0 individual(s) are excluded due to CR < 0.9
                 0 individual(s) are excluded due to HET outside ( 0.057 , 0.160 ) by Mean rule
        • 16586 marker(s) in the validated data
                 0 marker(s) are excluded due to duplicated name(s)
                 0 marker(s) are excluded due to missing/incorrect information about chromosome(s)
                 0 marker(s) are excluded due to missing/incorrect information about physical position(s)
          12157 marker(s) are excluded due to no polymorphism across all samples
                 0 marker(s) are excluded due to CR < 0
                 0 marker(s) are excluded due to a low minor allele frequency (MAF < 0.01)
            227 marker(s) are excluded due to the departure of HWE (pHWE < 0.05)
```



```
    • Note: The above quality control procedures exclude marker(s) independently (not hierarchically)
        ------------------------------------------------------------
        Marker                HOMO        pHWE
        ------------------------------------------------------------
        rs58108140                        0.0091
        ------------------------------------------------------------
        rs185940535           X
        rs190715172           X
        rs187884039           X
                ⋮
        ------------------------------------------------------------

[3] Data summary for p-value sequence 1 (Genotypic):
    •  15539 marker(s) are excluded due to missing p-value(s):: rs140628094, rs182473866, …
    •  0 marker(s) are excluded due to zero p-value(s)::
    •  1047 marker(s) are remained in the analysis
        --------------------------------------------------------------------------------
                               Chromosome-wise frequency of markers
         Chromosome    ------------------------------------------------------------------
                         SNPs
        --------------------------------------------------------------------------------
        1              1047
        --------------------------------------------------------------------------------
        Total          1047


End time: 2017-04-27 15:43:05
```

Table 10. WBAT_GENO_m2_PT10000_RESULT.txt

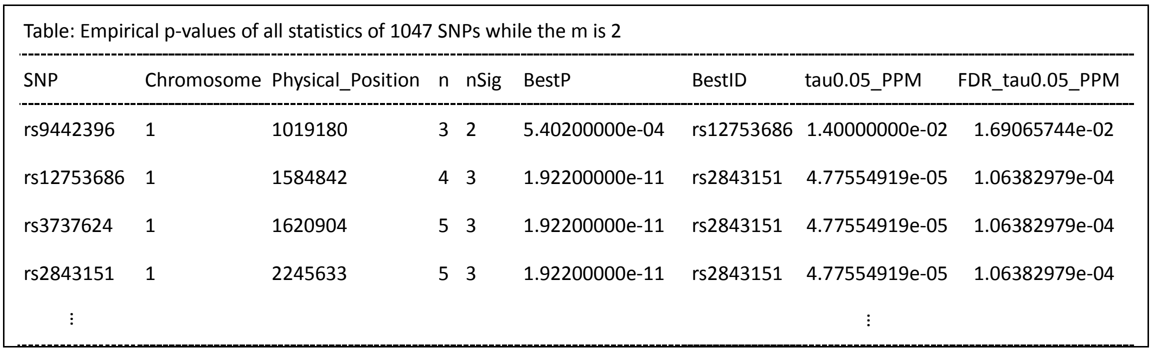


Table: Empirical p-values of all statistics of 1047 SNPs while the m is 2

SNP	Chromosome	Physical_Position	n	nSig	BestP	BestID	tau0.05_PPM	FDR_tau0.05_PPM
rs9442396	1	1019180	3	2	5.40200000e-04	rs12753686	1.40000000e-02	1.69065744e-02
rs12753686	1	1584842	4	3	1.92200000e-11	rs2843151	4.77554919e-05	1.06382979e-04
rs3737624	1	1620904	5	3	1.92200000e-11	rs2843151	4.77554919e-05	1.06382979e-04
rs2843151	1	2245633	5	3	1.92200000e-11	rs2843151	4.77554919e-05	1.06382979e-04
⋮							⋮	

The results file (WBAT_GENO_m2_PT10000_RESULT.txt) showed that 955 of the 1,047 variants were statistically significant (i.e., ancestry informative markers) after FDR correction (Table 10). On the basis of the identified ancestry informative variants, CEU and YRI samples were clearly separated in an allele frequency biplot (Figure 7). The construction of an allele frequency bioplot was reported previously [8].

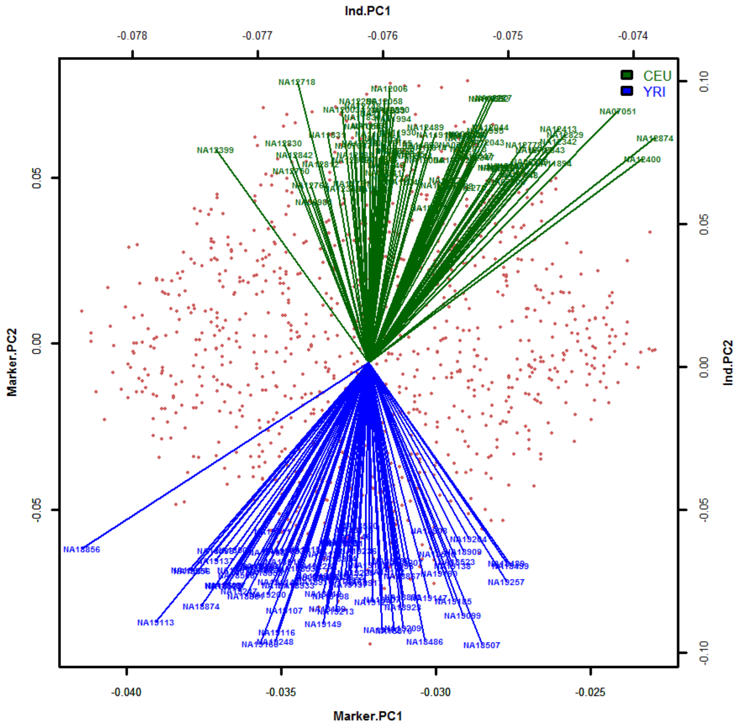


Figure 7. Biplot of 85 CEU and 88 YRI samples CEU and YRI samples are shown with green and blue arrows, respectively. Red points indicate the identified ancestry informative markers.

8. REFENCES

1. Eddelbuettel D, Francois R. Rcpp: Seamless R and C plus plus Integration, Journal of Statistical Software 2011;40:1-18.
2. Eddelbuettel D. Seamless R and C++ Integration with Rcpp. New York: Springer, 2013.
3. Eddelbuettel D, Sanderson C. RcppArmadillo: Accelerating R with high-performance C plus plus linear algebra, Computational Statistics & Data Analysis 2014;71:1054-1063.

4. Kane MJ, Emerson JW, Weston S. Scalable Strategies for Computing with Massive Data, Journal of Statistical Software 2013;55:1-19.

5. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, Nature 2007;447:661-678.

6. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research 2000;28:27-30.

7. Sudmant PH, Rausch T, Gardner EJ et al. An integrated map of structural variation in 2,504 human genomes, Nature 2015;526:75-81.

8. Yang HC, Lin HC, Huang MC et al. A new analysis tool for individual-level allele frequency for genomic studies, BMC Genomics 2010;11:415.