- RWAS & LRT version 0.5 (March 7th, 2012)

Following information is from http://genetics.cs.ucla.edu/rarevariants/. Hyperlinks in the text are not working (please visit the homepage).

1. Brief introduction to our methods

We developed two methods to detect associations of groups of rare variants. Those two methods are RWAS (Rare variant Weighted Aggregate Statistic), and LRT (Likelihood Ratio Test)

RWAS (Rare variant Weighted Aggregate Statistic) is a groupwise association test for identifying associations of groups of rare variants. RWAS groups variants and computes a weighted sum of differences in mutation counts between case and control individuals. Weights of RWAS are estimated from data to achieve nearly optimal power under a disease model in which all variants make an equally small contribution to population disease risk. For more information on RWAS, please refer to the following paper

Jae Hoon Sul, Buhm Han, Dan He, Eleazar Eskin. "An optimal weighted aggregated association test for identification of rare variants involved in common diseases." Genetics (In Press)

LRT (Likelihood Ratio Test) is a method that tries to identify which variants are causal by taking advantage of both prior information (of how likely each variant is functional) and data. LRT uses this information (of which variants are likely causal) to better detect associations of groups of rare variants. To identify causal variants, LRT assumes that some variants are causal and some are not (called "causal statuses of variants") and computes the likelihood of the data under every possible causal statuses. This allows LRT to compute likelihoods of null and alternative models where the null model is one that asserts no causal variants in a group while the alternative model asserts at least one causal variant. A statistic of LRT is a ratio between likelihoods of the two models, and the permutation test is performed to obtain the significance of the statistic. For more information on LRT, please refer to the following paper

Jae Hoon Sul, Buhm Han, Eleazar Eskin. "Increasing Power of Groupwise Association Test with Likelihood Ratio Test." In Proceedings of the Fifteenth Annual Conference on Research in Computational Biology (RECOMB-2011). Vancouver, Canada: March 28th-31st, 2011

2. Installing the software

1. Please make sure Java is installed on your machine. Our software is compatible with every system that supports java (e.g. Windows, Mac, Linux, etc). Open up terminal and type "java". If nothing happens or you see an error message, it means

that java is not probably installed on your machine (or java program may not be in your path). Please visit here for more information on checking whether Java is installed on your machine.

2. To install the latest version of Java,
Visit here
Download and install Java SE JRE for your platform (Windows, Linux)
For Mac users: Java in Mac OS should work without downloading the new Java. If you want to install the latest version of Java, you need to download the Java update from the Apple website (not from the above link). Visit Apple Support and search for "Java for Mac OS X 10.6 Update" (for Snow Leopard). The latest update is here

3. Unzip our software

4. Open up terminal, go to the directory where our software is unzipped, and type
java -jar RWAS.jar -h
If you do not see help messages, it means that Java is not correctly installed. Please try to install the latest version of Java. Also, make sure the directory where *.jar files are located contains "lib" directory. Our software needs the lib directory.

3. Input File Formats
Our software supports two types of input file format. One format contains information of only one gene ("Single Gene File Format") used by VT Test (Variable-Threshold test), and the other format contains information of multiple genes ("Multiple Genes File Format").
Important: There must be only a single white space (tab or space) between two columns in all input files. If there are multiple white spaces between two columns, it will give an error!

3.1 Single Gene File Format
There are 3 files for this format; 1) phenotype file, 2) SNP weight file, 3) genotype file.
1. Phenotype file has two columns on each line separated by a space; individual ID and phenotype value. Phenotype value must be 1 for controls and 2 for cases. Here is a sample phenotype file
ind1 1
ind2 1
ind3 2
ind4 2
ind5 2

2. SNP weight file has two columns on each line separated by a space; SNP ID and weight. Weight must be between 0 and 1. Note: This weight corresponds to c_i value in our paper, which we call prior information of variants. It indicates how likely the variant is deleterious or important. This weight IS NOT a weight (w_i) in our paper.

Our software automatically computes w_i, and this weight (c_i) is additional information about the SNP. If you do not know how important or deleterious a SNP is, try to use 1.0 for all SNPs.
snp1 0.5
snp2 0.3
snp3 0.9
snp4 0.2
snp5 0.5

3. Genotype file has three columns on each line separated by spaces; individual ID, SNP ID and genotype. Genotype is the number of minor alleles observed for this SNP, and it must be either 0, 1, 2, or 3 for missing (missing genotypes are considered as having no minor alleles). Note: if genotype is 0, its information can be omitted from the genotype file. In other words, if a pair of SNP ID and individual ID do not appear in the genotype file, they are assumed to have 0 genotype.
ind1 snp1 1
ind2 snp1 2
ind2 snp2 1
ind3 snp4 1
ind5 snp5 1

3.2 Multiple Genes File Format
There is only 1 file for this format
1. 1st line contains case-control statuses for all individuals (1 for controls and 2 or cases separated by spaces).

2. From 2nd line, each line contains information of each SNP.
1st column: Gene ID
2nd column: SNP weight
3rd column: genotype for individual 1 (0, 1, 2, *(3 for missing))
4th column: genotype for individual 2 (0, 1, 2, (3 for missing))
(n+2) column: genotype for individual n (0, 1, 2, (3 for missing))

*missing genotypes are considered as having no minor alleles

3. Here is a sample multiple genes file with 2 genes, 6 individuals, and 5 SNPs.
1 1 1 2 2 2
1 0.2 0 1 2 0 0 1
1 0.5 1 1 0 1 0 0
1 0.9 0 0 1 0 0 0
2 0.8 0 0 0 1 1 1
2 0.7 1 0 1 0 0 2

*We provide a utility program that converts a PLINK PED file into Multiple Genes File Format. For more information, click here.

4. Execution
Following information is also available by typing java -jar RWAS.jar [or LRT.jar] -h .

RWAS
1. For multiple genes input file format
java -jar RWAS.jar [-n] [-i input_file] [-o output_p-value_file] [-s start_index] [-e end_index] [-v] [-m MAF_threshold] [-t] [-p #_of_permutations] [-h]

Required parameters:
[-n]: indicates multiple genes input file format
[-i input_file]: file w/ phenotypes and genotypes
[-o output_p-value_file]: file w/ p-values [Gene_ID p-value] per line

Optional parameters:
[-s start_index]: starting index of gene ID (inclusive)
[-e end_index]: ending index of gene ID (exclusive)

Sample command:
java -jar RWAS.jar -n -i test.input.txt -o test.out.txt -v

2. For single gene input file format
java -jar RWAS.jar [-g] [-a phenotype_file] [-b snp_weight_file] [-c genotype_file] [-v] [-m MAF_threshold] [-t] [-p #_of_permutations] [-h]

Required parameters:
[-g]: indicates single gene input file format
[-a phenotype_file]: file w/ phenotypes of individuals: [Individual_ID phenotype] per line
[-b snp_weight_file]: file w/ weights of SNPs: [SNP_ID weight] per line
[-c genotype_file]: file w/ genotype: [Individual_ID SNP_ID genotype] per line

Sample command:
java -jar RWAS.jar -g -a test.input.ind -b test.input.snp -c test.input.geno -v

3. Optional parameters for both formats
[-v]: Turn on verbose option
[-m MAF_threshold]: Ignore SNP whose MAF is > MAF_threshold (default: all SNPs included)
[-t]: Perform permutation test to estimate p-value. Must specify -p option
[-p #_of_permutations]: # of permutations to perform. Must specify -t option
[-h]: Print this help

LRT
1. For multiple genes input file format
java -jar LRT.jar [-n] [-p #_of_permutations] [-i input_file] [-o output_p-value_file] [-s start_index] [-e end_index] [-v] [-m MAF_threshold] [-r hyper_MAF_threshold] [-h]

Required parameters:
[-n]: indicates multiple genes input file format
[-p #_of_permutations]: # of permutations to perform
[-i input_file]: file w/ phenotypes and genotypes
[-o output_p-value_file]: file w/ p-values [Gene_ID p-value] per line

Optional parameters:
[-s start_index]: starting index of gene ID (inclusive)
[-e end_index]: ending index of gene ID (exclusive)

Sample command:
java -jar LRT.jar -n -p 1000000 -i test.input.txt -o test.out.txt -v

2. For single gene input file format
java -jar LRT.jar [-g] [-p #_of_permutations] [-a phenotype_file] [-b snp_weight_file] [-c genotype_file] [-v] [-m MAF_threshold] [-r hyper_MAF_threshold] [-h]
Required parameters:
[-g]: indicates single gene input file format
[-p #_of_permutations]: # of permutations to perform
[-a phenotype_file]: file w/ phenotypes of individuals: [Individual_ID phenotype] per line
[-b snp_weight_file]: file w/ weights of SNPs: [SNP_ID weight] per line
[-c genotype_file]: file w/ genotype: [Individual_ID SNP_ID genotype] per line

Sample command:
java -jar LRT.jar -g -p 1000000 -a test.input.ind -b test.input.snp -c test.input.geno -v

3. Optional parameters for both formats
[-v]: Turn on verbose option
[-m MAF_threshold]: Ignore SNP whose MAF is > MAF_threshold (default: all SNPs included)
[-r hyper_MAF_threshold]: Hypergeometric dist. sampling for SNP whose MAF < threshold (default: 5%)
[-h]: Print this help

5. Important Notes
5.1 A permutation test is optional in RWAS while it is required in LRT. The original RWAS obtains a p-value from the standard normal distribution, and the default setting of RWAS does not perform the permutation test. If you want to perform the permutation for RWAS, please include "-t -p [# of permutations to perform]" options.

5.2 Phenotype values must be 1 for controls and 2 for cases in the phenotype file. Other values are not recognized and will give an error.

5.3 Our methods assume that rare variants are not in linkage disequilibrium with each other. However, if non-negligible LD is expected between variants, especially when common variants are in LD, please follow the below instructions to correctly handle LD.

RWAS: there are three options

1. Apply the permutation test as mentioned in the previous note (#1)
2. Set the weights of variants in LD to 0. Our software ignores all variants whose weights are equal to 0. For this option, users need to know which variants are in LD.
3. Remove all common variants from the analysis. You can remove them by using the option -m [MAF_threshold]. This will ignore every SNP whose MAF is greater than MAF_threshold. This might not be a very good solution since it removes all common variants that are not in LD. However, removing common variants from the analysis would not have dramatic effects on the results because they have very low weights in our method.

LRT: since the permutation test of LRT assumes independence among variants, the only way to handle LD in LRT is to remove variants in LD (the second and third options in RWAS).

5.4 Runtime of RWAS and LRT

Here are some rough estimates of runtime of our methods. The estimates are based on 1000 case and 1000 control individuals and 50 variants running on a 2.26 GHz Macbook.

RWAS with 1 million permutations: 9 ~ 10 minutes
LRT with 10 million permutations: 4 ~ 5 minutes

5.5 Handling missing genotypes

We consider missing genotypes as having no minor alleles. If this can cause problems, please remove SNPs that have high missing rate.

Please send any bugs, suggestions, and questions to jhsul@cs.ucla.edu (Jae Hoon Sul)