JAWAMix5 Users Manual r1.1.0

August 25, 2017

JAWAMix5 stands for HDF5-based JAva implementation of Whole genome Association studies using Mixed model. The motivation of developing JAWAMix5 is to provide a platform-independent toolkit for mixed model-based association mapping that is scalable for very large dataset. It also supports simulation-based power estimations We intend to offer long-term maintenance for JAWAMix5 and continue adding our new developments into it.

In its first release, we provide 9 functions.

1. Installation

The toolkit is a batteries-included executable, therefore no installation is needed. Just copy the executable, jawamix5.jar, and run it using the standard command for java packages:

```
java –Xmx2g –jar /path/to/jawamix5.jar
```

This will prompt a help message. If that does not happen, please be so kind as to send us an email.

2. Functions

We provide ten analytical functions in version r1.1.0, the first release of JAWAMix5:

(1) An approximation of the original mixed model (Kang, Zaitlen et al. 2008), i.e., EMMAX (Kang, Sul et al. 2010));
(2) Local variance component analysis by traditional point estimations (Yang, Benyamin et al. 2010), however jointly accounting for population structure;
(3) Local variance component analysis by Bayesian estimations;
(4) Rare variants analysis using collapsing test (Li and Leal 2008) with or without population structure controlled;
(5) Standard linear regression without mixed model;
(6) Standard stepwise regression without mixed model;
(7) Stepwise regression based on mixed model;
(8) Nested Association Mapping (NAM). (McMullen, Kresovich et al. 2009).
(9) Simulation-based Power Estimation for the design of Rare Disease sequencing studies

In addition to main analysis, we also provide assistant functions:

(1) Calculate kinship for the whole genome or particular regions;
(2) Import the input genome files (in CSV format) to HDF5 format;
(3) Change the ACGT coded genotypes into number-coded genotypes and, by the way, filter out some non-qualified variants.

3. Commands and options

All the functions are used as:

```
java –Xmx2g –jar /path/to/jawamix5.jar function <options>
```

The details of the options of each function are described below:

kinship: Compute IBS kinship matrix to assist the other analysis.
Mandatory parameters:

- -ig: input genotype file in HDF5 format
- -o: the output folder

Options:

- -w: tiling window size(bp), default is the whole genome
- -scale: maximal value of genotype coding, default is 2.0 (corresponding to the ordinary 0, 1, 2 coding for hom, het, hom genotypes)

import: Import genotype file from .CSV format into HDF5 encoded indexed files.
Mandatory parameters:

- -ig: input genotype file in plain text (.CSV format)
- -o: output in HDF5 in format

Options:

- -b: block size (number of variants per HDF5 block), default is 5,000 variants. This parameter should be tuned subject to the different sample size. When sample size is too large, one has to decrease the block size or increase the RAM.

char2num: Change the data coded by characters in original CSV file to numbers. (e.g., from A, C, G, T or Y, K, S, etc. to 0,1,2)
Mandatory parameters:

- -ig: input genotype file in plain text (.CSV format)
- -o: output (also a .CSV file with encoding changed: homozygote of major allele will be coded as 2, homozygote with minor allele will be coded as 0, and heterozygote will be coded as 1.)

emmax: Run single-marker mixed model EMMAX for phenotype(s)
Mandatory parameters:

- -ig: input genotype file in HDF5 format
- -ip: phenotype file
- -o: output folder
- -ik: kinship file generated by function "kinship" or other user defined method

Options:

- -p: P-value threshold *after* multiple correction. Variants that do not reach this threshold will not be written to the file. The default is 1,000.

- -index: the phenotype index (starting from zero) in the event that there are multiple phenotypes in the input file. If not specified, by default all the phenotypes will be analyzed sequentially.
- -min_size: minimal sample size below which the analysis will not be performed. Default setting is 100.
- -maf: threshold of minor allele frequency below which the result will not contribute the Manhattan plot (but still will be written to the .CSV text file). Default setting is 0.05.

emmax_stepwise: Run stepwise regression for phenotype(s) with population structure simultaneously controlled by mixed model.
Mandatory parameters:

- -ig: input genotype file in HDF5 format
- -ip: phenotype file
- -o: output folder
- -ik: kinship file generated by function "kinship" or other user defined method

Options:

- -r: total number of steps in the stepwise regression. Default is round=2.
- -p: P-value threshold *after* multiple correction. Variants that do not reach this threshold will not be written to the file. The default is 1000.
- -index: the phenotype index (starting from zero) in the event that there are multiple phenotypes in the input file. If not specified, by default all the phenotypes will be analyzed sequentially.
- -min_size: minimal sample size below which the analysis will not be performed. Default setting is 100.
- -maf: threshold of minor allele frequency below which the result will not contribute the the Manhattan plot (but still will be written to the .CSV text file). Default setting is 0.05.

lm: Run single-marker linear model analysis for phenotype(s) without population structure controlled.
Mandatory parameters:

- -ig: input genotype file in HDF5 format
- -ip: phenotype file
- -o: output folder

Options:

- -p: P-value threshold *after* multiple correction. Variants that do not reach this threshold will not be written to the file. The default is 1000.
- -index: the phenotype index (start from zero) in the event that there are multiple phenotypes in the input file. If not specified, by default all the phenotypes will be analyzed sequentially.
- -min_size: minimal sample size below which the analysis will not be performed. Default setting is 100.
- -maf: threshold of minor allele frequency below which the result will not contribute the the Manhattan plot (but still will be written to the .csv text file). Default setting is 0.05.

lm_stepwise: Run stepwise regression for phenotype(s) without population structure controlled.
Mandatory parameters:

- -ig: input genotype file in HDF5 format
- -ip: phenotype file
- -o: output folder

Options:

- -r: total number of steps in the stepwise regression. Default is round=2.
- -P: P-value threshold *after* multiple correction. Variants that do not reach this threshold will not be written to the file. The default is 1000.
- -index: the phenotype index (start from zero) in the event that there are multiple phenotypes in the input file. If not specified, by default all the phenotypes will be analyzed sequentially.
- -min_size: minimal sample size below which the analysis will not be performed. Default setting is 100.
- -maf: threshold of minor allele frequency below which the result will not contribute the the Manhattan plot (but still will be written to the .csv text file). Default setting is 0.05.

local: Run local variance component analysis for tiling windows throughout the genome. Both point estimation and Bayesian style posterior distribution will be done together.
Mandatory parameters:

- -ig: input genotype file in HDF5 format
- -ip: phenotype file
- -o: output folder
- -ik_l: folder contains local kinship files generated by function “kinship”. Since this method assumes the naming conventions of all kinship files, it is strongly recommended to use function “kinship” of JAWAMix5 to generate the kinship files.
- -w: tiling window size (it should be consistent with the kinship files and must therefore be the same as specified when running “kinship” function.)
- -ik_g: the global kinship file (generated by function “kinship” or any other user’s methods.)

Options:

- -index: the phenotype index (start from zero) in the event that there are multiple phenotypes in the input file. If not specified, by default all the phenotypes will be analyzed sequentially.
- -step: the size of searching grid. The smaller, the better resolution of the analysis, but also slower.

rare: Run rare variants analysis using both standard aggregate test and another aggregate test leveraging potential synthetic associations. Population structure will be controlled by mixed model.
Mandatory parameters:

- -ig: input genotype file in HDF5 format
- -ip: phenotype file

- -ik: kinship file generated by function “kinship” or any other user’s method
- -o: output folder
- -is: input list of potential synthetic associations, preferably generated by function “lm” or “emmax”.
- -ir: file listing regions that the rare variants in the same region should be collapsed.

Options:

- -index: the phenotype index (start from zero) in the event that there are multiple phenotypes in the input file. If not specified, by default all the phenotypes will be analyzed sequentially.
- -ld: LD threshold(D-prime). Default is 0.8
- -rare: MAF threshold to indentify rare variants. Default is 0.01.
- -dist: distance between the potential synthetic association and the regions in which the rare variants should be collapsed. Default is 100000 (which means 100kb).
- -syn_pvalue: significant level threshold for the nominal p-value of the potential synthetic associations.
- -syn_maf: MAF threshold for the potential synthetic associations. Default setting is 0.1.

nam: Nested association mapping for phenotype(s). (Imputation of genomes with RIL sample with founder genomes and marker information is included.)
Mandatory parameters:

- -p:list of RIL pedigreee files folder
- -c: Cross ID
- -fg: founder whole genome genotype file
- -ril: RIL phenotype
- -o: Output prefix

Options:

- -r: total number of steps in the stepwise regression. Default is round=2.
- -p: P-value threshold *after* multiple correction. The variants that do not reach this threshold will not be written to the file. The default is 1000.
- -index: the phenotype index (start from zero) in the event that there are multiple phenotypes in the input file. If not specified, by default all the phenotypes will be analyzed sequentially.
- -b: the block size of HDF5 file. Default setting is 5,000.
- -maf: threshold of minor allele frequency below which the result will not contribute the Manhattan plot (but still will be written to the .csv text file). Default setting is 0.05.

simperd: Simulation-based Power Estimation for the design of Rare Disease sequencing studies. Please refer the users manual for SimPERD for details.

4. Input/output file formats

Input genotype file. Excepting NAM, in which the imputation of recombinant inbred lines (RIL) is involved, all the input genotype files should be in comma separated values (CSV) format. The first two columns denote chromosome number and location, and the first row is the header line that contains the ids of all individuals. With that exception, each column contains the genome of an individual and each row contains the genotypes of all individuals for a variant. Here is the example header:

Chr,Loc,id_1,id_2,id3,…,id_n

The remaining rows encode genotypes in the following format:

5,456789,0,1,0,2,2,1,0,0,0,…,1,2

The first two columns indicate that the variant is at chromosome 5, location 456789. The remaining columns encode the genotypes. The users can use any numeric code to encode the genomes. The general convention is to use 0 and 2 for homozygote and 1 for heterozygote. It is assumed that 0 stands for reference allele or major allele and 2 stands for alternative allele. Another example is the copy number of genes. If one uses the A, C, T, G to encode the genomes, we provide a function **char2num** to convert. The chromosomes have to be coded as numbers as 1, 2, 3, … , 22, 23, 24, 25, instead of strings like "Chr1", "X" etc.

Input phenotype file. The phenotype file is a tab-separated text file. One can use one file for one phenotype or put multiple phenotypes into one file. The first row denotes the names of the phenotypes, while the other rows are the values of an individual for all the phenotypes. The first column records the ids of all relevant individuals, and the remaining columns contain the values of all individuals for one phenotype. If the phenotype of an individual has not been measured, we use "NA" to fill that cell.

Input/output kinship file. For a sample with *n* individuals, the kinship file is an *n x n* matrix. Different values in the same row are separated by comma. The order of individuals *has to* be consistent with the order in the genotype file (although not explicitly written in the kinship file).

Input region file. For rare variants analysis, the input regions file is a tab separated plain text file. There is no header, and each row denotes one region by the following four columns: name of the region, chromosome (again, a number, not a string), start location, and end location.

NAM related input files. There are quite a few input files for NAM since the pedigree information and marker genotypes as well as founder population genomes have to be specified. Their formats are:

(1) Information of crosses design: there is NO header in this file. Each line stands for a cross. There are five columns, from left to right: parent1 name, parent1 id, parent2 name, parent2 id, the pedigree id.
(2) Marker genotypes of all RILs: Standard file format of QTLNetwork.
(3) Founder genomes: the same as the genotype file for other analysis

(4) Phenotype file for the recombinant inbred lines (RILs): the same as phenotype files for other analysis, but with one more column denoting family IDs.

Output P-value file for single marker analysis. The output format is also CSV. For the single variant analysis, we provide one .CSV file for the each phenotype. Each line stands for the results of one variant. The columns are, from left to right: chromosome, location, p-value, adjusted R2, coefficient of regression, standard error, minor allele count. Results not passing the specified threshold will not be printed out. In addition to the text output, we also generate the Manhattan plot for the results, illustrating the logged P-value and variance explained.

Output text files for stepwise regression. For all functions in which the step-wise regression is involved (lm_stepwise, emmax_stepwise, nam), we provide an extra .CSV file reporting stepwise results for the best variants in each step, in addition to the CSV file recording the results of all significant results at the first round scan. The columns are, from left to right: chromosome, location, P-value, total adjusted R2 explained by the regression.

Output text files for local variant component analysis. The text output is also in CSV format. Except for the first two rows that are headers, each row stands for the result of a local region. There are two files: one is for the point estimates of variance component for all the regions; the other is for the Bayesian estimates.

(1) For the point estimates: from right to left, the columns are: chromosome, start location of the region, P-value, variance explained by the local region, variance explained by the whole genome (i.e. the random terms using global kinship as the variance-covariance matrix), and total variance explained.

(2) For the Bayesian estimation: chromosome, start location of the region, REML/(variance explained by global kinship) for all local potential values of variance explained by the focal local region.

Output plots. In general, Jawamix5 will draw standard Manhattan plots for all analysis. In addition, for Bayesian style local variance component analysis, a heatmap will be provided in which the darkness of the point denotes the weight of that value in the posteriori distribution.

Output P-value file for rare variants analysis For rare variants aggregate analysis, the four different algorithms, i.e., with or without leveraging potential synthetic association and with or without controlling population structure, will be done together. Therefore, four sets of results will be generated. The results without leveraging potential synthetic association will be the same as the standard output in this package (like **emmax** or **lm**); while the results with leveraging have one more column denoting which synthetic association the focal region is linked to.

Output of NAM analysis. The same as the stepwise linear regression output.

5. Information for Developers

This section contains information for developers who might want to look at the source code and modify it for different uses.

There are a few classes in the package, usually one class for each type of analysis, although there are a few small classes as constructors for data loading or very specific calculations (e.g., eigen values). The main class for storing data is VariantsDouble.java and VariantsByte.java, which are tightly linked to the low-level data storage. Usually one could use VariantsByte to encode genotypes but we also provide VariantsDouble since some times the genotype might be quantitative (e.g., methylation data or pooled sequencing).

The design of HDF5 hierarchy is described as follows: under the root, there is so far only one group called “genotype”. In the group, there are three attributes: “sample_size”: IntAttribute, “block_size”: IntAttribute, “num_chrs”: IntAttribute. There are three tables: “num_blocks_per_chr”: IntArray, and “sample_ids”: StringArray, and a compound array for annotations. For the moment, we have one example class of annotation using released 1000 Genomes Project Phase 1 annotation. Users can define their own annotation class(es).

Under the main group “genotype”, there are a number of subgroups named as “chr??”, where the “??” is a number denoting the chromosome. In each subgroup, there are two tables: “var_pos”: IntArray; and “var_mafc”: IntArray, storing respectively the position and minor allele counts. Finally, the actual genotypes are stored in the blocks named “position_fast”: DoubleMatrix if one uses VariantsDouble or ByteMatrix if one uses VariantsByte.

We have implemented a few functions to randomly load one variant (both by index in the variants list or by coordinate in the chromosome) or all variants in a specified region, from the genome. Therefore, in theory, the developer doesn’t need to look at the details of HDF5 blocks.

6. Contacts:

Quan Long, quan.long@gmi.oeaw.ac.at
Qingrun Zhang, zhangqr@gmail.com

7. Copy right license: (MIT)

Acknowledgements: We thank JHDF5 for Java HDF5 binding, JFreeChart for graphics, Apache Commons-Math for an excellent scientific computing library, and root search function in Michael Thomas Flanagan's Java Scientific Library (www.ee.ucl.ac.uk/~mflanaga). We are grateful to Oliver Stegle for discussions of Bayesian models, Thomas Friese for proof reading. This work was supported by GMI core funding from the Austrian Academy of Sciences.

References

Kang, H. M., J. H. Sul, et al. (2010). "Variance component model to account for sample structure in genome-wide association studies." Nat Genet **42**(4): 348-354.

Kang, H. M., N. A. Zaitlen, et al. (2008). "Efficient control of population structure in model organism association mapping." Genetics **178**(3): 1709-1723.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." Am J Hum Genet **83**(3): 311-321.

McMullen, M. D., S. Kresovich, et al. (2009). "Genetic properties of the maize nested association mapping population." Science **325**(5941): 737-740.

Yang, J., B. Benyamin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." Nat Genet **42**(7): 565-569.