BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 9 2013, pages 1220–1222
doi:10.1093/bioinformatics/btt122

Genetics and population analysis Advance Access publication March 11, 2013

JAWAMix5: an out-of-core HDF5-based java implementation of whole-genome association studies using mixed models

Quan Long*,†,‡, Qingrun Zhang†,‡, Bjarni J. Vilhjalmsson$, Petar Forai, Ümit Seren and Magnus Nordborg

Gregor Mendel Institute, Austrian Academy of Sciences, Vienna 1030, Austria

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: We present JAWAMix5, an out-of-core open-source toolkit for association mapping using high-throughput sequence data. Taking advantage of its HDF5-based implementation, JAWAMix5 stores genotype data on disk and accesses them as though stored in main memory. Therefore, it offers a scalable and fast analysis without concerns about memory usage, whatever the size of the dataset. We have implemented eight functions for association studies, including standard methods (linear models, linear mixed models, rare variants test, analysis in nested association mapping design and local variance component analysis), as well as a novel Bayesian local variance component analysis. Application to real data demonstrates that JAWAMix5 is reasonably fast compared with traditional solutions that load the complete dataset into memory, and that the memory usage is efficient regardless of the dataset size.

Availability: The source code, a 'batteries-included' executable and user manual can be freely downloaded from http://code.google.com/p/jawamix5/.

Contact: quan.long@gmi.oeaw.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 12, 2012; revised on February 7, 2013; accepted on February 18, 2013

1 INTRODUCTION

Next-generation sequencing (NGS) enables investigators to use whole-genome sequences for genotype–phenotype association mapping, but it brings with it the challenge of developing scalable tools to handle very large datasets. To do association studies, it is usually preferable to put data into random access memory (RAM); however, the sequence data generated by NGS are usually too large to be loaded into RAM. Therefore, analysts may have to use *ad hoc* methods to manipulate file reading. As the magnitude of sequencing projects goes up, this problem will become more and more pronounced.

To solve this problem of scalability, it would be ideal to have data stored on disk, but also provide a handy read/write protocol that users can use as if the data were stored in the main memory (an approach referred to as 'out-of-core' in computer science). This toolkit should offer transparency (i.e. hide the tedious implementation details) and high performance. Hierarchical Data Format (HDF5) (www.hdfgroup.org/HDF5) is a set of libraries designed to store and organize large datasets that was originally developed by the National Center for Supercomputing Applications. Because of its excellent performance and convenience, it has been widely used in many scientific computing communities, including storing NGS sequences (Mason *et al*., 2010). We developed JAWAMix5, a toolkit that uses HDF5 for storing and analyzing whole-genome genotypes for association mapping with various statistical models.

The linear mixed model has been considered an important framework in GWAS for controlling population structure (Atwell *et al*., 2010) and estimating genetic architecture (Yang *et al*., 2010), and it has recently been significantly improved (Listgarten *et al*., 2012; Segura *et al*., 2012). We implemented most functions in JAWAMix5 based on the mixed model. In addition, we provide standard functions without the mixed model as an alternative for users (e.g. stepwise regression and nested association mapping). Given that current implementations of mixed model are based on C/C++, R or Python, JAWAMix5 provides another alternative for researchers to use. Java programmers can contribute (Holland *et al*., 2008) based on the specifications described in our user manual.

2 FEATURES

We provide eight main functions in the first release of JAWAMix5: (i) GWAS in structured populations using the mixed model approach [EMMAX (Kang *et al*., 2010)]; (ii) local variance component analysis by traditional point estimations similar to (Hayes *et al*., 2010), and jointly accounting for population structure; (iii) local variance component analysis by Bayesian estimations (see motivation and descriptions in Supplementary Notes); (iv) rare variants analysis using aggregate test (Li and Leal, 2008), and an aggregate test jointly accounting for population structure by mixed model; (v) standard linear regression without mixed model; (vi) standard stepwise regression; (vii) stepwise regression based on mixed model; and (viii) imputation and regression analysis in the framework of nested association mapping (NAM) design (McMullen *et al*., 2009). The detailed formulations are presented in Supplementary Notes.

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[‡] Present address: Department of Genetics and Genomic Science, Mount Sinai School of Medicine, New York, NY, USA.

[$] Present address: Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.

Table 1. Runtime comparison between HDF5-based and RAM-based solution

	Standard regression	Imputation and stepwise regression for NAM	EMMAX algorithm	Calculate whole-genome IBS matrix	Local variance component (for 1500 regions)	Rare-variant tests
RAM	4 min	6 h	8 min	1 h	25 h	3 min
HDF5	5 min	7 h	10 min	20 min	25 h	2 min

Table 2. Comparison between JAWAMix5 to existing tools

	Calculate IBS matrix	Standard regression	EMMAX algorithm	Variance component
Software name	EMMAX	PLINK	EMMAX	GCTA
Runtime	33 min	16 min	12 min	20 h
JAWAMix5	20 min	5 min	10 min	25 h

To visualize the analytical results, we provide automatic plotting, e.g. Manhattan plots for logged *P*-values and variance explained, or heat-map for distributions from Bayesian analysis. An advantage of using HDF5 for storing the data is that users can use HDF5View, a GUI provided by the HDF5 group, to view the compressed raw genotype data. Users can then easily view details of interesting or suspicious results.

The core program and libraries are written in Java, so that users do not need to install any third-party library and can simply copy and run our executable, regardless of type of machine or operating system ('batteries-included' solution).

To facilitate GWAS using genotypes that are quantitative, e.g. copy number variants or methylation levels, in all functions of JAWAMix5, storing and analyzing genotypes as floating numbers are also supported.

3 PERFORMANCE

Given that the data are stored on disk instead of in RAM, one might expect slower runtimes compared with the traditional solution of putting data into RAM (in the event that one does have the resources). We tested the performance in the following two experiments: first, we compare our HDF5-based solution with our own RAM-based implementation to see what is the overhead brought by HDF5; second, we compare core functions of JAWAMix5 with other existing tools: EMMAX (Kang *et al.*, 2010), PLINK (Purcell *et al.*, 2007) and GCTA (Yang *et al.*, 2011) (GCTA is originally for estimating variance component of the whole chromosome. We make use of it for local region by generating .bed files using variants of focal region and code them with the same chromosome id.).

We analyzed our sequences data (~6 million SNPs times 400 individuals) in The 1001 Genomes Project (www.1001genomes.org) and found that performance is good: in the worse case, JAWAMix5 is ~20% slower than the RAM-based solutions, whereas in some cases, it is even faster. Detailed comparisons are listed in Table 1. Memory wise, JAWAMix5 uses <500 MB RAM regardless of the amount of the data, whereas the memory usage of memory-based solutions increases linearly proportional to the size of the data.

Then we run existing tools that generated same results. The comparisons between JAWAMix5 and existing tools are presented in Table 2. Details of how the tools are used are in Supplementary Notes.

In addition to testing it on *Arabidopsis* data, we also run JAWAMix5 on human data using Phase I data of The 1000 Genomes Project (www.1000genomes.org) with our simulated random phenotype. In all, it took 18 h, which is comparable with EMMAX that spent 21 h.

4 FUTURE WORK

Immediately planned extensions are gene–gene interaction analysis and the annotations based on existing known gene models and functions. Another extension we are working on is the HDF5 interface for storing methylation data. Additionally, more functions for RNA-Seq expression analysis will be added.

Although the novel Bayesian method in JAWAMix5 has revealed new biological insight in *Arabidopsis* data (Supplementary Notes), it has not been rigorously validated with extensive simulations. We plan to do it in the future work.

ACKNOWLEDGEMENTS

The authors 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). They are grateful to Oliver Stegle for discussions of Bayesian models and Thomas Friese for proof reading.

Funding: GMI core funding from the Austrian Academy of Sciences.

Conflict of interest: none declared.

REFERENCES

Atwell,S. *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Hayes,B.J. *et al.* (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet*., **6**, e1001139.

Holland,R.C. *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

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

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

Listgarten,J. *et al.* (2012) Improved linear mixed models for genome-wide association studies. *Nat. Methods*, **9**, 525–526.

Mason,C.E. *et al.* (2010) Standardizing the next generation of bioinformatics software development with BioHDF (HDF5). *Adv. Exp. Med. Biol*., **680**, 693–700.

McMullen,M.D. *et al.* (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737–740.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Segura,V. *et al.* (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet*., **44**, 825–830.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet*., **42**, 565–569.

Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet*., **88**, 76–82.