JAWAMix5: an out-of-core HDF5-based java implementation of whole-genome association studies using mixed models
Quan Long\textsuperscript{1,2}\#1, Qingrun Zhang\textsuperscript{1,2}\#1, Bjarni J. Viljalalmsson\textsuperscript{1,3}, Petar Fora\textsuperscript{1}, Ümit Seren\textsuperscript{1} and Magnus Nordborg\textsuperscript{1}

1. Gregor Mendel Institute, Austrian Academy of Sciences, Vienna 1030, Austria
2. Current address: Department of Genetics and Genomic Sci., Mount Sinai School of Medicine, New York, USA
3. Current address: Department of Biostatistics, Harvard School of Public Health, Boston, USA

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 (NAM) 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 to 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

1 INTRODUCTION

Next generation sequencing (NGS) enables investigators to use whole genome sequences for genotype-phenotype association mapping, but 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 is 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. Due to 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: (1) GWAS in structured populations using the mixed model approach (EMMAX (Kang, et al., 2010)); (2) local variance component analysis by traditional point estimations similar to (Hayes, et al., 2010), and jointly accounting for population structure; (3) local variance component analysis by Bayesian estimations (see motivation and descriptions in Supplementary Notes); (4) rare variants analysis using aggregate test (Li and Leal, 2008), and an aggregate test jointly accounting for population structure by mixed model; (5) standard linear regression without mixed model; (6) standard stepwise regression; (7) stepwise regression based on mixed model; (8) imputation and regression analysis in the framework of nested association mapping (NAM) design (McMullen, et al., 2009). The detailed formulations are presented in online Supplementary Notes.

\# Equal contribution

* To whom correspondence should be addressed.
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 (CNV) or methylation levels, in all functions of JAWAMix5, storing and analyzing genotypes as floating numbers are also supported.

3 PERFORMANCE

Given that the data is stored on disk instead of in RAM, one might expect slower runtimes compared to 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: firstly we compare our HDF5-based solution to our own RAM-based implementation to see what is the overhead brought by HDF5; secondly, we compare core functions of JAWAMix5 to other existing tools: EMMAX (Kang, et al., 2010), PLINK (Purcell, et al., 2007), and GCTA (Yang, et al., 2011).

We analyzed our sequences data (around 6 million SNPs times 400 individuals) in the 1,001 genomes project (www.1001genomes.org) and found that performance is good: in the worse case, JAWAMix5 is around 20% slower than the RAM based solutions, while in some cases it is even faster. Detailed comparisons are listed in Table 1. Memory wise, JAWAMix5 uses less than 500MB RAM regardless of the amount of the data, while the memory usage of memory-based solutions increase 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 1,000 Genomes Project (www.1000genomes.org) with our simulated random phenotype. Totally it took 18 hours, which is comparable to EMMAX that spent 21 hours.

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.

Conflict of interest: none declared.

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.ce.ucl.ac.uk/~mflanagan). 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


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.

