Comparisons of Modern GWA Algorithms on *Arabidopsis thaliana* Dataset

Julian Liber and Philip Ankley

**Abstract**

Genome wide association studies (GWAS) are a powerful tool for investigating the relation of single nucleotide polymorphisms to phenotypic traits. In this study, a comparison of the most modern GWA algorithms were evaluated by using a common GWAS dataset of *Arabidopsis thaliana*. Although a large number of issues arose while performing certain algorithms of interest, we conclude that the abundance and speed of modern algorithms can serve many differing needs when observing for associations.

**Introduction**

We conduct here a comparison of the most modern genome wide association (GWA) algorithms on a common dataset to evaluate results in terms of ease of use, accessibility, computational time and resources, and some indications of accuracy. The large number of algorithms published appear near indistinguishable to researchers conducting GWA studies, and we aim to help streamline the selection process. The algorithms used were all published in 2017, and thus represent the most modern algorithms available. While it is a potential result that an ensemble of algorithms provides the best result, avoiding the use of unfruitful algorithms offers reduced computational resources and faster identification of important single nucleotide polymorphisms (SNPs), as drug targets, therapeutic markers, or useful quantitative trait loci (QTL) for breeding.

Algorithms

The first algorithm under evaluation is Coal-Miner, which takes the form of a methodologic pipeline. The initial stages of the algorithm seeks to detect candidate loci and the subsequent stages test for association using a linear mixed model with multiple effects to account for sample relatedness within candidate loci and globally for the entire genome (Hejase et al. 2017).

BAL-VI is a bayesian adaptive lasso that uses variational inference. This allows for the detection of underlying variants by fitting all SNPs in a GWA study simultaneously which aims to speed up the process and accuracy of multi variant identification (Duan et al. 2017). MUSE, Multi-locus Sampling-based Epistasis Algorithm, is used to improve genetic trait prediction by conducting bidirectional sampling. It samples k-locus interactions from (k-1)-locus interactions and then decomposes the k-locus samples into multiple (k-1)-locus samples for further sampling (He and Parida 2017). Hierarchical association coefficient algorithm is a new method for measuring association between categories and observations. Its measure of association is based on degree of variance among the averages of all the categories in question. It uses two sorting extremes: observations being increasingly sorted into stratified ascending orders and observations being decreasingly sorted into stratified ascending orders, which allows the maximization of degree of variance among all averages. The algorithm looks to be an alternative to the F-test in GWAS (Kim 2017). LLR (latent-low rank) approach is used to improve power of identifying causal locus, by the use of a novel EM-path algorithm that reduces computational costs and allows the facilitation of parameter tuning. This approach was developed due to the complexity involved with polygenicity and identifying associated loci (Liu et al. 2017).

Iterative modified-Sure Independence Screening EM-Bayesian LASSO (ISIS EM-BLASSO) algorithm first uses an approach to reduce the number of markers, via correlation learning to achieve a moderate number. It then uses a moderate-scale variable selection method, SCAD, to select variables in the reduced model. Finally, the effects these two changes had on the variables is estimated by Estimation-Maximization (EM) Bayesian LASSO algorithm and tested by likelihood ratio statistic for true quantitative trait nucleotide (GTN) detection. The algorithm is shown to reduce computational time and expense (Tampa et al. 2017). LARS (least angle regression) algorithm is implemented to select SNPs most potentially associated with a trait, while empirical bayes is used to estimate the effects of the selected SNPs. All the nonzero effects are then further examined by likelihood ratio test to confirm true quantitative trait nucleotides. This entire method is referred to as pLARmEB (Zhang et al. 2017)

Our paper presents an independent evaluation on the accessibility, ease of use, and some information to the computational effectiveness of these variety of algorithms. Due to time constraints, the thoroughness of evaluation is not to the quality we desired, but we aim to provide a useful guide for others using the algorithms we evaluated.

**Methods**

Dataset

Our analysis was conducted on a GWAS dataset created by Atwell et al. in 2010, obtained from the R package p.exact as a gwaa.data object. It contains the phenotypic profiles of 199 genotypes of *Arabidopsis thaliana*, the most studied model plant and the first plant to have its genome published. It has a similar amount of linkage disequilibrium to humans, and therefore is expected to provide insights into effective strategies for GWAS in humans. It contains data on 107 phenotypes, allowing for multiple comparisons. *Arabidopsis thaliana* also has an extremely well annotated genome, allowing us to have a ground-truth for any SNPs suggested to be significantly associated with a phenotype. The dataset includes 250,000 SNPs for a genome size of around 120 million base pairs (bp), corresponding to about a SNP every 500 bp. Our analysis focused on the flowering trait ontology group, in correspondence with other studies conducting GWA.

Computation

All analysis was conducted using a computer with Intel Core i5-6200U CPU at 2.3 GHz and 8 GB DDR3 SDRAM. Some computational times are estimated because the computer could not be devoted to solely to running analysis. Scripts written in R were run in RStudio using R version 3.4.

**Results**

HA-Coefficient

*Ease of use*

The code was provided as supplementary material in the form of a word document, necessitating the copying of it into an R script file for running. The script involved both the generation and evaluation of a simulated dataset, without division of the script into discrete functions. While this made modification of the script to compartmentalize is necessary, it also made the input format of the genotypic and phenotypic data obvious. The author of the algorithm proposed a new metric of QTL significance, known as the HA-coefficient, which increases as the SNP becomes more significant. While this metric is analogous to a P-value, it is not as clearly understood how to interpret its meaning. Determining the necessary cut-off to call a SNP significant would take extensive model validation for this metric.

Plotting the results for HA-coefficient

*Prior validation*

This algorithm was presented in a paper providing only a simulated dataset, to our knowledge it had not been evaluated on real data.

*Time*

The dataset took a minimal amount of manipulation to make functional, due to its clarity of data input times. The runtime of the algorithm was near 5 minutes.

*Inputs*

Genotypic data is a data.matrix or numerical dataframe object with the number of rows equal to the number of individuals or samples, with the first column containing the ID of that individual. All later rows correspond to the genotype at each loci, as a numeric, such as 1 or 2. Phenotypic data is a numeric vector of length equal to number of individuals, containing values for a given phenotype for each individual. Missing data need not be removed.

ISIS EM-BLASSO

*Ease of use*

This algorithm was easily downloadable from source code on rdrr or as part of the mrMLM package from the Comprehensive R Archive Network (CRAN). However, we were unable to produce a functioning code from the package and instead depended on the source code. The documentation provides a poor description of the inputs, which instead needs to be implied from reading the source code. The variable names are not sufficiently descriptive to follow the code through without line-by-line reading. Due to numerous errors while running the functions, it was deemed more effective to run the code outside the function and debug with saved variable values. Extensive work was required to determine necessary inputs for the data.

*Prior Validation*

ISIS EM-BLASSO was evaluated on both simulated and real dataset (the same *Arabidopsis thaliana* dataset used here). The authors determined the algorithm to be much faster than conventional algorithms such as EMMA, by a factor of about 50, with comparable accuracy.

*Time*

A large amount of effort was invested in procuring a working algorithm, but the overall computation time was around 12 minutes. With adequate description of the input data and preformatting, computation is fast. While this is the case for all GWAS algorithms, converting the GENAbel GWAS class to a dataframe or data matrix is an extensive time commitment.

*Inputs*

Genotypic data is in the form of a data.matrix object, with columns corresponding to individuals and rows to SNPs, containing numerical data, typically 1s and 2s. Phenotypic data is a data frame, composed of one column and rows with phenotype values for each individual. A kinship matrix is required, which can be calculated using the ibs function of GENAbel on the snp.data object. Population structure is a data.matrix, composed of columns corresponding to the principal components of the genotype data. The number of columns appears to be user defined. The rows correspond to individuals, and the values are scores at each principal component for each individuals. This can be obtained using the prcomp function of the stats package in R, applied to a transposed version of the genotypic data. The raw genotype data is a dataframe with the first four columns corresponding to the SNP names, chromosome position, map position, and dominant genotype. Missing data is not tolerated, so data should only include individuals for which the phenotype is observed.

pLARmEB

*Ease of use*

The algorithm is available as part of the mrMLM package, like ISIS EM-BLASSO, and similarly we did not have functionality out-of-the-box. Source code was obtained from rdrr. The inputs are nearly identical to ISIS EM-BLASSO. Despite the similarity to ISIS EM-BLASSO is source and implementation, we were unable to produce an analysis after about 18 hours of computation time

*Prior validation*

The authors evaluated the algorithm’s effectiveness for both simulated data and the Atwell et al. 2010 dataset, focusing on flowering traits. Compared to LARmEB, EMMA, FASTmrEMMA, mrMLM and BhGLM, it was the fastest and had the best detection in the 0.01 to 0.001 P-value range for QTL. The authors concluded that it was the best of the tested algorithms.

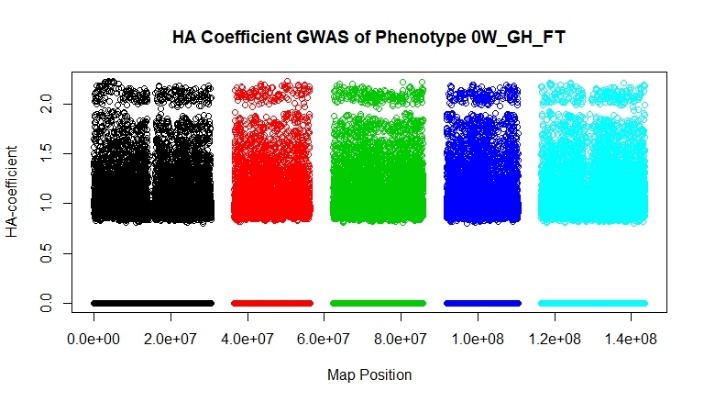
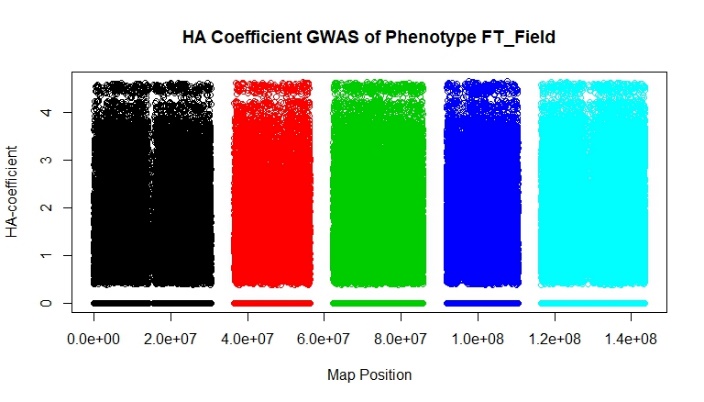
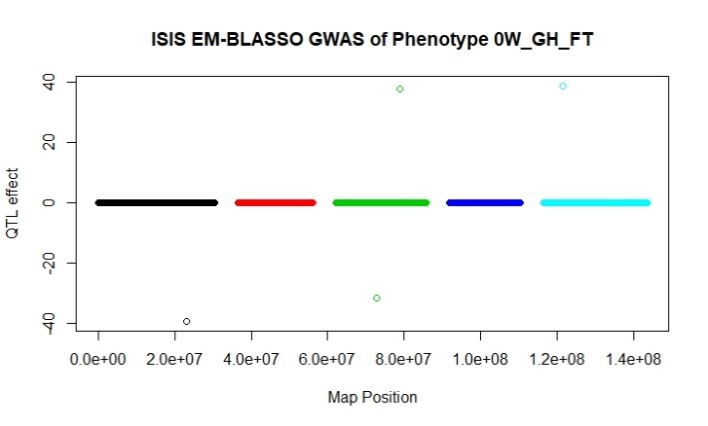
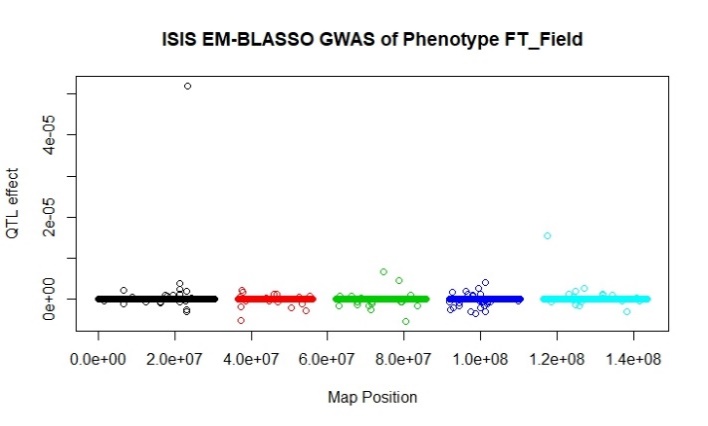
*Time*

The effort put into processing input data suitable for pLARmEB was the same as for ISIS EM-BLASSO, and therefore difficult and poorly documented. After running the algorithm for 18 hours, no output or error was produced to evaluate success of the algorithm. Greater computational resources or time may produce a result with pLARmEB.

*Inputs*

pLARmEB requires the same input data as ISIS EM-BLASSO.

Accuracy

While a number of SNPs were identified as linked to phenotype, connecting the SNPs to genomic positions posed an unexpected challenge. The dataset included map positions for the chromosome, from 0 to the length of the entire genome. However, the values used to find positions in the *Arabidopsis* genome, such as with the TAIR browser, are composed of chromosome number and site on that chromosome. It is possible that when the map positions in the dataset were created, the known size of the chromosomes was different than the best estimate today, resulting in shifted positions. Further analysis is necessary to find the correct locations of the implicated SNPs and relate their position to genes which may be involved in creating the phenotype of interest. Plots show the results for HA-Coefficient and ISIS EM-BLASSO analyses are show in Figure 1.

A.

C.

C.

D.

A.

B.

Figure 1. Resulting output of HA-Coefficient and ISIS EM-BLASSO algorithms. A and B display results for the phenotype FT Field, while C and D display the results for phenotype 01 GH FT. These phenotypes appear to differ in the number of associated strongly causal loci, with 0W GH FT showing fewer but more significant SNPs for the ISIS EM-BLASSO algorithm.

Notes on Other Algorithms

*Coal-Miner*

This GWAS pipeline was easily downloadable, however it required packages with Unix only versions. This was prohibitive for us the authors, who relied on Windows 10 devices to conduct the analyses.

*BGTools*

This package was downloadable from GitHub, but came with no documentation and was not installable in R.

*MUSE*

MUSE is a proposed algorithm, but no scripts are available to implement the code. If time allowed, it may have been possible to write original scripts from the paper.

*LLR*

While the LLR algorithm was easily downloaded from GitHub, it is written in Matlab, for which the authors did not have licenses. While better tools may be available in Matlab, if other authors cannot use the algorithm its utility is poor.

**Discussion**

While conducting this analysis, we encountered a large number of issues preventing completion of the main goal: creating an exhaustive analysis of the most recent genome wide association study algorithms for demonstrated effectiveness in their presentation and how useful of a tool the algorithms are to other researchers, in terms of accessibility, ease of use, computational time, and accuracy. It is some of these characteristics themselves which contributed to our difficulties, often because packages were not downloadable, could not be run within a reasonable time, or capable of running on our available hardware.

The algorithms which were able to be run, HA-coefficient and ISIS EM-BLASSO, appear to be generally fast and reasonably easy to run with good understanding of the requisite inputs. Interpretation is not necessarily as clear with HA-coefficient, because it lacks conventional statistics like p-value or LOD. With rigorous accuracy evaluation is could be determined which algorithm provides the best capture of important loci. It does appear, however, that ISIS EM-BLASSO implicates fewer and therefore likely more significant loci.

Our project was starting out strongly ambitious. The algorithms we were considering run using a variety of languages, using different genotypic formats, and multiple prerequisite packages and software. Due to the challenges of acquiring, installing, and running the software, our analysis was strongly limited. Earlier on in the project, gaining access to a computing core would have streamline the execution of the code, as well as overcoming issues with runtime and operating systems.

**Conclusion**

Modern GWAS algorithms provide faster and more accurate analysis than previous generations of algorithms. They employ combined and varied methodologies, giving authors the tools to experiment and find ones which fit their needs. While we attempted to demonstrate how well these algorithms work in a comprehensive review, more work will be required to make conclusions as to the best algorithms overall or for specific tasks.

**References**

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone

AM, Hu TT, Jiang R. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010 Jun;465(7298):627.

Duan W, Zhao Y, Wei Y, Yang S, Bai J, Shen S, Du M, Huang L, Hu Z, Chen F. A fast

algorithm for Bayesian multi-locus model in genome-wide association studies. Molecular Genetics and Genomics. 2017 Aug 1;292(4):923-34.

He D, Parida L. MUSE: A MULTI-LOCUS SAMPLING-BASED EPISTASIS ALGORITHM

FOR QUANTITATIVE GENETIC TRAIT PREDICTION. InPACIFIC SYMPOSIUM ON BIOCOMPUTING 2017 2017 (pp. 426-437).

Hejase HA, Vande Pol N, Bonito GM, Edger PP, Liu KJ. Coal-Miner: a coalescent-based

method for GWA studies of quantitative traits with complex evolutionary origins. bioRxiv. 2017 Jan 1:132951.

Kim B. Hierarchical Association Coefficient Algorithm: New Method for Genome-Wide

Association Study. Evolutionary Bioinformatics. 2017 Aug 22;13:1176934317713004.

Kooperberg C, LeBlanc M, Obenchain V. Risk prediction using genome‐wide association

studies. Genetic epidemiology. 2010 Nov 1;34(7):643-52.

Liu J, Wan X, Wang C, Yang C, Zhou X, Yang C. LLR: a latent low-rank approach to

colocalizing genetic risk variants in multiple GWAS. Bioinformatics. 2017 Aug 14;33(24):3878-86.

Tamba CL, Ni YL, Zhang YM. Iterative sure independence screening EM-Bayesian LASSO

algorithm for multi-locus genome-wide association studies. PLoS computational biology. 2017 Jan 31;13(1):e1005357.

Zhang J, Feng JY, Ni YL, Wen YJ, Niu Y, Tamba CL, Yue C, Song Q, Zhang YM. pLARmEB:

integration of least angle regression with empirical Bayes for multilocus genome-wide

association studies. Heredity. 2017 Jun;118(6):517.