JWAS version 2: Leveraging biological information and high-throughput phenotypes into genomic prediction and association

H. Cheng[1*], R. Fernando[2], D. Garrick[3], T. Zhao[1], J. Qu[1]

[1]Department of Animal Science, University of California Davis, 95616, Davis, US;
[2]Department of Animal Science, Iowa State University, 50010, Ames, US;
[3]School of Agriculture, Massey University, 4442, Palmerston North, New Zealand
[*]qtlcheng@ucdavis.edu

Abstract

The initial release of JWAS was developed as an open-source software tool for single-trait or multi-trait linear mixed models including those used for genome-enabled prediction or genome-wide association studies. We have now made extensive updates to JWAS to incorporate new features to accommodate high-throughput phenotypes and biological information useful in bridging the gap between DNA sequence and phenotypic information. The new features include but are not limited to: 1) Incorporating high-throughput phenotypes by simultaneous modeling of thousands of traits or through time-series (longitudinal trait) models; 2) Using functional annotation information as a-priori biological knowledge; 3) Extending mixed models to multi-layer neural networks that can accommodate intermediate omics data. This latest version of JWAS (version 2) is developed for researchers in both industry and academia to take full advantage of new information such as high-throughput phenotypes, functional annotations, and multi-omics data as they become increasingly available in this decade.

Introduction

The release of the USDA blueprint for animal genomics (2008-2017) triggered the development of many statistical methods to improve genetic progress as well as the implementation of these methods in software tools such as JWAS. The initial version of JWAS (Cheng *et al.*, 2018) was an open-source software tool for single-trait or multi-trait analyses accommodating those models used for genome-enabled prediction and genome-wide association studies, with either complete or incomplete genomic data, i.e., "single-step" methods (Legarra *et al.,* 2009; Fernando *et al.,* 2014; Fernando *et al.,* 2016). A portfolio of Bayesian regression models including variable selection and shrinkage estimation methods were implemented in JWAS to provide broad scope of analyses.

With the development of high-throughput sequencing and phenotyping technology, the USDA blueprint in animal genomics for the next decade (2018 - 2027) (Rexroad *et al.*, 2019) has identified several areas requiring more advanced usage of genomic information for animal production. Thus, there was a need to develop statistical methods and tools that could incorporate high-throughput phenotyping, multi-omics data, and biological information such as functional annotations for genomic analyses. Extensive additional updates to our JWAS package were required to accommodate the issues identified in the new blueprint for improved utilization of genomic information. The availability of powerful and efficient statistical methods implemented in extensible software tools that can accommodate the advancing state of the art would have impacts on a wide range of breeding programs. To fill this gap, we extensively upgraded JWAS and released JWAS version 2. The main objective of this paper is to introduce new features in JWAS for researchers in industry or academia, so they can take full advantage of new information such as high-throughput phenotypes, multi-omics data, and functional annotation of animal genomes as they become increasingly available in this decade.

Materials & Methods
A list of features in JWAS is shown in Table 1, and highlighted new features in JWAS version 2 are described below.

Table 1. A list of features in JWAS for single-trait or multi-trait analyses.

fixed class or covariate effects (e.g., age, sex)	
non-genomic random effects (e.g., litter, pen)	
use of pedigree information	
joint fitting of direct and maternal genetic effects	
permanent environmental effects	
use of genomic information	
complete genomic data[1]	Genomic BLUP Bayesian Alphabet
incomplete genomic data[2]	Genomic BLUP Bayesian Alphabet
biological information	multi-layer neural networks[3] multi-class models[3]
high-throughput phenotypes	sparse factor models[3] random regression models[3]

[1] "Complete genomic data" indicates that genotypes are available on all individuals.
[2] "Incomplete genomic data" indicates that genotypes are not available on all individuals ("single-step" analyses).
[3] These new features are described in the Materials and Methods section.

Mixed models extended to multi-layer neural networks including intermediate omics data. With the growing amount and diversity of intermediate omics data complementary to genomics (e.g., measures of DNA methylation, gene expression, and protein abundance), there is a need to develop methods to incorporate intermediate omics data into genome-enabled prediction and association studies to enhance conventional genomic evaluation. We have implemented a new method NN-LMM (Zhao *et al.*, 2021a; Zhao *et al.*, 2021b) to model the multiple layers of regulation from genotypes to intermediate omics features, then to phenotypes, by extending conventional linear mixed models to multi-layer artificial neural networks. NN-LMM incorporates intermediate omics features by adding middle layers between genotypes and phenotypes. Linear mixed models (e.g., pedigree-based BLUP, GBLUP, Bayesian Alphabet, single-step GBLUP, or single-step Bayesian Alphabet) can be used to sample marker effects or genetic values on intermediate omics features, and activation functions in neural networks are used to capture nonlinear possibly unknown relationships between intermediate omics features and phenotypes.

Using functional annotation information as a-priori biological knowledge. In addition to phenotypic, genomic, and pedigree data, massive amounts of biological information are being generated by the scientific community. For example, the Functional Annotation of Animal Genome (FAANG) consortium is functionally annotating genomic regulatory elements of domesticated animal species. However, such biological information is not generally considered in genomic prediction that routinely uses phenotypic information along with pedigree and SNP genotypes. We have implemented multi-class models (Wang *et al.*, 2021) in JWAS version 2 to allocate markers into different classes based on biological information and assigning separate priors to markers in these different classes. This is implemented for single-trait or multi-trait analyses.

Incorporating high-throughput phenotypes by simultaneous modeling of thousands of traits through sparse factor models. Large-scale phenotypic data are becoming increasingly accessible due to advances in high-throughput phenotyping platforms and technologies for multi-omics profiling. Although the incorporation of large-scale phenotypic data into genome-enabled analysis can enhance the power of prediction and association inference, genomic analyses of high-dimensional, highly correlated data are challenging. We have implemented a Bayesian sparse factor model (Runcie *et al.*, 2021) with different prior assumptions on marker effects to simultaneously analyze hundreds to thousands of traits for genomic prediction or GWAS. This Bayesian sparse factor model can effectively reduce a large multi-trait model into a set of parallel single-trait models by introducing the concept of latent traits, i.e., traits that are not directly observed, but can be inferred based on their effects on the observed traits. This model is based on some biologically reasonable assumptions including 1) a limited number of the latent traits control the majority of variation; 2) each latent trait controls only a subset of the observed traits; 3) each latent trait is controlled by genetics and the environment.

Incorporating high-throughput phenotypes through time-series (longitudinal trait) models. In animal and plant improvement, longitudinal traits may be recorded across several time points on a physiological cycle, and might be jointly analyzed to improve the performance relative to an analysis at one time point. The recent advancement in phenotyping platforms enables acquisition of large-scale non-destructive phenotypes measured at frequent intervals. In pedigree-based analyses, the use of random regression models (RRM) has been a typical approach employed for such analyses of longitudinal traits (Henderson, 1982); Laird and Ware, 1982). We have implemented Bayesian random regression models for marker effects that can accommodate variable selection for the analysis of longitudinal data that would accrue through high-throughput phenotyping.

Results

The JWAS software tool was and is being developed as a single-language software tool that is easy for both champions and novice community members to use, maintain, modify, or extend. In terms of speed, for the same BayesC analysis, JWAS is now faster than the C++ program, GenSel. Further, JWAS is relatively easy to extend: to accommodate categorical traits in JWAS required adding only about 40 lines of code; to extend conventional Bayesian regression methods to single-step analyses required adding one file composed of a few hundred of lines of code and a few minor modifications to the original code.

Discussion

Whole-genome analyses are usually computationally intensive even for data sets of moderate size. Although these analyses can be implemented with most dynamic languages such as R or Python with very readable code that is easy to understand, modify and extend, they are often

too slow for real data analyses. Therefore, for computational efficiency, compiled languages such as C, C++, or Fortran have historically been used to implement methods for genomic analyses. In order to make these more accessible, a dynamic language such as R may be used as a user interface to the underlying methods that are written in the compiled language. The problem of using compiled languages, however, is that they are usually hard to understand, modify, extend and maintain, in addition to being difficult to readily deploy across different operating systems. Julia (Bezanson *et al.*, 2017), a relatively new scientific programming language, approaches the computing speed of compiled languages, but retains the benefits of dynamic language.

References

Bezanson J., Edelman A., Karpinski S., and Shah V. B. (2017) SIAM Review 59(1):65-98. https://doi.org/10.1137/141000671

Cheng H., Garrick D.J., and Fernando R.L. (2018) Proc. of the 11th WCGALP, Auckland, New Zealand.

Fernando R.L., Dekkers J.C., and Garrick D.J. (2014) Genet. Sel. Evol. 46(1):1-13. https://doi.org/10.1186/1297-9686-46-50

Fernando R.L., Cheng H., Golden B.L., and Garrick D.J. (2016) Genet. Sel. Evol. 48(1):1-8. https://doi.org/10.1186/s12711-016-0273-2

Henderson C.R. (1982) Biometrics 38(3):623-640. https://doi.org/10.2307/2530044

Laird N.M., and Ware J.H. (1982) Biometrics 38(4):963-974. https://doi.org/10.2307/2529876

Legarra A., Aguilar I., and Misztal I. (2009) J. Dairy. Sci. 92(9):4656-4663. https://doi.org/10.3168/jds.2009-2061

Rexroad C., Vallet J., Matukumalli L.K., Reecy J., et al. (2019) Front. Genet. 10:327. https://doi.org/10.3389/fgene.2019.00327

Runcie D.E., Qu J., Cheng H., and Crawford L. (2021) Genome Biol. 22(1):1-25. https://doi.org/10.1186/s13059-021-02416-w

Wang Z., and Cheng H. (2021) Front. Genet. 12. https://doi: 10.3389/fgene.2021.717457

Zhao T., Fernando R., and Cheng H. (2021a) G3.11(10). https://doi.org/10.1093/g3journal/jkab228

Zhao T., Zeng J., and Cheng H. (2021b) BioRxiv preprint https://doi.org/10.1101/2021.12.10.472186